ICIBM 2019

INTERNATIONAL CONFERENCE ON INTELLIGENT BIOLOGY AND MEDICINE

JUNE 9 - 11, 2019 COLUMBUS, OHIO

CO-HOSTED BY:

THE INTERNATIONAL ASSOCIATION FOR INTELLIGENT BIOLOGY AND MEDICINE (IAIBM)
& **THE DEPARTMENT OF BIOMEDICAL INFORMATICS, THE OHIO STATE UNIVERSITY**

2019 International Conference on Intelligent Biology and Medicine (ICIBM 2019)

June 9-11, 2019
Columbus, OH, USA

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

TABLE OF CONTENTS

Welcome to ICIBM 2019!

On behalf of all our conference committees and organizers, we welcome you to the 2019 International Conference on Intelligent Biology and Medicine (ICIBM 2019). ICIBM is the official conference of The International Association for Intelligent Biology and Medicine (IAIBM, http://iaibm.org/), a non-profit organization whose mission is to promote the intelligent biology and medical science, through member discussion, network communication, collaborations, and education. This year, ICIBM 2019 is co-hosted by the Department of Biomedical Informatics at The Ohio State University.

The fields of bioinformatics, systems biology, and intelligent computing are continuing to evolve at a rapid pace and continue to have a strong impact in scientific research and medical innovations. With this in mind, we are pleased to provide a forum that fosters inter-disciplinary research and discussions, educational opportunities, and collaborative efforts among these ever growing and progressing fields. We are proud to have built on successes of previous years' conferences to provide an exciting program that provides a balanced mix spanning trainees and world-renown scientists, oral and poster presentations, workshops, tutorials, and plenty of built-in breaks for invaluable discussions.

This year, we have an exciting line-up for our keynote speakers, including world-renowned experts Drs. Jeremy Edwards, Peter Karp, Elaine Mardis, and T.M. Murali. Throughout the conference, we will also feature eminent scholar speakers, Drs. Bruce Aronow, Alla Karnovsky, Jeff Parvin, and Haixu Tang, and will be hosting four tutorials and workshops on the first day of the conference. In addition, talks will be given from faculty members, postdoctoral fellows, PhD students and trainee level awardees selected from 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 2019's program. We'd like to extend our thanks to our sponsors for making this event possible, including the National Science Foundation, BGI Americas, UTHealth, and Karger.

Last but not least, our sincerest thanks to members of all our ICIBM 2019 committees, and to our volunteers for their valuable efforts. Their dedication to making ICIBM 2019 a success is invaluable, and demonstrates the strength and commitment of our community.

On behalf of all of us, we hope that our hard work has provided a conference that is thought provoking, fosters collaboration and innovation, and is enjoyable for all of our attendees. Thank you for attending ICIBM 2019. We look forward to your participation in all our conference has to offer!

Sincerely,

Lang Li, PhD
ICIBM General Chair
Professor and Chair,
Department of
Biomedical Informatics
The Ohio State University

Ewy Mathé, PhD
ICIBM Program Co-Chair
Assistant Professor,
Department of
Biomedical Informatics
The Ohio State University

Chi Zhang, PhD
ICIBM Program Co-Chair
Assistant Professor of
Medical & Molecular
Genetics, School of
Medicine, Indiana University

ACKNOWLEDGEMENTS

General Chair
Lang Li, The Ohio State University, USA

Steering Committee
Zhongming Zhao, Co-Chair, The University of Texas Health Science Center at Houston, USA
Kai Wang, Co-Chair, Children's Hospital of Philadelphia, USA
Yidong Chen, The University of Texas Health Science Center at San Antonio, USA
Kun Huang, Indiana University, USA
Yi Xing, Children’s Hospital of Philadelphia, USA
Xiaohua Tony Hu, Drexel University, USA
Elmer Bernstam, The University of Texas Health Science Center at Houston, USA

Program Committee
Ewy Mathé, Co-Chair, The Ohio State University, USA
Chi Zhang, Co-Chair, Indiana University, USA
Yongsheng Bai, Indiana State University, USA
Guy Brock, The Ohio State University, USA
Sha Cao, Indiana University School of Medicine, USA
Yidong Chen, The University of Texas Health Science Center at San Antonio, USA
Jianlin Cheng, University of Missouri Columbia, USA
Lijun Cheng, The Ohio State University, USA
Mehmet M. Dalkilic, Indiana University Bloomington, USA
Youping Deng, University of Hawaii, USA
Abolfazl Doostparast, University of California at Santa Barbara, USA
Yan Guo, University of New Mexico, USA
Zhi Han, Indiana University School of Medicine, USA
Matthew Hayes, Xavier University of Louisiana, USA
Kun Huang, Indiana University School of Medicine, USA
Yang Huang, Kaiser Permanente, USA
Weichun Huang, NIH/NIEHS, USA
Sarath Janga, Indiana University - Purdue University Indianapolis, USA
Zhi-Liang Ji, Xiamen University, China
Peilin Jia, The University of Texas Health Science Center at Houston, USA
David Kline, The Ohio State University, USA
Jun Kong, Emory University, USA
Xiaokui Mo, The Ohio State University, USA
Tianbao Li, The University of Texas Health Science Center at San Antonio, USA
Aimin Li, Xi'an University of Technology, China
Lang Li, The Ohio State University, USA
Tao Li, Nankai University, China
Qiwei Li, Rice University, USA
Shantao Li, Stanford University, USA
Fuhai Li, The Methodist Hospital, Weill Medical College of Cornell University, USA
Li Liao, University of Delaware, USA
Xiaoming Liu, University of South Florida, USA
Yin Liu, The University of Texas Health Science Center at Houston, USA
Lu Liu, North Dakota State University, USA
Zhandong Liu, Baylor College of Medicine, USA

Qin Ma, The Ohio State University, USA
Mirjana Maletic-Savatic, Baylor College of Medicine, USA
Nitish Mishra, University of Nebraska Medical Center, USA
Tabrez Anwar Shamim Mohammad, Greehey Children's Cancer Research Institute (GCCRI), UTHSCSA, USA
Xia Ning, The Ohio State University, USA
Hatice Gulcin Ozer, The Ohio State University, USA
Jianhua Ruan, The University of Texas at San Antonio, USA
Xiaofeng Song, Nanjing University of Aeronautics and Astronautics, China
Manabu Torii, Kaiser Permanente, USA
Jun Wan, Indiana University School of Medicine, USA
Ying-Wooi Wan, Baylor College of Medicine, USA
Junbai Wang, Oslo University Hospital, Norway
Jiayin Wang, Xi'an Jiaotong University, China
Qingguo Wang, Lipscomb University, USA
Kai Wang, University of Philadelphia, USA
Yufeng Wang, The University of Texas at San Antonio, USA
Chaochun Wei, Shanghai Jiao Tong University, China
Lei Wei, The Ohio State University, USA
Yonghui Wu, University of Florida, USA
Junfeng Xia, Anhui University, China
Lei Xie, City University of New York, USA
Yi Xing, University of Philadelphia, USA
Hua Xu, The University of Texas School of Biomedical Informatics at Houston, USA
Yu Xue, Huazhong University of Science and Technology, China
Jingwen Yan, Indiana University - Purdue University Indianapolis, USA
Zhenqing Ye, Mayo Clinic, USA
Lianbo Yu, The Ohio State University, USA
Chi Zhang, Indiana University, USA
Xiaoli Zhang, The Ohio State University, USA
Yan Zhang, The Ohio State University, USA
Pengyue Zhang, The Ohio State University, USA
Ping Zhang, The Ohio State University, USA
Shaojie Zhang, University of Central Florida, USA
Han Zhang, Nankai University, China
Chi Zhang, Indiana University, School of Medicine, USA
Zhongming Zhao, The University of Texas Health Science Center at Houston, USA
Min Zhao, University of the Sunshine Coast, Australia
Jim Zheng, The University of Texas Health Science Center at Houston, USA
Yunyun Zhou, University of Mississippi Medical Center, USA

Publication Committee
Yan Guo, Co-Chair, The University of New Mexico, USA
Xia Ning, Co-Chair, The Ohio State University, USA

Workshop/Tutorial Committee
Michael Wagner, Co-Chair, Cincinnati Children’s Hospital Medical Center, USA
Jin Chen, Co-Chair, University of Kentucky, USA

Publicity Committee

Maciej Pietrzak, Co-Chair, The Ohio State University, USA
Licong Cui, Co-Chair, University of Kentucky, USA

Award Committee
Long Lu, Co-Chair, Cincinnati Children's Hospital Center, USA
Qiang (Shawn) Cheng, Co-Chair, University of Kentucky, USA
Qin Ma, The Ohio State University, USA

Trainee Committee
Tara Eicher, Co-Chair, The Ohio State University, USA
Astrid Manuel, Co-Chair, The University of Texas Health Science Center at Houston, USA

Local Organization Committee
Lai Wei, Chair, The Ohio State University, USA
Shasha Bai, The Ohio State University, USA

Industry/Sponsorship Committee
Zhongming Zhao, Co-Chair, The University of Texas Health Science Center at Houston, USA
Shiqiang Tao, Co-Chair, University of Kentucky, USA

International Conference on Intelligent Biology and Medicine Program-at-a-glance June 9-11, 2019

Sunday, June 9th

9:00	**Registration Opens** ***and Continental Breakfast***		
10:15-11:00	**Keynote Lecture (Room: Ballroom)** **Elaine Mardis, PhD** Co-executive Director of the Institute for Genomic Medicine at Nationwide Children's Hospital Nationwide Foundation Endowed Chair of Genomic Medicine Professor of Pediatrics at The Ohio State University College of Medicine **Title**: ***A System for Pediatric Precision Cancer Medicine***		
11:00-11:20	***Boxed Lunches Pickup for next session***		
CONCURRENT WORKSHOPS/TUTORIALS			
Room	**Ballroom**	**Pfahl 202**	**Pfahl 302**
11:20-1:30	Dr. Wenjin Zheng, UTH Dr. Yidong Chen, UTH Dr. Yufei Huang, UTSA **Data Driven Cancer Research: Data Science Research and Applications**		Dr. Yan Guo, UNM **Machine Learning Demystified**
1:30-1:45	***Break***		
1:45-3:45	Dr. Wenjin Zheng, UTH Dr. Yidong Chen, UTH Dr. Yufei Huang, UTSA **Data Driven Cancer Research: Tutorial on Deep Learning for Cancer Genomics**	Dr. Peter Karp SRI International **Tutorial for the BioCyc Microbial Genomes Web Portal**	Dr. Jianrong Wang, MSU **Epigenetics data analysis**
3:45-4:00	***Break***		
4:00-6:00	**Poster Session (Room: Ballroom)** ***Hors d'oeuvres Served***		

Monday, June 10th

7:30-8:30	**Registration Open** ***and Continental Breakfast***		
8:30-8:40	**Opening Remarks**		
8:40-9:30	**Keynote Lecture (Room: Ballroom)** **Peter Karp, PhD** Director, Bioinformatics Research Group Artificial Intelligence Center SRI International **Title**: ***BioCyc Tools for Metabolic Modeling and Omics Data Analysis***		
9:30-9:40	***Break***		
9:40-10:00	**Eminent Scholar Talk (Room: Ballroom)** **Jeffrey Parvin, MD, PhD** Professor, Department of Biomedical Informatics Associate Dean of Graduate Studies Co-Director, Biomedical Sciences Graduate Program The Ohio State University **Title**: ***The impact of sequence variants on protein function***		
10:00-10:10	***Break for parallel sessions***		
CONCURRENT SESSIONS			
Room	**Ballroom** **NGS & Tools** **Session Chair: 1**	**Pfahl 202** **General Genomics** **Session Chair: 2**	**Pfahl 302** **Bioinformatics** **Session Chair: 3**
10:10-10:30	**An ancestral informative marker panel design for individual ancestry estimation of Hispanic population using whole exome sequencing data** Li-Ju Wang, Catherine Zhang, Sophia Su, Hung-I Chen, Yu-Chiao Chiu, Zhao Lai, Hakim Bouamar, Francisco Cigarroa, Lu-Zhe Sun and Yidong Chen	**Association Analysis of Common and Rare SNVs using Adaptive Fisher Method to Detect Dense and Sparse Signals** Xiaoyu Cai, Lo-Bin Chang and Chi Song	**The Comparisons of Prognostic Power and Expression Level of Tumor Infiltrating Leukocytes in Hepatitis B- and Hepatitis C-related Hepatocellular Carcinomas** Yi-Wen Hsiao, Lu-Ting Chiu, Ching-Hsuan Chen, Wei-Liang Shih and Tzu-Pin Lu
10:30-10:50	**normGAM: An R package to remove systematic biases in genome architecture mapping data** Tong Liu and Zheng Wang	**An integrative, genomic, transcriptomic and network-assisted study to identify genes associated with human cleft lip with or without cleft palate** Fangfang Yan, Yulin Dai, Junichi Iwata, Zhongming Zhao and Peilin Jia	**SigUNet: signal peptide recognition based on semantic segmentation** Jhe-Ming Wu, Yu-Chen Liu and Tien-Hao Chang

10:50-11:10	**Sparse Convolutional Denoising Autoencoders for Genotype Imputation** Junjie Chen and Xinghua Shi	**Association between ALS and retroviruses: Evidence from bioinformatics analysis** Jon Klein, Zhifu Sun and Nathan Staff	**Integrated metabolomics and transcriptomics study of traditional herb Astragalusmembranaceu sBge.var. mongolicus (Bge.) Hsiao reveals global metabolic profile and novel phytochemical ingredients** Xueting Wu, Xuetong Li, Wei Wang, Yuanhong Shan, Cuiting Wang, Mulan Zhu, Qiong La, Yang Zhong, Ye Xu, Peng Nan and Xuan Li
11:10-11:20	***Coffee/Tea Break***		
11:20-11:40	**High dimensional model representation of log likelihood ratio: Binary classification with SNP data** Ali Foroughi Pour, Maciej Pietrzak, Lara E. Sucheston-Campbell, Ezgi Karaesmen, Lori A. Dalton and Grzegorz A. Rempala	**ManiNetCluster: A novel manifold learning approach to reveal the functional links between gene networks** Nam Nguyen, Ian Blaby and Daifeng Wang	**BayesMetab: Treatment of Missing Values in Metabolomic Studies using a Bayesian Modeling Approach** Jasmit Shah, Guy Brock and Jeremy Gaskins
11:40-12:00	**Decoding regulatory structures and features from epigenomics profiles: a Roadmap-ENCODE Variational Auto-Encoder (RE-VAE) model** Ruifeng Hu, Guangsheng Pei, Peilin Jia and Zhongming Zhao	**Human protein-RNA interaction network is highly stable across vertebrates** Aarthi Ramakrishnan and Sarath Chandra Janga	**Dense module searching for gene networks associated with multiple sclerosis** Astrid Manuel, Yulin Dai, Leorah Freeman, Peilin Jia and Zhongming Zhao
12:00-12:20	**A unified STR profiling system across multiple species with whole genome sequencing data** Liu Yilin, Xu Jiao and Li Shuaicheng	**Differential co-expression analysis reveals early stage gene dis-coordination in Alzheimer's disease** Yurika Upadhyaya, Linhui Xie, Paul Salama, Sha Cao, Kwagnsik Nho, Andrew Saykin and Jingwen Yan	**Expression correlation attenuates within and between key signaling pathways in CKD progression** Hui Yu, Danqian Chen, Olufunmilola Oyebamiji, Yan Guo and Ying-Yong Zhao
12:20-1:30	***Lunch Break****-Boxed Lunches*		

1:30-1:50	**Eminent Scholar Talk (Room: Ballroom)** **Alla Karnovsky, PhD** Research Associate Professor of Computational Medicine & Bioinformatics Assistant Director, Masters Program University of Michigan **Title**: ***Identifying biologically relevant modules in metabolomics and lipidomics data with Differential Network-based Enrichment Analysis (DNEA)***		
1:50-2:00	***Short Break***		
2:00-2:50	**Keynote Lecture (Room: Ballroom)** **T.M. Murali, PhD** Professor, Department of Computer Science Co-director, ICTAS Center for Systems Biology of Engineered Tissues Virginia Tech **Title**: ***Pathways on Demand: Automated Reconstruction of Human Signaling Networks***		
2:50-3:00	***Break for parallel sessions***		
CONCURRENT SESSIONS			
Room	**Ballroom** **NGS & Tools** **Session Chair: 4**	**Pfahl 202** **Bioinformatics** **Session Chairs: 5**	**Pfahl 302** **Cancer Genomics** **Session Chairs: 6**
3:00-3:20	**Investigating Skewness to Understand Gene Expression Heterogeneity in Large Patient Cohorts** Benjamin Church, Henry Williams and Jessica Mar	**M3S: A comprehensive model selection for multi-modal single-cell RNA sequencing data** Yu Zhang, Changlin Wan, Pengcheng Wang, Wennan Chang, Yan Huo, Jian Chen, Qin Ma, Sha Cao and Chi Zhang	**Comparative evaluation of network features for the prediction of breast cancer metastasis** Nahim Adnan, Zhijie Liu, Tim Huang and Jianhua Ruan
3:20-3:40	**Clonal reconstruction from time course genomic sequencing data** Wazim Mohammed Ismail and Haixu Tang	**Multi-objective optimized fuzzy clustering for detecting cell clusters from single cell expression profiles** Saurav Mallik and Zhongming Zhao	**Highly robust model of transcription regulator activity predicts breast cancer overall survival** Chuanpeng Dong, Jiannan Liu, Steven X. Chen, Tianhan Dong, Guanglong Jiang, Yue Wang, Huanmei Wu, Jill L. Reiter and Yunlong Liu
3:40-4:00	**CNV detection from circulating tumor DNA in late stage non-small cell lung cancer patients** Hao Peng, Qiangsheng Dai, Zisong Zhou, Xiaochen Zhao, Dadong Zhang, Kejun Nan, Zhu-An Ou, Fugen Li, Hua Dong, Lei Tian, Yu Yao	**Network-based single-cell RNA-seq data imputation enhances cell type identification** Maryam Zand and Jianhua Ruan	**Pseudogene-gene functional networks are prognostic of patient survival in breast cancer** Sasha Smerekanych, Travis Johnson, Kun Huang and Yan Zhang

4:00-4:10	***Coffee/Tea Break***		
4:10-4:30	**A Protocol to Evaluate RNA Sequencing Normalization Methods** Zachary Abrams, Travis Johnson, Kun Huang, Philip Payne and Kevin Coombes	**The landscape of alternative splicing in HIV-1 infected CD4 T-cells** Seyoun Byun, Seonggyun Han, Yue Zheng, Vicente Planelles and Younghee Lee	**Deep Learning-based Cancer Survival Prognosis from RNA-seq Data: Approaches and Evaluations** Zhi Huang, Travis Johnson, Zhi Han, Bryan Helm, Sha Cao, Chi Zhang, Paul Salama, Maher Rizkalla, Christina Yu, Jun Cheng, Shunian Xiang, Xiaohui Zhan, Jie Zhang and Kun Huang
4:30-4:50	**Fully Moderated T-statistic in Linear Modeling of Mixed Effects for Differential Expression Analysis** Lianbo Yu, Jianying Zhang, Guy Brock and Soledad Fernandez	**Biological Representation of Chemicals Using Latent Target Interaction Profile** Mohamed Ayed, Hansaim Lim and Lei Xie	**Transcription factor expression as a predictor of colon cancer prognosis: A machine learning practice** Jiannan Liu, Chuanpeng Dong, Guanglong Jiang, Xiaoyu Lu, Yunlong Liu and Huanmei Wu
4:50-5:10	**SMaSH: Sample Matching using SNPs in Humans** Maximillian Westphal, David Frankhouser, Carmine Sonzone, Peter G. Shields, Pearlly Yan and Ralf Bundschuh	**Long non-coding RNA expression levels modulate cell-type specific splicing patterns by altering their interaction landscape with RNA-binding proteins** Felipe Wendt Porto, Swapna Vidhur Daulatabad and Sarath Chandra Janga	**A pan-cancer study of class-3 semaphorins as therapeutic targets in cancer** Xiaoli Zhang, Brett Klamer, Jin Li, Soledad Fernandez and Lang Li
5:10-5:30	**BISR-RNAseq: An efficient and scalable RNAseq analysis workflow with interactive report generation** Venkat Sundar Gadepalli, Hatice Gulcin Ozer, Ayse Selen Yilmaz, Maciej Pietrzak and Amy Webb	**Super Clustering Approach for Fully Automated Single Particle Picking in Cryo-EM** Adil Al-Azzawi, Anes Ouadou and Jianlin Cheng	**Predicting Re-admission to Hospital for Diabetes Treatment: A Machine Learning based Solution** Satish Mahadevan Srinivasan, Yok-Fong Paat, Philmore Halls, Ruth Kalule and Thomas E. Harvey.
5:30	**Cocktail Hour (Conference Center Lobby)**		
6:30	**Shuttle Bus Transportation provided to Park of Roses**		
7:00	**BANQUET (Park of Roses- 3901 N High St, Columbus, OH 43214)**		

Tuesday, June 11th

7:30-8:40	**Registration Open** ***and Continental Breakfast***		
8:40-9:30	**Keynote Lecture (Room: Ballroom)** **Jeremy Edwards, PhD** Professor, Department of Chemistry & Chemical Biology The University of New Mexico **Title**: ***Technologies for Human Genome and Transcriptome Sequencing***		
9:30-9:40	***Break***		
9:40-10:00	**Eminent Scholar Talk (Room: Ballroom)** **Bruce Aronow, PhD** Professor, UC Department of Pediatrics Co-director, Computational Medicine Center Cincinnati Children's Hospital Medical Center **Title:** ***ToppCell: A Workbench for the Analysis, Modeling and Prediction of the Molecular Basis of Development and Function of Cells and Tissues based on Single Cell Atlas Datasets***		
10:00-10:10	***Break for parallel sessions***		
CONCURRENT SESSIONS			
Room:	**Ballroom** **Cancer Genomics** **Session Chair: 7**	**Pfahl 202** **Scientific Databases** **Session Chair: 8**	**Pfahl 302** **Computational Drug Discovery** **Session Chair: 9**
10:10-10:30	**Identify rewired pathways between primary breast cancer and liver metastatic cancer using transcriptome data** Limei Wang, Jin Li, Enze Liu, Garrett Kinnebrew, Yang Huo, Zhi Zeng, Wanli Jiang, Lijun Cheng, Hongchao Lv, Weixing Feng and Lang Li	**MIRIA: a webserver for statistical, visual and meta-analysis of RNA editing data in mammals** Xikang Feng, Zishuai Wang, Hechen Li and Shuaicheng Li	**A deep learning-based method for drug-target interaction prediction based on long short-term memory neural network** Yanbin Wang, Zhuhong You, Shan Yang, Haicheng Yi, Zhanheng Chen and Kai Zheng
10:30-10:50	**Kinetic modeling of DUSP regulation in Herceptin-resistant HER2-positive breast cancer** Petronela Buiga, Ari Elson, Lydia Tabernero and Jean-Marc Schwartz	**dbMTS: a comprehensive database of putative human microRNA target site SNVs and their functional predictions** Chang Li, Michael Swartz, Bing Yu and Xiaoming Liu	**Mining and visualizing high-order directional drug interaction effects using the FAERS database** Xiaohui Yao, Tiffany Tsang, Sara Quinney, Pengyue Zhang, Xia Ning, Lang Li and Li Shen

10:50-11:10	**Gene co-expression networks restructured by gene fusion in rhabdomyosarcoma cancers** Bryan Helm, Xiaohui Zhan, Zhi Han, Dong Ni, Jie Zhang and Kun Huang	**A harmonized neurodegenerative transcriptome database to nominate mouse models for functional follow-up and validation of Alzheimer's gene networks** Rami Al-Ouran, Ying-Wooi Wan and Zhandong Liu	**SCNrank: Spectral Clustering for Network-based target Ranking to reveal potential drug targets and its application in pancreatic ductal adenocarcinoma** Enze Liu, Xiaoqi Liu, Zhuangzhuang Zhang, Xiaolin Cheng, Murray Korc and Lijun Cheng
11:10-11:20	***Coffee/Tea Break***		
11:20-11:40	**Convolutional neural network models for cancer type prediction based on gene expression** Milad Mostavi, Yu-Chiao Chiu, Yufei Huang and Yidong Chen	**Forming Big Datasets through Latent Class Concatenation of Imperfectly Matched Databases Features** Christopher Bartlett, Brett Klamer, Steven Buyske, Stephen Petrill and William Ray	**Network as a biomarker: A novel network-based sparse Bayesian machine for pathway-driven drug response prediction** Lei Frank Huang, Hongting Liu, Yi Zheng and Richard Lu
11:40-12:00	**Integrative Network Analysis Identifies Potential Targets and Drugs for Ovarian Cancer** Tianyu Zhang, Liwei Zhang and Fuhai Li	**Challenges in proteogenomics: a comparison of analysis methods with the case study of the DREAM Proteogenomics Sub-Challenge** Tara Eicher, Andrew Patt, Esko Kautto, Raghu Machiraju, Ewy Mathe and Yan Zhang	**Computational Drug Repositioning for Precision Cancer Medicine Based on Cancer Cells Screening** Abhishek Majumdar, Shaofeng Wu and Yaoqin Lu
12:00-12:10	**A Novel Graph Regularized Non-negative Matrix Factorization based on Error Weight Matrix for High Dimensional Biomedical Data Clustering** Meijun Zhou, Xianjun Shen, Limin Yu, Xingpeng Jiang, Jincai Yang and Yujian Yang	**PATH: An interactive web platform for analysis of time-course high-dimensional genomic data** Yuping Zhang, Yang Chen and Zhengqing Ouyang	**Development of a RNA-Seq based Prognostic Signature for Colon Cancer** Bjarne Bartlett, Yong Zhu, Mark Menor, Vedbar Khadka, Jicai Zhang, Jie Zheng, Bin Jiang and Youping Deng
12:10-12:20	**Skyhawk: An Artificial Neural Network-based discriminator for reviewing clinically significant genomic variants** Ruibang Luo, Tak-Wah Lam and Michael Schatz	**LCLE: a web portal for comprehensive gene distance analysis for correlation networks in liver cancer** Xiuquan Wang, Xiaoqian Zhu, Keli Xu, Junqing Wang and Yunyun Zhou	**Machine Learning Distilled Metabolite Biomarkers for Early Stage Renal Injury** Yan Guo, Dianqian Chen, Hui Yu and Ying-Yong Zhao

12:20-1:35	***Lunch Break*** *- Boxed Lunches*
1:35-1:55	**Eminent Scholar Talk (Room: Ballroom)** **Haixu Tang, PhD** Professor of Informatics and Computing Director, Data Science Academic Programs Adjunct Professor of Biology Indiana University **Title**: ***Prediction, Searching and Clustering of Tandem Mass Spectra of Peptides***
1:55-2:10	***Award presentation (Room: Ballroom)***
2:10-2:20	***Coffee/Tea Break***

CONCURRENT SESSIONS					
Ballroom **International PI Talk** **Session Chair: 10**		**Pfahl 202** **General Bioinformatics** **Session Chair: 11**		**Pfahl 302** **Cancer and Medical Genomics** **Session Chair: 12**	
2:20-2:40	**Development of predictive models to distinguish metals from non-metal toxicants, and individual metal from one another** Zongtao Yu, Yong Zhu, Junmei Ai, Jicai Zhang, Bin Jiang, Youping Deng and Bjarne Bartlett	2:20-2:30	**On the analysis of the human immunome via an information theoretical approach** Maciej Pietrzak, Gerard Lozanski, Michael Grever, Jeffrey Jones, Leslie Andritsos, James Blachly, Kerry Rogers and Michal Seweryn.	2:20-2:30	**Rapid Evolution of Expression Levels in Hepatocellular Carcinoma** Fan Zhang and Kuo Michael D. Kuo
		2:30-2:40	**Elimination of DNase nucleotide-specific bias to enhance recognition of DNA-binding proteins** Weixing Feng, Chongchong Luo, Duojiao Chen, Weixin Xie, Ruida Cong, Chengkui Zhao, Bo He and Yunlong Liu	2:30-2:40	**Identifying Interaction Clusters for MiRNA and MRNA Pairs in TCGA Network** Xinqing Dai, Lizhong Ding, Hui Jiang, Samuel Handelman and Yongsheng Bai
2:40-3:00	**DNA methylation markers for pan-cancer prediction by deep learning** Biao Liu, Yulu Liu, Mengyao Li, Shuang Yang, Shuai Cheng Li and Xingxin Pan	2:40-2:50	**RNASeqR: an R package for automated two-group RNA-Seq analysis workflow** Kuan-Hao Chao, Yi-Wen Hsiao, Yi-Fang Lee, Chien-Yueh Lee, Liang-Chuan Lai, Mong-Hsun Tsai, Tzu-Pin Lu and Eric Y. Chuang	2:40-2:50	**Generating Simulated CGH and Sequencing Data to Assess Genomic Segmentation Algorithms** Mark Zucker and Kevin Coombes

		2:50-3:00	**The Minimum Weight Clique Partition Problem and its Application to Structural Variant Calling** Matthew Hayes and Derrick Mullins	2:50-3:00	**Mapping genes and pathways to age-associated psychological changes in humans using latent semantic analysis** Pankaj Dholaniya, Vikram Naik and Baby Kumari
3:00-3:20	**Molecular evolution of circadian clock genes in spotted gar (Lepisosteus oculatus)** Yi Sun, Chao liu, Xiaolong Liu, Xudong Pan, Moli Huang, Jian Huang, Changhong Liu, Jiguang Zhang, John H. Postlethwait and Han Wang	3:00-3:10	**GPU Empowered Pipelines for Calculating High-Dimensional Kinship Matrices and Facilitating 1D and 2D GWAS** Wenchao Zhang, Xinbin Dai, Shizhong Xu and Patrick Zhao	3:00-3:10	**Cross - species Conserved Proteins Complex Identification and Exploration of Species Functional Evolution** Xianjun Shen, Meijun Zhou, Limin Yu, Li Yi, Cuihong Wan, Xingpeng Jiang and Tingting He
3:20-3:40	**DeepShape: Estimating Isoform-Level Ribosome Abundance and Distribution with Ribo-seq data** Hongfei Cui, Hailin Hu, Jianyang Zeng and Ting Chen				
3:40	**Adjourn**				

Keynote Speaker
Elaine Mardis, PhD
Sunday, June 9, 2019
10:15 – 11:00 AM
Ballroom

Bio

Elaine Mardis, PhD is co-Executive Director of the Institute for Genomic Medicine at Nationwide Children's Hospital and the Nationwide Foundation Endowed Chair of Genomic Medicine. She also is Professor of Pediatrics at The Ohio State University College of Medicine. Dr. Mardis joined Nationwide Children's Hospital in 2016.

