INDIANA UNIVERSITY SCHOOL OF MEDICINE

CENTER FOR COMPUTATIONAL BIOLOGY AND BIOINFORMATICS

CCBB Annual Retreat (Virtual)

22 October 2021

Please register to obtain the ZOOM link:
https://iu.zoom.us/meeting/register/tZcuc-uhrDMqGdzb3iVHOmCdouimGc-spJmW

***9:00-9:10 am* – Opening**

***9:10-9:30 am* – State of the CCBB**

Dr. Yunlong Liu
Professor of Medical and Molecular Genetics, Biostatistics, and BioHealth Informatics
T. K. Li Chair for Medical Research
Director, Center for Computational Biology and Bioinformatics (CCBB)
Director, Center for Medical Genomics (CMG)
Indiana University School of Medicine

9:30-10:15 am* – Keynote talk I - *Genetic variation of RNA processing and modification in human tissues

Dr. Yi Xing
Professor & Executive Director, Department of Biomedical and Health Informatics
Director, Center for Computational and Genomic Medicine
The Children's Hospital of Philadelphia
Perelman School of Medicine, University of Pennsylvania

Dr. Yi Xing is the Francis West Lewis Endowed Chair and Founding Director of the Center for Computational and Genomic Medicine, as well as the Executive Director of the Department of Biomedical and Health Informatics at the Children's Hospital of Philadelphia (CHOP). Dr. Xing is also a Professor of Pathology and Laboratory Medicine at the University of Pennsylvania (Penn). Prior to his appointment at CHOP and Penn, Dr. Xing was a Professor of Microbiology, Immunology, and Molecular Genetics at UCLA, and served as Program Director of UCLA's Bioinformatics Interdepartmental Ph.D. Program. Dr. Xing received his BS in Molecular and Cellular Biology and BE in Computer Science and Technology from the University of Science and Technology of China (2001). He completed his PhD training in Bioinformatics with Dr. Christopher Lee at UCLA (2001-2006), and his postdoctoral training with Drs. Wing Hung Wong and Matthew Scott at the Stanford University (2006-2007). Dr. Xing has an extensive publication record in bioinformatics, genomics, and RNA biology. His work has provided fundamental insights into the function, regulation, and evolution of post-transcriptional RNA processing in mammals. His current research merges the fields of computational biology, biomedical data science, RNA genomics, human genetics, precision medicine, and immuno-oncology.

***10:15 am-11:15 am* – Concurrent sessions of CCBB trainee oral presentations**

	MAIN ZOOM LINK		BREAKOUT ZOOM ROOM	
TIME	**Speaker (PI)**	**Title**	**Speaker (PI)**	**Title**
10:15 am	Rudong Li (Yunlong Liu)	Investigating inheritability of alternative splicing in alcohol use disorders	Menghao Huang (Charlie Dong)	SIRT6 controls hepatic lipogenesis by suppressing LXR, ChREBP, and SREBP1
10:20 am	Linhui Xie (Jinwen Yan)	Integrative -omics for discovery of network-level disease biomarkers for Alzheimer's disease	Parth Kothiya (Sara Quinney)	IPDB: Integrated Pregnancy Database with clinical and omics data
10:25 am	Xiaoyu Lu (Sha Cao)	Detecting cell-type specific variations of within pathway gene interaction in AD using covariance regression	Quoseena Mir (Sarath Janga)	Identification of splice isoforms in T helper (Th) cells using Nanopore-based cDNA sequencing
10:30 am	Neil McCracken (Amber Mosley)	Needles in a stack of needles: Finding significant thermal shifts in thermal proteome profiling with inflect statistical analysis and bioinformatics	Melanie Cheung See Kit (Ian Webb)	Cross-linking reagents as molecular calipers to probe protein structure via gas-phase ion/ion reactions coupled to ion mobility/time-of-flight mass spectrometry
10:35 am	***Q&A***		**Q&A**	
10:45 am	Wenrong Chen (Xiaowen Liu)	Evaluation of deep learning models for retention and migration time prediction in top-down mass spectrometry	Xiaoqing Huang (Jie Zhang)	Understanding the uncoupling of tauopathy and dementia through comparative analysis of subgroups of atypical Alzheimer's disease patients
10:50 am	Abdul Rehman Basharat (Yong Zang)	A software tool for proteoform feature detection from top-down mass spectrometry data	Avery Runnebohm (Amber Mosley)	Thermal proteome profiling of triple negative breast cancer cells treated with IB-DNQ and Rucaparib
10:55 am	Garrick Chang (Jing Liu)	Image analysis pipeline for spatial and translational properties of RNA molecules in β-Cells of type 1 diabetes	Rebecca Cain (Ian Webb)	Time-resolved electrospray ionization combined with electron capture dissociation for characterization of protein folding intermediates
11:00 am	Mansi Srivastava (Sarath Janga)	Protein Occupancy Profile-sequencing (POP-seq) for global mapping of protein RNA interactions across human cell lines	Kailing Li (Jun Wan)	The updated version of global evaluation of SARS-CoV-2/nCoV-19 sequences (GESS) database
11:05 am	***Q&A***		**Q&A**	

***11:15-1:15 pm – Lunch break* & Online poster session (see abstracts and poster #'s below)**

GATHERTOWN link: https://gather.town/app/p72CjElJWxWYLo04/CCBB%20Retreat%20Room

***1:15-2:15 pm* – Best paper presentations** (back to Main ZOOM link)

Time	*Speaker (PI)*	*Title*
1:15 pm	Wennan Chang (Sha Cao & Chi Zhang)	A graph neural network model to estimate cell-wise metabolic flux using single-cell RNA-seq data. Genome Res. 2021; 31:1867-1884, doi:10.1101/gr.271205.120
1:30 pm	Hyeong Geug Kim (Charlie Dong)	Sestrin proteins protect against lipotoxicity-induced oxidative stress in the liver via suppression of C-Jun N-terminal kinases. *Cell Mol Gastroenterol Hepatol.* 2021;12(3):921-942, doi: 10.1016/j.jcmgh.2021.04.015.
1:45 pm	Chuanpeng Dong (Yunlong Liu)	Intron retention-induced neoantigen load correlates with unfavorable prognosis in multiple myeloma. *Oncogene* 2021; *In Press*, https://doi.org/10.1038/s41388-021-02005-y
2:00 pm	Tongxin Wang (Kun Huang)	MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. *Nat. Comm*. 2021 12, 3445, |doi:10.1038/s41467-021-23774.w

***2:15-3:45 pm* – Works in progress: 2020 CCBB Pilot grant awardees**

Time	*Speaker*	*Title*
2:15 pm	Dr. Evan Cornett (PI)	Toward global proteomic analysis of lysine methylation in triple-negative breast cancer
2:30 pm	Dr. Cristian Lasagna-Reeves (PI)	Identification of key genes/proteins in subpopulations of Alzheimer's disease patients with uncoupled tauopathy phenotype
2:45 pm	Yujing Li from the lab of Dr. Xiongbin Lu (PI)	Targeting Tregs in TNBC immunotherapy
3:00 pm	Dr. Ian Webb (PI)	Developing next generation tools for badly behaving proteins
3:15 pm	Dr. Lei Yang (PI)	SARS-CoV-2 gene-induced cardiac injury and dysfunction of human cardiomyocytes
3:30 pm	Dr. Jie Zhang (PI)	Understanding the transcriptional regulation of pathogenic microglia in AD

***3:45-4:00 pm* – Break**

4:00-4:45 pm* – Keynote talk II - *In-cell protein footprinting coupled with mass spectrometry for structural biology across the proteome

Dr. Lisa Jones
Associate Professor, Department of Pharmaceutical Sciences
University of Maryland School of Pharmacy

The Jones Lab is a structural proteomics group that uses biochemical, analytical, and biophysical approaches to study protein interactions important in biological processes. Our research focuses on the use of protein foot printing methods coupled with mass spectrometry to identify these interactions. We are especially interested in the use of the footprinting method fast photochemical oxidation of proteins. FPOP uses a pulsed excimer laser at 248 nm for photolysis of hydrogen peroxide of form hydroxyl radicals. These radicals them oxidize the side chains of amino acids. The power of the method lies in oxidatively modifying the samples in 2 states, such as ligand-bound and ligand-free. The modification differences between states provide information on ligand binding sites, protein interaction sites, and protein conformational change. We use FPOP to identify protein interactions and conformational change of various protein systems. We are especially interested in studying large, complex protein systems such as virus capsids and antibodies. A major focus of the lab is extending FPOP as an in cell method for monitoring proteins in their native cellular environment. This method would be especially useful for membrane proteins, the largest class of drug targets, which are challenging to study in vitro owing to the difficulty of purifying these proteins. Currently, we are developing in cell FPOP using the chimeric protein GCamP2 as a model system. The Jones Lab also utilizes protein footprinting methods coupled with mass spectrometry to characterize protein-protein interactions. We specifically use the hydroxyl radical footprinting method fast photochemical oxidation of proteins (FPOP) to study protein structure. In FPOP, hydroxyl radicals are generated by photolysis of hydrogen peroxide via an excimer laser.