Educated at the University of Oklahoma with a B.S. in Zoology and a Ph.D. in Chemistry and Biochemistry, Dr. Mardis did postgraduate work in industry at BioRad Laboratories. She was a member of the faculty of Washington University School of Medicine from 1993-2016.

Dr. Mardis has authored over 350 articles in prestigious peer-reviewed journals and has written book chapters for several medical textbooks. She serves as an associate editor for three peer-reviewed journals (Disease Models and Mechanisms, Molecular Cancer Research, and Annals of Oncology) and is Editor-in-Chief of Molecular Case Studies, published by Cold Spring Harbor Press. Dr. Mardis has given lectures at scientific meetings worldwide and was awarded the Morton K Schwartz award from the American Association for Clinical Chemistry in 2016. She has been listed since 2013 as one of the most highly cited researchers in the world by Thompson Reuters. Dr. Mardis has been a member of the American Association for Cancer Research (AACR) since 2007, was the program committee chair for the 2018 AACR Annual Meeting and is the current AACR President. She was elected a Fellow of the AACR Academy in 2019.

Title: A System for Pediatric Precision Cancer Medicine

Through large-scale discovery genomics, we and others have made inroads to characterize the breadth of different genomic events leading to the development and progression of pediatric cancers, including both somatic alterations and germline predisposition. An ongoing study across the breadth of different pediatric tumor types seen in our hospital has revealed new insights and evidence supporting the clinical benefit obtained from NGS-based characterization of DNA and RNA. The implementation of our cancer protocol, the computational aspects of characterizing each patient's genomic data toward therapeutic decision-making, proper diagnosis, and prognostic evidence will be described, as will specific example cases

Keynote Speaker
Peter Karp, PhD
Monday, June 10, 2019
8:40 – 9:30 AM
Ballroom

Bio

Peter D. Karp is the director of the Bioinformatics Research Group within the Artificial Intelligence Center at SRI International. Dr. Karp has authored 170 publications in bioinformatics and computer science in areas including metabolic pathway bioinformatics, computational genomics, scientific visualization, and scientific databases. He is a Fellow of the American Association for the Advancement of Science and of the International Society for Computational Biology. He received the Ph.D. degree in Computer Science from Stanford University in 1989, and was a postdoctoral fellow at the NIH National Center for Biotechnology Information.

Title: BioCyc Tools for Metabolic Modeling and Omics Data Analysis

BioCyc (https://biocyc.org/) is an extensive web portal for microbial genomes and metabolic pathways. BioCyc contains 14,500 microbial genomes including 2,400 organisms from the Human Microbiome Project. Two BioCyc databases are noteworthy as bacterial references: the EcoCyc database for Escherichia coli K-12 has been curated from 36,000 publications, and the BsubCyc database for Bacillus subtilis 168 has been curated from 4,000 publications. In addition, the MetaCyc database is a multi-organism metabolic pathway database describing 2,700 experimentally elucidated pathways from all domains of life. The Pathway Tools software behind BioCyc provides extensive tools for computational genomics, pathway bioinformatics, regulatory bioinformatics, and omics data analysis.

This presentation will explore two aspects of Pathway Tools in detail. The software can compute a qualitative metabolic reconstruction for an organism given its sequenced genome, and can convert that reconstruction into a quantative metabolic model. Using flux-balance analysis that model can be used to predict gene essentiality and predict growth or no growth under different nutrient conditions. A bottleneck in constructing metabolic models is filling gaps in the metabolic network that arise due to incompleteness in the genome annotation. We present a new algorithm for taxonomic gap filling with significantly improved accuracy.

Our Omics Dashboard is a novel interactive tool for exploration and analysis of transcriptomics and metabolomics datasets through a hierarchy of cellular systems and subsystems. This highly visual tool enables the user to survey their data at a very high level of abstraction to view the integrated responses of a large array of cellular systems such as amino-acid biosynthesis, DNA repair, and locomotion. The user can drill down to graph data at the subsystem, pathway, and gene levels. For example, the data graph for amino-acid biosynthesis can be expanded to graphs for each individual biosynthetic pathway. The Dashboard organization is defined by Gene Ontology and by the MetaCyc pathway ontology

Keynote Speaker
T.M. Murali, PhD
Monday, June 10, 2019
2:00 – 2:50 PM
Ballroom

Bio

T. M. Murali is a professor in the Department of Computer Science at Virginia Tech. He is the associate director for the Computational Tissue Engineering interdisciplinary graduate education program. Murali's research group develops phenomenological and predictive models dealing with the function, behavior, and properties of large-scale molecular interaction networks in the cell. He received his undergraduate degree in computer science from the Indian Institute of Technology, Madras and his Sc. M. and Ph. D. degrees from Brown University. Murali is a Distinguished Scientist of the Association for Computing Machinery (ACM).

Title: Pathways on Demand: Automated Reconstruction of Human Signaling Networks

Signaling pathways are a cornerstone of systems biology. Several databases store high-quality representations of these pathways that are amenable for automated analyses. However, high-quality curation of the proteins and interactions in signaling pathways is slow and painstaking. As a result, many experimentally-detected interactions are not annotated to any pathways. A natural question that arises is whether or not it is possible to automatically leverage existing pathway annotations to identify new interactions for inclusion in a given pathway. We present a series of algorithms that can automatically reconstruct the interactions in a signaling pathway of interest and dissect crosstalk among pathways.

PathLinker, the first approach, efficiently computes multiple short paths from the receptors to transcription factors (TFs) in a pathway within a background protein interaction network. It does not require any information on which interactions are in the pathway. The second approach, called RegLinker, is applicable when some interactions in a pathway are already known. The key idea underlying this approach is the use of regular-language constraints to control the number of non-pathway interactions that are present in the computed paths. Finally, we discuss XTalk, the first path-based approach for identifying pairs of pathways that may crosstalk. XTalk starts from the biological definition of crosstalk: when the stimulation of one pathway's receptors triggers a response downstream of the transcription factors of a different pathway. By design, XTalk reports the precise network of interactions and mechanisms that support the identified crosstalk.

We evaluate each algorithm systematically and demonstrate its superiority to several state-of-the-art approaches. We discuss experimental validation of a novel pathway in Wnt/β-catenin signaling discovered by PathLinker. We use RegLinker to propose new extensions to the pathways and discuss the literature that supports the inclusion of these proteins in the pathways. These results show the broad potential of automated analysis for prioritizing proteins and interactions for experimental study and attenuating difficulties of traditional manual inquiry of signaling pathways. We conclude by presenting the challenges that persist in this field and suggest strategies for overcoming them.

Keynote Speaker
Jeremy Edwards, PhD
Tuesday, June 11, 2019
8:40 – 9:30 AM
Ballroom

Bio

Dr. Edwards has worked at the interface of biology, bioinformatics, and engineering since the beginning of my scientific career. His graduate advisor was Dr. Bernhard Palsson, and he was the first person to take genome sequence information and develop predictive mathematical models of bacterial metabolism. His research started a significant global effort and many papers from his graduate work have over 800 citations. His graduate work sparked an intense interest in genomics technology, and thus he worked with Dr. George Church at Harvard Medical School for his post-doctoral studies. Dr. Edwards has worked on the development of genome technologies since that time. Now, his laboratory is in the NCI designated Cancer Research and Treatment Center at the University of New Mexico Health Sciences Center. He has a very active group of engineers, biologists, and chemists, all working together toward the development of ultrahigh-throughput DNA sequencing technology and computational biology. The move to the UNM in 2005 was ideal since he was born and raised in Albuquerque, NM.

Title: Technologies for Human Genome and Transcriptome Sequencing

Recent advances in whole genome sequencing have significantly driven down the associated costs of sequencing and increased the throughput and availability of genetic information to the biomedical research field. Currently, massive amounts of highly accurate sequencing data can be acquired for less than $1,000 per human genome. However, despite the astonishing advances, the read length of the individual reads has only increased modestly. In this presentation, I will describe a new technology to sequence and assemble complex, repetitive regions of the genome and, that also enables haplotype-resolved whole genome sequencing, a critical element necessary to establish context in disease-association studies, need to be further developed. These efforts hold the potential for a paradigm shift in our understanding of the human genome and complex disease states.

Eminent Scholar Talk
Jeffrey Parvin, MD, PhD
Monday, June 10, 2019
9:40 – 10:00 AM
Ballroom

Bio

The Parvin lab has a history of innovation in gene expression and DNA repair studies. As a graduate student studying influenza viruses, Jeff Parvin developed the first robust in vitro transcription system for the influenza viral RNA-dependent RNA polymerase on a synthetic RNA template. As a postdoctoral fellow at the MIT Center for Cancer Research (now called the Koch Institute for Integrative Cancer Research), he characterized the function of proteins that regulate gene expression. He was an Assistant and then Associate Professor of Pathology at Harvard Medical School where he established a successful research program with multiple NIH and American Cancer Society Awards with a major focus on BRCA1 and breast cancer. Dr. Parvin discovered that transcription elongation on chromatin templates was dependent on topoisomerase activities relieving positive supercoils that accumulate in front of the elongating RNA polymerase. Dr. Parvin studied the BRCA1 protein, initially as a transcription factor but pursued other functions of BRCA1. He developed biochemical assays dependent on the BRCA1 E3 ubiquitin ligase in regulating transcription and centrosome dynamics. Dr. Parvin's group was among the first to use gene expression data to derive a network for breast cancer based on co-expression. Dr. Parvin developed innovative methods for studying the DNA repair function of BRCA1 and variants of BRCA1. Recently, Dr. Parvin published the first article in a multi-year effort to establish high-throughput functional assays for BRCA1 and BARD1 DNA repair factors. Dr. Parvin joined The Ohio State University Comprehensive Cancer Center (OSUCCC) and department of Biomedical Informatics in 2007. Dr. Parvin is a Director of the Biomedical Sciences Graduate Program, the umbrella graduate program of the OSU College of Medicine, and he is the Associate Dean for Graduate Studies.

Title: The impact of sequence variants on protein function

The sequencing revolution has uncovered a high degree of variation in the genome, and the functional impact of this variation is mostly unknown. When changes of the sequence of a protein-coding gene disrupt the reading frame, truncating the encoded protein, it is usually clear that they would harm gene function. However, sequence changes often do not disrupt the reading fame but result in missense substitutions, changing the amino acid sequence of the encoded protein. The change that gives rise to the missense variants is often subtle and its impact difficult to predict based on sequence information alone. To understand the impact of these variants we need experimental functional data. Using the breast and ovarian cancer tumor suppressor gene, BRCA1, as a test case, we have developed a high-throughput functional assay for analyzing specific missense variants in the BRCA1 protein for function in DNA repair. Using this multiplexed approach, we have analyzed over 1,000 missense variants from the amino terminus of the BRCA1 protein. Our results are, for the most part, consistent with other assays for functional impact. Comparing our results against variants with known effects on cancer predisposition, pathogenic mutations were consistently loss-of-function, and benign variants were functional for DNA repair. We anticipate that this assay can be used to functionally characterize BRCA1 missense variants at scale and provide information for variants that would otherwise have an unknown effect on disease predisposition.

Eminent Scholar Talk
Alla Karnovsky, PhD
Monday, June 10, 2019
1:30 – 1:50 PM
Ballroom

Bio

Dr. Karnovsky got her Ph.D. in cell and developmental biology from the Russian Academy of Sciences. She did her postdoctoral work at the University of Colorado at Boulder, followed by the nine years of bioinformatics work in Pharmaceutical industry. In 2007 she returned to academia. Currently she is a Research Associate Professor of Computational Medicine and Bioinformatics at the University of Michigan. Her research interests involve the analysis of high throughput omics data, focusing primarily on metabolomics, and the development of computational methods and tools for the analysis and integration of metabolomics data with other types of genomic data.

Title: Identifying biologically relevant modules in metabolomics and lipidomics data with Differential Network-based Enrichment Analysis (DNEA)

Metabolomics and lipidomics datasets are becoming increasingly large and complex, requiring powerful statistical and bioinformatics tools. A common approach to interpreting the results of metabolomics and lipidomics experiments is to map and visualize experimentally measured metabolites in the context of known biochemical pathways. A number of tools for performing this type of analysis have been developed including our tool Metscape (http://metscape.med.umich.edu/). Some of the existing tools have adopted Functional Enrichment Testing methods developed for gene expression data for the analysis of metabolomics data. However, the scope of their application has been limited to known compounds from large, well-annotated pathways, which are often occupied by a small portion of the experimentally measured metabolome.

An alternative to knowledge-based data analysis is to infer meaningful associations between metabolites/lipids from experimental data and build data-driven metabolic networks to help generate biological insights. We developed a new Differential Network Enrichment Analysis method (DNEA) that uses joint structural sparsity estimation to build partial correlation networks from the data (for two or more experimental conditions), performs consensus clustering to identify highly connected network components (subnetworks), and uses Network-based Gene Set Analysis (NetGSA) to identify the differentially enriched subnetworks. We used DNEA to analyze a number of publically available metabolomics and lipidomics datasets from a variety of diseases and found that DNEA can help identify alterations in both network structure and expression levels of interacting biomolecules that impact disease phenotypes.

Eminent Scholar Talk
Bruce Aronow, PhD
Tuesday, June 11, 2019
9:40 – 10:00 AM
Ballroom

Bio

Bruce Aronow is the John J Hutton MD Professor of Biomedical Informatics and Pediatrics at the Cincinnati Children's Hospital Research Foundation and University of Cincinnati where he has been since doing his post-doctoral fellowship there starting in 1985 and receiving tenure in 1995. Bruce grew up in the Stanford University environment where his father was a founding member of its School of Medicine. His training is in Chemistry, Physics, Biology, Genetics and Medicine, and has for many years been dedicated to the advancement of Computational Biology and its application to all fields of Biomedical Research.

Title: ToppCell: A Workbench for the Analysis, Modeling and Prediction of the Molecular Basis of Development and Function of Cells and Tissues based on Single Cell Atlas Datasets

Alexis Mitelpunkt, Jake Wang, Kang Jin, Balaji Iyer, Saif Alimohamed, Scott Tabar, Eric Bardes, Bruce Aronow
[1]Cincinnati Children's Hospital Medical Center, [2]University of Cincinnati Electrical Engineering and Computer Science, Cincinnati, OH;

Single Cell Atlases have been proposed as a common platform to address fundamental challenges of biology and medicine such as understanding the molecular and cellular basis of differentiation, organogenesis, and physiology as well as disease processes that subvert these. To improve our ability to learn from single cell atlas data, for example about how diverse tissues achieve specific physiological functions or affected by disease, we are developing methods to represent tissues as ensembles of cell types and states, each of which exhibit a specific gene expression pattern (and or other molecular cellular characteristic). Following normalization and clustering, gene signatures from differential expression analyses are used to derive sets of gene modules from each dataset that define celltype-specific signatures that can be further contextualized based on metadata attributes of the sample of origin, technology, etc. as well other differences of which can be used to further define and distinguish the classes and subclasses of celltypes present in each dataset. Doing this has allowed us to construct an open access web database ToppCell (http://toppcell.cchmc.org/) that allows single cell genomic datasets to be explored, mined, and compared to reference geneset associations with respect to functional and structural gene and protein features, as well as correlation or overlaps with gene signatures from other similar cells, tissues, or perturbations. The ToppCell portal provides access to these signatures that can then be used as computable gene modules that can help unravel mechanisms of cell subtype function. ToppCell allows individual modules or parent-child module pairs to be used to create networks of multiple genes and pathways that can explain the shared functional associations of a high fraction of a given cell type's top overexpressed genes (eg 40-60% of top 400 genes can be connected typically to 4-8 high level lineage and cell subtype-specific functions). Additional genes of each functional network as well as connections between networks can be inferred (or hypothesized) using machine learning-based Fuzzy multidimensional functional or structural associations as well as Social Network-like Page Rank

algorithms. In addition, intercellular signaling or shared functional tasks that are accomplished between lineages, can also be can also be analyzed by a new modified version of toppCluster[1] that has been adapted to identify candidate tissue level intercellular networks based on known protein-protein interactions of surface or secreted proteins from each lineage. This novel ability enables exploration of targeted pathway and networks as well as discovery of underlying differentiation processes and identification of previously unknown cell subtypes. Examples to be shown will include the development and maturation of the brain and lung

Eminent Scholar Talk
Haixu Tang, PhD
Tuesday, June 11, 2019
1:35 – 1:55 PM
Ballroom

Bio

Dr. Haixu Tang is a Professor and a Grant Thornton Scholar in the Department of Computer Science and the Director of Data Science Academic Programs in the School of Informatics, Computing and Engineering at Indiana University. He received his Ph.D from Shanghai Institute of Biochemistry, Chinese Academy of Sciences in 1998, and conducted post-doctoral research in University of Southern California and University of California, San Diego before joining Indiana University in 2004, where he was promoted to Professor in 2015. He was a recipient of the NSF CAREER Award in 2007 and the Outstanding Junior Faculty Award from Indiana University in 2009. His is interested in algorithmic and statistical problems arising in bioinformatics, in particular in genomics and proteomics.

Title: Prediction, Searching and Clustering of Tandem Mass Spectra of Peptides

The availability of the large number of MS datasets (sometimes collected along with genomic/transcriptomic data on matched samples) poses great opportunities for bioinformaticians. We are developing efficient computational tools to address the Clustering, Searching and Prediction of peptide tandem mass spectra for mining the massive MS/MS datasets. We are developing machine learning models for full-spectrum prediction of peptide MS/MS spectra solely from their peptide sequences without presumed fragmentation rules. Furthermore, to speed up the spectra clustering and searching process, we propose a series of algorithmic techniques, including a run-length encoding (RLE) scheme to compress MS/MS spectra and the locality-sensitive hashing (LSH) technique for indexing the spectra. Combining these methods, computational interpretation of high-throughput proteomic data will become more accurate and efficient.

Concurrent-Workshops
Sunday, June 9, 2019
11:20 – 1:30 PM
Ballroom

Data Driven Cancer Research: Data Science Research and Applications

Dr. Wenjin Zheng, The University of Texas Health Science Center at Houston
Dr. Yidong Chen, The University of Texas Health Science Center at Houston
Dr. Yufei Huang, The University of Texas Health Science Center at San Antonio

The vast amount of available data has shifted the paradigm of cancer research from data generation to data analysis (Zhu and Zheng 2018). In this workshop, we will discuss about data drive cancer research at all scales ranging from electronic health record mining to genomic data analysis, and provide a tutorial on deep learning for cancer genomics (Zhang *et al*, 2019).

In this first session, we will discuss about how to build infrastructure to support data drive cancer research. We will use live research projects to explain different data drive research models, including an ongoing deep learning based approach to help with clinical decision support. We will also present three data driven projects on harnessing open genomic data and artificial intelligence to discover new cancer therapeutics; predicting synergistic drug combination by integrating multi-omics data in deep learning models; and predicting N6-methyladenosine (m^6A) disease association using deep learning.

Concurrent-Workshops
Sunday, June 9, 2019
11:20 – 1:30 PM
Pfahl 302

Machine Learning Demystified

Dr. Yan Guo, The University of New Mexico

Abstract

The concept of machine learning has existed for decades. With the blooming of high throughput genomic technology, machine learning methods have been frequently applied to high throughput genomic data to assist biological researches. In this workshop, we will introduce the concept of machine learning and how it can be used to support biomedical researches. The following traditional machine learning methods will be briefly discussed: hieratical clustering, principal component analysis, decision tree, and random forest. Neural networks based deep learning concept will also be introduced.

Concurrent-Workshops
Sunday, June 9, 2019
1:45 – 3:45 PM
Ballroom

Data Driven Cancer Research: Tutorial on Deep Learning for Cancer Genomics

Dr. Wenjin Zheng, The University of Texas Health Science Center at Houston
Dr. Yidong Chen, The University of Texas Health Science Center at Houston
Dr. Yufei Huang, The University of Texas Health Science Center at San Antonio

The vast amount of available data has shifted the paradigm of cancer research from data generation to data analysis (Zhu and Zheng 2018). In this workshop, we will discuss about data drive cancer research at all scales ranging from electronic health record mining to genomic data analysis, and provide a tutorial on deep learning for cancer genomics (Zhang *et al*, 2019).

In this section, we will provide a comprehensive survey on how deep learning models are built for "omics" data analysis and drug response prediction. The goal is to educate audience about the deep learning basics and its applications to genomics data so that they can apply them in cancer genomics research. The first part of the tutorial will cover the deep learning basics and important deep learning models. Next, we will survey existing work on deep learning models for genomics. In the third part, we will focus on discussing the deep learning models for cancer type classification and drug response prediction.

Concurrent-Workshops
Sunday, June 9, 2019
1:45 – 3:45 PM
Pfahl 202

BioCyc Microbial Genomes Web Portal

Dr. Peter Karp, SRI International

Abstract

BioCyc (https://biocyc.org/) is an extensive user-friendly genome informatics portal containing 14,700 microbial genomes. BioCyc combines computational inferences (such as predicted metabolic pathways and operons) with information from multiple microbial databases, and with literature curated information, and offers a large suite of bioinformatics tools. This 90-minute three-part tutorial will show you how to use many aspects of BioCyc.

Introduction to BioCyc
o Database selection
o Search operations
o Gene, metabolite, reaction, and pathway pages
o BLAST search, sequence pattern search, extracting sequences

Transcriptomics and Metabolomics data analysis
o Omics Dashboard
o Metabolic network browser and omics viewer
o Displaying high-throughput data on individual pathways and metabolic network diagrams
o Pathway collages
o Pathway covering

SmartTables and comparative analysis
o Using SmartTables to store, share, and analyze lists of genes and metabolites
o Multi-organism search
o Orthologs in BioCyc
o Genome browser and comparative genome browser
o Comparative analysis tables

Concurrent-Workshops
Sunday, June 9, 2019
1:45 – 3:45 PM
Pfahl 302

Machine learning for epigenomics data integration and gene regulation

Dr. Jianrong Wang, Michigan State University

Abstract

Machine learning is becoming the driving force for efficient and robust discoveries in biological and biomedical research, given the vast amount of functional genomics and epigenomics datasets generated by high-throughput techniques. These heterogeneous panels of epigenomics data provide the unique opportunity to systematically annotate the functional roles of specific genomic locations in diverse cellular contexts, differentiation stages and environmental conditions, which is a critical step to decode gene regulation systems. In this workshop, we will introduce a suite of machine learning algorithms and associated software that can integrate high-dimensional epigenomics data, along with transcriptomics and genomics information, in order to 1) predict different families of regulatory elements, 2) segment large-scale chromatin domains, 3) identify target genes regulated by distal regulatory elements, and 4) prioritize the genetic variants that may have causal effects on epigenetics and gene regulation. The introductions will include brief overview of the mathematical foundations for different algorithms, comparative discussion of software performance, and examples of biological applications.