***4:45-5:00 pm* – Announcements of CCBB Trainee Best Poster/Presentation Awards**

***5:00-6:00 pm* – Virtual social time**

GATHERTOWN link: https://gather.town/app/p72CjElJWxWYLo04/CCBB%20Retreat%20Room

Online poster session
11:15-1:15 pm

GATHERTOWN link: https://gather.town/app/p72CjEIJWxWYLo04/CCBB%20Retreat%20Room

Poster Numbers and Abstracts

11:15 – 12:15 pm Presentation/Discussion of P01 – P16
12:15 -1:15 pm Presentation/Discussion of P17 – P33

ID	Presenter	PI	Title
P01	Blessed W. **Aruldhas**	Sara Quinney	Relationship of the Metabolic Ratios of R and S-methadone with QT Lengthening
P02	Dominique A. **Baldwin**	Amber Mosley	Isobaric Trigger Channel-based Analysis of RNA Polymerase II Transcription Elongation Dynamics
P03	Katlyn Hughes **Burriss**	Amber Mosley	Disruption Compensation (DisCo) Analysis of Complex Protein Networks
P04	Abdul Rehman **Basharat**	Yong Zang	A Software Tool for Proteoform Feature Detection from Top-down Mass Spectrometry Data
P05	Rebecca **Cain**	Ian Webb	Time-Resolved Electrospray Ionization Combined with Electron Capture Dissociation for Characterization of Protein Folding Intermediates
P06	Garrick **Chang**	Jing Liu	Image Analysis Pipeline for Spatial and Translational Properties of RNA Molecules in β-Cells of Type 1 Diabetes
P07	Wennan **Chang**	Chi Zhang	Computational Modeling of Neurotransmitter Biosynthesis and Synapse by using Single-Cell, Tissue and Spatial Transcriptomics Data
P08	Andy B. **Chen**	Yunlong Liu	Functional Screening of 3'-UTR Variants Combined with Genome-wide Association Identifies Causal Regulatory Mechanisms Impacting Alcohol Consumption
P09	Wenrong **Chen**	Xiaowen Liu	Evaluation of Deep Learning Models for Retention and Migration Time Prediction in Top-down Mass Spectrometry
P10	Melanie **Cheung See Kit**	Ian Webb	Cross-linking Reagents as Molecular Calipers to Probe Protein Structure via Gas-phase Ion/Ion Reactions Coupled to Ion Mobility/Time-of-flight Mass Spectrometry

P11	Pengtao **Dang**	Sha Cao	Statistical Modeling of Spatial Dependent Functional Variations by using Spatial Transcriptomics Data
P12	Swapna Vidhur **Daulatabad**	Sarath Chandra Janga	RNA Structure Elucidation at Single Molecule Resolution using an Integrated Framework of Long Read Sequencing and SHAPE-seq
P13	Hunter **Gill**	Sarath Chandra Janga	RAZOR: A Central Database of PCR Primers Targeting Human Respiratory Viruses
P14	Ankita **Gurav**	Ian Webb	Charge Inversion Ion/Ion Reactions for Enhancing Ion Mobility-Mass Spectrometry Separations of Oligosaccharide Anions
P15	Dwight William **Hall**	Sarath Chandra Janga	SARS-CoV-2 Diversity and Phylogeny in Indiana COVID-19 Patients
P16	Doaa **Hassan**	Sarath Chandra Janga	Penguin: A Tool for Predicting Pseudouridine Sites in Direct RNA Nanopore Sequencing Data
P17	Menghao **Huang**	X. Charlie Dong	SIRT6 Controls Hepatic Lipogenesis by Suppressing LXR, ChREBP, and SREBP1
P18	Xiaoqing **Huang**	Jie Zhang	Understanding the Uncoupling of Tauopathy and Dementia Through Comparative Analysis of Subgroups of Atypical Alzheimer's Disease Patients
P19	Fadil **Iqbal**	Jing Liu	Observing Chromatin Dynamics at High Imaging Speeds using Deep Learning
P20	Guanglong **Jiang**	Yunlong Liu	The Genetic Regulation of the Biogenesis of Human isomiRs
P21	Parth G **Kothiya**	Sara Quinney	IPDB: Integrated Pregnancy Database with Clinical and Omics Data
P22	Alexander **Krohannon**	Sarath Chandra Janga	CASowary: CRISPR-Cas13 Guide RNA Predictor for Transcript Depletion
P23	Kailing **Li**	Jun Wan	The Updated Version of Global Evaluation of SARS-CoV-2/nCoV-19 Sequences (GESS) Database
P24	Rudong **Li**	Yunlong Liu	Investigating Inheritability of Alternative Splicing in Alcohol Use Disorders
P25	Enze **Liu**	Brian A. Walker	The Alternative Splicing Landscape In Multiple Myeloma Is Determined by IGH Translocations and Mutations of RNA Processing Genes
P26	Xiaoyu **Lu**	Sha Cao	Detecting Cell-type Specific Variations within Pathway Gene Interaction in AD using Covariance Regression
P27	Neil **McCracken**	Amber Mosley	Needles in a Stack of Needles: Finding Significant Thermal Shifts in Thermal Proteome

			Profiling with Inflect Statistical Analysis and Bioinformatics
P28	Quoseena **Mir**	Sarath Chandra Janga	Identification of Splice Isoforms in T helper (Th) Cells using Nanopore-based cDNA Sequencing
P29	Avery **Runnebohm**	Amber Mosley	Thermal Proteome Profiling of Triple Negative Breast Cancer Cells Treated with IB-DNQ and Rucaparib
P30	Mansi **Srivastava**	Sarath Chandra Janga	Protein Occupancy Profile-sequencing (POP-seq) for Global Mapping of Protein RNA Interactions Across Human Cell Lines
P31	Linhui **Xie**	Jingwen Yan	Integrative -omics for Discovery of Network-level Disease Biomarkers for Alzheimer's Disease
P32	Shaoyang **Huang**, Kevin **Hu**, Noah **Meroueh**, Jonathan **Yang**, Grace **Yang** (Carmel High School students)	Sha Cao, Chi Zhang	Computational Modeling of Metabolic Flux of Branched Chain Amino Acids in Cancer and Alzheimer's Disease
P33	Grace **Yang**, Jonathan **Yang**, Noah **Meroueh**, Kevin **Hu**, Shaoyang **Huang** (Carmel High School students)	Sha Cao, Chi Zhang	Computational Modeling of Cell Type Specific Metabolic Rate of Glucose Flow and Glutaminolysis in Cancer Microenvironment

P01
Relationship of the Metabolic Ratios of R and S-methadone with QT Lengthening

Blessed W. Aruldhas[1,2], Heather A. Jaynes[1], Brian R. Overholser[1,3], Sara K. Quinney[1], James E Tisdale[1,3]

[1]Indiana University School of Medicine, Indianapolis, Indiana, USA
[2]Christian Medical College, Vellore, India
[3]College of Pharmacy, Purdue University, Indianapolis, Indiana, USA

BACKGROUND:
Methadone-induced QT interval prolongation is mainly caused by its S-enantiomer, with some contribution from R-methadone. The metabolite ethylidene-1,5-demethyl-3,3-diphenylpyrrolidine (EDDP) only weakly inhibits hERG current. We assessed the relationship between QT prolongation and metabolic ratios [R, S and racemic methadone] and determined a risk threshold for the metabolic ratios.

METHODS:
In this prospective study, an electrocardiogram (ECG) was performed before initiating maintenance methadone therapy in 43 adults. A steady state ECG and blood sample were subsequently obtained. Bazett's-corrected QT (QTc) was measured manually. Metabolic ratios were calculated from the enantiomeric concentrations of methadone and EDDP. Receiver operating characteristics (ROC) curves were used to find a threshold for R, S and racemic metabolic ratios to predict QTc lengthening >30ms from baseline.

RESULTS:
There was an inverse relationship between the extent of QT lengthening and metabolic ratios on the 2D-densitogram. ROC analysis for predicting QTc>30 ms from baseline are presented in the Table.

Metabolic ratio	Metabolic ratio threshold	AUC under ROC curve	Specificity	Sensitivity	Accuracy	Precision
Racemic EDDP: racemic methadone	11.5	0.782	0.65	0.88	0.74	0.63
R-EDDP: R-methadone	8.3	0.779	0.77	0.82	0.79	0.70
S-EDDP: S-methadone	12.3	0.755	0.69	0.71	0.70	0.60

CONCLUSION:
Methadone metabolic ratios are reasonably accurate for identifying patients at risk for QT prolongation. Larger studies are required to evaluate methadone metabolic ratios for risk assessment.

P02

Isobaric Trigger Channel-based Analysis of RNA Polymerase II Transcription Elongation Dynamics

Dominique A. Baldwin[1], Katlyn H. Burriss[1], Whitney R. Smith-Kinnaman[1], I-Ju Yeh[1], Edward A. Motea[1], Amber L. Mosley[1,2]

[1] Department of Biochemistry and Molecular Biology, Indiana University School of Medicine, Indianapolis, Indiana 46202, United States

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

Data dependent acquisition (DDA) is a mass spectrometry in which the highest abundance peptide ions in a given time are selected for MS^2/MS^n based peptide sequencing. This approach has facilitated deep coverage of proteomes especially when coupled with biochemical fractionation approaches. However, variability in peptide elution/ co-elution and the large dynamic range of proteomes has led to challenges with shotgun proteomics for the detection of low abundance and/or heterogeneous protein species including valuable targets such as: sequence specific transcription factors, post-translationally modified proteins, and specific protein sequence alleles.