An ancestral informative marker panel design for individual ancestry estimation of Hispanic population using whole exome sequencing data

Li-Ju Wang[1], Catherine W. Zhang[1], Sophia C. Su[1], Hung-I H. Chen[1], Yu-Chiao Chiu[1], Zhao Lai[1,2], Hakim Bouamar[3], Francisco G. Cigarroa[4], Lu-Zhe Sun[3], Yidong Chen[1,5§]

[1]Greehey Children's Cancer Research Institute,

[2]Department of Molecular Medicine,

[3]Department of Cell Systems and Anatomy,

[4]Department of Surgery, and

[5]Department of Epidemiology and Biostatistics, University of Texas Health Science Center at San Antonio, San Antonio, TX 78229 14

[§]Corresponding Author

Background

Europeans and American Indians were major genetic ancestry of Hispanics in the U.S. In those ancestral groups, it has markedly different incidence rates and outcomes in many types of cancers. Therefore, the genetic admixture may cause biased genetic association study with cancer susceptibility variants specifically in Hispanics. The incidence rate and genetic mutational pattern of liver cancer have been shown substantial disparity between Hispanic, Asian and non-Hispanic white populations. Currently, ancestry informative marker (AIM) panels have been widely utilized with up to a few hundred ancestry-informative single nucleotide polymorphisms (SNPs) to infer ancestry admixture. Notably, current available AIMs are predominantly located in intron and intergenic regions, while the whole exome sequencing (WES) protocols commonly used in translational research and clinical practice do not contain these markers, thus, the challenge to accurately determine a patient's admixture proportion without subject to additional DNA testing.

Results

Here we designed a bioinformatics pipeline to obtain an AIM panel. The panel infers 3-way genetic admixture from three distinct continental populations (African (AFR), European (EUR), and East Asian (EAS)) constraint within evolutionary-conserved exome regions. Briefly, we extract ~1 million exonic SNPs from all individuals of three populations in the 1000 Genomes Project. Then, the SNPs were trimmed by their linkage disequilibrium (LD), restricted to biallelic variants only, and then assembled as an AIM panel with the top ancestral informativeness statistics based on the In-statistic. In the three training populations, the optimally selected AIM panel with 250 markers, or the UT-AIM250 panel, was performed with better accuracy as one of the published AIM panels when we tested with 3 ancestral populations (Accuracy: 0.995 ± 0.012 for AFR, 0.997 ± 0.007 for EUR, and 0.994 ± 0.012 for EAS). We demonstrated the utility of UT-AIM250 panel on the admixed

American (AMR) of the 1000 Genomes Project and obtained similar results (AFR: 0.085± 0.098; EUR: 0.665 ± 0.182; and EAS 0.250 ± 0.205) to previously published AIM panels (Phillips-AIM34: AFR: 0.096 ± 0.127, EUR: 0.575 ± 0.29; and EAS: 0.330 ± 0.315; Wei-AIM278: AFR: 0.070 ± 0.096, EUR: 0.537 ± 0.267, and EAS: 0.393 ± 0.300) with no significant difference (Pearson correlation, P < 10-50, n = 347 samples). Subsequently, we applied UT-AIM250 panel to clinical datasets of self-reported Hispanic patients in South Texas with hepatocellular carcinoma (26 patients). Our estimated admixture proportions from adjacent non-cancer liver tissue data of Hispanics in South Texas is (AFR: 0.065 ± 0.043; EUR: 0.594 ± 0.150; and EAS: 0.341 ± 0.160), with smaller variation due to its unique Texan/Mexican American population in South Texas. Similar admixture proportion from the corresponding tumor tissue we also obtained.

Conclusions

Taken together, we demonstrated the feasibility of using evolutionary-conserved exome regions to distinguish genetic ancestry descendants based on 3 continental-ancestry proportion, provided a robust and reliable control for sample collection or patient stratification for genetic analysis.

---

normGAM: An R package to remove systematic biases in genome architecture mapping data

Tong Liu[1] and Zheng Wang[1,*]

[1] Department of Computer Science, University of Miami, 1365 Memorial Drive, P.O. Box 248154, Coral Gables, FL, 33124, USA

* To whom correspondence should be addressed. Tel: +1 (305) 284-3642; Fax: (305) 284-2264; Email: zheng.wang@miami.edu
TL: tong.liu@miami.edu
ZW: zheng.wang@miami.edu

Abstract

Background: The genome architecture mapping (GAM) technique can capture genome-wide chromatin interactions. However, besides the known systematic biases in the raw GAM data, we have found a new type of systematic bias. It is necessary to develop and evaluate effective normalization methods to remove all systematic biases in the raw GAM data.

Results: We have detected a new type of systematic bias, the fragment length bias, in the genome architecture mapping (GAM) data, which is different from the bias of window detection frequency previously mentioned in the paper introducing the GAM method but is similar to the bias of distances between restriction sites existing in raw Hi-C data. We have found that the normalization method (a normalized variant of the linkage disequilibrium) used in the GAM paper is not able to effectively eliminate the new fragment length bias at 1 Mb resolution (slightly better at 30 kb resolution). We have developed an R package named normGAM for eliminating the new fragment length bias and the other three biases existing in raw GAM data, which are the biases related to window detection frequency, mappability, and

GC content. Five normalization methods have been implemented and included in the R package including Knight-Ruiz 2-norm (KR2, newly designed by us), normalized linkage disequilibrium (NLD), vanilla coverage (VC), sequential component normalization (SCN), and iterative correction and eigenvector decomposition (ICE).

Conclusions: Based on our evaluations, the five normalization methods can eliminate the four biases in raw GAM data, with VC and KR2 performing better than the others. We have observed that the KR2-normalized GAM data have a higher correlation with the KR-normalized Hi-C data on the same cell samples indicating that the KR-related methods are better than the others for keeping the consistency between the GAM and Hi-C experiments. Compared with the raw GAM data, the normalized GAM data are more consistent with the normalized distances from the fluorescence in situ hybridization (FISH) experiment. The source code of normGAM can be freely downloaded from http://dna.cs.miami.edu/normGAM/.

_

Sparse Convolutional Denoising Autoencoders for Genotype Imputation

Junjie Chen and Xinghua Shi[*]

Abstract

Background: Genotype imputation is an essential tool in genomic analysis where missing data can be computationally imputed to improve the power and performance of various types of genomic analysis ranging from genome wide associations to phenotype prediction. Traditional genotype imputation methods typically require a reference panel with densely genotyped markers in populations of samples and are usually computationally expensive. Hence, genome imputation is methodologically challenging for those species without high resolution reference genomes and computationally challenging for those with large genomes. Deep learning based methods have been recently reported to nicely address the missing data problems in various fields. To explore the performance of deep learning for genotype imputation, in this study we propose a deep learning model called a Sparse Convolutional Denoising Autoencoder (SCDA) to impute missing genotypes.

Results: We developed the SCDA model and optimized its architecture by learning on a yeast genotype datset. To comprehensively evaluate the performance of the SCDA model, we then simulated three missing scenarios for genotype imputation based on the yeast data, and run 10 times repeats for each imputation method on every missing scenario. Our results showed that SCDA achieved average of accuracy of 0.9976 and standard deviation of 1.1E-4 in average results of all missing scenarios, significantly outperformed three popular imputation methods based on statistical inference including mean value, earest neighbors and singular-value decomposition.

Conclusions: SCDA outperforms other statistical inference methods for imputing missing data in a genoype matrix, and shows strong robustness on different missing scenarios. Such performance is benefit from convolutional layer that can extract various LD patterns in the genotype and spare weight matrix that is a result of L1 regularization. This study thus points to

another novel application of deep learning models in missing data imputation in genomic studies.

Keywords: genotype imputation; convolutional neural network; autoencoder; deep learning; genomic analysis.

_

High dimensional model representation of log likelihood ratio: Binary classification with SNP data

Ali Foroughi pour[1,2], Maciej Pietrzak[3,5], Lara E. Sucheston-Campbell[6], Ezgi Karaesmen[6], Lori A. Dalton[1] and Grzegorz A. Rempa-la[3,4*]

Abstract

Developing binary classification rules given SNP observations has been a major challenge for many modern bioinformatics, e.g., predicting risk of a future disease event, applications when studying complex diseases such as cancer. Small-sample high-dimensional nature of data, the large number of potentially disease associated SNPs, weak effect of each SNP on the outcome, i.e., class labels, and highly non-linear SNP interactions are several key factors exacerbating developing prediction rules with high accuracy for SNP data. Additionally, when reported in dosage, SNPs take a finite number of integer values which may be best understood as linket or categorical variables; however many classification rules treat them as real numbers, which may result in imposing certain properties on the dataset. To address these issues, we use the theory of high dimensional model representation to build appropriate low dimensional glass-box models computing the log-likelihood ratios given categorical observations while accounting for the effects of single SNPs and pairwise SNP interactions. Additionally, the theory can be used to detect significant pairwise SNP interactions. We apply the developed classifier to a synthetic data generated using the HAPGEN2 project. We observed the proposed classifier enjoys superior accuracy compared with many popular algorithms used for SNP classification, such generalized linear models.

Additionally, we detect many pairwise SNP interactions across SNPs that affect the log likelihood ratio, i.e., significantly affect the risk of being associated with the "high risk" class. The results suggest the proposed method might be an interesting approach to analyzing SNP data, in particular studying pairwise SNP interactions

_

Decoding regulatory structures and features from epigenomics profiles: a Roadmap-ENCODE Variational Auto-Encoder (RE-VAE) model

Ruifeng Hu1, Guangsheng Pei1, Peilin Jia1,*, Zhongming Zhao1, 2, 3,*
1Center for Precision Health, School of Biomedical Informatics, The University of Texas Health
Science Center at Houston, Houston, TX 77030, USA
2Human Genetics Center, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA

3Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN
37203, USA

Abstract

Background: The development of chromatin immunoprecipitation (ChIP) with massively parallel DNA sequencing (ChIP-seq) technologies has promoted generation of large-scale epigenomics data, providing us unprecedented opportunities to explore the landscape of epigenomic profiles at scales across both histone marks and tissue types. In addition to many tools directly for data analysis, advanced computational approaches, such as deep learning, have recently become promising to deeply mine the data structures and identify important regulators from complex functional genomics data.

Methods: We implemented a neural network framework, a Variational Auto-Encoder (VAE) model, to explore the epigenomic data from the Roadmap Epigenomics Project and the Encyclopedia of DNA Elements (ENCODE) project. Our model is applied to 935 reference samples, covering 28 tissues and 12 histone marks. We used the enhancer and promoter regions as the annotation features and ChIP-seq signal values in these regions as the feature values. Through a parameter sweep process, we identified the suitable hyperparameter values and built a VAE model to represent the epigenomics data and to further explore the biological regulation.

Results: The resultant Roadmap-ENCODE VAE (RE-VAE) model contained data compression and feature representation. Using the compressed data in the latent space, we found that the majority of histone marks were well clustered but not for tissues or cell types. Tissue or cell specificity was observed only in some histone marks (e.g., H3K4me1 and H3K27ac) and could be characterized when the number of tissue samples is large (e.g., blood and brain). In blood, the contributive regions and genes identified by RE-VAE model were confirmed by tissue-specificity enrichment analysis with an independent tissue expression panel. Finally, we demonstrated that RE-VAE model could detect cancer cell lines with similar epigenomics profiles.

Conclusion: We introduced and implemented a VAE model to represent large-scale epigenomics data. The model could be used to explore classifications of histone modifications and tissue/cell specificity and to classify new data with unknown sources.

_

A unified STR profiling system across multiple species with whole genome sequencing data

Yilin Liu [1y], Jiao Xu[1y] and Shuaicheng Li[2*]

Abstract

Background: Short tandem repeats (STRs) are adopted as genetic markers in forensic scenes, due to their high polymorphism in eukaryotic genomes. A variety of STRs profiling systems have been developed for species including human, dog, cat, cattle, *etc*. To maintain these systems simultaneously can be costly. These mammals share many high similar regions along their genomes. With the availability of the huge amount of the whole genomics data from these species, it is possible to develop a unified STR profiling system. In this study, our objective is to propose and develop a unified set of STR loci that could be simultaneously applied to multiple species.

Result: To find the unified STR set, we collected the whole genome sequence data of the concerned species and mapped them to the human genome reference (hg19) [1]. Then we expected the STR loci acrosses the species. From these loci, we proposed an algorithm which selected a subset of loci by incorporating the optimized combined power of discrimination. Our results show that the unified set of loci have high combined power of discrimination ($> 1\ 10^{-9}$) for both individual species and the mixed population, as well as the random-match probability, less than 10^{-7} for all the involved species, indicating that the identified set of STR loci could be applied to multiple species.

Conclusions: For the forensic scenes, a unified STR profiling system is possible. The system can be applied to the individual identification or paternal test of each of the ten common species which are *Sus scrofa* (pig), *Bos taurus* (cattle), *Capra hircus* (goat), *Equus caballus* (horse), *Canis lupus familiaris* (dog), *Felis catus* (cat), *Ovis aries* (sheep), *Oryctolagus cuniculus* (rabbit), and *Bos grunniens* (yak), and *Homo sapiens* (human). To generate such a unified STR profiling system, a loci selection algorithm using a greedy strategy is proposed here, which can be applied under different forensic parameters and on the different number of species.

Concurrent Session - General Genomics
Monday, June 10, 2019
10:10 AM – 12:20 PM
Pfahl 202

Association Analysis of Common and Rare SNVs using Adaptive Fisher Method to Detect Dense and Sparse Signals

Xiaoyu Cai1, Lo-Bin Chang1, and Chi Song_2

1Department of Statistics, The Ohio State University
2College of Public Health, Division of Biostatistics, The Ohio StateUniversity

Correspondence to: Chi Song, College of Public Health, Division of Biostatistics, The Ohio State University, 1841 Neil Ave., 208E Cunz Hall, Columbus, OH 43210. E-mail:song.1188@osu.edu

Abstract

The development of next generation sequencing (NGS) technology and genotype imputation methods enabled researchers to measure both common and rare variants in genome-wide association studies (GWAS). Statistical methods have been proposed to test a set of genomic variants together to detect if any of them is associated with the phenotype or disease. In practice, within the set of variants, there is an unknown proportion of variants truly causal or associated with the disease. Because most developed methods are sensitive to either the dense scenario, where a large proportion of the variants are associated, or the sparse scenario, where only a small proportion of the variants are associated, there is a demand of statistical methods with high power in both scenarios. In this paper, we propose a new association test (weighted Adaptive
Fisher, wAF) that can adapt to both the dense and sparse scenario by adding weights to the Adaptive Fisher (AF) method we developed before. Using both simulation and the Genome-Wide Association Study of Schizophrenia data, we have shown that the new method enjoys comparable or better power to popular methods such as sequence kernel association test (SKAT and SKAT-O) and adaptive SPU (aSPU) test.

_

An integrative, genomic, transcriptomic and network-assisted study to identify genes associated with human cleft lip with or without cleft palate

Fangfang Yan[1,#], Yulin Dai[1,#], Junichi Iwata[2, 3], Zhongming Zhao[1, 4, 5,*], Peilin Jia[1,*]

[1]Center for Precision Health, School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA

[2]Department of Diagnostic and Biomedical Sciences, School of Dentistry, The University of Texas Health Science Center at Houston, Houston, TX 77054, USA

[3]Center for Craniofacial Research, The University of Texas Health Science Center at Houston, Houston, TX 77054, 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

# These authors contributed equally.

* To whom correspondence should be addressed:
Zhongming Zhao, Ph.D.
Center for Precision Health School of Biomedical Informatics The University of Texas Health Science Center at Houston 7000 Fannin St. Suite 600, Houston, TX 77030
Phone: 713-500-3631
Email: zhongming.zhao@uth.tmc.edu

Peilin Jia, Ph.D.
Center for Precision Health School of Biomedical Informatics The University of Texas Health Science Center at Houston 7000 Fannin St. Suite 600, Houston, TX 77030
Email:peilin.jia@uth.tmc.edu

Email addresses:
FY: Fanfang.Yan.1@uth.tmc.edu YD: Yulin.Dai@uth.tmc.edu
JI: Junichi.Iwata@uth.tmc.edu
ZZ: Zhongming.Zhao@uth.tmc.edu PJ: Peilin.Jia@uth.tmc.edu

Abstract

Background: Cleft lip with or without cleft palate (CL/P) is one of the most common congenital human birth defects. A combination of genetic and epidemiology studies has contributed to a better knowledge of CL/P-associated candidate genes and environmental risk factors. However, the etiology of CL/P remains not fully understood. In this study, to identify new CL/P-associated genes, we conducted an integrative analysis using our in-house network tools, dmGWAS and EW_dmGWAS, in a combination with Genome-Wide Association Study (GWAS) data, the human protein-protein interaction (PPI) network, and differential gene expression profiles.

Methods: In total, 1,907 case-parent trios, including 878 trios from European ancestry and 1,029 trios from Asian ancestry, were downloaded from the Genotype and Phenotype database (dbGaP) that contained CL/P cases from different populations. Gene-based p-values were calculated by using the Pathway scoring algorithm (Pascal). dmGWAS and EW_dmGWAS were applied to identify dense modules from the human PPI network. Enrichment analysis was conducted using WebGestalt.

Results: A total of 87 genes were consistently detected in both European and Asian ancestries in dmGWAS. There were 31.0% (27/87) showed nominal significance with CL/P (gene-based $p < 0.05$), with three genes showing strong association signals, including *KIAA1598*, *GPR183*, and *ZMYND11* ($p < 1□10^{-3}$). In EW_dmGWAS, we identified 253 and 245 module genes associated with CL/P for European ancestry and the Asian ancestry, respectively. Functional enrichment analysis demonstrated that these genes were involved in cell adhesion, protein localization to the plasma membrane, the regulation of the apoptotic signaling pathway, and other pathological conditions. A small proportion of genes (5.1% for

European ancestry; 2.4% for Asian ancestry) had prior evidence in CL/P as annotated in CleftGeneDB database. Our analysis highlighted nine novel CL/P candidate genes (*BRD1*, *CREBBP*, *CSK*, *DNM1L, LOR*, *PTPN18*, *SND1, TGS1*, and *VIM*) and 17 previously reported genes in the top modules.

Conclusion: The genes identified through superimposing GWAS signals and differential gene expression profiles onto human PPI network greatly advances our understanding of the etiology of CL/P. We predicted nine novel CL/P candidate genes through the integrative and multi-omics analyses.

Keywords: cleft lip, cleft palate, dense module search, genome-wide association studies (GWAS), network

---

Association between ALS and retroviruses: Evidence from bioinformatics analysis

Jon P. Klein[1]; Zhifu Sun[2]; Nathan P. Staff[1]

1. Department of Neurology, Mayo Clinic, Rochester, MN
2. Department of Health Science Research, Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, MN. 200 First St. SW, Rochester, MN 55905
Tel. 507-284-8491 Fax 507-284-4074

Correspondence to Dr. Nathan P. Staff
Running Head: ALS retroviral pathways

Disclosures
Jon Klein reports no relevant disclosures
Zhifu Sun reports no relevant disclosures
Nathan Staff reports no relevant disclosures

Abstract

Objective: Emerging evidence suggests retroviruses play a role in the pathophysiology of amyotrophic lateral sclerosis (ALS). Specifically, activation of ancient viral genes embedded in the human genome is theorized to lead to motor neuron degeneration. We explore whether connections exist between ALS and retroviruses through protein interaction networks (PIN) and pathway analysis, and consider the potential roles in drug target discovery.

Methods: Protein database and pathway/network analytical software including Ingenuity Pathway BioProfiler, STRING, and CytoScape were utilized to identify overlapping protein interaction networks and extract core cluster(s) of retroviruses and ALS.

Results: Topological and statistical analysis of the ALS-PIN and retrovirus-PIN identified a shared, essential protein network and a core cluster with significant connections with both networks. The identified core cluster has three interleukin molecules IL10, Il-6 and IL-1B, a central apoptosis regulator TP53, and several major transcription regulators including MAPK1, ANXA5, SQSTM1, SREBF2, and FADD. Pathway enrichment analysis showed that this core cluster is associated with the glucocorticoid receptor singling and neuroinflammation signaling pathways. For confirmation purposes, we applied the same methodology to the West Nile and Polio virus, which demonstrated trivial connectivity with ALS, supporting the unique connection between ALS and retroviruses.

Conclusions: Bioinformatics analysis provides evidence to support pathological links between ALS and retroviral activation. The neuroinflammation and apoptotic regulation pathways are specifically implicated. The continuation and further analysis of large scale genome studies may prove useful in exploring genes important in retroviral activation and ALS, which may help discover new drug targets.

_

ManiNetCluster: A novel manifold learning approach to reveal the functional links between gene networks

Nam D Nguyen[1], Ian K Blaby[2*] and Daifeng Wang[3,4*]

Abstract

Background: The coordination of genomic functions is a critical and complex process across biological systems such as phenotypes or states (e.g., time, disease, organism, environmental perturbation). Understanding how the complexity of genomic function relates to these states remains a challenge. To address this, we have developed a novel computational method, ManiNetCluster, which simultaneously aligns and clusters gene networks (e.g., co-expression) to systematically reveal the links of genomic function between different conditions. Specifically, ManiNetCluster employs manifold learning to uncover and match local and non-linear structures among networks, and identifies cross-network functional links.

Results: We demonstrated that ManiNetCluster better aligns the orthologous genes from their developmental expression profiles across model organisms than state-of-the-art methods (p-value < $2.2\ 10^{-}16$). This indicates the potential non-linear interactions of evolutionarily conserved genes across species in development. Furthermore, we applied ManiNetCluster to time series transcriptome data measured in the green alga *Chlamydomonas reinhardtii* to discover the genomic functions linking various metabolic processes between the light and dark periods of a diurnally cycling culture. We identified a number of genes putatively regulating processes across each lighting regime.

Conclusions: ManiNetCluster provides a novel computational tool to uncover the genes linking various functions from different networks, providing new insight on how gene functions coordinate across different conditions. ManiNetCluster is publicly available as an R package at https://github.com/namtk/ManiNetCluster.

Keywords: Manifold Learning; Manifold Regularization; Clustering; Multiview Learning; Functional Genomics; Comparative Network Analysis; Comparative Genomics; Biofuel

_

Human protein-RNA interaction network is highly stable across vertebrates

Aarthi Ramakrishnan[1] and Sarath Chandra Janga[1,2,3,*]

[1]Department of Bio Health Informatics, School of Informatics and Computing, Indiana University Purdue University, Indianapolis, Indiana 46202, United States of America

[2]Centre for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, Indiana 46202, United States of America

[3]Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, Indiana 46202, United States of America

* Corresponding author:
Sarath Chandra Janga
E-mail: scjanga@iupui.edu
719 Indiana Avenue Ste 319, Walker Plaza Building Indianapolis, Indiana - 46202
Tel: +1-317-278-4147, Fax: +1-317-278-9201

Abstract

RNA-binding proteins (RBPs) are crucial in modulating RNA metabolism in eukaryotes thereby controlling an extensive network of RBP-RNA interactions. In this study, we employ CLIP-seq datasets for 60 human RBPs and demonstrate that most binding sites for a third of these RBPs are conserved in at least 50% of the studied vertebrate species. Across the studied RBPs, binding sites were found to exhibit a median conservation of 58%, ~20% higher than random genomic locations, suggesting a significantly higher preservation of RBP-RNA interaction networks across vertebrates. RBP binding sites were highly conserved across primates with weak conservation profiles in birds and fishes. We also note that phylogenetic relationship between members of an RBP family does not explain the extent of conservation of their binding sites across species. Multivariate analysis to uncover features contributing to differences in the extents of conservation of binding sites across RBPs revealed RBP expression level and number of post-transcriptional targets to be the most prominent factors. Examination of the location of binding sites at the gene level confirmed that binding sites occurring on the 3' region of a gene are highly conserved across species with 90% of the RBPs exhibiting a significantly higher conservation of binding sites in 3' regions of a gene than those occurring in the 5'. Gene set enrichment analysis on the extent of conservation of binding sites to identify significantly associated human phenotypes revealed an enrichment for multiple developmental abnormalities, suggestive of the importance of lineage-specific developmental events in post-transcriptional regulatory network evolution.

_

Differential co-expression analysis reveals early stage gene dis-coordination in Alzheimer's disease

Yurika Upadhyaya[1†], Linhui Xie[2†], Paul Salama[2], Sha Cao[3], Kwagnsik Nho[4], Andrew J. Saykin[4], Jingwen1,4*and For the ADNI

Abstract

Background: Alzheimer's disease (AD) is one of the leading cause of death in the US and there is no validated drugs to stop, slow or prevent AD. Despite tremendous effort on biomarker discovery, existing findings are mostly individual biomarkers and provide limited insights into the discoordination of genes underlying AD. We propose to explore the gene co-expression patterns in multiple AD stages, including cognitively normal (CN), early mild cognitive impairment (EMCI), late MCI and AD. We modified traiditonal joint graphical lasso to model our asusmption that the co-expression networks in consecutive disease stages are largely similar with critical differences. In addition, we performed subsequent network comparison analysis for identification of stage specific dis-coordination between genes. We focused our analysis on top AD-enriched pathways.

Results: The network level clustering coefficient remains stable from CN to LMCI and then decreases significantly when progressing to AD. When evaluating edge level differences, we identified eight gene modules with continuously decreasing or increasing patterns during AD progression. Five of them shows significant changes from CN to EMCI and thus have the potential to serve system biomarkers for early screening of AD.

Keywords: gene co-expression network; stage-specific co-expression changes; Alzheimer's disease

Concurrent Sessions-Bioinformatics
Monday, June 10, 2019
10:10 AM – 12:20 PM
Pfahl 302

The Comparisons of Prognostic Power and Expression Level of Tumor Infiltrating Leukocytes in Hepatitis B- and Hepatitis C-related Hepatocellular Carcinomas

Yi-Wen Hsiao[1], Lu-Ting Chiu[2], Ching-Hsuan Chen[2], Wei- Liang Shih[2], Tzu-Pin Lu[2]

Bioinformatics and Biostatistics Core Lab, Center of Genomic and Precision Medicine, National Taiwan University
Institute of Epidemiology and Preventive Medicine, Department of Public Health, National Taiwan University

Abstract

Background: Tumor infiltrating lymphocytes (TILs) are immune cells surrounding tumor cells, and several studies have shown that TILs are potential survival predictors in different cancers. However, challenge arises; few studies have been performed for dissecting the differences between hepatitis B- and hepatitis C-related hepatocellular carcinoma (B-HCC and C-HCC). Therefore, we aim to determine whether the expression levels of the TILs are potential predictors for survival outcomes in hepatocellular carcinomas and which TILs are the most significant ones. 41

Methods: Two bioinformatics algorithms including ESTIMATE and CIBERSORT were utilized to analyze the gene expression profiles from 6 datasets. The ESTIMATE algorithm examined the total expression level of the TILS whereas the CIBERSORT algorithm reported the expression levels of 22 different TILs. Both subtypes of hepatocellular carcinoma including B-HCC and C-HCC were analyzed accordingly. 47

Results: The results indicated that the total expression level of TILs was higher in the non-tumor part regardless of the HCC types. Alternatively, the significant TILs associated with survival outcome and recurrence pattern varied from subtypes to subtypes. For example, in B-HCC, plasma cells (hazard ratio [HR]=1.05; 95% CI 1.00- 52 1.10; p=0.034) and activated dendritic cells (HR=1.08; 95% CI 1.01-1.17; p=0.03) were significantly associated with the overall survival, whereas in C-HCC disease, monocyte (HR=1.13) were significantly associated with the overall survival. Furthermore, for the recurrence-free survival (RFS), CD8+ T cells (HR=0.98) and M0 macrophages (HR=1.02) were potential biomarkers in B-HCC, whereas neutrophil (HR=1.01) was an independent predictor in C-HCC. Lastly, in HCC including B-HCC and C-HCC, CD8+ T cells (HR=0.97) and activated dendritic cells (HR=1.09) have significant association with OS; gamma delta T cells (HR=1.04), monocytes (HR=1.05), M0 macrophages (HR=1.04), M1 macrophages (HR=1.02) and activated dendritic cells (HR=1.15) are highly associated with RFS. 62

Conclusions: These findings demonstrated that the TILs are potential survival predictors in HCC and different kinds of TILs are observed according to the virus types. Therefore, further investigations are warranted to elucidate the etiology of TILs in HCC, which may improve the immunotherapy outcomes

Keywords: hepatocellular carcinoma, hepatitis B virus, hepatitis C virus, tumor-infiltrating lymphocytes, immune cell, ESTIMATE, CIBERSORT 70

_

SigUNet : signal peptide recognition based on semantic segmentation

Jhe-Ming Wu1, Yu-Chen Liu2 , Darby Tien-Hao Chang1*

1 Department of Electrical Engineering, National Cheng Kung University, Tainan, Taiwan

* Corresponding author
E-mail: darby@mail.ncku.edu.tw

Abstract

Protein sorting indicates the mechanism of transporting proteins to their destination, where signal peptides play an important role. Recognition of signal peptides is an important first step to understand the active locations and functions of proteins. Many computational methods have been proposed to help signal peptide recognition. In recent years, the development of deep learning methods had a significant advance in many research fields. However, most of existing neural network models for signal peptide recognition are relatively simple without leveraging the developments of other fields.