We have recently developed a new approach to tackle this problem using isobaric trigger channels. Specifically, we have shown that the quantitative analysis of all subunits of the cleavage and polyadenylation factor (CPF) can be achieved through use of a purified CPF isobaric trigger channel (Peck Justice, et. al, Analytical Chemistry, 2021). In this study, we found that reproducible detection/quantitation could be achieved for post-translationally modified peptides at specific residues. These data suggest that isobaric trigger channels could be used to provide sensitive detection of peptide sequences of interest in DDA studies. We are continuing the trajectory of this study through the development of synthesized peptide libraries with the objective of creating proteome specific isobaric trigger channels that boost detection and quantification of specific peptides, modifications, and protein regions regulated by differential processing at the protein/mRNA level in DDA MS workflows.

Current work explores the utility of isobaric trigger channels to provide insights into the protein abundance and protein modification status of the human RNA Polymerase II transcription elongation machinery in a WT system compared to a system treated with the Cdk9-kinase inhibitor flavopiridol. This involves the design of a literature curated synthesized peptide library for the trigger channel for targets of interest. Additionally, MS analysis was conducted alongside a phosphor-proteomics analysis to determine the effectiveness of different analytical methods for detection/quantification of targets. Phospho-proteomic analysis showed many dose-dependent changes in phosphorylation due to CDK9 inhibition, however; quantification and detection were not achieved for most of the selected targets showcasing the utility for the development of our trigger peptide-based technique. Utilizing the isobaric trigger channel, 21 of the 28 target peptides/phospho-peptides were detected in our initial experiments and we anticipate this number will increase with optimization for our targets. The development of this analytical resource will facilitate quantitative analysis of clinically important Cdk9-kinase substrates with analytical precision exceeding those of traditional approaches.

P03
Disruption Compensation (DisCo) Analysis of Complex Protein Networks

Katlyn Hughes Burriss[1], Guihong Qi[1], Whitney Smith-Kinnaman[1], Amber Mosley[1,2]

[1] Department of Biochemistry and Molecular Biology, Indiana University School of Medicine, Indianapolis, Indiana 46202, United States

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

Essential cellular processes are regulated through intricate networks of proteins, protein complexes, protein-protein interactions (PPIs), and post-translational modification (PTM) of these proteins and their corresponding complexes. RNA Polymerase II (RNAPII) transcription is a prime example of an essential cellular process that has been the subject of decades of research to illuminate its regulatory mechanisms. While data from genetic screens has aided in early efforts to identify many proteins that aid in RNAPII transcription, the underlying biochemical mechanisms of how these actors are regulated through PPIs and PTMs remain elusive.

To interrogate the protein-level mechanisms that regulate transcription, we developed an affinity-purification mass spectrometry (AP-MS) method, termed disruption-compensation (DisCo) network analysis. DisCo, in conjunction with isotopic label-based (TMT) MS quantitation, measures specific alterations in the RNAPII protein interactome. Measuring the alterations in PPIs that result from genetic perturbation of RNAPII-interactors provides insight into the wide variety of biochemical mechanisms that regulate transcription. We performed RNAPII DisCo in RNAPII phosphatase mutant *Saccharomyces cerevisiae* strains, *ssu72-2* and *fcp1-1*, as well as an isogenic parental strain BY4741, to gain insight into how dynamic phosphorylation of RNAPII regulates PPIs during transcription. Based on current literature knowledge of phosphatases Ssu72 and Fcp1, we hypothesized that *ssu72-2* and *fcp1-1* would result in termination-defective RNAPII and the accumulation of termination machinery at the site of transcription.

However, though DisCo analysis we were able to observe that *ssu72-2* effects the ability of RNAPII to proceed through elongation, causing proteasome-mediated degradation of RNAPII, while *fcp1-1* was observed to primarily alter interactions of transcription initiation factors. Global proteomic experiments were also performed with these strains to validate DisCo's ability to measure changes in protein interactions, not global abundance changes. Furthermore, Multi-OMIC analysis of DisCo and global protein data, in conjunction with previous genetic screens in the literature, provide insights into how these yeast cells are attempting to compensate in these mutants, further detailing the role the phosphatases play in transcription regulation.

While TMT DisCo provides a novel level of depth of coverage for the RNAPII transcription interactome, there are still challenges to quantitating the less direct, more dynamic RNAPII interactors, such as the cleavage and polyadenylation machinery. Currently, our lab is developing an adaptation of TMT DisCo using a trigger channel to boost signal of these elusive regulators and enable accurate quantitation of less abundant proteins, while not compromising the quantitation quality for the rest of the protein sample.

P04

A Software Tool for Proteoform Feature Detection from Top-down Mass Spectrometry Data

Abdul Rehman Basharat[1], Yong Zang[2], Xiaowen Liu[* 1, 3]

[1]School of Informatics and Computing, Indiana University - Purdue University Indianapolis, Indianapolis, IN; [2]Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, IN; [3]School of Medicine, Tulane University, New Orleans, LA

In top-down mass spectrometry (MS), feature detection[1] aims to find all isotopic envelopes of a proteoform in an LC-MS map and reports the monoisotopic mass, abundance, charge state range, and retention time (RT) range of the proteoform. Accurate feature detection is essential for improving the accuracy in proteoform characterization and quantification. However, the complexity of top-down MS data makes it a challenging computational problem. Here we present a tool for detecting proteoform features from top-down MS data and introduce a logistic regression model for evaluating features.

The methods in MS-Deconv[2] are used to report isotopic envelopes of proteoforms from single MS1 spectra in top-down MS data. The envelopes from all MS1 spectra are ranked using intensities to obtain an envelope list. We use a greedy method to identify proteoform features. Ineach round, the best envelope in the list is used as a seed to find all isotopic envelopes of the proteoform, which are then removed from the list. To evaluate proteoform features, a logistic regression model was trained using three proteoform feature attributes: the EnvCNN[3] score, the charge state range, and the RT range. To train the model and evaluate the performance of the proposed methods, ovarian tumor[4] (OT) and breast cancer[5] (BC) datasets were used. Both datasets were analyzed using a Thermo Orbitrap Elite mass spectrometer. The OT and BC datasets contained ten and eighteen technical replicates, respectively.

We identified features from each replicate of the two datasets. A feature was labeled positive if it appeared in at least five replicates and negative if it appeared in only one replicate. Otherwise, the feature was removed. The first replicate of the OT dataset was used to train and evaluate the model. We randomly divided the features for training and testing with a 70:30 split. Our model achieved the AUC ROC value of 97.75% on the OT test dataset. We used the first replicate of the BC dataset to assess the performance of the model trained on the OT dataset. The AUC ROCvalue of the logistic regression model (92.1%) outperformed the EnvCNN method (90.6%). We also used the frequency of each feature across multiple replicates to benchmark our tool's performance against ProMex[4] (Figure).

In conclusion, our method achieved high accuracy in identifying proteoform features. Furthermore, in comparison with ProMex, our tool reported a significantly large number offeatures in all replicates for both datasets.

References

1. Schaffer, L. V.; Millikin, R. J.; Miller, R. M.; Anderson, L. C.; Fellers, R. T.; Ge, Y.; Kelleher, N. L.; LeDuc, R. D.; Liu, X.; Payne, S. H., Identification and quantification of proteoforms by mass spectrometry. *Proteomics* **2019,** *19* (10), 1800361.

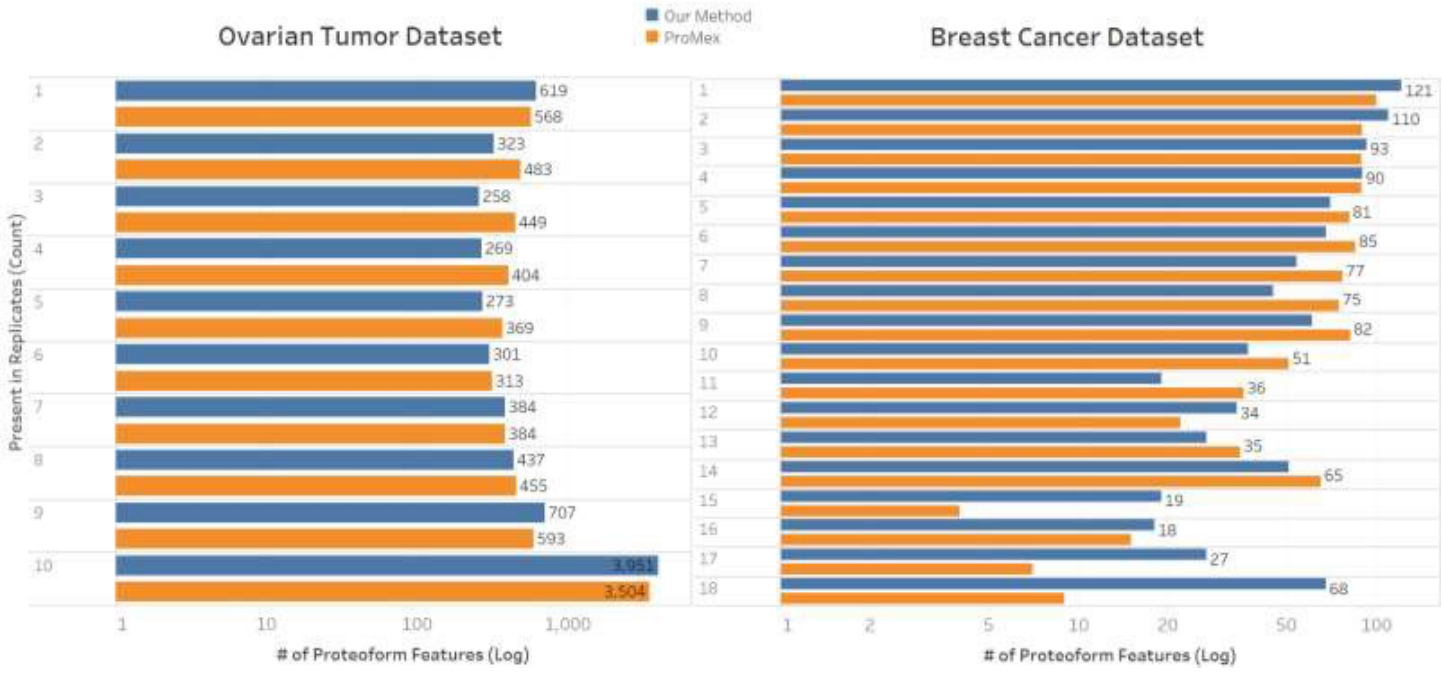
2. Liu, X.; Inbar, Y.; Dorrestein, P. C.; Wynne, C.; Edwards, N.; Souda, P.; Whitelegge,

3. J. P.; Bafna, V.; Pevzner, P. A., Deconvolution and database search of complex tandem mass spectra of intact proteins. *Molecular & Cellular Proteomics* **2010,** *9* (12), 2772-2782.

4. Basharat, A. R.; Ning, X.; Liu, X., EnvCNN: A Convolutional Neural Network Model for Evaluating Isotopic Envelopes in Top-Down Mass-Spectral Deconvolution. *Analytical Chemistry* **2020,** *92* (11), 7778-7785.

5. Park, J.; Piehowski, P. D.; Wilkins, C.; Zhou, M.; Mendoza, J.; Fujimoto, G. M.; Gibbons, B. C.; Shaw, J. B.; Shen, Y.; Shukla, A. K.; Moore, R. J.; Liu, T.; Petyuk, V. A.; Tolić, N.; Paša-Tolić, L.; Smith, R. D.; Payne, S. H.; Kim, S., Informed-Proteomics: open- source software package for top-down proteomics. *Nature Methods* **2017,** *14* (9), 909-914.

6. Ntai, I.; LeDuc, R. D.; Fellers, R. T.; Erdmann-Gilmore, P.; Davies, S. R.; Rumsey, J.; Early, B. P.; Thomas, P. M.; Li, S.; Compton, P. D., Integrated bottom-up and top-down proteomics of patient-derived breast tumor xenografts. *Molecular & Cellular Proteomics* **2016,** *15* (1), 45-56.


Our Method
ProMex
Ovarian Tumor Dataset
Breast Cancer Dataset
Present in Replicates (Count)
# of Proteoform Features (Log)
# of Proteoform Features (Log)
1
2
3
4
5
6
7
8
9
10
619
568
323
483
258
449
269
404
273
369
301
313
384
384
437
455
707
593
3,951
3,504
1
10
100
1,000
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
121
110
93
90
81
85
77
75
82
51
36
34
35
65
19
18
27
68
1
2
5
10
20
50
100

P05

Time-Resolved Electrospray Ionization Combined with Electron Capture Dissociation for Characterization of Protein Folding Intermediates

Cain, Rebecca[1], Webb, Ian[1,2]

[1]Department of Chemistry and Chemical Biology, Indiana University Purdue University Indianapolis, Indianapolis, IN,46202

[2]Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, IN, 46202

Protein folding/unfolding mechanisms are essential for understanding protein structure and function, but elusive unfolding intermediates are short lived and difficult to measure, thus complicating the complete characterization of these mechanisms. Different conformations of protein unfolding intermediates can be identified by their different charge states. Tightly folded proteins have less available protonated sites and thus have lower charge states compared to unfolded conformations. Protein unfolding studies utilizing a mixing device controls the amount of unfolding by varying the concentration of denaturant used prior to analysis with the mass spectrometer. This technique is combined with electron capture dissociation (ECD) to look at fragmentation produced by different unfolding stages to fully characterize the mechanisms of protein unfolding. Model protein solutions were made at 10 μM in water and mixed with acetic acid concentrations at pHs 1.96, 2.19, 2.35, and 2.57. A mixing apparatus was assembled using 70 cm long infusion lines for both the protein and acid solutions that were connected to a three way mixing tee. Both solutions in the infusion lines were infused at 5 ul/min each. The final line in the mixing tee sprayed the solution mixture directly to electrospray source for analysis with a Waters Synapt G2-Si HDMS. At each acid concentration, different charge states for the various unfolding stages were isolated and subjected to ECD fragmentation. Our preliminary data for the model protein cytochrome c illustrates higher concentrations of acid correspond to more unfolding of the protein. The acetic acid concentration at pH 1.96 produces unfolded higher charge states up to 15+ compared to intermediate or folded charge states produced in the lower acid concentration at pH 2.57. Preliminary ECD data compares fragments of the higher and lower charge states. At the higher charge states, more ion fragments are seen which corresponds to more unfolded states that have more exposed sites, while the lower charge states produce less fragmentation due to more protected sites. Comparing the ECD fragments of folded, unfolded, and intermediates allows us to identify which residues in the protein are most dynamic and are exposed first during the unfolding process. Time-resolved unfolding coupled with ECD fragmentation allows for insight into characterizing folding intermediates.

P06

Image Analysis Pipeline for Spatial and Translational Properties of RNA Molecules in β-Cells of Type 1 Diabetes

Garrick Chang[1], Hua Lin[1], Syed Farooq[2], Wenting Wu[3], Carmella Evans-Molina[4], Jing Liu[1]

[1]Department of Physics, IUPUI, Indianapolis, IN, USA.

[2]Department of Pediatrics, IUPUI, Indianapolis, IN, USA.

[3]Department of Medical and Molecular Genetics, IUPUI, Indianapolis, IN, USA.

[4]IU Center for Diabetes and Metabolic Diseases, Indianapolis, IN, USA.

We developed a computational pipeline with the emphasis on analyzing RNA localization and translation kinetics of β-cells in human and mouse tissue sections that are associated with the type 1 diabetes (T1D). The pipeline consists of cell segmentation, object detection and localization analysis of mRNA. By implementing our pipeline with several research projects on T1D, we investigated a group of RNA molecules, including mRNA, miRNA, and mRNA splicing variant, along with stress granules by analyzing colocalization between mRNA and binding proteins. Results of the pipelines from each projects show significant differences of spatial distribution of RNA in nucleus/cytoplasm between T1D and control samples, agreeing with the hypothesizes for the respective projects and proofing the effectiveness of our pipeline. With the additional analysis by tracking mRNA on consecutive frames and studying its colocalization with ribosome particles, translation process can be interpreted by examining the rate and duration of intensity change of ribosome and mRNA particles, allowing deeper understanding on functionality and reproductivity of β-cells associated with T1D and effects of different medical treatments. Overall, the pipeline provides a systematic and effective method to analyze and discover the translation properties and functionally of β-cells associated with T1D.

P07

Computational Modeling of Neurotransmitter Biosynthesis and Synapse by using Single-Cell, Tissue and Spatial Transcriptomics Data

Wennan Chang[1,2], Norah Alghamdi[1,3], Pengtao Dang[1,4], Xiaoyu Lu[1,3], Changlin Wan[1,2], Sha Cao[1,5*], Chi Zhang[1,6*]

[1]Center for Computational Biology and Bioinformatics, [5]Department of Biostatistics, [6]Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, Indianapolis, IN, USA; [2]Department of Electric Computing Engineering, Purdue University, West Lafayette, IN, USA; [3]Department of Biohealth Informatics, [4]Department of Electric Computing Engineering, Indiana University Purdue University Indianapolis, Indianapolis, Indianapolis, IN, USA. Contact e-mail: Chi Zhang, czhang87@iu.edu

We recently developed the first computational model to estimate cell-wise metabolic flux by using scRNA-seq data. In this study, we extended the capability to functional analysis of neurotransmitter biosynthesis and degradation, synapse, and downstream effects by using transcriptomics data. We manually collected and reconstructed biosynthesis and degradation modules of glutamate, GABA, acetylcholine, histamine, dopamine and serotonin, as well as their receptors and other synapse-related pathways. To the best of our knowledge, this is the first capability to systematically model neurotransmission and synapse by using omics data. We applied the method to single cell, tissue and spatial transcriptomics data of Alzheimer's disease and normal brain samples and identified significant variations in neurotransmission in AD.

P08

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

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

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

Genome-wide association studies (GWAS) have identified loci associated with alcohol consumption, but many are in non-coding regulatory regions, where additional information is needed to evaluate their function. Our objective is to understand how variants at these regulatory loci are functionally involved in contributing to alcohol consumption and alcohol use disorder.