This work proposes a convolutional neural network architecture without fully connected layers, which is an important architecture improvement in computer vision. For processing protein data, this work proposes several modifications, such as information enhancement of pooling and trainable network architecture. The experiments results that the proposed model outperforms conventional neural networks on eukaryotic signal peptides. However, the advantage is not observed on bacterial signal peptides because of the small data size. This work also discusses the model reduction and data augmentation issues to solve the problem. This work has three contributions: (1) developing an accurate signal peptide recognizer, (2) demonstrating the potential of leveraging advanced networks from other fields and (3) important modifications while using advanced networks on signal peptide recognition**.**

_

Integrated metabolomics and transcriptomics study of traditional herb *AstragalusmembranaceusBge.var.mongolicus (Bge.) Hsiao* reveals global metabolic profile and novel phytochemical ingredients

Xueting Wu[1,#], Xuetong Li[1,2,#], Wei Wang[3,#], Yuanhong Shan[1], Cuiting Wang[1], Mulan Zhu[4], Qiong La[5], Yang Zhong[3, 5], Ye Xu[6,*], Peng Nan[3,*], 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]Ministry of Education Key Laboratory for Biodiversity Science and Ecological Engineering, School of Life Sciences, Fudan University, Shanghai 200438, China

[4]Shanghai Chenshan Plant Science Research Center, Shanghai Chenshan Botanical

Garden, Shanghai 201602, China
[5]Research Institute of Biodiversity & Geobiology, Department of Life Science, Tibet University, Lhasa, China 850000, China
[6]Department of Colorectal Surgery, Fudan University Shanghai Cancer Center, Shanghai, China

*Corresponding author
Email:wuxt@sibs.ac.cn(XT.Wu),
xtli@sibs.ac.cn(XT.Li),tracy0811@vip.qq.com(W.Wang),
yhshan@sibs.ac.cn(YH.Shan),
ctwang@sibs.ac.cn(CT.Wang)mlzhu@sippe.ac.cn(ML.Zhu),
lhagchong@utibet.edu.cn(Q.La), yangzhong@fudan.edu.cn(Y.Zhong),
xu.shirley021@gmail.com(Y.Xu), nanpeng@fudan.edu.cn(P.Nan),
lixuan@sippe.ac.cn(X.Li).
#These authors contributed equally to this work

Abstract

Background: *Astragalus membranaceus Bge. var. mongolicus (Bge.) Hsiao* (AMM) is one of the most common herbs widely used in South and East Asia, to enhance people's health and reinforce vital energy. Despite its prevalence, however, the knowledge about pyhtochemical compositions and metabolite biosynthesis in AMM is very limited.

Results: An integrated metabolomics and transcriptomics analysis using state-of-the-art UPLC-Q-Orbitrap mass spectrometer and advanced bioinformatics pipeline was conducted to study global metabolic profiles and phytochemical ingredients/biosynthesis in AMM. A total 5,435 metabolites were detected in AMM, from which 2,190 were annotated, representing an order of magnitude increase over previously known. Metabolic profiling of AMM tissues found contents and synthetic enzymes for phytochemicals were significant higher in leaf and stem in general, whereas the contents of the main bioactive ingredients were significantly enriched in root, underlying the value of root in herbal remedies. Using integrated metabolomics and transcriptomics data, we illustrated the complete pathways of phenylpropanoid biosynthesis, flavonoid biosynthesis, and isoflavonoid biosynthesis, in which some are first reported in the herb. More importantly, we discovered novel flavonoid derivatives using informatics method for neutral loss scan, in addition to inferring their likely synthesis pathways in AMM.

Conclusion: The current study represents the most comprehensive metabolomics and transcriptomics analysis on traditional herb AMM. We demonstrated our integrated metabolomics and transcriptomics approach offers great potentials in discovering novel metabolite structure and associated synthesis pathways. This study provides novel insights into the phytochemical ingredients, metabolite biosynthesis, and complex metabolic network in herbs, highlighting the rich natural resource and nutritional value of traditional herbal plants.

_

BayesMetab: Treatment of Missing Values in Metabolomic Studies using a Bayesian Modeling Approach

Jasmit Shah1, Guy N. Brock2, and Jeremy Gaskins3*

1 Department of Internal Medicine, The Aga Khan University, Nairobi, Kenya 2 Department of Biomedical Informatics, The Ohio State University, Columbus, OH 43210, USA
3 Department of Bioinformatics and Biostatistics, University of Louisville, Louisville, KY 40202, USA

*Corresponding author
Email addresses:
JS: jasmit.shah@aku.edu
GNB: guy.brock@osumc.edu
JG: jeremy.gaskins@louisville.edu

Abstract

Background

With the rise of metabolomics, the development of methods to address analytical challenges in the analysis of metabolomics data is of great importance. Missing values (MVs) are pervasive, yet the treatment of MVs can have a substantial impact on downstream statistical analyses. The MVs problem in metabolomics is quite challenging and can arise because the metabolite is not biologically present in the sample, or is present in the sample but at a concentration below the lower limit of detection (LOD), or is present in the sample but undetected due to technical issues related to sample pre-processing steps. The former is considered missing not at random (MNAR) while the latter is an example of missing at random (MAR). Typically such MVs are substituted by a minimum value, which may lead to severely biased results in downstream analyses.

Methods

We develop a Bayesian model that systematically accounts for missing values based on a Markov chain Monte Carlo (MCMC) algorithm that incorporates data augmentation by allowing MVs to be due to either truncation below the LOD or other technical reasons unrelated to its abundance.. A key piece of building the model that can accommodate MV imputation is the choice of the structure of the dependence. We adapt the sparse Bayesian infinite factor model for the consideration of a flexible, lower-dimensional choice for the covariance matrix. Statistical inference may be performed using either the posterior samples of the parameters or by using the data sets imputed during MCMC.

Results

Based on a variety of performance metrics (power for detecting differential abundance, area under the curve, bias and MSE for parameter estimates) our simulation results indicate that our Bayesian method outperformed other imputation algorithms when there is a mixture of missingness due to MAR and MNAR. Further, our approach was competitive with other methods tailored specifically to MNAR in situations where missing data were completely MNAR. Applying our approach to an analysis of metabolomics data from a mouse myocardial infarction revealed several statistical significant metabolites not previously identified that were of direct biological relevance to the study.

Conclusions

Our findings demonstrate that our Bayesian method has improved performance in imputing the missing values and performing statistical inference compared to other current methods when missing values are due to a mixture of MNAR and MAR. Analysis of real metabolomics data strongly suggests this mixture is likely to occur in practice and thus it is

important to consider an imputation model that accounts for a mixture of missing data types.

_

Dense module searching for gene networks associated with multiple sclerosis

Astrid M Manuel1, Yulin Dai1, Leorah A. Freeman2, Peilin Jia1,*, Zhongming Zhao1, 3, 4,*

1Center for Precision Health, School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA
2Department of Neurology, McGovern Medical School, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA
3Human Genetics Center, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA
4Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN 37203, USA

Abstract

Background: Multiple sclerosis (MS) is a complex disease in which the immune system attacks the central nervous system. The molecular mechanisms contributing to the etiology of MS remain poorly understood. Genome-wide association studies (GWAS) of MS have identified a small number of genetic loci significant at the genome level, but they are mainly non-coding variants. Network-assisted analysis may help better interpret the functional roles of the variants with association signals.

Materials and methods: The Dense Module Searching of GWAS tool (dmGWAS version 2.4) developed in our team is applied to two MS GWAS datasets (GeneMSA and IMSGC GWAS) using the human protein interactome as the reference network. Network modules derived from independent analysis of each GWAS dataset were then conjunctively studied by the dual evaluation function implemented in this new version of dmGWAS. Top network modules were assessed by gene set enrichment analysis and potential MS drug targets.

Results: Approximately 7,500 significant network modules were identified for each independent GWAS dataset, and 20 significant modules were identified from the dual evaluation. The top modules included *GRB2, HDAC1*, *JAK2*, *MAPK1,* and *STAT3* as central genes. Top module genes were enriched with functional terms such as “regulation of glial cell differentiation” (adjusted p-value = 2.58×10-3), “T-cell costimulation” (adjusted p-value = 2.11×10-6) and “virus receptor activity” (adjusted p-value = 1.67×10-3). Interestingly, top gene networks included several MS FDA approved drug target genes *HDAC1*, *IL2RA*, *KEAP1*, and *RELA,*

Conclusions: Our dmGWAS network analyses highlighted several genes (*GRB2, HDAC1, IL2RA, JAK2, KEAP1, MAPK1, RELA* and *STAT3*) in top modules that are promising to interpret GWAS signals and link to MS drug targets. The genes enriched with glial cell differentiation are important for understanding neurodegenerative processes in MS and for remyelination therapy investigation. Importantly, our identified genetic signals enriched in T cell costimulation and viral receptor activity supported the viral infection onset hypothesis for MS.

Keywords: GWAS, multiple sclerosis, dmGWAS, network module, gene set enrichment analysis, drug target

_

Expression correlation attenuates within and between key signaling pathways in CKD progression

Hui Yu1, Danqian Chen2, Olufunmilola Oyebamiji1, Ying Yong Zhao2*, Yan Guo1*

1Department of Internal Medicine, University of New Mexico, Albuquerque, NM, 87131, USA
2Key Laboratory of Resource Biology and Biotechnology in Western China, School of Life Sciences, Northwest University, Xi'an, Shaanxi 710069, China

* Corresponding authors

Abstract: Differential coexpression represents a unique and complementary perspective into transcriptome data, which has been employed to dissect transcriptomes of various human diseases. Total RNA-seq was performed on kidney tissue samples from 140 patients with chronic kidney disease (CKD). We applied a variety of differential coexpression oriented methods to analyze the transcriptome transition in CKD from the early, mild phase to the late, severe kidney damage phase. We discovered a global expression correlation attenuation in CKD progression, with pathway *Regulation of nuclear SMAD2/3 signaling* demonstrating the most remarkable intra-pathway correlation rewiring. Moreover, the pathway *Signaling events mediated by focal adhesion kinase* displayed significantly weakened crosstalk with seven pathways, including *Regulation of nuclear SMAD2/3 signaling*. Well-known relevant genes, such as *ACTN4*, were characterized with widespread correlation disassociation with partners from a wide array of signaling pathways. Altogether, our analysis results presented an important resource of vanishing hub genes and disrupted correlations within and between key signaling pathways underlying the pathophysiological mechanisms of CKD progression.
Keywords: chronic kidney disease; differential co-expression; correlation attenuation; pathway crosstalk

Investigating Skewness to Understand Gene Expression Heterogeneity in Large Patient Cohorts

Benjamin V. Church[1,2], Henry T. Williams[1,2], Jessica C. Mar[1,3,4,*]

[1]Department of Systems and Computational Biology, Albert Einstein College of Medicine, 1300 Morris Park Avenue, Bronx, NY, 10461, USA

[2]Columbia University, New York, USA.

[3]Department of Epidemiology and Population Health, Albert Einstein College of Medicine, 1300 Morris Park Avenue, Bronx, NY, 10461, USA

[4]Australian Institute for Bioengineering and Nanotechnology, The University of Queensland, QLD 4072, Australia

*Corresponding author

BVC: bvc2105@columbia.edu HTW: htw2116@columbia.edu
JCM: j.mar@uq.edu.au

Abstract

Skewness is an under-utilized statistical measure that captures the degree of asymmetry in the distribution of any dataset. This study applied a new metric based on skewness to identify regulators or genes that have outlier expression in large patient cohorts. We investigated whether specific patterns of skewed expression were related to the enrichment of biological pathways or genomic properties like DNA methylation status. Our study used publicly available datasets that were generated using both RNA-sequencing and microarray technology platforms. For comparison, the datasets selected for this study also included different samples derived from control donors and cancer patients. When comparing the shift in expression skewness between cancer and control datasets, we observed an enrichment of pathways related to immune function that reflect increases towards positive skewness in the cancer relative to control datasets. Significant correlation was also detected between expression skewness and differential DNA methylation occurring in the promotor regions for four TCGA cancer cohorts. Our results indicate the expression skewness can reveal new insights into transcription based on outlier and asymmetrical behaviour in large patient cohorts.

Keywords:
Skewness, gene expression, non-Normality, TCGA, cancer genomics.

Clonal reconstruction from time course genomic sequencing data

Wazim Mohammed Ismail* and Haixu Tang

Abstract

Background: Bacterial cells during many replication cycles accumulate spontaneous mutations, which result in the birth of novel clones. As a result of this *clonal expansion*, an evolving bacterial population has different clonal composition over time, as revealed in the long-term evolution experiments (LTEEs). Accurately inferring the *haplotypes* of novel clones as well as the clonal frequencies and the clonal evolutionary history in a bacterial population is useful for the characterization of the evolutionary pressure on multiple correlated mutations instead of that on individual mutations.

Results: In this paper, we study the computational problem of reconstructing the haplotypes of bacterial clones from the *variant allele frequencies* observed from an evolving bacterial population at multiple time points. We formalize the problem using a maximum likelihood function, which is defined under the assumption that mutations occur spontaneously, and thus the likelihood of a mutation occurring in a specific clone is proportional to the frequency of the clone in the population when the mutation occurs. We developed a series of heuristic algorithms to address the maximum likelihood inference, and showed through simulation experiments that the algorithms are fast and achieve near optimal accuracy that is practically plausible under the maximum likelihood framework. We also validate our method using experimental data obtained from a recent study on long-term evolution of Escherichia coli.

Conclusion: We developed efficient algorithms to reconstruct the clonal evolution history from time course genomic sequencing data. Our algorithm can also incorporate clonal sequencing data to improve the reconstruction results when they are available. Based on the evaluation on both simulated and experimental sequencing data, our algorithms can achieve satisfactory results on the genome sequencing data from long-term evolution experiments.

Availability: The program (ClonalTREE) is available as open-source software on GitHub at https://github.com/COL-IU/ClonalTREE

Keywords: clonal reconstruction; time course; maximum likelihood; long-term evolution experiment

---

_

CNV detection from circulating tumor DNA in late stage non-small cell lung cancer patients

Hao Peng[1*], Qiangsheng Dai[2*], Zisong Zhou[3*], Xiaochen Zhao[4], Dadong Zhang[5], Kejun Nan[6], Zhu-An Ou[7], Fugen Li[3], Hua Dong[3#], Lei Tian[8#], Yu Yao[6#]

[1]The First People's Hospital of Yunnan Province, Yunnan, China

[2]Department of Oncology, The First Affiliated Hospital, Sun Yet-sen University, Guangdong, China

[3]The Bioinformatics Department, 3DMed Inc., Shanghai, China

[4]The Medical Department, 3DMed Inc., Shanghai, China

[5]The Translational Science Department, 3DMed Inc., Shanghai, China

[6]Department of medical Oncology, the First Affiliated Hospital of Xi'an Jiaotong University, Shanxi, China

[7]Department of Thoracic Surgery, Guangzhou General Hospital of PLA, Guangdong, China

[8]Department of Thoracic Surgery, The First Affiliated Hospital of Chongqing Medical University, Chongqing, China

Abstracts

Background: While methods for detecting SNV and indel in circulating tumor DNA (ctDNA) with hybridization capture-based next generation sequencing (NGS) has been available, algorithm for calling copy number variation(CNV) is more challenging. Here we present a method enabling CNV detection from a 150 gene panel in very low amount of ctDNA.

Results: Frist of all, a read depth based copy number estimation method without paired blood sample was developed and cfDNA sequencing data from 10healthy people was used to build a background panel of normal (PoN) model. Then in- silicon and in-vitro simulations were performed to define the limit of detection (LOD) for EGFR, ERBB2 and MET. For in-silicon spike-in, the LOD of ctDNA fraction is about 1% for EGFR (2.2 copy), 0.3% for ERBB2 (2.2 copy), 5% for MET (2.5 copy). For in-vitro spike-in, the LOD of ctDNA fraction is about 3% for EGFR (2.6 copy), 1% for ERBB2 (2.6 copy), 5% for MET (2.5 copy). Compared with 48 samples' WES CNV results, the concordance rate for EGFR, ERBB2 and MET CNV is 78%, 89.6% and 92.4%, respectively. In another large independent cohort profiled with 150 gene panel from 4009 lung cancer ctDNA sample, we detected 42 HER2 amplification (1.05%), 168 EGFR amplification (4.19%) and 63 MET amplification samples (1.57%). One lung adenocarcinoma patient with MET amplification detected by our method after acquired resistance by EGFR TKI, accepted Crizotinib and reached partial response.

Conclusions: These data support our blood based CNV detection assay was well validated in both technique and clinical perspective. It can detect CNV in low amount of ctDNA with high specificity and concordance, which enable CNV calling in a noninvasive way for cancer patients.

_

A Protocol to Evaluate RNA Sequencing Normalization Methods

Zachary B. Abrams, PhD1; Travis Johnson, MS1; Kun Huang, PhD2; Philip R.O. Payne, PhD3; Kevin Coombes, PhD1;

1. Dept. Biomedical Informatics, Ohio State University, 250 Lincoln Tower, 1800 Cannon Dr. Columbus, Ohio, 43210
2. Div. Hematology/Oncology, Dept. Medicine, Indiana University School of Medicine Indianapolis, IN 46202
3. Dept. Biomedical Informatics, Washington University 4444 Forest Park Ave, Suite 6318 Campus Box 8102 St. Louis, MO 63108-2212

Key words: RNASeq, Normalization, Standardization, Biological Variability

Abstract
Background: RNA sequencing technologies have allowed researchers to gain a better understanding of how the transcriptome affects disease. However, sequencing technologies often unintentionally introduce experimental error into RNA sequencing data. To counteract this, normalization methods are standardly applied with the intent of reducing the non-biologically derived variability inherent in transcriptomic measurements. However, the comparative efficacy of the various normalization techniques has not been tested in a standardized manner. Here we propose tests that evaluate numerous normalization techniques and applied them to a large-scale standard data set. These tests comprise a protocol that allows researchers to measure the amount of non-biological variability which is present in any data set after normalization has been performed, a crucial step to assessing the biological validity of data following normalization.

Results: In this study we present two tests to assess the validity of normalization methods applied to a large-scale data set collected for systematic evaluation purposes. We tested various RNASeq normalization procedures and concluded that transcripts per million (TPM) was the best performing normalization method based on its preservation of biological signal as compared to the other methods tested.

Conclusion: Normalization is of vital importance to accurately interpret the results of genomic and transcriptomic experiments. More work, however, needs to be performed to optimize normalization methods for RNASeq data. The present effort helps pave the way for more systematic evaluations of normalization methods across different platforms. With our proposed schema researchers can evaluate their own or future normalization methods to further improve the field of RNASeq normalization.

Fully Moderated T-statistic in Linear Modeling of Mixed Effects for Differential Expression Analysis

Lianbo Yu*, Jianying Zhang, Guy Brock and Soledad Fernandez

Abstract
Background: Gene expression profiling experiments with few replicates lead to great variability in the estimates of gene variances. Toward this end, several moderated t-test methods have been developed to reduce this variability and to increase power of tests of differential expression. Most of these moderation methods are based on linear models with fixed effects where residual variances are smoothed under a hierarchical Bayes framework. However, they are inadequate for designs with complex correlation structures, therefore application of moderation methods to linear models with mixed effects are needed for differential expression analysis.
Results: We demonstrated the implementation of the fully moderated t-statistic method for linear models with mixed effects, where both residual variances and variance estimates of random effects are smoothed under a hierarchical Bayes framework. We compared the proposed method with two current moderation methods and show that the proposed method

can control the expected number of false positives at the nominal level, while the two current moderation methods fail.

Conclusions: We proposed an approach for testing differential expression under complex correlation structures while providing variance shrinkage. The proposed method is able to improve power by moderation and control the expected number of false positives properly at the nominal level.

Keywords: Fully Moderated T-statistic; Linear Mixed-Effects Model; Variance Shrinkage; Expected Number of False Positives

_

SMaSH: Sample Matching using SNPs inHumans

Maximillian Westphal[1], David Frankhouser[2,3], Carmine Sonzone[4], Peter G. Shields[4,5,6], Pearlly Yan[5,6] and Ralf Bundschuh[1,5,7,8,9*]

Abstract

Background: Inadvertent sample swaps are a real threat to data quality in any medium to large scale omics studies. While matches between samples from the same individual can in principle be identified from a few well characterized single nucleotide polymorphisms (SNPs), omics data types often only provide low to moderate coverage, thus requiring integration of evidence from a large number of SNPs to determine if two samples derive from the same individual or not.

Results: We present a selection of about six thousand SNPs in the human genome and a Bayesian framework that is able to robustly identify sample matches between next generation sequencing data sets. We validate our approach on a variety of data sets. Most importantly, we show that our approach can establish identity between different omics data types such as Exome, RNA-Seq, and MethylCap-Seq. We demonstrate how identity detection degrades with sample quality and read coverage, but show that twenty million reads of a fairly low quality RNA-Seq sample are still sufficient for reliable sample identification.

Conclusion: Our tool, SMASH, is able to identify sample mismatches in next generation sequencing data sets between different sequencing modalities and for low quality sequencing data.

Keywords: sample swap; next generation sequencing data; identity matching

_

BISR-RNAseq: An efficient and scalable RNAseq analysis workflow with interactive report generation

Venkat Sundar Gadepalli,[1,2,3] Hatice Gulcin Ozer,[1,2,3] Ayse Selen Yilmaz,[1,2,3] Maciej Pietrzak,[1,2,3]
Amy Webb,[1,2,3]

Biomedical Informatics, The Ohio State University, Columbus, OH.
The James Comprehensive Cancer Center, The Ohio State University, Columbus, OH.
Bioinformatics Shared Resource Group, The Ohio State University, Columbus, OH.

Abstract

RNA sequencing has become an increasingly affordable way to profile gene expression patterns. Here we introduce a workflow implementing several open-source softwares that can be run on a high performance computing environment. The workflow allows for the analysis (alignment, QC, gene-wise counts generation) of raw RNAseq data and seamless integration of quality analysis and differential expression results into a configurable R shiny web application. Developed as a tool by the Bioinformatics Shared Resource Group (BISR) at the Ohio State University, we have applied the pipeline to a few publicly available RNAseq datasets downloaded from GEO in order to demonstrate the feasibility of this workflow. Source code is available here: workflow: https://github.com/MPiet11/BISR-RNAseq and shiny: https://code.bmi.osumc.edu/gadepalli.3/BISR_RNASeq_ICIBM19. Example dataset is demonstrated here: https://dataportal.bmi.osumc.edu/RNA_Seq/.

Concurrent Sessions-Bioinformatics
Monday, June 10, 2019
3:00PM – 5:20 PM
Pfahl 202

M3S: A comprehensive model selection for multi-modal single-cell RNA sequencing data

Yu Zhang[1,2*+], Changlin Wan[2,3+], Pengcheng Wang[4], Wennan Chang[2,3], Yan Huo[2,5], Jian Chen[6], Qin Ma[7] Sha Cao[2,8], Chi Zhang[2,3,9*]

[1]Colleges of Computer Science and Technology, Jilin University, Changchun,130012, China,

[2]Center for Computational Biology and Bioinformatics, [8]Department of Biostatistics, Indiana University, School of Medicine, [9]Department of Medical and Molecular Genetics, Indianapolis, IN,46202, USA.

[3]Department of Electronic Computer Engineering, Purdue University

[4]Department of Computer Science, Indiana University-Purdue University Indianapolis, Indianapolis, IN,46202, USA.

[5]Department of Computer Science, China Medical University, Shen Yang, 110001, China

[6]Shanghai pulmonary hospital, Shanghai, China, 200082

[7]Department of Biomedical Informatics, the Ohio State University, Columbus, OH, 43210, USA,

*To whom correspondence should be addressed: +1 317-278-9625; Email: czhang87@iu.edu. Correspondence is also addressed to Yu Zhang: Email: zy26@jlu.edu.cn

Abstract

Summary: Multi-Modal Model Selection (M3S) is an R package for gene-wise selection of the most proper multi-modality statistical model and downstream analysis, useful in a single-cell or large scale bulk tissue transcriptomic data. M3S is featured with (1) gene-wise

selection of the most parsimonious model among 11 most commonly utilized ones, that can best fit the expression distribution of the gene, (2) parameter estimation of a selected model, and (3) differential gene expression test based on multi-modality model.

Availability and implementation: M3S is an open source package and is available through GitHub at ps://github.com/zy26/M3S.

Multi-objective optimized fuzzy clustering for detecting cell clusters from single cell expression profiles

Saurav Mallik[1*y*] and Zhongming Zhao[1,2*y*]

Abstract

Background: Rapid advance in single cell RNA sequencing (scRNA-seq) allows to measure the expression of genes at single-cell resolution in complex diseases or tissues. Detection of correct cell clusters is currently a main challenging task for scRNA-seq analysis. Many methods have been developed to predict cell clusters from the scRNA-seq data. However, this challenge currently remains.

Methods: We proposed multi-objective optimization based fuzzy clustering approach for detecting cell clusters from the single cell expression data. We first conducted initial filtering and SCnorm normalization, respectively. We considered nine case studies by selecting different cluster numbers (cl = 2,3,...,10), and applied fuzzy c-means clustering algorithm individually. From each case, we evaluated the scores of four cluster validity index measures, Partition Entropy (*PE*), Partition Coefficient (*PC*), Modified Partition Coefficient (*MPC*) and Fuzzy Silhouette Index (*FSI*). Next, we set the first measure as minimization objective ( ) and remaining three as maximization objectives ( ), and then applied a multi-objective decision making technique, TOPSIS, to identify the best optimal solution. The best optimal solution (case study) that had the highest TOPSIS score was selected as final optimal clustering. Finally, we obtained differentially expressed genes using Limma through the comparison of expression of the samples between each resultant cluster and the remaining clusters.