Using a massively parallel reporter assay (MPRA) in neuroblastoma and microglia cell lines, we evaluated the functional activity of variants in the 3' untranslated regions (3'-UTR) of genes associated with neurological disorders. We then analyzed the heritability enrichment of the functionally relevant variants identified by MPRA in two GWAS of alcohol use: drinks per week from GWAS & Sequencing Consortium of Alcohol and Nicotine use (GSCAN); and alcohol use disorder (AUD) from the Million Veteran Program (MVP). We integrated the estimated effects from MPRA results with GWAS results from GSCAN drinks per week and MVP alcohol use disorders identification test-consumption (AUDIT-C) to identify genes in which the activity of SNPs in the 3'-UTR were associated with alcohol consumption.

Of the 13,515 SNPs tested, 400 (neuroblastoma) and 657 (microglia) significantly impacted gene expression. We found that functionally active variants explained a higher proportion of heritability than other variants tested in both cell lines. A pathway analysis of these differentially expressed genes identified several inflammation response pathways, including IL-17, NF-kappa B, and TNF signaling.

Because these genes were identified by their genetic components, these results suggest that variation in response to inflammation contributes to the onset of increased alcohol consumption. In this study, integrating MPRA and GWAS results allowed us to gain insight into genetic mechanisms affecting alcohol consumption, and using this framework can help elucidate these relationships for genes within GWAS loci in other traits.

P09

Evaluation of Deep Learning Models for Retention and Migration Time Prediction in Top-down Mass Spectrometry

Wenrong Chen[1], Elijah N. McCool[2], Liangliang Sun[2], Xia Ning[3,4,5], Xiaowen Liu[6]

[1]Department of BioHealth Informatics, Indiana University-Purdue University Indianapolis, Indianapolis, IN, [2]Department of Chemistry, Michigan State University, East Lansing, MI, [3]Department of Biomedical Informatics, The Ohio State University, Columbus, Ohio, [4]Department of Computer Science and Engineering, The Ohio State University, Columbus, Ohio, [5]Translational Data Analytics Institute, The Ohio State University, Columbus, Ohio, [6]Deming Department of Medicine, Tulane University, New Orleans, LA

Introduction Reverse-phase liquid chromatography (RPLC) and Capillary zone electrophoresis (CZE) are two popular proteoform separation methods in mass spectrometry (MS)-based top-down proteomics. Accurate prediction of the retention/migration time of proteoforms is important for increasing the accuracy of proteoform identification and quantification. There is still a lack of methods for the problem in top-down MS. We evaluated the performance of retention/migration time prediction with the deep learning models including fully connected neural network (FNN), DeepRT (+), Prosit and DeepDIA models for retention/migration time prediction of proteoforms in top-down MS.

Methods Several deep neural network models are used to predict the retention/migration time of proteoforms in top-down MS. The FNN model contains an input layer, several hidden layers with dropout, and a fully connected output layer. The DeepRT (+) model is composed of two convolution layers and two capsule layers which are the variance of convolution layers. The Prosit model consists of an embedding layer, a bidirectional GRU layer, a one-directional GRU layer, an attention layer, and two dense layers. The DeepDIA model contains a convolutional, a max pooling layer, a bidirectional LSTM layer, and three dense layers.

Results We evaluated the models on an ovarian tumor (OT) RPLC-MS data set and a SW480 CZE-MS data set. After database searching, we filtered proteoform identifications with an E-value cutoff of 1e-5 and removed those with variable or unexpected modifications. We identified 610 proteoforms (188 proteins) with their retention time from the RPLC-MS data set and then randomly split proteoforms with protein groups into a training set (437) and a test set (173) with a ratio of 7:3 approximately. Similarly, we identified 1230 proteoforms (470 proteins) with their migration time from the CZE-MS data set and further split them into a training set (878) and a test set (352).

For the proteoforms identified from the CZE-MS data, we used a semi-empirical model to normalize the migration times of proteoforms to remove the variances among fractions. For the OT and SW480 dataset, we evaluated our deep neural network models with 5-fold cross validation of training set for hyperparameter tuning. With the best setting of each model, we trained the models on the training set and evaluated with test set (Table 1).

Discussion According to the results in table 1, the FNN model and Prosit model showed better performance and robustness than DeepRT (+) and DeepDIA model for two kinds of top-down MS proteoforms.

Table 1: Comparison of deep learning models in retention time/migration time prediction for top-down MS datasets

Model	Dataset: OT (7:3)		Dataset: SW480 (7:3)	
	R^2	MSE	R^2	MSE
FNN	0.8123	0.006979	**0.9042**	**0.001320**
DeepRT+	0.7586	0.008519	0.8208	0.002470
Prosit	**0.8545**	**0.005479**	0.9021	0.001350
DeepDIA	0.7449	0.009602	0.5371	0.006381

P10

Cross-linking Reagents as Molecular Calipers to Probe Protein Structure via Gas-phase Ion/Ion Reactions Coupled to Ion Mobility/Time-of-flight Mass Spectrometry

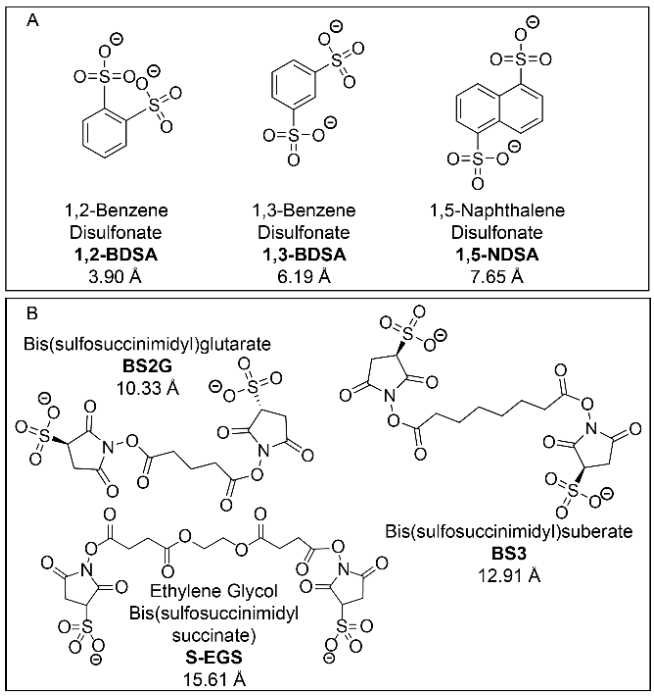
Melanie Cheung See Kit[1], Ian K. Webb[1, 2]

[1]Department of Chemistry and Chemical Biology, Indiana University Purdue University Indianapolis, Indianapolis, IN, 46202

[2]Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, IN, 46202

Proteins are an important component of the cellular machinery and their function highly correlates to structure and dynamics. Combining ion mobility, native mass spectrometry (IM/MS)and gas-phase ion/ion reactions provides a sensitive and relatively quick approach to study protein structure. Unlike solution phase labeling, gas-phase ion/ion reactions are achieved in- line under controlled conditions to improve reaction efficiency and resulting products are easily separated by IM and analyzed by tandem MS. Reacting proteins with cross-linking reagents of various spacer lengths and flexibility thus provides detailed residue-level information pertaining to protein folding and dynamics.

Cross-linking reagents used to map the structure of cytochrome c



For this study, all compounds contained two sulfonate groups, which facilitate electrostatic attachment to accessible protonated residues from the protein to form a product that can be analyzed by tandem MS. One scaffold type 'A' included phenyl ring(s) with sulfonates attached while type 'B' consisted of a more flexible spacer linker between two sulfonated N-hydroxysuccinimide (sulfo-NHS) ester reactive groups. Native cytochrome c +7 and −2 sulfonated ions reacted in the trap cell and the reactant and charge-reduced product ions were separated by ion mobility, followed by fragmentation by electron- capture dissociation (ECD) and additional energy in the transfer cell, which improved sequence coverage. With the sulfonate groups in the ortho position, 1, 2-BDSA only modified one residue, namely Lys55. In contrast, 1, 3-BDSA, with the sulfonates in the meta position, reacted at Arg38 and Lys55 while 1, 5-NDSA modified Arg38 and Lys72. The increased distance between the reactive sulfonates for the type 'B' cross-linkers allowed the reagents to span a wider distance, reflected by differentpairs of labeled sites. BS2G and BS3 both cross-linked Arg38 and Lys72 while S-EGS modified Arg38 and Lys88. The results are consistent with the solution phase structural data and support the application of a gas phase ion/ion reaction approach combined with IM/MS/MS to provide structural insights into protein conformations. Additionally, the modified sites will be used as distance constraints during molecular dynamics simulation of cytochrome c structure in the gas phase.

P11

Statistical Modeling of Spatial Dependent Functional Variations by using Spatial Transcriptomics Data

Pengtao Dang[1,2], Wennan Chang[1,3], Xiaoyu Lu[1,4], Changlin Wan[1,3], Chi Zhang[1,5], Sha Cao[1,6*]

[1]Center for Computational Biology and Bioinformatics, [5]Department of Medical and Molecular Genetics, [6]Department of Biostatistics, Indiana University School of Medicine, Indianapolis, Indianapolis, IN, USA; [2]Department of Electric Computing Engineering, [4]Department of Biohealth Informatics, Indiana University Purdue University Indianapolis, Indianapolis, Indianapolis, IN, USA; [3]Department of Electric Computing Engineering, Purdue University, West Lafayette, IN, USA. Contact e-mail: Sha Cao, shacao@iu.edu

We proposed a Spatial Robust Mixture Regression model to investigate the relationship between a response variable and a set of explanatory variables over the spatial domain, assuming that the relationships may exhibit complex spatially dynamic patterns that cannot be captured by constant regression coefficients. Our method integrates the robust finite mixture Gaussian regression model with spatial constraints, to simultaneously handle the spatial nonstationarity, local homogeneity, and outlier contaminations. Compared with existing spatial regression models, our proposed model assumes the existence of a few distinct regression models that are estimated based on observations that exhibit similar response-predictor relationships. As such, the proposed model not only accounts for nonstationarity in the spatial trend but also clusters observations into a few distinct and homogenous groups. This provides an advantage on interpretation with a few stationary sub-processes identified that capture the predominant relationships between response and predictor variables. Moreover, the proposed method incorporates robust procedures to handle contaminations from both regression outliers and spatial outliers. By doing so, we robustly segment the spatial domain into distinct local regions with similar regression coefficients and sporadic locations that are purely outliers. A rigorous statistical hypothesis testing procedure has been designed to test the significance of such segmentation. Experimental results on many synthetic and real-world datasets demonstrate the robustness, accuracy, and effectiveness of our proposed method, compared with other robust finite mixture regression, spatial regression, and spatial segmentation methods.

P12

RNA Structure Elucidation at Single Molecule Resolution using an Integrated Framework of Long Read Sequencing and SHAPE-seq

Swapna Vidhur Daulatabad, Quoseena Mir, Sarath Chandra Janga

Indiana University - Purdue University Indianapolis, Department of Biohealth Informatics, Indianapolis, IN,

RNA molecules have a wide functional diversity, one of the key attributes governing such ability stems from the dynamic nature of RNA structure. Despite rapid progress in identifying novel functions of RNAs and novel classes of RNAs, we still lack a comprehensive understanding of the RNA structure. With the advancements in next generation sequencing, chemical probing methods and their efficient integration have enabled us to characterize RNA structure more clearly than ever. One such protocol that not only dissects the RNA structure but also does with great throughput is SHAPE-seq. However, common limitations of SHAPE-seq and similar protocols, include the use of short read sequencing as a downstream step, which provides incomplete information of the RNA molecule and lack of isoform level specificity of the transcript, restricting from developing a wholistic and comprehensive paradigm for RNA structure problem. Therefore, we propose an efficient integration of Nanopore based single molecule direct RNA long read sequencing and advanced machine learning frameworks to infer and predict RNA structure accurately. The SHAPE-seq protocol aims to modify the RNA once per molecule ("single-hit" kinetics) and Nanopore enables full-length and strand specific direct RNA sequencing at single molecule resolution. Combining both the approaches helps us reduce the biases induced by reverse transcriptase truncation and non-specificity of the transcripts. Traditional chemical probing combined with short read sequencing and NMR validation have enabled us to develop better model RNAs like the HepC IRES, 338nt. Hence, in this study, *in-vitro* transcribed HepC IRES RNA molecule was probed with the 1M7 reagent and each control probed with DMSO. Polyadenylated RNA molecules were sequenced using MinION from Oxford Nanopore Technologies, to generate 605k reads for HepC IRES. We generated a Nanopore direct RNA-seq based reactivity measure that quantifies the SHAPE reagent introduced miscalling abundance at each base. This reactivity score was used to predict the secondary structure of HepC-IRES RNA, which was approximately 70% similar to accepted structure. Furthermore, we observed the raw direct RNA-sequencing data of these molecules, to develop de-novo RNA structure predicting machine learning model at single molecule resolution, which benchmarked at 94% accuracy. Such accurate RNA-structure prediction models and corresponding methods can help us gain insights on the role of the structure in modulating the post-transcriptional regulatory processes.

P13

RAZOR: A Central Database of PCR Primers Targeting Human Respiratory Viruses

Hunter Gill, Sarath Chandra Janga

Department of BioHealth Informatics, School of Informatics and Computing, Indiana University Purdue University, Indianapolis, Indiana

The Coronavirus Disease 2019 (COVID-19) pandemic demonstrates Polymerase Chain Reaction (PCR) diagnostic testing is a gold standard for respiratory virus detection. Although reliance on PCR testing is high, there are few online resources for cataloging high-quality PCR primers against respiratory viruses. The RAZOR database is introduced to fill this gap and provide a volume of PCR primers for 21 human respiratory viruses including SARS-CoV-2. The PCR primers are designed with Primer3 for Real-Time PCR and conventional PCR experiments and are sorted into distinct specificity categories depending on the number of targeted viral genomes. Results are presented in an event driven IGV interface with several options for filtering and downloading data. RAZOR also supports community-driven collaboration in which experimental groups can submit validations of predicted PCR primers for all users to view. These features combine to make RAZOR an engaging public resource for selecting PCR primers and developing testing approaches for respiratory viruses.
URL: https://sysbio.informatics.iupui.edu/primer_project/razor/

P14

Charge Inversion Ion/Ion Reactions for Enhancing Ion Mobility-Mass Spectrometry Separations of Oligosaccharide Anions

Ankita Gurav[1], Ian Webb[1,2]

[1]Department of Chemistry and Chemical Biology, Indiana University Purdue University Indianapolis, [2]Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, IN, 46202

Objective/Hypothesis:

Oligosaccharides can exist in many isomeric forms making their structural determination and differentiation difficult. Ion mobility-MS (IM-MS) separations are fast and separate ions based on their shape-to-charge ratio, useful for isomer separation. We have reacted alkali earth cations complexed with tris-1,10-phenanthroline in the 2^+ charge state with singly deprotonated oligosaccharides to obtain unique drift time fingerprints for each isomer.

Methods:

The individual oligosaccharide solutions were made at 100mM each and diluted to 10µM in MilliQ H_2O. An equimolar mixture of the same was also made at 10µM. These were sprayed as anions using a negative mode nanoelectrospray. The cation solution, 1mM calcium tris-1,10-phenanthroline ($Ca(phen)_3$), was electrosprayed from a separate emitter. A doubly charged $Ca(phen)_3$ cationic complex was reacted with each negatively charged carbohydrate isomer and a mixture of all isomers. Ion/ion reactions were conducted using ETD acquisition mode in Waters Synapt G2-Si HDMS instrument and thereafter analyzed using MassLynx software. Collisionally induced dissociation to facilitate the loss of the phenanthroline ligands was performed prior to the IM separation by increasing the energy into the helium cell.

Data/Discussion:

We have demonstrated our ion/ion charge inversion IM-MS strategy with trisaccharides-raffinose, melezitose and kestose which are commonly used to demonstrate oligosaccharide separations by IM/MS. When electrosprayed as anions, they are identical in both *m/z* and mobility (given the resolving power of our instrument) overlapping completely in IM and *m/z* space. When reacted with $Ca(phen)_3^{2+}$, a long-lived electrostatically bound complex is formed with a net single positive charge. These products (*m/z* 1083) do not increase the separation between the three isomers. However, as injection energy into the mobility separator is increased, the complex $[M+Ca]^+$ is formed (*m/z* 543), the arrival time distributions for which show distinct features for each isomer. For raffinose, we repeatably observed prominent IM peaks at 6.6-6.7 ms and 8.3-8.5 ms. For melezitose, we observed a peak at range 7.2-7.3 ms. For kestose, we observed a peak at 6.78 ms. When charge inverted, the mixture of the trissacharides shows IM fingerprints that distinguish the three sugars. Interestingly, the same reactions with $Mg(phen)_3^{2+}$ do not show any differences in the arrival time distributions. We are currently working on charge inversion reactions with IM-MS separations with other different length carbohydrate isomers.

Originality:

Demonstrating the application of charge inversion ion/ion reactions to enable IM-MS separation of isomeric oligosaccharides.

P15
SARS-CoV-2 Diversity and Phylogeny in Indiana COVID-19 Patients

Dwight William Hall[1 £], Quoseena Mir[1 £], Rajneesh Srivastava[1 £], Guang-Sheng Lei[3], David James Tauriainen[2], John-Paul Lavik[3], Ryan F. Relich[3], Sarath Chandra Janga[1,4,5*]

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

[2] School of Informatics and Computing, Indianapolis, Indiana

[3] Department of Pathology and Laboratory Medicine, Indiana University School of Medicine, Indianapolis, Indiana

[4] Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Health Information and Translational Sciences (HITS), Indianapolis, Indiana

[5] Department of Medical and Molecular Genetics, Indiana University School of Medicine, Medical Research and Library Building, Indianapolis, Indiana

SARS-CoV-2, the virus of COVID-19, emerged and spread quickly to every country in late 2019. It has exposed numerous weaknesses in public health and emerging disease response infrastructures. Despite the availability of vaccines in many parts of the world, SARS-CoV-2 continues to circulate. Coincident with its ongoing spread is the emergence of variant viruses, including increased transmissibility and virulence. In this study, we sequenced, genotyped, and mapped the diversity of SARS-CoV-2 in human clinical samples obtained from COVID-19 patients in Indiana, USA, to characterize the evolutionary and transmission dynamics for SARS-CoV-2 variants circulating in Indiana during the early days of the COVID-19 pandemic and the current data. Our approach included SARS-CoV-2 genomic RNA extraction from 40 swab specimens, sequenced using a MinION Sequencer, and the data analysis refined using ARTIC Network protocols. Consensus sequences of complete viral genomes were analyzed using phylogenetic trees, a geographical map, and genotype diversity using Nextstrain's open-source software. Genomic diversity was used to infer mutation sites among the samples. Geographical location was used to determine which countries had the most similar sequences to the samples from Indiana. The resulting phylogenetic tree and transmission map can be viewed at https://covid19-indiana.soic.iupui.edu/Indiana. We analyzed the mutation event frequency in Indiana Samples, and we found three reoccurring mutations T265I, Q57J, and D614G. Our initial analysis of the D614G mutation on the global phylogenetic tree indicated that 39 of the Indiana samples belonged to the G (glycine) group. In contrast, one Indiana sample corresponded to the D (aspartic acid) group. The geographical analysis of the Indiana samples collected early in the pandemic revealed that within the US, the SARS-CoV-2 sequences were most similar to sequences from Virginia and Michigan, while outside the US, most similar to sequences from Victoria, Australia. Finally, we collected 14,000 SARS-CoV-2 sequences and analyzed the emerging lineages and clades, which show the divergence from our samples from a year ago. Our proposed end-to-end framework for whole genome sequencing and tracking SARS-CoV-2 variants in a geographical context could be a model for real-time understanding of viruses' evolution and transmission dynamics in remote locations.