Results: We applied our approach to a scRNA-seq dataset of "*Whole Organoid Replicate 1* " for the rare intestinal cell type of *mus musculus* (NCBI GEO ID: GSE62270), which had 23,630 features (genes) and 288 cells initially. First, we performed prefiltering analysis and filtered out some low non-zero count cells as well as low non-zero count genes. We obtained optimal cluster result (TOPSIS optimal score= 0.8584) that comprised of two clusters. One cluster had 115 cells and the other had 91 cells. The evaluated scores of the four cluster validity indices, *FSI*, *PE*, *PC* and *MPC* for the optimized fuzzy clustering were 0.4818, 0.5784, 0.6073 and 0.2145, respectively. We further identified 1,240 differentially expressed genes (cluster 1 vs cluster 2) using Limma. The top ten gene markers were *Rps21, Slc5a1, Crip1, Rpl15, Rpl3, Rpl27a, Khk, Rps3a1, Aldob* and *Rps17* (ranked by Bonferroni corrected p-value). In this list, *Khk* (encoding ketohexokinase) was found to be a novel marker for the rare intestinal cell type.

Conclusions: Our method identified multi-objective optimized clusters and potential gene markers for each resultant cluster from the dataset. This method may be useful to detect cell clusters from future scRNA-seq data.

Keywords: Single cell sequencing; multi-objective optimization; TOPSIS; fuzzy clustering; cluster validity indices; Limma

---

_

Network-based single-cell RNA-seq data imputation enhances cell type identification

Maryam Zand and Jianhua Ruan[*]

Abstract

Background: Single-cell RNA sequencing is a powerful technology for obtaining transcriptome at single cell resolution. However, it suffers from dropout events (i.e., excess zero counts) since only a small fraction of transcripts get sequenced in each cell during sequencing process. This inherent sparsity of expression profiles hinders further characterizations at cell/gene-level such as cell type identification and downstream analysis.

Results: To alleviate this dropout issue we introduce a network-based method, netImpute, by leveraging the hidden information in gene co-expression networks to recover real signals. netImpute employs Random Walk with Restart (RWR) to adjust the gene expression level in a given cell by borrowing information from its neighbors in a gene co-expression network. Performance evaluation and comparison with existing tools on simulated data and seven real datasets show that netImpute substantially enhances clustering accuracy and data visualization clarity, thanks to its effective treatment of dropouts. While the idea of netImpute is general and can be applied with other types of networks such as cell co-expression network or protein-protein interaction (PPI) network, evaluation results show that gene co-expression network is consistently more beneficial, presumably because PPI network usually lacks cell type context, while cell co-expression network can cause information loss for rare cell types.

Conclusion: Evaluation results on several biological datasets show that netImpute can more effectively recover missing transcripts in scRNA-seq data and enhance the identification and visualization of heterogeneous cell types than existing methods.

Keywords: scRNA-seq data; Data imputation; Co-expression network; Graph random walk; Clustering

---

_

The landscape of alternative splicing in HIV-1 infected CD4 T-cells

Seyoun Byun[1,†], Seonggyun Han[1,†], Yue Zheng[2], Vicente Planelles[2] and Younghee Lee[1,3,*]=

[1]*Department of Biomedical Informatics, University of Utah School of Medicine, Salt Lake City, Utah, USA*

[2]*Department of Pathology, University of Utah School of Medicine, Salt Lake City, Utah, USA*

[3]*Huntsman Cancer Institute, University of Utah School of Medicine, Salt Lake City, Utah, USA*

[†] Contributed equally

Corresponding Authors:
Corresponding Authors: Younghee Lee, Ph.D.
Corresponding authors' address: Younghee Lee
Department of Biomedical Informatics
University of Utah School of Medicine
Salt Lake City, Utah, USA

Corresponding authors' e-mail address: younghee.lee@utah.edu

Running title
Alternative splicing in CD4 T-cells harboring HIV-1

Abstract

Background: Elucidating molecular mechanisms that are altered during HIV-1 infection may provide a better understanding of the HIV-1 life cycle and how it interacts with infected T-cells. One such mechanism is alternative splicing (AS), which has been studied for HIV-1 itself, but no systematic analysis has yet been performed on infected T-cells. We hypothesized that AS patterns in infected T-cells may illuminate the molecular mechanisms underlying HIV-1 infection and identify candidate molecular markers for specifically targeting infected T-cells.
Methods: We downloaded previously published raw RNA-seq data obtained from HIV-1 infected and non-infected T-cells. We estimated percent spliced in (PSI) levels for each AS exon, then identified differential AS events in the infected cells (FDR < 0.05, PSI difference > 0.1). We performed functional gene set enrichment analysis on the genes with differentially expressed AS exons to identify their functional roles. In addition, we used RT-PCR to validate differential alternative splicing events in cyclin T1 (*CCNT1*) as a case study.
Results: We identified 427 candidate genes with differentially expressed AS exons in infected T-cells, including 20 genes related to cell surface, 35 to kinases, and 121 to immune-related genes. In addition, protein-protein interaction analysis identified six essential subnetworks related to the viral life cycle, including Transcriptional regulation by TP53, Class I MHC mediated antigen, G2/M transition, and late phase of HIV life cycle. *CCNT1* exon 7 was more frequently skipped in infected T-cells, leading to loss of the key Cyclin_N motif and affecting HIV-1 transcriptional elongation.
Conclusions: Our findings may provide new insight into systemic host AS regulation under HIV-1 infection and may provide useful initial candidates for the discovery of new markers for specifically targeting infected T-cells.

Keywords: Alternative splicing, *CCNT1*, HIV-1, *CD46*, CD4 T-cell

---

_

Biological Representation of Chemicals Using Latent Target Interaction Profile

Mohamed Ayed
Ph.D. Program in Computer Science The Graduate Center,
The City University of New York, USA
mhady7@gmail.com

Hansaim Lim
Ph.D. Program in Biochemistry, The Graduate Center,
The City University of New York USA
hansaim.lim41@myhunter.cuny.edu

Lei Xie*
Department of Computer Science, Hunter College, & The Graduate Center, The City University of New York, USA
lei.xie@hunter.cuny.edu

*To whom correspondence should be addressed: Lei Xie, Email: lei.xie@hunter.cuny.edu

Supplementary information: Supplementary data are available at *BMC* online

Abstract

Background: Computational prediction of a phenotypic response upon the chemical perturbation on a biological system plays an important role in drug discovery, and many other applications. Chemical fingerprints are a widely used feature to build machine learning models. However, the fingerprints that are derived from chemical structures ignore the biological context, thus, they suffer from several problems such as the activity cliff and curse of dimensionality. Fundamentally, the chemical modulation of biological activities is a multi-scale process. It is the genome-wide chemical-target interactions that modulate chemical phenotypic responses. Thus, the genome- scale chemical-target interaction profile will more directly correlate with *in vitro* and *in vivo* activities than the chemical structure. Nevertheless, the scope of direct application of the chemical-target interaction profile is limited due to the severe incompleteness, biasness, and noisiness of bioassay data.

Results: To address the aforementioned problems, we developed a novel chemical representation method: Latent Target Interaction Profile (LTIP). LTIP embeds chemicals into a low dimensional continuous latent space that represents genome-scale chemical-target interactions. Subsequently LTIP can be used as a feature to build machine learning models. Using the drug sensitivity of cancer cell lines as a benchmark, we have shown that the LTIP robustly outperforms chemical fingerprints regardless of machine learning algorithms. Moreover, the LTIP is complementary with the chemical fingerprints. It is possible for us to combine LTIP with other fingerprints to further improve the performance of bioactivity prediction.

Conclusions: Our results demonstrate the potential of LTIP in particular and multi-scale modeling in general in predictive modeling of chemical modulation of biological activities.

Keywords: Machine learning, genome-wide target binding, chemical embedding, fingerprint, bioactivity, LTIP

_

Long non-coding RNA expression levels modulate cell-type specific splicing patterns by altering their interaction landscape with RNA-binding proteins

Felipe Wendt Porto[1], Swapna Vidhur Daulatabad[1], Sarath Chandra Janga[1, 2, 3, *]

[1]Department of BioHealth Informatics, School of Informatics and Computing, IUPUI, Indianapolis, IN, USA.

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

[3]Centre for Computational Biology and Bioinformatics Indiana University School of Medicine,
Indianapolis, IN, USA. Corresponding author*

Email addresses:
FWP: fporto@iu.edu
SVD: swapdaul@iupui.edu
*SCJ: scjanga@iupui.edu

Abstract

Background: Recent developments in our understanding of the interactions between long non- coding RNAs (lncRNAs) and cellular components have improved treatment approaches for various human diseases including cancer, vascular diseases, and neurological diseases. Although investigation of specific lncRNAs revealed their role in the metabolism of cellular RNA, our understanding of their contribution to post-transcriptional regulation is relatively limited. In this study, we explore the role of lncRNAs in modulating alternative splicing and their impact on downstream protein-RNA interaction networks.

Results: Analysis of alternative splicing events across 39 lncRNA knockdown and wildtype RNA- sequencing datasets from three human cell lines: HeLa (Cervical Cancer), K562 (Myeloid Leukemia), and U87 (Glioblastoma), resulted in high confidence (fdr < 0.01) identification of 11630 skipped exon events and 5895 retained intron events, implicating 759 genes to be impacted at post-transcriptional level due to the loss of lncRNAs. We observed that a majority of the alternatively spliced genes in a lncRNA knockdown were specific to the cell type, in tandem, the functions annotated to the genes affected by alternative splicing across each lncRNA knockdown also displayed cell type specificity. To understand the mechanism behind this cell-type specific alternative splicing patterns, we analyzed RNA binding protein (RBP)-RNA interaction profiles across the spliced regions, to observe cell type specific alternative splice event RBP binding preference.

Conclusions: Despite limited RBP binding data across cell lines, alternatively spliced events detected in lncRNA perturbation experiments were associated with RBPs binding in proximal intron-exon junctions, in a cell type specific manner. The cellular functions affected by alternative splicing were also affected in a cell type specific manner. Based on the RBP binding profiles in HeLa and K562 cells, we hypothesize that several lncRNAs are likely to exhibit a sponge effect in disease contexts, resulting in the functional disruption of RBPs, and their downstream functions. We propose that such lncRNA sponges can extensively rewire the post-transcriptional gene regulatory networks by altering the protein-RNA interaction landscape in a cell-type specific manner.

Keywords

Long non-coding RNA, cell type specific, alternative splicing, functional enrichment, RNA-binding proteins, protein binding sponges, secondary RNA structure, and cancer.

_

Super Clustering Approach for Fully Automated Single Particle Picking in Cryo-EM

Adil Al-Azzawi[1], Anes Ouadou[1], and Jianlin Cheng[1,2*]

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

[2]Informatics Institute, University of Missouri, Columbia, MO 65211, USA

* Corresponding author
Email addresses:
AA: aaadn5@mail.missouri.edu
AO: aomqc@mail.missouri.edu
JC:chengji@missouri.edu

Abstract

Background

Structure determining of complex proteins and macromolecular in the Cryo-EM (Cryo-electron microscopy) still a big challenge which requires substantial human intervention, labor-intensive and time-consuming. For the preparation stage, the researcher must indicate, detect, and select hundreds of thousands of good input single-particle examples for cryo-EM reconstruction. The performance of the existing tools still does not meet the requirements of the researcher in this filed according to the variety of particles shapes and the quality of micrographs. Some cryo-EM images have very complex (irregular) protein shape and extremely low signal-to-noise ratio (SNR) which some existing automated particle-selection methods still required a large number of manually high-quality particles to identify and detect them. To address this issue, we propose a fully automated single particle picking method (SuperCryoEMPicker) based on the idea of the super clustering using unsupervised learning.

Methods

We design a fully automated, unsupervised approach for single particle picking in cryo-EM micrographs that focus on identify, detect, and pick the complex and irregular protein shapes in the extremely low signal-to-noise micrographs. To adjust the low SNR micrographs, our model has two preprocessing stages. First, the original three-dimensional grid of voxels cryo-EM (MRC file format) is converted to another graphic file and the global intensity level s adjusted using a suite scientific cryo-EM image processing tools EMAN2. Second, the new micrographs format (PNG file) allows us to apply some advanced image processing tools to in case of improving the quality of the cryo-EM images. Second, binary mask image is generating from each individual cryo- EM using particle clustering in terms of localized and identify the particles in the cryo- EM image. Two different clustering approaches have been used. The first one is the regular clustering using k-means, fuzzy c-means (FCM), and the intensity-based clustering (ICB). The second approach is the super clustering approach. Super clustering approach is mainly based on generate an intermedia micrograph map using the simple linear iterative clustering (SLIC) and the original clustering algorithms.

Results

Experimental results show that the super particle clustering and picking from the intermedia micrograph map using super clustering SP-k-means, SP-FCM, and SP-ICB have a more robust detection and selection than those directly selected from the original noisy micrographs.

Conclusions

SuperCryoEMPicker can automatically and effectively identify and recognizes very complex particle-like objects from an extremely low-SNR micrographs condition. As a fully automated particle detection and selection method, the proposed method, can help the researchers from the laborious work for manually particles identification election work, also without the need of labeled training data and human intervention, therefore is a useful tool for cryo-EM protein structure determination.

Keywords

Super clustering, Intensity Based Clustering (IBC), micrograph, cryo-EM, singe particle pickling, protein structure determination, k-means, Fuzzy-c-means.

_

Comparative evaluation of network features for the prediction of breast cancer metastasis

Nahim Adnan[1], Zhijie Liu[2], Tim H.M. Huang[2] and Jianhua Ruan[1*]

Abstract

Background: Discovering a highly accurate and robust gene signature for the prediction of breast cancer metastasis from gene expression profiling of primary tumors is one of the most challenging tasks to reduce the number of deaths in women. Due to the limited success of gene-based features in achieving satisfactory prediction accuracy, many methodologies have been proposed in recent years to develop network-based features by integrating network information with gene expression. However, evaluation results are inconsistent to confirm the effectiveness of network-based features, because of many confounding factors involved in classification model learning process, such as data normalization, dimension reduction, and feature selection. An unbiased comparative evaluation is essential for uncovering the strength of network-based features.

Results: In this study, we compared several types of network-based features obtained from protein-protein interaction network and gene co-expression network for their ability in predicting breast cancer metastasis using gene expression data from more than 10 patient cohorts. While network-based features are usually statistically more significant than gene-based features, a consistent improvement of prediction performance using network-based features requires a substantial number of patients in the dataset. In addition, contrary to many previous reports, no evidence was found to support the robustness of network-based features and we argue some of the robustness may be due to the inherent bias associated with node degree in the network. In addition, different types of network features seem to cover different pathways and are complementary to each other. Consequently, an ensemble classifier combining different network features was proposed and was found to significantly outperform classifiers based on gene features or any single type of network features.

Conclusions: Network-based features and their combination show promise for improving the prediction of breast cancer metastasis but may require a large amount of training data. Robustness claim of network-based features needs to be re-examined with network node degree and other confounding factors in consideration.

Keywords: breast cancer; metastasis prediction; network; gene expression analysis

---

_

Highly robust model of transcription regulator activity predicts breast cancer overall survival

Chuanpeng Dong[1,2,*], Jiannan Liu[2,*], Steven X. Chen[1], Tianhan Dong[3], Guanglong Jiang[1,2], Yue Wang[1], Huanmei Wu[2], Jill L. Reiter[1], Yunlong Liu[1 ,2]

[1] Department of Medical and Molecular Genetics, Center for Computational Biology and Bioinformatics, Indiana University School of Medicine;

[2] Department of BioHealth Informatics, School of Informatics and Computing, Indiana University-Purdue University Indianapolis

[3] Department of Pharmacology and Toxicology, Indiana University School of Medicine;

[*] Chuanpeng Dong and Jiannan Liu contributed equally to this work.

To whom correspondence should be addressed：
Yunlong Liu, Department of Molecular and Medical Genetics, Indiana University School of Medicine, Indianapolis, IN 46202, USA. yunliu@iu.edu;
Jill L. Reiter, Department of Molecular and Medical Genetics, Indiana University School of Medicine, Indianapolis, IN 46202, USA. jireiter@iupui.edu

Abstract
While several multigene signatures are available for predicting breast cancer prognosis, particularly in early stage disease, effective molecular indicators are needed especially for triple-negative carcinomas to improve treatments and predict diagnostic outcomes. The objective of this study was to identify transcriptional regulatory networks to better understand mechanisms giving rise to breast cancer development and incorporate this information into a model for predicting clinical outcomes. Gene expression profiles from 1097 breast cancer patients were retrieved from The Cancer Genome Atlas (TCGA). Breast cancer-specific transcription regulatory information was identified by considering the binding site information from ENCODE and the top co-expressed targets in TCGA using a nonlinear approach. We used this information to build a multi-regulator linear model to predict breast cancer patient survival. This model was validated in more than 5000 breast cancer patients from the Gene Expression Omnibus (GEO) databases. Our findings demonstrate that transcriptional regulator activities correlate with PAM50 gene expression profiles of the basal-like tumor subtype. We also report that transcription regulator activity model can predict overall breast cancer patient survival regardless of tumor subtype and is associated with cell cycle and PLK-signaling pathways. This finding provides additional biological insights into the mechanisms of breast cancer progression.

Keywords:
Breast cancer, transcription regulators, prognostic model

Pseudogene-gene functional networks are prognostic of patient survival in breast cancer

Sasha Smerekanych[1,2¶], Travis S Johnson[2¶], Kun Huang[3,4], Yan Zhang[2,5]*

[1] Kenyon College, Gambier, OH 43022, United States

[2] Department of Biomedical Informatics, College of Medicine, The Ohio State University, Columbus, OH 43210, United States

[3] School of Medicine, Indiana University, Indianapolis, IN 46202, USA

[4] School of Informatics and Computing, Indiana University, Indianapolis, IN 46262, USA
[5] The Ohio State University Comprehensive Cancer Center (OSUCCC – James), Columbus, OH 43210, United States

¶ These authors contributed equally to this work.
* Correspondence: yan.zhang@osumc.edu

Abstract

Given the vast range of molecular mechanisms giving rise to breast cancer, it is unlikely universal cures exist. However, by providing a more precise prognosis for breast cancer patients through integrative models, treatments can become more individualized, resulting in more successful outcomes. Specifically, we combine gene expression, pseudogene expression, miRNA expression, clinical factors, and pseudogene-gene functional networks to generate these models for breast cancer prognostics. Establishing a LASSO-generated molecular gene signature revealed that the increased expression of genes STXBP5, GALP and LOC387646 indicate a poor prognosis for a breast cancer patient. We also found that increased CTSLP8 and RPS10P20 and decreased HLA-K pseudogene expression indicate poor prognosis for a patient. Perhaps most importantly we identified a pseudogene-gene interaction, GPS2-GPS2P1 (improved prognosis) that is prognostic where neither the gene nor pseudogene alone is prognostic of survival. Besides, miR-3923 was predicted to target GPS2 using miRanda, PicTar, and TargetScan, which imply modules of gene-pseudogene-miRNAs that are potentially functionally related to patient survival.

In our LASSO-based model, we take into account features including pseudogenes, genes and candidate pseudogene-gene interactions. Key biomarkers were identified from the features. The identification of key biomarkers in combination with significant clinical factors (such as stage and radiation therapy status) should be considered as well, enabling a specific prognostic prediction and future treatment plan for an individual patient. Here we used our PseudoFuN web application to identify the candidate pseudogene-gene interactions as candidate features in our integrative models. We further identified potential miRNAs targeting those features in our models using PseudoFuN as well. From this study, we present an interpretable survival model based on LASSO and decision trees, we also provide a novel feature set which includes pseudogene-gene interaction terms that have been ignored by previous prognostic models. We find that some interaction terms for pseudogenes and genes are significantly prognostic of survival. These interactions are cross-over interactions, where the impact of the gene expression on survival changes with pseudogene expression and vice versa. These may imply more complicated regulation mechanisms than previously understood. We recommend these novel feature sets be considered when training other types of prognostic models as well, which may provide more comprehensive insights into personalized treatment ecisions.

_

Deep Learning-based Cancer Survival Prognosis from RNA-seq Data: Approaches and Evaluations

Zhi Huang1,2,4, Travis S. Johnson2,3, Zhi Han2, Bryan Helm2, Sha Cao5, Chi Zhang5,4, Paul Salama4, Maher Rizkalla4, Christina Y. Yu2,3, Jun Cheng2,6, Shunian Xiang5,7, Xiaohui Zhan2,7, Jie Zhang5, Kun Huang2,4,*

1School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, 47907, USA
2Department of Medicine, Indiana University School of Medicine, Indianapolis, IN, 46202, USA
3Department of Biomedical Informatics, The Ohio State University, Columbus, OH, 43210, USA
4Department of Electrical and Computer Engineering, Indiana University - Purdue University Indianapolis, Indianapolis, IN, 46202, USA
5Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, IN, 46202, USA
6National-Regional Key Technology Engineering Laboratory for Medical Ultrasound, School of Biomedical Engineering, Health Science Center, Shenzhen University, Shenzhen, 518060, China
7Guangdong Key Laboratory for Biomedical Measurements and Ultrasound Imaging, School of Biomedical Engineering, Shenzhen University, Shenzhen, 518060, China

*To whom correspondence should be addressed

Abstract
Recent advances in kernel-based Deep Learning models have introduced a new era in medical research. Originally designed for image processing, Deep Learning models are now applied to survival prognosis of cancer patients. Specifically, Deep Learning versions of the Cox proportional hazards models are trained with transcriptomic data to predict survival outcomes in cancer patients. In this study, a broad analysis is performed on TCGA cancers using a variety of Deep Learning-based models, including Cox-nnet, DeepSurv, and a method proposed by our group named AECOX (AutoEncoder with Cox regression network). Concordance index and p-value of the log-rank test are used to evaluate the model performances. All models show competitive results across 12 cancer types. The last hidden layers of the Deep Learning approaches are lower dimensional representations of the input data that can be used for feature reduction and visualization. Furthermore, the prognosis performances reveal a negative correlation between model accuracy and tumor mutation burden (TMB), sug-gesting an association between TMB and survival prognosis accuracy.
Contact: kunhuang@iu.edu
Keywords: Deep Learning, Cancer Prognosis, Survival Analysis, Tumor Mutation Burden, Cox Regression

---

Transcription factor expression as a predictor of colon cancer prognosis: A machine learning practice

Jiannan Liu[1*], Chuanpeng Dong[1,2*], Guanglong Jiang[1,2,3], Xiaoyu Lu[1,2], Yunlong Liu[2,3], Huanmei Wu[1]

[1]Depart of BioHealth Informatics, School of Informatics and Computing, Indiana University-Purdue University Indianapolis, Indianapolis, Indiana

[2]Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, Indiana.

[3]Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, Indiana

[*]Jiannan Liu and Chuanpeng Dong contributed equally to this work.

Corresponding Author:
Huanmei Wu, Depart of BioHealth Informatics, School of Informatics and Computing, Indiana University-Purdue University Indianapolis, Indianapolis, Indiana, Tel +1 317-278-1692, Fax +1 317-278-1692, Email hw9@iupui.edu

Abstract

Background: Colon cancer is one of the leading causes of cancer deaths in the USA and around the world. Molecular level characters, such as gene expression levels and mutations, may provide profound information for precision treatment apart from pathological indicators. Transcription factors function as critical regulators in all aspects of cell life, but transcription factors-based biomarkers for colon cancer prognosis were still rare and necessary.

Methods: We implemented an innovative process to select the transcription factors variables and evaluate the prognostic prediction power by combining the Cox PH model with the Random Forest algorithm. We picked five top-ranked transcription factors and built a prediction model by using Cox PH regression. Using Kaplan-Meier analysis, we validated our predictive model on four independent publicly available datasets (GSE39582, GSE17536, GSE37892, and GSE17537) from the GEO database, consisting of 1584 colon cancer patients.

Results: A five-transcription-factors based predictive model for colon cancer prognosis has been developed by using TCGA colon cancer patient data. Five transcription factors identified for the predictive model is HOXC9, ZNF556, HEYL, HOXC4 and HOXC6. The prediction power of the model is validated with four GEO datasets consisting of 1,584 patient samples. Kaplan-Meier curve and log-rank tests were conducted on both training and validation datasets, the difference of overall survival time between predicted low and high-risk groups can be clearly observed. Gene set enrichment analysis was performed to further investigate the difference between low and high-risk groups in the gene pathway level. The biological meaning was interpreted. Overall, the results of our prove our prediction model has a strong prediction power on colon cancer prognosis.

Conclusions: Transcription factors can be used to construct colon cancer prognostic signatures with strong prediction power. The variable selection process used in this study has the potential to be implemented in the prognostic signature discovery of other cancer types. Our five TF-based predictive model would help with understanding the hidden relationship between colon cancer patient survival and transcription factor activities. It will also provide more insights into the precision treatment of colon cancer patients from a genomic information perspective.

Key words: colon cancer, transcription factor, machine learning, cancer prognosis

_

A pan-cancer study of class-3 semaphorins as therapeutic targets in cancer

Xiaoli Zhang*, Brett Klamer, Jin Li, Soledad Fernandez, and Lang Li

Department of Biomedical Informatics, The Ohio State University, Columbus, OH, 43201.

***Corresponding author:**
Dr. Xiaoli Zhang, Department of Biomedical Informatics, College of Medicine, The Ohio State University. 320B Lincoln Tower, 1800 Cannon Dr., Columbus, Oh, 43210,
Email: Xiaoli.zhang@osumc.edu

Abstract

Background

Initially characterized as axon guidance factors, semaphorins also have been implicated to have critical roles in multiple physiological and developmental functions, including the regulation of immune responses, angiogenesis, organ formation, and the etiology of multiple forms of cancer. Moreover, their contribution in immunity and the regulation of tumour microenvironment is becoming increasingly recognized. Here, we provide a comprehensive analysis of class-3 semaphorins, the only secreted family of genes among veterbrate semaphorins, in terms of their expression profiles and their association with patient survival. We also relate their role with immune subtypes, tumour microenvironment, and drug sensitivity using a pan-cancer study.

Results

Expression profiles of class-3 semaphorins (SEMA3s) and their association with patient survival and tumour microenvironment were studied in 31 cancer types using the TCGA pan-cancer data. The expression of SEMA3 family varies in different cancer types with striking inter- and intra-cancer heterogeneity. In general, our results show that SEMA3A, SEMA3C, and SEMA3F are primarily upregulated in cancer cells, while the rest of SEMA3s are mainly down-regulated in the tested tumours. The expression of SEMA3 family members was frequently associated with patient overall survival. However, the direction of the association varied with regards to the particular SEMA3 isoform queried and the specific cancer type tested. More specifically, SEMA3A and SEMA3E primarily associate with a poor prognosis of survival, while SEMA3G typically associates with survival advantage. The rest of SEMA3s show either survival advantage or disadvantage dependent on cancer type. In addition, all SEMA3 genes show significant association with immune infiltrate subtypes, and they also correlate with level of stromal cell infiltration and tumour cell stemness with various degrees. Finally, our study revealed that SEMA3 genes, especially SEMA3C and SEMA3F may contribute to drug induced cancer cell resistance.

Conclusions

Our systematic analysis of class-3 semaphorin gene expression and their association with immune infiltrates, tumour microenvironment and cancer patient outcomes highlights the need to study each SEMA3 member as a separate entity within each specific cancer type. Also our study validated the identification of class-3 semaphorin signals as promising therapeutic targets in cancer.

Keywords

Class-3 semaphorins, gene expression, survival, tumour suppression, tumour promotion, immune subtype, tumour microenvironment, drug sensitivity

_

Predicting Re-admission to Hospital for Diabetes Treatment: A Machine Learning based Solution

Satish M. Srinivasan1, Yok-Fong Paat2, Philmore Halls1, Ruth Kalule1, Thomas E. Harvey1

1School of Graduate Professional Studies, Penn State Great Valley, Malvern, PA 19355
2College of Health Science, The University of Texas at El Paso, El Paso, TX 79968
§ Corresponding author:

Email address:
§ SMS: sus64@psu.edu
YFP: ypaat@utep.edu 2

Abstract

Background: Predictive analytics embraces an extensive range of techniques including but are not limited to statistical modeling, Machine Learning, Artificial Intelligence and Data Mining. It has profound usefulness in different application areas such as business intelligence, public health, disaster management and response, as well as many other fields. The technique is well-known as a practice for identifying patterns within data to predict future outcomes and trends. The objective of this study is to design and implement a predictive analytics system that can be used to forecast the likelihood that a diabetic patient will be readmitted to the hospital.

Results: Upon extensively cleaning the Diabetes 130-US hospitals dataset containing patient records spanning 10 years from 1999 till 2008, we modeled the relationship between the predictors and the response variable using the Random Forest classifier. Upon performing hyperparameter optimization for the Random Forest, we obtained a maximum AUC of 0.684 with a precision and recall of 46% and 60% respectively and an F1 Score of 52.07%. Our study reveals that attributes such as *number of inpatient visits*, *discharge disposition*, *admission type*, and *number of laboratory tests* are strong predictors for the response variable (*i.e.* re-admission of patients).

Conclusion: Findings from this study can help hospitals design suitable protocols to ensure that patients with a higher probability of re-admission are recovering well and possibly reducing the risk of future re-admission. In the long run, not only will our study improve the life quality of diabetic patients, it will also help in reducing the medical expenses associated with re-admission.

Keywords: Random Forest, Data Cleaning, Predictive Analytics, Hyperparameter tuning, optimization

Concurrent Session - Cancer Genomics
Tuesday, June 11, 2019
10:10 AM – 12:20 PM
Ballroom

Identify rewired pathways between primary breast cancer and liver metastatic cancer using transcriptome data

Limei Wang1,2,3,#, Jin Li2,3,#, Enze Liu4, Garrett Kinnebrew2, Yang Huo4, Zhi Zeng2,5, Wanli Jiang2,6, Lijun Cheng2, Hongchao Lv3, Weixing Feng1,*, Lang Li2,*

1. College of Automation, Harbin Engineering University, Harbin, China, 150001
2. Department of Biomedical Informatics, College of Medicine, The Ohio State University, Columbus, OH, 43210
3. College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, China, 150086
4. The Center for Computational Biology and Bioinformatics, School of Medicine, Indiana University, Indianapolis, IN, 46202
5. Department of Pathology, Renmin Hospital of Wuhan University, Wuhan, China, 430060
6. Department of Thoracic Surgery, Renmin Hospital of Wuhan University, Wuhan, China, 430060

#Equal contributors

*Corresponding Authors: Weixing Feng, fengweixing@hrbeu.edu.cn
Lang Li, Lang.Li@osumc.edu
Emails for other authors
Limei Wang: wanglm@hrbeu.edu.cn
Jin Li: Jin.Li@osumc.edu
Enze Liu: enzeliu@iu.edu
Garrett Kinnebrew: gkinnebrew@gmail.com
Yang Huo: yanghuo@iu.edu
Zhi Zeng: zhi.zeng@osumc.edu
Wanli Jiang: jiang_wanli@163.com
Lijun Cheng: Lijun.Cheng@osumc.edu
Hongchao Lv: Lvhongchao@gmail.com

Abstract

Background: Pathway analysis was widely used in cancer and metastasis research, however, pathways under different biological conditions, such as primary cancer and metastatic cancer, had different active genes, topologies or active sub-pathways, and these pathways were named as rewired pathways. Several biological experiments observed rewired pathways in cancers, but the concept had not been fully explored in the computational cancer system biology. Here we proposed a rewired pathway identification method using transcriptome data and applied to primary breast cancer and breast cancer liver metastasis.

Results: Several pathways were significantly enriched among primary breast cancer and breast cancer liver metastasis separately. Interestingly, some pathways, such as *cytokine-*

cytokine receptor interaction (hsa04060) and *Calcium signaling* (hsa04020), were significantly enriched under both conditions. After the pathway networks topology analysis, *TGF beta signaling* was found to be a hub pathway. Based on these 3 pathways, we recognized 3 distinctive types of rewired pathways. In the *cytokine-cytokine receptor interaction* pathway, there were several active alteration patterns in each cytokine-cytokine receptor pair. Thirteen cytokine-cytokine receptor pairs with both genes' activity changes inversely pattern were verified by literature. The results showed that most of them were related to breast cancer or metastasis. The second kind of rewired pathway was that some sub-pathways were active in only one condition. In the *calcium signaling* pathway, the sub-pathway TnC, PHK, CAMK, NOS, ADCY, FAK2, IP3-3K was only active in the breast cancer liver metastasis. In the third type of rewired pathway, there were nodes that were significantly active in both conditions, but the active genes in these nodes were different. In *calcium signaling pathway* and *TGF beta signaling pathway*, Node E2F4/5 can be E2F5 or E2F4 which were significantly active in primary breast cancer and metastasis separately. These E2F4 or E2F5 rewired pathways were identified and verified by previous researches.
Conclusions: It was the first time to use transcriptome data to identify rewired pathways in cancer metastasis. The results showed the proposed method was valid and effective and could be helpful for future research on breast cancer metastasis mechanisms and developing drugs that may prevent cancer metastasis.
Keywords: rewired pathway, breast cancer, liver metastasis, microarray, gene active status

Kinetic modeling of DUSP regulation in Herceptin-resistant HER2-positive breast cancer

Petronela Buiga1,2, Ari Elson2, Lydia Tabernero1, Jean-Marc Schwartz1,*

1 School of Biological Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester, UK.
2 Department of Molecular Genetics, Weizmann Institute of Science, Rehovot, Israel.

* Corresponding author: jean-marc.schwartz@manchester.ac.uk

Abstract

Background: HER2 positive breast cancer is an aggressive type of breast cancer characterized by overexpression of the receptor-type protein tyrosine kinase HER2 or amplification of the *HER2* gene. It is commonly treated by the drug trastuzumab (Herceptin), but resistance to its action frequently develops and limits its therapeutic benefit. Dual specificity phosphatases (DUSPs) were previously highlighted as central regulators of HER2 signaling, therefore understanding their role is crucial to designing new strategies to improve the efficacy of Herceptin treatment.
Results: We investigate whether inhibiting certain DUSPs re-sensitizes Herceptinresistant breast cancer cells to the drug. We built a series of kinetic models incorporating the key players of HER2 signaling pathways and simulating a range of inhibition intensities. The simulation results were compared to live tumor cells in culture and showed good agreement with the experimental analyses. In particular, we observed that Herceptin-resistant DUSP16-

silenced breast cancer cells became more responsive to the drug when treated for 72 hours with Herceptin, showing a decrease in resistance in agreement with the model predictions.
Conclusions: Overall, we showed that kinetic modeling of signaling pathways is able to generate predictions that assist experimental research in the identification of potential targets for cancer treatment.

Keywords
Kinetic model, Breast cancer, Herceptin, Dual-specificity phosphatases

_

Gene co-expression networks restructured by and gene fusion in rhabdomyosarcoma cancers

Bryan R. Helm[1], Xiaohui Zhan[1], Pankita H. Pandya[2], Mary E. Murray[2], Karen E. Pollok[2,3], Jamie L. Renbarger[2], Michael J. Ferguson[2], Mark S. Marshall[2], Zhi Han[1], Dong Ni[4], Jie Zhang[1], Kun Huang[1]

1. Department of Medical Science, Indiana University, 340 West 10th St., Fairbanks Hall, Suite 6200, Indianapolis, IN 46202-3082

2. Department of Pediatrics, School of Medical Science, Indiana University, 340 West 10th St., Fairbanks Hall, Suite 6200, Indianapolis, IN 46202-3082
3. Department of Pharmacology and Toxicology, School of Medical Science, Indiana University, 340 West 10th St., Fairbanks Hall, Suite 6200, Indianapolis, IN 46202-3082
4. National-Regional Key Technology Engineering Laboratory for Medical Ultrasound, Guangdong Key Laboratory for Biomedical Measurements and Ultrasound Imaging, School of Biomedical Engineering, Health Science Center, Shenzhen University, Shenzhen, China

Abstract

In alveolar-type rhabdomyosarcoma, a recurrent chromosome translocation results in a gene fusion that is comprised of "forkhead homeobox 1" (FOXO1) and either paired-homeobox 3 (PAX3) or paired-homeobox 7 (PAX7). The fusion protein (FOXO1-PAX3/7) retains both binding domains and becomes a novel and potent transcriptional regulator. Many studies have identified genes that have differential expression between fusion negative and positive rhabdomyosarcoma; however, these analyses have not addressed the networks of interactions among genes. We examined gene co-expression networks in microarray data (NCBI GEO, GSE #66533, Sun et al., 2015) for alveolar rhabdomyosarcoma with 25 fusion negative, 26 FOXO1-PAX3 positive, and 7 FOXO1-PAX7 positive samples. Gene network modules were calculated using "local maximum Quasi-Clique Merger" (lmQCM) on all groups combined and for fusion negative and positive data separately. lmQCM identified 41 gene co-expression modules, of which 17/41 showed statistically signification variation in gene expression with respect to sample fusion status. 109 gene co-expression modules were identified fusion negative samples, whereas only 53 co-expression modules were observed in fusion positive. Deeper analysis of fusion positive co-expression modules revealed 17/53 had differential expression with respect to PAX3 or PAX7 samples. We submitted co-expression modules related to fusion status for gene list enrichment analysis. Overall, we observed substantial restructuring of gene co-expression networks relative to fusion status and type in alveolar rhabdomyosarcoma.

Keywords: rhabdomyosarcoma; gene fusion; gene co-expression analysis; quasi-clique merger; copy number variation

Convolutional neural network models for cancer type prediction based on gene expression

Milad Mostavi[1,2], Yu-Chiao Chiu[1], Yufei Huang[2,3§], Yidong Chen[1,3§]

[1]Greehey Children's Cancer Research Institute, University of Texas Health Science Center at San Antonio, San Antonio, TX 78229, USA

[2]Department of Electrical and Computer Engineering, University of Texas at San Antonio, San Antonio, TX 78249, USA

[3]Department of Epidemiology and Biostatistics, University of Texas Health Science Center at San Antonio, San Antonio, TX 78229, USA

Abstract

Background

Precise prediction of cancer types is vital for cancer diagnosis and therapy. Important cancer marker genes can be revealed through the prediction model interpretation. Several studies have attempted to build machine learning models for this task but did not take into consideration the effect of the tissue of origin that can bias the identification of cancer markers.

Results

We developed Convolutional Neural Networks (CNNs) models that take unstructured gene expression inputs to classify tumors into their designated cancer types or as normal. Based on different designs of gene embeddings and convolution schemes, we implemented three CNN models: 1D-CNN, 2D-Vanilla-CNN, and 2D-Hybrid-CNN. The models were trained and tested on combined 10,340 samples of 33 cancer types and 731 matched normal tissues of The Cancer Genome Atlas (TCGA). Our models achieved excellent prediction accuracies (93.9-95.0%) among 34 classes (33 cancers and normal). We interpreted the 1D-CNN model with a guided saliency technique and identified a total of 2,090 cancer markers (108 per class). We confirmed the concordance of differential expression of these markers between the cancer type they represent and others. In breast cancer, our model identified well-known markers, such as *GATA3* and *ESR1*. Finally, we extended the 1D- CNN model for a prediction of breast cancer subtypes and achieved an average accuracy of 88.42% among 5 subtypes.

Conclusions

Here we present novel CNN designs for accurately and simultaneously cancer/normal and cancer type prediction based on gene expression profiles, and unique model interpretation scheme to elucidate biologically relevant cancer marker genes that demonstrated the benefit to eliminating the effects of tissue-of-origin. The proposed model had light hyperparameters to be trained and thus was easily applied to classify breast cancer subtypes. Overall, the study presents several accurate predictor machines for cancer types that may facilitate cancer diagnosis in the future.

Keywords: deep Learning; convolutional neural networks, The Cancer Genome Atlas; cancer type prediction; cancer gene markers; breast cancer subtype prediction

_

Integrative Network Analysis Identifies Potential Targets and Drugs for Ovarian Cancer

Tianyu Zhang[1,3], Liwei Zhang[3], Fuhai Li[1,2*]

[1]Institute for Informatics (I2), and [2]Department of Pediatrics, Washington University School of Medicine, Washington University in St. Louis, St. Louis, Missouri, United States; [3]Dalian University of Technology, Dalian, China;
*Corresponding author: Fuhai.Li@wustl.edu

Abstract

Though the ovarian cancer accounts for 2.5% of all cancers in female, the death rate of ovarian cancer is high, which is the fifth leading cause of cancer death (5% of cancer death) in female, and the 5-year survival rate is less than 50%. The oncogenic molecular signaling of ovarian cancer are complicated and remain unclear. And there is a lack of effective targeted therapies for ovarian cancer treatment. In this study, we propose to explore the activated signaling pathways of individual ovarian cancer patients and sub-groups, and identify the potential targets and drugs that are able to disrupt the core signaling pathways. In specific, we first identify the up-regulated genes of individual cancer patients using Markov chain Monte Carlo (MCMC), and then identify the potential activated transcription factors. After dividing ovarian cancer patients into three sub- groups sharing common transcription factors, we uncover the up-stream signaling pathways of activated transcription factors in each sub-group. For example, the PI3K/AKT, WNT, TP53, and MTOR signaling are uncovered to play important roles in ovarian cancer. These up-stream signaling pathways could be therapeutic targets to disrupt the activation of multiple activated transcription factors. We mapped all FDA approved drugs targeting on the upstream signaling targets, which also indicate the potential synergistic drug combinations for ovarian cancer treatment.

_

A Novel Graph Regularized Non-negative Matrix Factorization based on Error Weight Matrix for High Dimensional Biomedical Data Clustering

Meijun Zhou[1], Xianjun Shen[1*], Limin Yu[1], Xingpeng Jiang[1*], Jincai Yang[1], Xiaohua Hu[1, 2]

[1]School of Computer, CentralChinaNormalUniversity, Wuhan, China [2]College of Computing and Informatics, DrexelUniversity, Philadelphia,USA
xjshen@mail.ccnu.edu.cn; xpjiang@mails.ccnu.edu.cn

Abstract

Nonnegative Matrix Factorization (NMF) as a prevalent technique for finding parts-based representations of nonnegative data and interpretability of corresponding results has been widely applied in a wide range of applications. On the other hand, the data is ordinarily sampled from a low dimensional manifold embedded in the higher dimensional space. One hopes to have a compact representation, which abide by the intrinsic geometric structure and simultaneously reveals the hidden semantics. However, it is difficult to accurate representation the inherent manifold of data space in the case of high-dimensional and large data sets. In this paper, we propose a novel algorithm, called error weight matrix graph regularized non-negative matrix factorization (EGNMF), for this purpose. In EGNMF, error weight matrix is constructed to restrict the representation accuracy of before and after dimensionality reduction. Experiments results on different real-word datasets have validated that the proposed method is superior to the latest extension algorithms of GNMF.

Keywords—Non-negative Matrix Factorization; Manifold Learning; Graph Laplacian ; Error Weight Matrix

_

Skyhawk: An Artificial NeuralNetwork based discriminator for reviewing clinically significant genomic variants

Ruibang Luo[1,2,*], Tak-Wah Lam[1], Michael C. Schatz[2]

[1] Department of Computer Science, The University of Hong Kong, Hong Kong

[2] Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA

[*] Correspondence should be addressed to rbluo@cs.hku.hk

Abstract

Motivation: Many rare diseases and cancers are fundamentally diseases of the genome. In the past several years, genome sequencing has become one of the most important tools in clinical practice for rare disease diagnosis and targeted cancer therapy. However, variant interpretation remains the bottleneck as is not yet automated and may take a specialist several hours of work per patient. On average, one-fifth of this time is spent on visually confirming the authenticity of the candidate variants.

Results: We developed Skyhawk, an artificial neural network-based discriminator that mimics the process of expert review on clinically significant genomics variants. Skyhawk runs in less than one minute to review ten thousand variants, and about 30 minutes to review all variants in a typical whole-genome sequencing sample. Among the false positive singletons identified by GATK HaplotypeCaller, UnifiedGenotyper and 16GT in the HG005 GIAB sample, 79.7% were rejected by Skyhawk. Worked on the Variants with Unknown Significance (VUS), Skyhawk marked most of the false positive variants for manual review and most of the true positive variants no need for review.

Availability: Skyhawk is easy to use and freely available at
https://github.com/aquaskyline/Skyhawk

MIRIA: a webserver for statistical, visual and meta-analysis of RNA editing data in mammals

Xikang Feng[1†], Zishuai Wang[1†], Hechen Li[1] and Shuaicheng Li[1,*]

[1] Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong

* To whom correspondence should be addressed. Tel: +(852)-3442-9412; Fax: +(852)-3442-0503; Email: shuaicli@cityu.edu.hk.
†These authors contributed to this work as first authors.

Abstract

Background

Adenosine-to-inosine RNA editing can markedly diversify the transcriptome, leading to a variety of critical molecular and biological processes in mammals. Over the past several years, researchers have developed several new pipelines and software packages to identify RNA editing sites with a focus on downstream statistical analysis and functional interpretation.

Results

Here, we developed a user-friendly public webserver named MIRIA that integrates statistics and visualization techniques to facilitate the comprehensive analysis of RNA editing sites data identified by the pipelines and software packages. MIRIA is unique in that provides several analytical functions, including RNA editing type statistics, genomic feature annotations, editing level statistics, genome- wide distribution of RNA editing sites, tissue-specific analysis and conservation analysis. We collected high-throughput RNA sequencing (RNA-seq) data from eight tissues across seven species as the experimental data for MIRIA and constructed an example result page.

Conclusion

MIRIA provides both visualization and analysis of mammal RNA editing data for experimental biologists who are interested in revealing the functions of RNA editing sites. MIRIA is freely available at https://mammal.deepomics.org.

_

dbMTS: a comprehensive database of putative human microRNA target site SNVs and their functional predictions

Chang Li1†, Michael D. Swartz2, Bing Yu1, Xiaoming Liu1†*

1Human Genetics Center and Department of Epidemiology, Human Genetics and Environmental Sciences, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX

2Department of Biostatistics, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX
†Current address: USF Genomics, College of Public Health, University of South Florida, Tampa, FL
***Correspondence:**
Xiaoming Liu, xiaomingliu@health.usf.edu
Full list of author emails is available at the end of the article

Abstract

microRNAs (miRNAs) are short non-coding RNAs that can repress the expression of protein coding messenger RNAs (mRNAs) by binding to the 3'UTR of the target. Genetic mutations such as single nucleotide variants (SNVs) in the 3'UTR of the mRNAs can disrupt this regulatory effect. In this study, we presented dbMTS, the database for miRNA target site (MTS) SNVs, that include all potential MTS SNVs in the 3'UTR of human genome along with hundreds of functional annotations. This database can help studies easily identify putative SNVs that affect miRNA targeting and facilitate the prioritization of their functional importance. dbMTS is freely available at: https://sites.google.com/site/jpopgen/dbNSFP.

_

A harmonized neurodegenerative transcriptome database to nominate mouse models for functional follow-up and validation of Alzheimer's gene networks Rami

Al-Ouran [1,2], Ying-Wooi Wan [1,2], Zhandong Liu [1,2]

1 Department of Pediatrics, Baylor College of Medicine, Houston, TX 77030, USA
2 Jan and Dan Duncan Neurologic Research Institute, Texas Children's Hospital, Houston, TX 77030, USA
Email addresses:
RA: rami.al-ouran@bcm.edu YW: yingwoow@bcm.edu ZL: zhandonl@bcm.edu

Abstract

Background

Target nomination for Alzheimer's disease (AD) drug development has been a major challenge in the path to finding a cure for AD. Extensive transcriptome profiles of healthy and AD brains have revealed many gene changes, and functional validation on these changes is a critical next step. Model organisms are utilized to improve our understanding of the disease and to help in screening of genes for therapy development. Several model organisms exist for studying AD and selecting the best model organism for validation and functional follow-up represents a challenge. To help in nominating the best AD mouse model, we processed transcriptomic data for hundreds of mouse models through a unified pipeline and made it available through a web interface for users.

Results

Here we present a web interface to processed mouse transcriptomic datasets related to neurological disorders where users can examine genes across hundreds of mouse model studies to help in selecting the optimal model for further investigations. We processed 96 mouse studies related to AD, Parkinson's disease (PD), Huntington's disease (HD), Amyotrophic Lateral Sclerosis (ALS), Spinocerebellar ataxia (SCA) and models related to

aging. The reprocessed studies were made available through a web interface where a user can input a list of genes of interest and quickly check the gene expression across the reprocessed mouse studies.

Conclusions

The Mouse neurodegenerative database is available at: https://alouranbcm.shinyapps.io/m_db_1/. Users can quickly learn about different behaviors of genes across different mouse models and help in nominating the optimal mouse model for experimental validation.

_

Forming Big Datasets through Latent Class Concatenation of Imperfectly Matched Databases Features

Christopher W Bartlett[1,2§], Brett G Klammer [1*], Steven Buyske[3], Stephen A Petrill[4], William C Ray[1,2]

[1]Battelle Center for Mathematical Medicine, The Abigail Wexner Research Institute, Nationwide Children's Hosptial, Columbus, Ohio, USA

[2]Department of Pediatrics, College of Medicine, The Ohio State University, Columbus, Ohio, USA

[3]Departments of Statistics and Genetics, Rutgers University, Piscataway, NJ, USA

[4]Department of Psychology, College of Arts and Sciences, The Ohio State University, Columbus, Ohio, USA

[§]Corresponding author
*Current Affiliation: Center for Biostatistics, The Ohio State University, Columbus, Ohio, USA

Email addresses:
CWB: Christopher.Bartlett@NationwideChildrens.org
BGK: Brett.Klamer@osumc.edu
SAP: petrill.2@osu.edu
WCR: willray@mac.com

Abstract

Background

Researchers often need to combine data from many different sources to increase statistical power and study more subtle or complicated effects than is possible using a single dataset. Perfect overlap of datasets is rare in academic studies since virtually every dataset is collected for a unique purpose and without coordination across parties not-at-hand (such as companies or other universities). When measurements across datasets are similar prima facie, there is a desire to combine the data for a single analysis. We developed a statistical method for combining datasets that measure the same underlying construct, though in similar but not identical ways. The advantage of combining datasets based on an empirically derived set of common metrics is simplicity and statistical power of a single joint analysis of all the data.

Results

Here we present our method, ROSETTA, to concatenate datasets based on empirically derived latent traits. Two simulation studies show the performance of ROSETTA concatenating imperfectly matched datasets to a condition of full information. The first study examined a wide range of correlation while the second study was modeled after the observed correlations in a well-characterized clinical, behavioral cohort.
Both studies consistently show correlations > 0.94, often >0.96 indicating the robustness of the method and validating the general approach.

Conclusions

Incomplete concordance of measurements across datasets poses a major challenge for researchers seeking to combine similar databases from different sources. In any given field, some measurements are fairly standard, but every organization collecting data makes unique decisions on instruments, protocols, and methods of processing thedata. This typically denies literal concatenation of the raw data, but mixing metrics can greatly reduce the sensitivity of the downstream analysis. We provide one alternative to a meta-analysis by developing a method that statistically equates similar but distinct manifest metrics into a set of empirically derived metrics that can be used for analysis across all datasets.

_

Challenges in proteogenomics: a comparison of analysis methods with the case study of the DREAM Proteogenomics Sub-Challenge

Tara Eicher[1], Andrew Patt[2], Esko Kautto[2], Raghu Machiraju[1,2]*, Ewy Mathé[2]*, Yan Zhang[2]*§

[1]Department of Computer Science and Engineering, The Ohio State University, Columbus, 43210, United States of America

[2]Department of Biomedical Informatics, The Ohio State University, Columbus, 43210, United States of America

§Corresponding Author

*These authors contributed equally to this work.

E-mail addresses:
TE: eicher.33@osu.edu
AP: patt.14@osu.edu EK: kautto.1@osu.edu
RM: machiraju.1@osu.edu EM:ewy.mathe@osumc.edu
YZ: yan.zhang@osumc.edu

Abstract

Proteomic measurements, which closely reflect phenotypes, provide insights into gene expression regulations and mechanisms underlying altered phenotypes. Further, integration of data on proteome and transcriptome levels can validate gene signatures associated with a phenotype. However, proteomic data is not as abundant as genomic data, and it is thus beneficial to use genomic features to predict protein abundances when matching proteomic samples or measurements within samples are lacking. Here we evaluate and compare three data-driven models for prediction of proteomic data from mRNA in breast and ovarian cancers using the 2017 DREAM Proteogenomics Challenge data. Our results show that Bayesian network and random forest approaches can predict protein

abundance levels with median ground truth - predicted correlation values between 0.29 and 0.55. Logic-based predictors are not as accurate overall, but perform well for a subset of proteins across multiple cross-validations. In this study, we not only benchmarked several machine learning approaches for predicting proteomic data, but also discussed the challenges and potential solutions in state-of-the-art proteogenomic analyses.

____
_

PATH: An interactive web platform for analysis of time-course high-dimensional genomic data

Yuping Zhang 1,3,6_, Yang Chen 2 and Zhengqing Ouyang 2,3,4,5_

1Department of Statistics, University of Connecticut, Storrs, CT, USA and
2The Jackson Laboratory for Genomic Medicine, Farmington, CT, USA and
3Institute for Systems Genomics, University of Connecticut, Storrs, CT, USA and
4Department of Biomedical Engineering, University of Connecticut, Storrs, CT, USA and
5Department of Genetics and Genome Sciences, University of Connecticut, Farmington, CT, USA and
6Center for Quantitative Medicine, University of Connecticut, Farmington, CT, USA

Abstract

Discovering patterns in time-course genomic data can provide insights on the dynamics of biological systems in health and disease. Here, we present a Platform for Analysis of Time-course High-dimensional data (PATH) with applications in genomics research. This web application provides a user-friendly interface with interactive data visualization, dimension reduction, pattern discovery, feature selection based on the Principal Trend Analysis. Furthermore, the web application enables interactive and integrative analysis of timecourse high-dimensional data based on the Joint Principal Trend Analysis. The utilities of PATH are demonstrated through simulated and real examples. PATH is freely accessible at http://ouyanglab.jax.org/path/.