P16

Penguin: A Tool for Predicting Pseudouridine Sites in Direct RNA Nanopore Sequencing Data

Doaa Hassan[1,4], Daniel Acevedo[1,5], Swapna Vidhur Daulatabad[1], Quoseena Mir[1], Sarath Chandra Janga[1,2,3]

[1]Department of BioHealth Informatics, School of Informatics and Computing, Indiana University Purdue University, 535 West Michigan Street, Indianapolis, Indiana 46202

[2]Department of Medical and Molecular Genetics, Indiana University School of Medicine, Medical Research and Library Building, 975 West Walnut Street, Indianapolis, Indiana, 46202

[3]Centre for Computational Biology and Bioinformatics, Indiana University School of Medicine, 5021 Health Information and Translational Sciences (HITS), 410 West 10th Street, Indianapolis, Indiana, 46202

[4]Computers and Systems Department, National Telecommunication Institute, Cairo, Egypt.

[5]Computer Science Department, University of Texas Rio Grande Valley

Pseudouridine is one of the most abundant RNA modifications, occurring when uridines are catalyzed by Pseudouridine synthase proteins. It plays an important role in many biological processes and also has an importance in drug development. Recently, the single-molecule sequencing techniques such as the direct RNA sequencing platform offered by Oxford Nanopore technologies enable direct detection of RNA modifications on the molecule that is being sequenced, but to our knowledge this technology has not been used to identify RNA Pseudouridinesites. To this end, in this paper, we address this limitation by introducing a tool called Penguin thatintegrates several developed machine learning (ML) models (i.e., predictors) to identify RNA Pseudouridine sites in Nanopore direct RNA sequencing reads. Penguin extracts a set of features from the raw signal measured by the Oxford Nanopore and the corresponding basecalled k-mer. Those features are used to train the predictors included in Penguin, which in turn, is able to predictwhether the signal is modified by the presence of Pseudouridine sites. We have included various predictors in Penguin including Support vector machine (SVM), Random Forest (RF), and Neuralnetwork (NN). The results on the two benchmark data sets show that Penguin is able to identify Pseudouridine sites with a high accuracy of 93.38% and 92.61% using SVM in random split testing and independent validation testing respectively. Thus, Penguin outperforms the existing Pseudouridine predictors in the literature that achieved an accuracy of 76.0 at most with an independent validation testing. A GitHub of the tool is accessible at https://github.com/Janga- Lab/Penguin.

P17

SIRT6 Controls Hepatic Lipogenesis by Suppressing LXR, ChREBP, and SREBP1

Chaoyu Zhu[1,2,#], Menghao Huang[1,#], Hyeong-Geug Kim[1], Kushan Chowdhury[1], Jing Gao[1,3], Sheng Liu[4,5], Jun Wan[4,5], Li Wei[2,*], and X. Charlie Dong[1,5,*]

[#]Contributing equally to this work.

[1]Department of Biochemistry and Molecular Biology, Indiana University School of Medicine, 635 Barnhill Drive, Indianapolis, IN 46202, USA

[2]Department of Endocrinology and Metabolism, Shanghai Jiao Tong University Affiliated Sixth People's Hospital, 600 Yishan Road, Shanghai 200233, China

[3]College of Food Science and Nutritional Engineering, China Agricultural University, 17 Qinghua Donglu, Beijing 100083, China

[4]Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, IN 46202, USA

[5]Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, IN 46202, USA

Fatty liver disease is the most prevalent chronic liver disorder, which is manifested by hepatic triglyceride elevation, inflammation, and fibrosis. Sirtuin 6 (Sirt6), an NAD^+-dependent deacetylase, has been implicated in hepatic glucose and lipid metabolism; however, the underlying mechanisms are incompletely understood. The aim of this study was to identify and characterize novel players and mechanisms that are responsible for the Sirt6-mediated metabolic regulation in the liver. We generated and characterized Sirt6 liver-specific knockout mice regarding its role in the development of fatty liver disease. We used cell models to validate the molecular alterations observed in the animal models. Biochemical and molecular biological approaches were used to illustrate protein-protein interactions and gene regulation. Our data show that Sirt6 liver-specific knockout mice develop more severe fatty liver disease than wild-type mice do on a Western diet. Hepatic Sirt6 deficiency leads to elevated levels and transcriptional activities of carbohydrate response element binding protein (ChREBP) and sterol regulatory element binding protein 1 (SREBP1). Mechanistically, our data reveal protein-protein interactions between Sirt6 and liver X receptor α (LXRα), ChREBP, or SREBP1c in hepatocytes. Moreover, Sirt6 suppresses transcriptional activities of LXRα, ChREBP, and SREBP1c through direct deacetylation. In conclusion, this work has identified a key mechanism that is responsible for the salutary function of Sirt6 in the inhibition of hepatic lipogenesis by suppressing LXR, ChREBP, and SREBP1.

P18

Understanding the Uncoupling of Tauopathy and Dementia Through Comparative Analysis of Subgroups of Atypical Alzheimer's Disease Patients

Xiaoqing Huang[1], Nur Jury Garfe[2], Jiannan Liu[3], Cristian Lasagna Reeves[2] and Jie Zhang[4]

[1]Department of Biostatistics and Health Data Science, [2]Department of Anatomy, Cell Biology & Physiology, [3]Department of BioHealth Informatics, [4]Department of Medical & Molecular Genetics, Indiana University, School of Medicine

Background: Typical Alzheimer's disease (AD) patient brains show characteristic neurofibrillary tangles (NFTs), which protein tau aggregates to form bundles of microtubule-associated neurofibrillary tangle (NFT). A major focus of AD research has been the understanding of the roles of pathological tau in neurons and NFTs (tauopathy) in the disease initiation and progression, which severity can be measured with Braak staging. A subpopulation of patients also exists, whose brain showed typical AD-like high NFTs, but with no or low cognitive deterioration (Asymptomatic AD, aka AsymAD) or present severe cognitive impairment but low NFTs (low-NFT AD).

Method: We performed differential expression analyses on both RNA and protein expression data from the atypical groups of MSBB as well as ROSMAP using typical AD patients as reference, then combined with pathway analysis to search for enriched pathway and upstream regulators which may play the neuron-protective or neuron-vulnerable role in the two atypical AD subgroups.

Result: We identified several oppositely regulated pathways as compared with typical AD group for the two atypical groups in MSBB cohort, such as oxidative phosphorylation pathway. Several enriched pathways related to immune response in the brain in the low-NFT AD group are absent from AsymAD group. In ROSMAP cohort, we identified higher expression of actin-binding proteins and ROS eliminating genes in AsymAD group, and higher expression of immune-related proteins/mRNAs in the low-NFT AD group.

Conclusion: Through the comparative analyses on transcriptomic and proteomic data, we identified several pathways/genes that may potentially render neuron-protection or neuron-vulnerability to AD for subgroups of atypical AD patients, including oppositely regulated oxidative phosphorylation and mitochondrial functions, elevated actin binding signals in AsymAD, and elevated inflammatory and PP1 regulatory signals in low-NFT AD group. Further bioinformatic analysis will be performed to identify candidate genes/proteins and prioritize them for future *in vivo* animal model experiments. This work will not only improve precision medicine in AD, but also provide insights to prevent or slow down cognitive impairment during the long disease progression.