*To whom correspondence should be addressed.
Email: yuping.zhang@uconn.edu; zhengqing.ouyang@jax.org.

____
_

LCLE: a web portal for comprehensive gene distance analysis for correlation networks in liver cancer

Xiuquan Wang#,1, Xiaoqian Zhu#,2, Yunyun Zhou*,2

1 Department of Mathematics and Computer Science, Natural Science Division, Tougaloo College, 500 W. County Line Rd, Jackson, MS, 39174, United States
2 Department of Data Science, John D. Bower School of Population Health, University of Mississippi Medical Center, 2500 North State Street, Jackson, MS, 39216, United States

* Corresponding to: yzhou@umc.edu;
# The first two shared the co-first authors.

Abstract

Most of the currently available co-expression network analysis method only can capture linear correlation among genes; however, ignore the non-linear dependent correlations. Accurately and easily getting the distance values among genes are of significant importance in clustering genes which are shared in the same biological functions. We developed an online tool, LCLE, which is able to systematically analyze gene expression data to identify more comprehensive relationships among lncRNAs and protein-coding genes (PCGs) from five different distances metrics. Our simulation results demonstrated that the selection of an appropriate distance method could help to identify novel important genes from networks. Users can download and visualize figures, and distance tables analyzed from publically available RNAseq data such as TCGA and GTEx or upload their own data for analysis. Overall, our web portal will benefit for biologists or clinicians without programming background in identifying novel co-regulation relations for lncRNAs and PCGs.

Keywords: adjacent matrix, co-expression network, correlation, non-coding RNA, liver cancer

Concurrent Session - Computational Drug Discovery
Tuesday, June 11, 2019
10:10 AM – 12:20 PM
Pfahl 302

A deep learning-based method for drug-target interaction prediction based on long short-term memory neural network

Yan-Bin Wang[1,2, #] , Zhu-Hong You[1 #,*], Shan Yang[1], Hai-Cheng Yi[1,2], Zhan Heng Chen[1,2], Kai Zheng[1]

[1]Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Science, Urumqi 830011, China.

[2]Department of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 100049, China.
*Correspondence: zhuhongyou@ms.xjb.ac.cn (Z.Y.); Tel.: +86 18160622862(Z.Y.);
#These authors contributed equally to this work.

Email addresses:

YBW: wangyanbin15@mails.ucas.ac.cn
ZHY: zhuhongyou@ms.xjb.ac.cn
YS: yangshan16@mails.ucas.ac.cn
HCY: yihaicheng17@mails.ucas.ac.cn
ZHC: chenzhanheng17@mails.ucas.ac.cn
KZ: zhengkai@ms.xjb.ac.cn

Abstract

Background：The key to modern drug discovery is to find, identify and prepare drug molecular targets. However, due to the influence of throughput, precision and cost, traditional experimental methods are difficult to be widely used to infer these potential Drug-Target Interactions (DTIs). Therefore, it is urgent to develop effective computational methods to validate the interaction between drugs and target.
Results: In this paper, we developed a deep neural network model with memory and Turing completeness for DTIs prediction using drug substructure and protein sequence. More specifically, the proteins evolutionary features are extracted via Position Specific Scoring Matrix (PSSM) and Legendre Moment (LM) and associate it with drugs molecular substructure fingerprints to form feature pairs of drug-target. Then we utilize the Sparse Principal Component Analysis (SPCA) to reduce the feature dimension and eliminate the noise. Lastly, the deep long short-term memory (DLSTM) neural network with memory and Turing completeness is constructed for carrying out prediction. A significant improvement in DTIs prediction performance can be observed on experimental results, with AUC of 0.9951, 0.9705, 0.9951, 0.9206, respectively, on four classes important drug-target datasets.
Conclusion: The results demonstration that the proposed approach is competitive and have

a great advantage over state-of-the-art drug-target prediction techniques. To the best of our knowledge, this study first tests the potential of deep learning method with memory and Turing completeness in DTIs prediction.

Keywords: Drug-target, Deep learning, Legendre moment, Sparse principal component analysis

---

Mining and visualizing high-order directional drug interaction effects using the FAERS database

Xiaohui Yao[1,†], Tiffany Tsang[2,†], Sara Quinney[3], Pengyue Zhang[4], Xia Ning[4], Lang Li[4] and Li Shen[1*]

Abstract

Background: Adverse drug events (ADEs) often occur as a result of drug-drug interactions (DDIs). The use of data mining for detecting effects of drug combinations on ADE has attracted growing attention and interest, however, most studies focused on analyzing pairwise DDIs. Recent efforts have been made to explore the directional relationships among high-dimensional drug combinations and have shown effectiveness on prediction of ADE risk. However, the existing approaches become inefficient from both computational and illustrative perspectives when considering more than three drugs.

Results: We proposed an efficient approach to estimate the directional effects of high-order DDIs through frequent itemset mining, and further developed a novel visualization method to organize and present the high-order directional DDI effects involving more than three drugs in an interactive, concise and comprehensive manner. We demonstrated its performance by mining the directional DDIs associated with myopathy using a publicly available FAERS dataset. Directional effects of DDIs involving up to seven drugs were reported. Our analysis confirmed previously reported myopathy associated DDIs including interactions between fusidic acid with simvastatin and atorvastatin. Furthermore, we uncovered a number of novel DDIs leading to increased risk for myopathy, such as the co-administration of zoledronate with different types of drugs including antibiotics (ciprofloxacin, levofloxacin) and analgesics (acetaminophen, fentanyl, gabapentin, oxycodone). Finally, we visualized directional DDI findings via the proposed tool, which allows one to interactively select any drug combination as the baseline and zoom in/out to obtain both detailed and overall picture of interested drugs.

Conclusions: We developed a more efficient data mining strategy to identify high-order directional DDIs, and designed an scalable tool to visualize high-order DDI findings. The proposed method and tool have the potential to contribute to the drug interaction research and ultimately impact patient health care.

SCNrank: Spectral Clustering for Network-based Ranking to reveal potential drug targets and its application in pancreatic ductal adenocarcinoma

Enze Liu[1], Xiaoqi Liu[3], ZhuangZhuang Zhang[3], Xiaolin Cheng[4], Murray Korc[5, 6*] and Lijun Cheng[2*]

3 School of Informatics and Computing, Indiana University, Indianapolis, IN 46202

4 Department of Biomedical informatics, College of medicine, the Ohio State University, Columbus, OH 43210
5 Department of Toxicology and Cancer Biology, College of Medicine, University of Kentucky. Lexington, KY 40536.
6 College of Pharmacy, Division of Medicinal Chemistry and Pharmacognosy, the Ohio State University, Columbus, OH 43210
7 Department of Medicine, Biochemistry and Molecular Biology, School of Medicine, Indiana University, Indianapolis, IN, 46202
8 Pancreatic Cancer Signature Center, School of Medicine, Indiana University, Indianapolis, IN, 46202
*Corresponding author: Lijun Cheng, Lijun.cheng@osumc.edu and Murray Korc, mkorc@iu.edu

Contact information: Enze Liu, enzeliu@iu.edu;
Xiaoqi Liu, Xiaoqi.Liu@uky.edu;
ZhuangZhuang Zhang, Zhuangzhuang.Zhang@uky.edu; Murray Korc, mkorc@iu.edu;
Xiaolin Cheng, cheng.1302@osu.edu
Lijun Cheng, Lijun.Cheng@osumc.edu

Abstract

Background

Pancreatic ductal adenocarcinoma (PDAC) is the most common pancreatic malignancy. Due to its wide heterogeneity, PDAC acts aggressively and responds poorly to most chemotherapies, causing an urgent need for developing new therapeutic strategies. Cell lines have been used as the foundation for drug development and disease modeling. CRISPR-Cas9 plays as a key tool for every step-in drug discovery: from target identification and validation to preclinical cancer cell testing. Using cell-line models and CRISPR-Cas9 technology together makes drug targets prediction feasible. However, there is still a big gap between prediction results and actionable targets in real tumor. Biological network models provide great modus to mimic genetic interactions in real biological systems, which can benefit gene perturbation study and potential targets identification for treating PDAC. Nevertheless, building a network model that takes cell-line data and CRISPR-Cas9 as input to accurately predict potential targets that will respond well on real tissue remains unsolved.

Methods

We developed a novel algorithm 'Spectral Clustering for Network-based target Ranking' (SCNrank) that systematically integrate three types of data: expression profiles from tumor tissue, normal tissue and cell-line PDAC, protein-protein interaction network (PPI) and CRISPR-Cas9 data to prioritize potential drug targets for PDAC. The whole algorithm can be classified into three steps: 1. using STRING PPI network skeleton, SCNrank constructs tissue integrated networks with tumor and normal tissue PDAC expression profile; 2. With the same network skeleton, SCNrank constructed cell-line integrated networks using cell-line PDAC expression profile and CRISPR- Cas 9 data. 3. SCNrank applies a novel spectral clustering approach to reduce data dimension and generate gene clusters that carry common features from both networks. Finally, SCNrank applies a scoring scheme called 'Target Influence score' (**TI**), which estimates a given target's influence towards the cluster it belongs to, for scoring and ranking each drug targets.

Results

We applied SCNrank to analyze data 263 gene expression profiles, CRPSPR-Cas9 data from 22 different pancreatic cancer cell-lines and STRING protein-protein interaction (PPI) network. With SCNrank, we successfully constructed an integrated tissue PDAC network and an integrated cell- line PDAC network, both of which contains 4,414 selected genes that are overexpressed in tumor tissue samples. After clustering, 4,414 genes are distributed in 198 clusters, which include 367 targets of FDA approved drugs. These drug targets are all scored and ranked by their TI scores, which we defined to measure their influence towards the network. We validated top-ranked targets in three aspects: Firstly, mapping them onto the existing clinical drug targets of PDAC to measure the concordance. Secondly, we performed enrichment analysis to these drug targets and the clusters there are within, to reveal functional associations between clusters and PDAC; Thirdly, we performed survival analysis for the top-ranked targets to connect targets with clinical outcomes. Survival analysis reveals that overexpression of three top-ranked genes, PGK1, HMMR and POLE2, significantly increase the risk of death in PDAC patients.

Conclusion

SCNrank is an unbiased algorithm that systematically integrates multiple types of omics data to do potential drug target selection and ranking. SCNrank shows great capability in predicting drug targets for PDAC. Candidate pancreatic cancer-associated genes predicted by our SCNrank bapproach have the potential to guide genetics-based anti-pancreatic drug discovery.

____

Network as a biomarker: A novel network-based sparse Bayesian machine for pathway-driven drug response prediction

Lei Frank Huang1,2,3*, Hongting Liu1, Yi Zheng1,2, Richard Lu1,2

1Department of Pediatrics, Brain Tumor Center, Division of Experimental Hematology and Cancer Biology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH 45229, USA.
2Department of Pediatrics, College of Medicine, University of Cincinnati, Cincinnati, OH 45229, USA.
3Department of Information Science, School of Mathematical Sciences and LMAM, Peking University, Beijing, 100871.
*Corresponding author, Lei.Huang@cchmc.org

Abstract:

Background: With the advances in different biological networks, including gene regulation, gene co-expression, protein-protein interaction networks, and advanced approaches for network reconstruction, analysis, and interpretation, it is possible to discover reliable and accurate molecular network-based biomarkers for monitoring cancer treatment. Such efforts will also pave the way toward the realization of biomarker-driven personalized medicine against cancer.

Result: Previously, we have reconstructed disease-specific driver signaling networks using multi-omics profiles and cancer signaling pathway data. In this study, we developed a network-based sparse Bayesian machine (NBSBM) approach, using previously derived disease-specific driver signaling network to predict the cancer cell response to drugs. We have compared the proposed method with network-based support vector machine (NBSVM) approaches. We found the NBSBM approach can achieve much better accuracy than the other two NBSVM methods. The gene modules selected from the disease-specific driver networks for predicting the drug response might be directly involved in drug sensitivity or resistance.

Conclusions: This work provides a disease-specific network-based drug sensitivity prediction approach and it can uncover the potential mechanisms of action of drugs by selecting the most predictive sub-networks from the disease-specific network.

Availability: The source code is available at https://github.com/Roosevelt-PKU/

Computational Drug Repositioning for Precision Cancer Medicine Based on Cancer Cells Screening

Majumdar Abhishek[1#], Shaofeng Wu [1#], Yaoqin Lu[2#], Enze Liu[3], Tao Han[1], Yang Huo[3] and Lijun Cheng[1*]

1 Department of Biomedical Informatics, College of Medicine, The Ohio State University, Columbus, OH 43210
2 Department of occupational and Environmental Health, School of Public Health, XinJiang Medical University, Urumqi 830011, Xinjiang Uygur Autonomous Region, China
3 School of Informatics and Computing, Indiana University, Indianapolis, IN 46202

#Co-first authors with equal contributions

***Corresponding author**: Lijun Cheng (lijun.cheng@osumc.edu)
Abhishek Majumdar, abhishek.maj08@gmail.com; Abhishek.Majumdar@osumc.edu
Shaofeng Wu, wu.2946@buckeyemail.osu.edu
Yaoqin Lu, lyq_superior@163.com
Enze Liu, enzeliu@iu.edu
Tao Han, Tao.Han@osumc.edu Yang Huo, yanghuo@iu.edu
Lijun Cheng, Lijun.cheng@osumc.edu

Abstract

Background: Cancer cell lines are frequently used to research as in-vitro tumor models. Genomic data and large-scale drug screening have accelerated and facilitated drug repositioning for cancer research. However, the issue of how to utilize these genome mRNA data with drug response from cancer cells to predict drug reposition at the level of single patient has not been addressed. In this study, two important questions are focused on: (1) How to identify optimum cancer cell lines as individual tumor model by transcriptome genomic comparison systematically; (2) How to identify optimum therapeutic interventions for the single patient from a large scale of cancer cells' drug screening.

Methods: Molecular data was collected from Cancer Cell Line Encyclopedia (CCLE), which included 1097 gene expression profiles with 23,316 gene. The common 610 cancer cell lines distributed in 29 tissues were used for further data analysis. Cancer Therapeutics Response Portal (CTRP) provides 481 drug screening for sensitivity responses across the 610 cancer cell lines. A new computational method for drug repositioning was developed to predict the best fitting cancer cells for optimal drug efficacy in precision cancer medicine. Firstly, a *K*-means clustering with an optimum *t*-Distributed Stochastic Neighbor Embedding (*t*-SNE) is used to detect the similarity between cancer cells and a single tumor. The novelty integrating technology can map samples' similarity from a high-dimensional space to a lower-dimensional visualization space faithfully to preserve a measure of similarity, and then seeking the consistence clusters of the optimal cancer cells for the single patient. Secondly, by ranking the average of AUC (Area Under drug response Curve) values of each drug across these similar cell lines, we prioritized these top efficacy drugs for the single patient.

Results: A hundred patients with triple negative breast cancer were observed their associated cancer drugs by the calculation platform. The average 30 cancer cells were selected as the most similar cells with the tumor in mRNA by our integrating method of K-means with updated t- SNE. The average ten optimum drugs out of 481 were recommended for the single patient by our AUC ranking system. Comparing with the Cancer Genome Atlas (TCGA) and clinical trial records, for the drugs paclitaxel (Taxol), and taxotere (Docetaxel), 92.59% and 85.71% of the results (respectively) kept accordance with the current research and observation.

Conclusion: A new computational method was developed to predict the drug repositioning for precision cancer medicine. The unique algorithm was based on cancer cells screening and systematic analysis: K-means concordance clustering with an updated t-SNE is developed to detect the similarity between cancer cells and a single tumor in a reduce dimension visualization space. A ranking system for drug response is conducted to deliver potential efficacy drugs for a single patient by ranked AUCs from a large-scale screening of cancer cells. The research provides a novelty useful tool to detect efficacy drugs for the right patient by genomic omics data.

_

Development of a RNA-Seq based Prognostic Signature for Colon Cancer

Bjarne Bartlett1,2, Yong Zhu3, Mark Menor1, Vedbar S. Khadka1, Jicai Zhang4, Jie Zheng5, §, Bin Jiang3§ , Youping Deng1,2§

1. Bioinformatics Core, Department of Complementary and Integrative Medicine, University of Hawaii, Honolulu, HI 96813 USA
2. Molecular Biosciences & Bioengineering, University of Hawaii, Honolulu, HI 96822 USA
3. National Medical Centre of Colorectal Disease, The Third Affiliated Hospital of Nanjing University of Chinese Medicine, Nanjing, China.
4. Department of Laboratory Medicine, Shiyan Taihe Hospital, College of Biomedical Engineering, Hubei University of Medicine, Shiyan, Hubei 442000, P.R. China
5. Department of Pathology, Jinmen First Hospital, Jinmen, Hubei, P.R. China

§Corresponding Author
Email addresses:
bjarne@hawaii.edu
zhuyong839@sina.com
mmenor@hawaii.edu
vedbar@hawaii.edu
jicaizhang@taihehospital.com

Zhengjie9803@163.com
jbfirsth@aliyun.com
dengy@hawaii.edu (corresponding)

Abstract

Background

RNA-Seq data has recently been used to successfully develop prognostic signatures to predict cancer patients who will have a worse prognosis. We designed a study to ascertain whether a prognostic model, based on RNA-Seq data, would have clinical utility for predicting survival in patients with colon adenocarcinomas (COAD). Of particular interest are early stage COAD patients, for whom the benefits of adjuvant therapy are unclear.

Methods

Data from 468 COAD patients from The Cancer Genome Atlas (TCGA) were obtained and divided into two datasets: training (n=312) and validation (n=156). The training cohort was used to develop a prognostic signature by using univariate cox analysis to assess the prognostic potential of each gene and subsequently building a prognostic model using multivariate cox analysis. Patient survival was compared between different risk groups based on predictions from our model and cancer stage.

Results

In the training cohort, univariate cox analysis identified 15 genes ($p<0.05$) with prognostic potential. From this group, multivariate cox analysis generated a 5 gene signature ($p<0.05$) that included 2 long noncoding RNAs (lncRNAs). A score threshold (risk score > 0.432) representing the top 25% of the training cohort was used to identify high-risk patients with shorter survival times. High-risk patients predicted by our 5-gene, RNA-Seq signature had significantly shorter survival in both the training ($p=0.00$) and test ($p=0.003$) cohorts.

Conclusions

Here we present an RNA-Seq prognostic signature that can identify high-risk COAD patients with shorter survival. This signature would have clinical utility as part of an RNA-seq screening program, particularly for identifying early stage COAD patients who could be recommended adjuvant therapy after resection based on the results of this prognostic signature.

____

_

Machine Learning Distilled Metabolite Biomarkers for Early Stage Renal Injury

Yan Guo1*, Hui Yu1, Danqian Chen2, Ying-Yong Zhao2*,

1Department of Internal Medicine, University of New Mexico, Albuquerque, NM, USA
2Key Laboratory of Resource Biology and Biotechnology in Western China, Ministry of Education, School of Life Sciences, Northwest University, No. 229 Taibai North Road, Xi’an, Shaanxi 710069, China
Corresponding authors: Yan Guo, Ying-Yong Zhao

Abstract

Early prediction and accurate monitoring of chronic kidney disease (CKD) may improve care and decrease the frequent progression to end-stage renal disease. There is an urgent demand to discover specific biomarkers that allow for monitoring of early-stage CKD, and response to treatment. To discover such biomarkers, shotgun high throughput was applied to the detection of serum metabolites biomarker discovery for early stages of CKD from 703 participants. Ultra performance liquid chromatography coupled with high-definition mass spectrometry (UPLC-HDMS)-based metabolomics was used for the determination of 703 fasting serum samples from five stages of CKD patients and age-matched healthy controls. We discovered a set of metabolite biomarkers using a series of classic and neural network based machine learning techniques. This set of metabolites can separate early CKD stage patents from normal subjects with high accuracy. Our study illustrates the power of machine learning methods in biomarker study.

Keywords: Chronic kidney disease, glomerular filtration rate, metabolite, machine learning, deep learning

Development of predictive models to distinguish metals from non-metal toxicants, and individual metal from one another

Zongtao Yu1,*, Yong Zhu2,*, Junmei Ai3, Bjarne Bartlett4, Jicai Zhang1,#, Bin Jiang2,#, Youping Deng4,#

1Department of Laboratory Medicine, Shiyan Taihe Hospital, College of Biomedical Engineering， Hubei University of Medicine，Shiyan 442000, Hubei, China.
2National Center of Colorectal Disease, Nanjing Municipal Hospital of Chinese Medicine, the Third Affiliated Hospital, Nanjing University of Chinese Medicine, Nanjing, China. 210001
3Departments of Internal Medicine, Rush University Medical Center, Chicago, IL 60612.
4 Department of Complementary & Integrative Medicine, University of Hawaii John A. Burns School of Medicine, Honolulu, HI 96813, USA

*These authors contributed equally to this paper.
#**Corresponding Author:** Youping Deng, Ph. D., Department of Complementary & Integrative Medicine, University of Hawaii John A. Burns School of Medicine, 651 Ilalo Street, Honolulu, Hawaii 96813,Tel: 808.692.1664; Fax: 808.692.1984 Email: dengy@hawaii.edu

Abstract

Background: Toxic heavy metals are well-known environmental pollutants. Microarray classifier analysis has shown promise in the toxicogenomics area in identifying biomarkers to predict unknown samples and help understand toxic mechanisms. We aim to identify gene markers and build predictive models to distinguish metals from non-metal toxicants and individual metal from one another.

Methods: Gene expression profiles were generated from cultured rat primary hepatocytes that were treated with 105 distinct compounds including 9 heavy metals (selenium, chromium, arsenic, lead, cadmium, nickel, zinc, copper and tungsten) and their respective vehicle controls. Microarray classifier analyses were conducted by comparing different feature types, sizes, and two feature selection methods based on the classification algorithm of LibSVM.

Results: Based on an independent dataset test, we found that using only 15 gene markers, we were able to distinguish metals from non-metal toxicants with 100% accuracy. Of these, 6 and 9 genes were commonly down- and up-regulated respectively by most of the metals. More than half (8) of these 15 genes are membrane protein-coding genes. Genes in the list include ADORA2B, ARNT, S100G, and DIO3. We also identified a 10-gene marker list that can discriminate an individual metal from one another with 100% accuracy. We could find a specific gene marker for each metal in the 10-gene marker list. Genes in this list include GSTM2, HSD11B, AREG, and C8B.

Conclusions: Our results demonstrate that in using a microarray classifier for analysis can create diagnostic classifiers for predicting an exact metal contaminant from a large scale of contaminant pool with high prediction accuracy, but can also identify valuable biomarkers to help understand the common and underlying toxic mechanisms induced by metals.

Keywords: Biomarker, Microarray, Toxic heavy metals, Classification

DNA methylation markers for pan-cancer prediction by deep learning

Biao Liu1,2,3*, Yulu Liu1*, Xingxin Pan1, Mengyao Li2,3, Shuang Yang1§, Shuai Cheng Li4§

1BGI Education Center, University of Chinese Academy of Sciences, Shenzhen, China
2Shenzhen Key Laboratory of Forensics, BGI-Shenzhen, Shenzhen, China
3Forensic Genomics International (FGI), BGI-Shenzhen, Shenzhen, China
4Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong
*These authors contributed equally to this work
§Corresponding author

Email addresses:
Biao Liu: liubiao2@genomics.cn
Yulu Liu: liuyulu@genomics.cn
Xingxin Pan: panxingxin@genomics.cn
Mengyao Li: limengyao@genomics.cn
Shuang Yang: yangsh@genomics.cn
Shuai Cheng Li: shuaicli@gmail.com

Abstract

Background

For cancer diagnosis, many DNA methylation markers have been identified. However, few studies have tried to find DNA methylation markers to diagnose diverse cancer types simultaneously, i.e., pan-cancers. In this study, we tried to identify DNA methylation markers to differentiate cancer samples from the respective normal samples in pan-cancers.

Methods

We collected whole genome methylation data of 27 cancer types containing 10,140 cancer samples and 3,386 normal samples, and divided all samples into five data sets, including one training data set, one validation data set and three test data sets. We applied machine learning to identify DNA methylation markers, and specifically, we constructed diagnostic prediction models by deep learning.

Results

We identified two categories of markers: 12 CpG markers and 13 promoter markers. Three of 12 CpG markers and four of 13 promoter markers locate at cancer-related genes. With the CpG markers, our model achieves an average sensitivity and specificity on test data sets as 92.8% and 90.1%, respectively. For promoter markers, the average sensitivity and specificity on test data sets were 89.8% and 81.1%, respectively. Furthermore, in cell-free DNA methylation data of 163 prostate cancer samples and two unrelated normal samples, the CpG markers achieve both the sensitivity and specificity as 100%, and the promoter markers achieve 92% and 100% respectively. And for both marker types, specificity of normal whole blood is 100%.

Conclusions

To conclude, we identified methylation markers to diagnose pan-cancers, which might be applied to the liquid biopsy of cancers.

Keywords: biomarker, methylation, pan-cancer, deep learning, CpG, promote

DeepShape: Estimating Isoform-Level Ribosome Abundance and Distribution with Ribo-seq data

Hongfei Cui[1], Hailin Hu[2], Jianyang Zeng[3,*] and Ting Che

Abstract

Motivation: Ribosome profiling brings insight to the process of translation. A basic step in profile construction at transcript level is to map Ribo-seq data to transcripts, and then assign a huge number of multiple-mapped reads to similar isoforms. Existing methods either discard the multiple mapped-reads, or allocate them randomly, or assign them proportionally according to transcript abundance estimated from RNA-seq data.

Results: Here we present ***DeepShape***, an RNA-seq free computational method to estimate ribosome abundance of isoforms, and simultaneously compute their ribosome profiles using a deep learning model. Our simulation results demonstrate that ***DeepShape*** can provide more accurate estimations on both ribosome abundance and profiles when compared to state-of-the-art methods. We applied ***DeepShape*** to a set of Ribo-seq data from PC3 human prostate cancer cells with and without PP242 treatment. In the four cell invasion/metastasis genes that are translationally regulated by PP242 treatment, different isoforms show very different characteristics of translational efficiency and regulation patterns. Transcript level ribosome distributions were analyzed by "Codon Residence Index (CRI)" proposed in this study to investigate the relative speed that a ribosome moves on a codon compared to its synonymous codons. We observe consistent CRI patterns in PC3 cells. We found that the translation of several codons could be regulated by PP242 treatment. In summary, we demonstrate that ***DeepShape*** can serve as a powerful tool for Ribo-seq data analysis.

Availability and implementation: A software package in Python called ***DeepShape*** is freely available at
https://github.com/cuihf06/DeepShape.

Contact: tingchen@tsinghua.edu.cn, zengjy321@mail.tsinghua.edu.cn

On the analysis of the human immunome via an information theoretical approach.

Maciej Pietrzak[1], Gerard Lozanski[2], Michael Grever[2], Jeffrey Jones[2], Leslie Andritsos[2], James Blachly[2], Kerry Rogers[3] and Michal T Seweryn[4*]

Abstract

Background: Recent advances in the technology of flow cytometry allows to precisely identify hundreds of cell population using a limited number of antibodies. The markers expressed by immune cells carry significant information on the function and activation status of a cell. This information is crucial both for accurate description of the steady state behaviour of the immune system, as well as identification of deviations from this state towards disease phenotypes. Therefore, deep phenotyping of the cellular components of the immune system (the immunome) enables to gain new insight and decompose the multilayer immune network both in health and disease. The complexity of data produced by the flow cytometric analysis of the human immunome requires computational approaches that allow to detect not only the large-scale changes in the dominant components of the immunome, but, what seems more important, consistent differences in the non-abundant components and relations between them. At the same time, given that the analysis of the immunome is in most cases a ’small n, large p’ problem, it is crucial that the proposed method avoids producing false positive results with high likelihood.
Results: In this note, we build upon an approach of the authors developed in the context of T-cell antigen receptor repertoire analysis and develop an algorithm that scores cell populations by quantifying the amount of information that it carries about the case/control status in the context of the entire immunome. Using this approach we are able to measure the distance of a single ’case’ immunome for a set of ’controls’ without using any clustering/grouping approach and linkage function. We show that the information-based similarity measures we use are able to detect overlap between rare cell populations in the immunomes. We show that our feature selection algorithm is at least as sensitive to signal as other machine learning tools that are used for the analysis of high-dimensional flow cytometry data. At the same tim, ou methods, together with appropriate post-processing, retains low level of false positives. We also demonstrate, that we are able to identify a set of positive controls in a real-life immunome data from Hairy Cell Leukemia patients and detect other, biologically relevant cell populations in this context.
Conclusion: Our information-based feature selection algorithm for human immunome data enables to detect subtle differences between case and control immunomes with low false discovery rate. This is achieved by the measuring the amount of information that is encoded in a single population of cells in the context of the entire immunome.

Elimination of DNase nucleotide-specific bias to enhance recognition of DNA-binding proteins

Weixing Feng1, Chongchong Luo1, Duojiao Chen1, Weixin Xie1, Ruida Cong1, Chengkui Zhao1, Bo He1§, Yunlong Liu1,2§

1 Institute of Intelligent System and Bioinformatics, College of Automation, Harbin Engineering University, Harbin, Heilongjiang, 150001, China
2 Center for Computational Biology and Bioinformatics, School of Medicine, Indiana University, 410 W. 10th St., Suite 5000, Indianapolis, IN 46202, USA
§ Corresponding author

Abstract

Background

Recognition of DNA-binding proteins is essential in research of genomic regulation mechanisms relative to disease risk and treatment response. Along with many elaborate signals found within DNase footprint regions, the researchers attempted to identify DNA-binding proteins through such DNase signals. However, as a structural protein, the DNase enzyme has a nucleotide-specific binding tendency, which leads to bias in the signals and severely affects the recognition result.

Results

Through analyzing the distribution of the surrounding nucleotides, we found that nucleotides in ten consecutive positions around the DNase cutting point show distinctive distribution ratios from the others. To cope with a large number of possible nucleotide combinations, we built a deep recurrent neural network model to simulate the DNase nucleotide-specific binding tendency, and accordingly developed a method to eliminate the bias. In the end, based on the bias-eliminated DNase signals, we discriminated two typical DNA-binding proteins and obtained much better results.

Conclusions

Elimination of DNase nucleotide-specific bias obviously enhances the recognition of DNA-binding proteins. Besides DNase signals, the proposed method is also available for other similar signals, such as ATAC-seq signals (assay for transposase-accessible chromatin sequencing).

Keywords: DNase, Bias, Elimination, DNA, Protein, Binding, Recognition

---

RNASeqR: an R package for automated two-group RNA-Seq analysis workflow

Kuan-Hao Chao[1], Yi-Wen Hsiao[2], Yi-Fang Lee[3], Chien-Yueh Lee[2], Liang-Chuan Lai[2,4], Mong-Hsun Tsai[2,5,6], Tzu-Pin Lu[2,7,#], Eric Y. Chuang[1,2,3,#]

[1]Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan

[2]Bioinformatics and Biostatistics Core, Center of Genomic and Precision Medicine, National Taiwan University, Taipei, Taiwan

[3]Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, Taipei, Taiwan

[4]Graduate Institute of Physiology, National Taiwan University, Taipei, Taiwan

[5]Institute of Biotechnology, National Taiwan University, Taipei, Taiwan

[6]Center for Biotechnology, National Taiwan University, Taipei, Taiwan
[7]Institute of Epidemiology and Preventive Medicine, Department of Public Health, National Taiwan University, Taipei, Taiwan

#These authors contributed equally
Correspondence:
Tzu-Pin Lu
Institute of Epidemiology and Preventive Medicine, Department of Public Health, National Taiwan University
No. 17 Xu-Zhou Road, Taipei, Taiwan 10055
TEL: +886-2-3366-8042
FAX: +886-2-3322-4179
Email: tplu@ntu.edu.tw
Eric Y. Chuang
Graduate institute of Biomedical Electronics and Bioinformatics
National Taiwan University
No.1 Sec. 4 Roosevelt Road, Taipei, Taiwan 10617
TEL: +886-2-3366-3660
FAX: +886-2-3322-4170
Emai:chuangey@ntu.edu.tw
Emails:
Kuan-Hao Chao: b05901180@ntu.edu.tw
Yi-Wen Hsiao: y.w.hsiao9419@gmail.com
Yi-Fang Lee: r05945010@ntu.edu.tw
Chien-Yueh Lee: d00945006@ntu.edu.tw
Liang-Chuan Lai: llai@ntu.edu.tw
Mong-Hsun Tsai: motion@ntu.edu.tw
Tzu-Pin Lu: tplu@ntu.edu.tw
Eric Y. Chuang: chuangey@ntu.edu.tw

The Minimum Weight Clique Partition Problem and its Application to Structural Variant Calling

Matthew Hayes1 and Derrick Mullins1

1 Xavier University of Louisiana, New Orleans, LA 70125, USA
2 fmhayes5,dmullinsg@xula.edu

Abstract

The calling of genomic structural variants (SV) in high- throughput sequencing data necessitates prior discovery of abnormally aligned discordant read pair clusters that indicate candidate SVs. Some methods for SV discovery collect these candidate variants by heuristically searching for maximal cliques in an undirected graph, with nodes representing discordant read pairs and edges between vertices indicates that the read pairs overlap. In this paper, we consider the Minimum Weight Clique Partition Problem and its application to the problem of discordant read pair clustering. Our results demonstrate that methods which approximate or heuristically solve this problem can enhance the predictive abilities of structural variant calling algorithms.

Keywords: clique, structural variant, minimum weight clique, mini- mum weight clique partition

---

GPU Empowered Pipelines for Calculating High-Dimensional Kinship Matrices and Facilitating 1D and 2D GWAS

1. Wenchao Zhang[1], Xinbin Dai[1], Shizhong Xu[2], and Patrick X. Zhao[1*]*

[1]Noble Research Institute, LLC, 2510 Sam Noble Parkway, Ardmore, OK 73401, USA.

[2]Department of Botany and Plant Sciences, University of California, Riverside, CA 92521, USA

Wenchao Zhang, wezhang@noble.org;
Xinbin Dai, xdai@noble.org;
Shizhong Xu, shizhong.xu@ucr.edu;
Patrick X. Zhao, pzhao@noble.org;

[*]Corresponding Authors
Patrick X. Zhao, pzhao@noble.org; Phone: +1-580-224-6725
Shizhong Xu, shizhong.xu@ucr.edu; phone: +1-951-827-5898;

Abstract

Background: The high-throughput next generation sequencing (NGS) technology has enabled researchers to sequence and discover huge numbers of single nucleotide polymorphisms (SNPs) and to further explore diversities within species by constructing haplotype maps and conducting genome-wide association studies (GWAS). A typical GWAS study may deal with genotypic data for millions of SNPs. The first step of GWAS analysis is computing a kinship matrix, which describes the relationships among individuals. Conventional kinship matrix calculation mathematically has a complexity of $(\boldsymbol{mn^2})$, where $\boldsymbol{m}$ is the marker number and $\boldsymbol{n}$ is the number of individuals. In epistasis analyses, marker pairs are involved for kinship matrix calculation. For $\boldsymbol{m}$ markers, there are $\boldsymbol{c^2} = (\boldsymbol{m} - \boldsymbol{1})/\boldsymbol{2}$ marker pairs and calculation of an epistatic kinship matrix has a complexity of $(\boldsymbol{m^2n^2})$. When the sample and the number of markers are very large, calculation of various kinship matrices presents a huge computation burden. Fast computational strategies are urgently needed. Recently, a graphics processing unit (GPU) with multiple (>1,000) hardware processor cores has been used as a standard high-performance computing (HPC) solution system for large-scale computing. Especially, the GPU-empowered HPC platform is particularly essential for large-scale matrix operation. Kinship matrix calculation involves pure matrix operations and thus is suitable for GPU-employed parallel structure.

Methods: We thoroughly investigated the properties of various kinship matrices, including the main effect (additive and dominance) and epistatic effect kinship, which can be used for the 1D (main effect) and 2D (epistatic effect) GWAS analysis. We found that kinship matrix calculation is linear and thus we can divide the high-dimensional markers (main effect) and marker pairs (epistatic effect) into successive blocks. We then calculate the kinship matrix for each block and merge together the block-wise kinship matrices to form the genome-wide kinship matrix. All the matrix operations thus can be parallelized by GPU kernels.

Results: We developed two GPU-empowered pipelines: KMC1D and KMC2D for main effect kinship matrix calculation and epistatic effect kinship matrix calculation, respectively. A typical epistatic kinship matrix calculation with about 131 million marker pairs from 16,190 SNP markers and 1,390 individuals originally need several weeks in sequential model but only half an hour in KMC2D. Our simulation results for KMC1D and KMC2D show that the calculation speed can be improved 100 to 400 times over the conventional calculation without GPU parallel computing.

Conclusions: We believe that our GPU empowered pipelines have the capability to calculate the high dimensional kinship matrices, facilitating the main 1D and epistatic 2D GWAS analysis.

Concurrent Session - Cancer and Medical Genomics
Tuesday, June 11, 2019
2:20 PM – 3:10 PM
Pfahl 302

Rapid Evolution of Expression Levels in Hepatocellular Carcinoma

Fan Zhang*1,2, Michael D. Kuo*3

1 Vermont Genetics Network, University of Vermont, Burlington, Vermont 05405 USA
2 Department of Biology, University of Vermont, Burlington, Vermont 05405 USA
3 Departments of Radiological Sciences and Pathology, David Geffen School of Medicine at UCLA, Los Angeles, CA 90095-9717, USA

*Corresponding authors
Email addresses:
Fan Zhang: fan.zhang@uvm.edu
Michael D. Kuo: michaelkuo@mednet.ucla.edu

Abstract

Background
The human evolution and cancer evolution both have been researched for several years, but little is known about the molecular similarities between human and cancer evolution. One interesting and important question when comparing and analyzing human evolution and cancer evolution is whether cancer susceptibility is related to human evolution. There are a few microarray studies on human evolution or cancer development. Yet, to date, no microarray studies have been performed with both.

Results
Since cancer is an evolution on a small time and space scale, we compared and analyzed liver gene expression data among orangutan, chimpanzee, human, nontumor tissue, and primary cancer using linear mixed model, Analysis of Variance (ANOVA), Gene Ontology (GO), and Human Evolution Based Cancer Gene Expression Analysis.

Conclusions
Our results revealed not only rapid evolution of expression levels in hepatocellular carcinoma relative to the gene expression evolution rate of human, but also the correlation between hepatocellular carcinoma specific gene expression and human specific gene expression and between hepatocellular carcinoma conserved gene expression and human conserved gene expression. We also found a strong statistical relationship between gene function and expression pattern.

Identifying Interaction Clusters for MiRNA and MRNA Pairs in TCGA Network

Xinqing Dai2*, Lizhong Ding1*, Hui Jiang3, Yongsheng Bai1,4§

1Department of Biology, 2Department of Mathematics and Computer Science, Indiana State University, Terre Haute, IN 47809, U.S.A

3Department of Biostatistics, 4Department of Internal Medicine, University of Michigan, Ann Arbor, MI 48109, U.S.A

Xinqing Dai: dai.xinqing@outlook.com
Lizhong Ding: lding@sycamores.indstate.edu
Hui Jiang: jianghui@umich.edu
*These authors contributed equally to this work
§Correspondence should be addressed to
Yongsheng Bai: ybai@umich.edu

Abstract

Existing methods often fail to recognize the conversions for the biological roles of the pairs of genes and miRNAs between the tumor and normal samples. We have developed a novel cluster scoring method to identify mRNA-miRNA interaction pairs and clusters while considering tumor and normal samples simultaneously. Our method has identified 54 significant clusters for 15 cancer types selected from The Cancer Genome Atlas project. We also determined the shared 2 clusters across tumor types and/or subtypes. In addition, we compared gene and miRNA overlap between lists identified in our Liver Hepatocellular Carcinoma (LIHC) study and ones reported from human and rat nonalcoholic fatty liver disease studies (NAFLD). Finally, we analyzed biological functions for the identified significant cluster in LIHC and uncovered a significantly enriched pathway (Phospholipase D Signaling Pathway) for six genes.

Keywords
miRNA, mRNA, Liver Hepatocellular Carcinoma, Gene regulation, Clustering algorithm, The Cancer Genome Atlas

---

Generating Simulated CGH and Sequencing Data to Assess Genomic Segmentation Algorithms

Mark Zucker1 and Kevin Coombes1

1Department of Biomedical Informatics, Wexner Medical Center, The Ohio State University

Abstract

Background: In order to validate methods for the analysis of high throughput data, it is necessary to obtain data for which the underlying truth is known, so one can verify the accuracy of inferences made by the method and thus quantify the confidence with which it can make inferences. Knowing the ground truth can be extraordinarily difficult in biology, since one can essentially never knows, even in highly controlled conditions, what proportion of cells have what aberrations in a bulk cell sample, particularly in populations of aberration-prone cancer cells. For this reason, the ability to simulate CGH array and DNA sequencing data that recapitulates the variance structure and population complexity of real biological samples would be very useful in assessing the accuracy of – and comparing – bioinformatics algorithms. In particular, we discuss here the use of segmentation algorithms to identify breakpoints and copy number variation in CGH array or sequencing data.

Results: We developed a tool, implemented in an R package, to simulate both 'ground truth' and realistic CGH array and/or SNV data. We present this tool and apply it to the assessment of several different approaches to segmentation of copy number data from CGH arrays, with a particular interest in detecting CNVs in cancer samples. We demonstrate that DNAcopy, an algorithm using circular binary segmentation, generally performs best, which is an agreement with previous research. We further determine the conditions under which it and other methods break down. In particular, we assess how characteristics such as clonal heterogeneity, the presence of nested CNVs, and the type of aberration affect algorithm accuracy.

Conclusion: The simulations we generated proved to be useful in determining not just the comparative overall accuracy of different algorithms, but also in determining how their efficacy is affected by the biological characteristics of samples from which the data was generated.

Mapping genes and pathways to age-associated psychological changes in humans using latent semantic analysis

Pankaj Singh Dholaniya*, Vikram Teja Naik, Baby Kumari

Department of Biotechnology and Bioinformatics, School of Life Sciences, University of Hyderabad, Telangana, India

Abstract:

The role of genes as the genetic marker provides a distinctive characteristic analogy to identify and pinpoint the cause and effect of various diseases. Thorough research done by the researchers on various diseases at the genome level gave exponential rise to scientific information, which is available on various databases. In the present study, we have performed the concept mining using latent semantic analysis on the literature abstracts discussing the various age-dependent psychological changes in humans. Prominent keywords are then taken from the derived concepts to search the relevant pathways in the KEGG database. The expression patterns of the genes belonging to these pathways were analyzed from the published age-dependent expression datasets and the potential genes have been identified.

Cross - species Conserved Proteins Complex Identification and Exploration of Species Functional Evolution

Xianjun Shen[1*], Meijun Zhou[1*], Limin Yu[1], Li Yi[1], Cuihong Wan[1], Tingting He[1], Xiaohua Hu[1, 2]

[1]School of Computer, CentralChinaNormalUniversity, Wuhan, China [2]College of Computing and Informatics, Drexel University, Philadelphia,USA xjshen@main.ccnu.edu.cn; zhmjunjun@163.com

Abstract

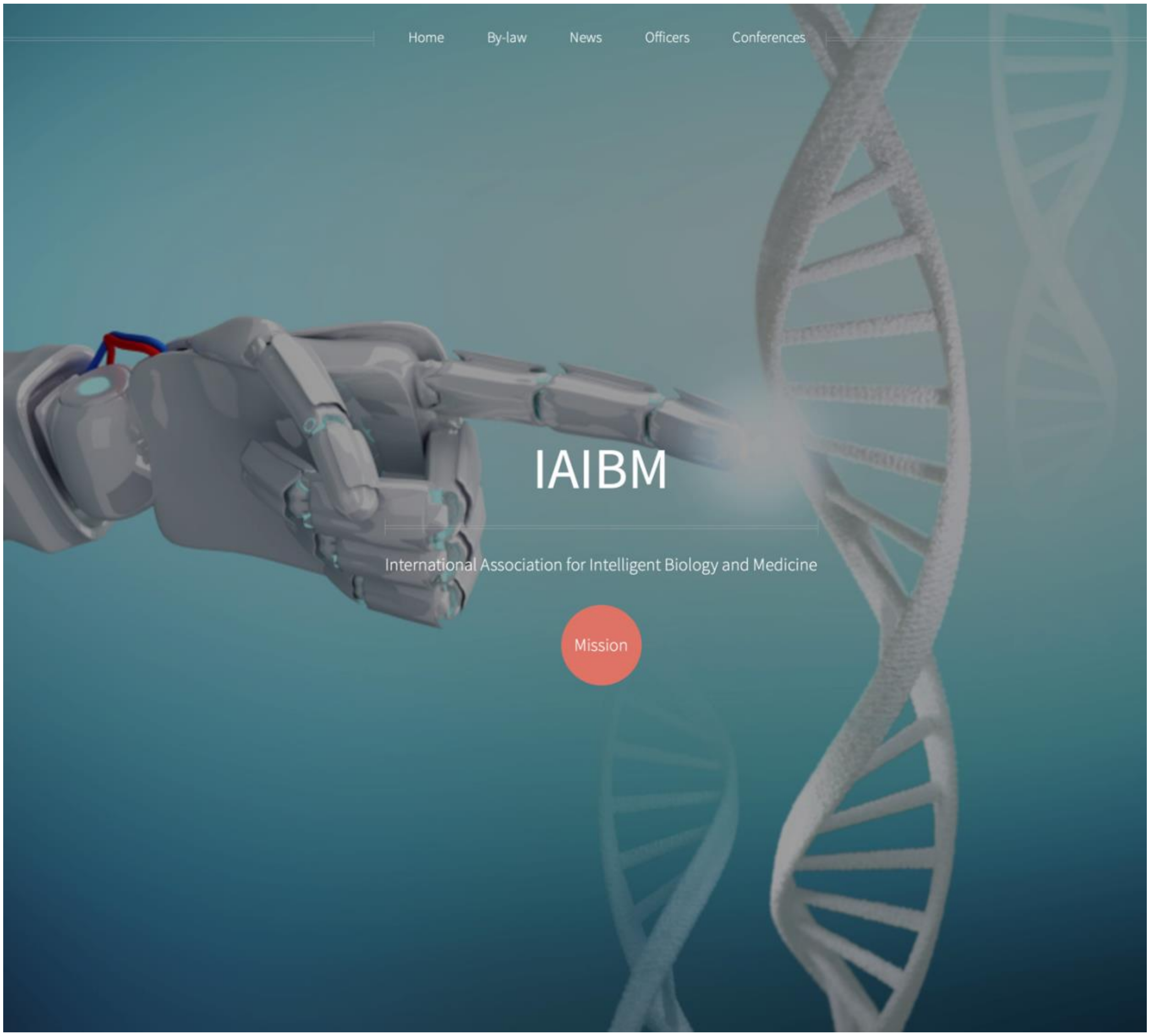
Traditionally, the network of protein interactions for single species, while helping to explore the cellular mechanisms of the organization, neglects the correlation between species. However, protein complexes of different species have functional commonality and individuality, so it is significant to study the interaction between species and species. In this paper, we propose a novel method to study species evolution at the protein level by the added components of protein complexes in single species. It was found that the specific components added in the evolution of species play an important role in the organism, such as SWI/SNF protein complex inhibits tumorigenesis in Mus musculus, but some components of this protein complex are newly added during evolution, which make Mus musculus distinguish from other species. In addition, the experimental results indicates that the specific newly added components have significantly enrichment of specific biological functions and do play an important role in the individual species.

Keywords: cross-species; conserved proteins complex; species functional evolution; functional enrichment

NOTES

About IAIBM

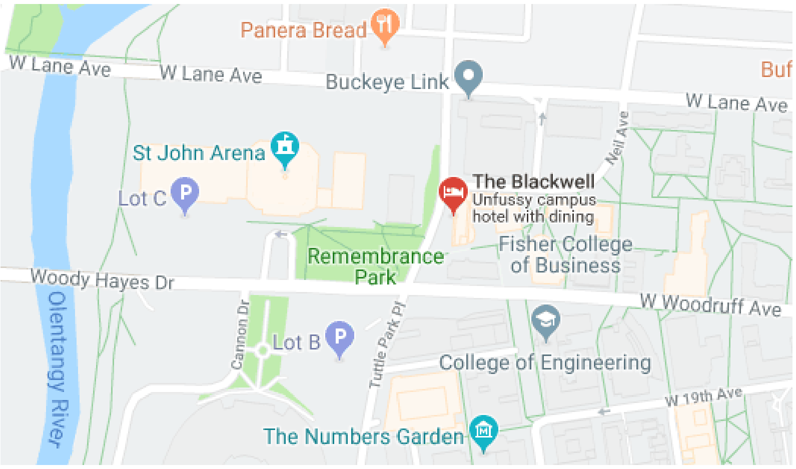
The International Association for Intelligent Biology and Medicine (IAIBM) is a non-profit organization. It was formed on January 19, 2018 with its founding president Dr. Zhongming Zhao. Its mission is to promote the intelligent biology and medical science, including bioinformatics, systems biology,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


Home
By-law
News
Officers
Conferences
IAIBM
International Association for Intelligent Biology and Medicine
Mission

Conference Location

The Blackwell Inn and Pfahl Conference Center
2110 Tuttle Park Place Columbus, Ohio 43210
Phone: (614) 247-4000

Impress your guests at our state-of-the-art conference and event space. The Blackwell Inn and Pfahl Conference Center, located on Ohio State's campus, is the ideal venue for a corporate event, training seminar, banquet, wedding, or conference. We offer about 20,000 square feet of meeting and event space throughout multiple event rooms, complete with cutting-edge technology and an expert staff to assist you in planning and executing the perfect event. Additionally, our hotel has 151 guestrooms and suites to accommodate all your event attendees.



Airport Information

John Glenn Columbus International Airport - located about 10 miles from Hotel, 15-20 minute drive time. Approximate one-way Uber/Lyft fare: $25-40.

Transportation

The Blackwell Inn has all your transportation needs covered. **We proudly offer complimentary, prearranged airport transportation (based on availability) to and from John Glenn Columbus International Airport (CMH) for our hotel guests**. Contact our Concierge or Guest Services Representative at (614) 247-4050 or (866) 247-4003 to schedule your airport transportation at least 24 hours in advance. You may also schedule your airport transportation using our e-Concierge.

Special Acknowledgments

We are grateful for the help from the following volunteers:

Megan BeVier
Jaquan Blue
Anita Bratcher
Judith Buster
Vishal Dey
Leticia Flores
Michelle Freeman
Gabrielle Kokanos
Mikayla McCormic
Andrew Patt
Susan Rojas
Jasmine Scott
Kyle Spencer
Xiao Zha

MANY THANKS TO OUR SPONSORS!

UTHealth®
The University of Texas
Health Science Center at Houston
School of Biomedical Informatics

UTHealth®
The University of Texas
Health Science Center at Houston

NSF
National Science Foundation
WHERE DISCOVERIES BEGIN

BGI
GLOBAL GENOMIC SERVICES

Humanizing Genomics
macrogen

Karger

CAP ACCREDITED
COLLEGE of AMERICAN PATHOLOGISTS

Clinical Laboratory Improvement Amendments

Whole Genome

Exome and Targeted

Transcriptome/RNA

Humanizing Genomics
macrogen

Epigenome

Metagenome

Sanger Sequencing

Sample Preparation

Bioinformatics Service

End-to-End Sequencing Solutions

Macrogen Corp.
inquiry@macrogenlab.com
(301) 251 - 1007

UTHealth®
The University of Texas
Health Science Center at Houston

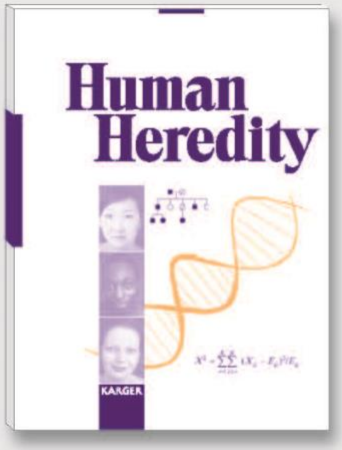
UTHealth

Established in 1972 by The University of Texas System Board of Regents, The University of Texas Health Science Center at Houston (UTHealth) is Houston's Health University and Texas' resource for health care education, innovation, scientific discovery and excellence in patient care. The most comprehensive academic health center in The UT System and the U.S. Gulf Coast region, UTHealth is home to schools of biomedical informatics, biomedical sciences, dentistry, nursing and public health and the John P. and Kathrine G. McGovern Medical School. UTHealth includes The University of Texas Harris County Psychiatric Center and a growing network of clinics throughout the region. The university's primary teaching hospitals include Memorial Hermann-Texas Medical Center, Children's Memorial Hermann Hospital and Harris Health Lyndon B. Johnson Hospital. As a comprehensive health science university, the mission of The University of Texas Health Science Center at Houston is to educate health science professionals, discover and translate advances in the biomedical and social sciences, and model the best practices in clinical care and public health.

CPH
Center for
Precision Health

Center for Precision Health

The Center for Precision Health (CPH) is a joint enterprise bridging School of Biomedical Informatics (SBMI) and School of Public Health (SPH) at UTHealth. CPH was established in January 2016, with the funds approved by the Texas Legislature to enhance biomedical and health informatics education and research in the State of Texas in the era of big data and precision medicine. Dr. Zhao was recruited from Vanderbilt University Medical Center to serve as its founding director. A synergizing entity, CPH will build upon the core strengths of SBMI (i.e., informatics programs, centers and clinical resources) and SPH (i.e., cohort-based research, statistical genomics, computational biology, environmental, behavioral and policy research programs) and develop both independent and collaborative research programs, as well as precision health resources, for UTHealth and other Texas Medical Center (TMC) institutions. CPH faculty will actively participate in educational programs and curriculum development at SBMI and SPH. In addition, CPH faculty will actively promote informatics and population health technology development. CPH currently has four high priority research areas: (1) Population-based Genomics for Precision Health, (2) Cancer Precision Medicine, (3) Translational Bioinformatics, and (4) Smart Clinical Trials.


Human Heredity
KARGER