P19

Observing Chromatin Dynamics at High Imaging Speeds using Deep Learning

Fadil Iqbal[1], Paul Kefer[2], Mengdi Zhang[1,3], Josh Lawrimore[4], Maëlle Locatelli[5], Clayton Seitz[1], Kerry Bloom, Keith Bonin[2], Pierre-Alexandre Vidi[5], and Jing Liu[1]

[1]Department of Physics, Indiana University-Purdue University Indianapolis, IN 46202, USA;

[2]Department of Physics, Wake Forest University, Winston-Salem, NC 27109, USA;

[3]Harbin Medical University, Harbin, Heilongjiang, China;

[4] Department of Biology, University of North Carolina, Champ Hill, North Carolina, USA;

[5]Department of Cancer Biology, Wake Forest School of Medicine, Winston-Salem, NC 27157, USA,

The goal of Single Particle Tracking experiments is to determine the interaction between the particle and its environment – in our case, the Chromatin structure. It is important to know the exact position of these particles to construct their trajectories and do the imaging at high speeds. Going for high imaging speeds requires us to reduce the exposure time of the camera, but this comes at the cost of losing image quality or signal to noise ratio (SNR). Deep Learning (DL) is a new branch of Artificial Intelligence which has shown promise in restoring low SNR images. The question arises whether these techniques can be used in our experiments: to restore low SNR movies and recover their trajectories. Our study has verified the capability of two DL packages which are capable of restoration in two ways: supervised and unsupervised. We applied them to the following examples: 1. A synthetic bead model where noise can be added artificially – we were able to restore low SNR movies and restore their individual trajectories by both supervised and unsupervised methods. 2. Experimentally we applied these techniques on our nucleosomes data to verify if deep learning can restore stationary and dynamic data. It was noted that the supervised approach delivered better results. However, there are instances where these algorithms produce deceiving results, despite ‘looking good’ to the naked eye. We anticipate broad application of this approach to critically evaluate artificial intelligence solutions for quantitative microscopy.

P20

The Genetic Regulation of the Biogenesis of Human isomiRs

Guanglong Jiang, Jill L. Reiter, Yunlong Liu

Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, IN, 46202

Background: MicroRNA plays a critical role in post-transcriptional gene regulation. MicroRNA isoforms, or isomiRs, differ in sequence from their canonical counterparts as a consequence of imprecise cleavage, nucleotide substitution or addition. IsomiRs generated from the same precursor sequence may work competitively with the canonical miRNA, or target different mRNAs and thus expand their scope in post-transcriptional regulations. Recent studies indicated that isomiRs exhibit a population or gender-dependent expression manner. However, mechanisms underlying the dynamic regulations are largely unknown.

Method: The small RNA-seq reads for 452 unrelated human lymphoblastoid cell lines from 1000 genomes project individuals were aligned to human precursor miRNA sequence using isomiRID. To correct for the alignment bias towards reference sequences, the customized sequences incorporating all combinations of SNP alleles were generated for mapping of small RNA-seq reads. The genetic association study is conducted between cis-acting SNPs in the region of precursor miRNA, and 5' isomiR composition, defined as the ratio of a subtype of 5'-isomiR variants (substitution, trimming, extension or non-template addition) against all isomiRs identified for a specific mature miRNA.

Results: Totally 839 associations were conducted between SNPs and isomiR variants for 5'- substitutions, trimmings, extensions or additions, among which 10, 75, 38 and 13 SNP-isomiR pairs, respectively, were significantly associated using a false discovery rate (FDR) < 0.05. Rs6505162 was one of the SNPs significantly associated with 5'-extension of hsa-miR-423-3p (FDR = 3.0×10^{-21}) and 5'-trimming of hsa-miR-423-5p (FDR = 2.4×10^{-17}). Differential expression between tumor and normal samples in TCGA BRCA dataset found that the 5'-extension isomiR of hsa-miR-423-3p was significantly higher in tumor compared to normal (t-test p = 0.00023).

Conclusions: This study investigated the isomiR expression and cis-regulation of the biogenesis of isomiRs in human lymphoblastoid cell lines. Numerous SNPs were found significantly associated with 5' isomiRs including substitution, trimming, extension and addition. Empirical evidence was observed in supporting that genetic variant contributes to the composition of 5'- isomiRs by altering the sequence of precursor miRNA, and therefore the secondary structure of the hairpin and enzyme cleavage or transcript editing of isomiRs afterwards. The genetic regulation of isomiR maturation may shed light on our understanding of the miRNA post-transcriptional regulation.

P21
IPDB: Integrated Pregnancy Database with Clinical and Omics Data

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

[1] Department of BioHealth Informatics, Indiana University, Indiana University–Purdue University Indianapolis, IN [2] Department of Health Services Administration and Policy, Temple University College of Public Health, Philadelphia, PA [3] Department of Obstetrics & Gynecology, Indiana University School of Medicine, Indianapolis, IN
Presenter's email address: pkothiya@iupui.edu; Presenter's phone number: 5516894953

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

Resources and methods:
The IPDB source data include (i) Redcap data manually extracted from EHR of two local hospitals that comprise various clinical and demographic data (~600 clinical variables), including vital signs, past and current diagnoses, family and social history, medications, laboratory values, and maternal and neonatal outcomes; (ii) biobank samples and resulting omics data and other laboratory tests; and (iii) longitudinal data, such as smoking and drinking before and during pregnancy and clinical visits. Data cleaning, preprocessing, reformatting, normalization, and standardizing are first performed before data merging.

The IPDB was built with a multi-modal concept using a relational database design to link the heterogeneous information. The core database tables are pregnancy, mother, clinical_visit, biospecimen, and neonatal_outcomes. Special entity and relationship tables are carefully designed for complicated situations, such as one pregnancy with multiple babies, multiple pregnancies, abnormal birth outcomes, and connecting EHR data with biosample analysis. The other design principle is the extensibility of the database, which can easily incorporate additional clinical variables, extra lab tests, and new analytical results. Yet database design considerations include scalability to allow new datasets, intensive query processing, data access controls, and system recovery.

Results:
We have developed a prototype of the IPDB using a MySQL database that integrates longitudinal clinical information, biosample inventories, and omics data analyses (metabolomics, genomics, etc.). Users can access and manipulate data through a graphical user interface (GUI) implemented by R-shiny. Users can conveniently query information over multiple pregnancies and evaluate various outcomes. Moreover, it empowers users to obtain postpartum information and laboratory values. A visualization dashboard has been developed with descriptive statistics and illustrations over selected data on maternal age, gestational age, race, health factors, and outcomes.

Summary:
The cross-modality feature of IPDB integrates multiple data streams, from structured EHR data to genomics and metabolomics, in a format facilitating future bioinformatics analyses. It can significantly improve the performance of phenotyping and prediction algorithms, enabling knowledge discovery at the patient and population level.

P22

CASowary: CRISPR-Cas13 Guide RNA Predictor for Transcript Depletion

Alexander Krohannon[1], Mansi Srivastava[1], Simone Rauch[2], Rajneesh Srivastava[1], Bryan C. Dickinson[2], Sarath Chandra Janga[1,3,4*]

[1]Department of BioHealth Informatics, School of Informatics and Computing, Indiana University Purdue University, Indianapolis, Indiana 46202

[2]Department of Chemistry, The University of Chicago, Chicago, Illinois, USA

[3]Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, Indiana 46202

[4]Department of Medical and Molecular Genetics, Indiana University School of Medicine, Medical Research and Library Building, Indianapolis, Indiana, 46202

Recent discovery of the gene editing system - CRISPR (Clustered Regularly Interspersed Short Palindromic Repeats) associated proteins (Cas), has resulted in its widespread use for improved understanding of a variety of biological systems. Cas13, a lesser studied Cas protein, has been repurposed to allow for efficient and precise editing of RNA molecules. The Cas13 system utilizes base complementarity between a crRNA/sgRNA (crispr RNA or single guide RNA) and a target RNA transcript, to preferentially bind to only the target transcript. Unlike targeting the upstream regulatory regions of protein coding genes on the genome, the transcriptome is significantly more redundant, leading to many transcripts having wide stretches of identical nucleotide sequences. Transcripts also exhibit complex three-dimensional structures and interact with an array of RBPs (RNA Binding Proteins), both of which may impact the effectiveness of transcript depletion of target sequences. However, our understanding of the features and corresponding methods which can predict whether a specific sgRNA will effectively knockdown a transcript is very limited. Here we present a novel machine learning and computational tool, CASowary, to predict the efficacy of a sgRNA. We used publicly available RNA knockdown data from Cas13 characterization experiments for 555 sgRNAs targeting the transcriptome in HEK293 cells, in conjunction with transcriptome-wide protein occupancy information on RNA. Our model utilizes a Decision Tree architecture with a set of 112 sequence and target availability features, to classify sgRNA efficacy into one of four classes, based upon expected level of target transcript knockdown. After accounting for noise in the training data set, the noise-normalized accuracy exceeds 70%. Additionally, highly effective sgRNA predictions have been experimentally validated using an independent RNA targeting Cas system – CIRTS, confirming the robustness and reproducibility of our model's sgRNA predictions. Utilizing transcriptome wide protein occupancy map generated using POP-seq in HeLa cells against publicly available protein-RNA interaction map in Hek293 cells, we show that CASowary can predict high quality guides for numerous transcripts in a cell line specific manner. Application of CASowary to whole transcriptomes should enable rapid deployment of CRISPR/Cas13 systems, facilitating the development of therapeutic interventions linked with aberrations in RNA regulatory processes.

P23

The Updated Version of Global Evaluation of SARS-CoV-2/nCoV-19 Sequences (GESS) Database

Kailing Li[1], Audrey Wang[2], Alex Lu[2], Sheng Liu[3], Shuyi Fang[1], Lei Yang[4,5], Kai Yang[4,5], Jun Wan[1,3,5]

[1]Department of BioHealth Informatics, School of Informatics and Computing, Indiana University Purdue University, Indianapolis, Indiana 46202

[2]Park Tudor School, Indianapolis, Indiana 46240

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

[4]Department of Pediatrics, Indiana University School of Medicine, Indianapolis, Indiana 46202

[5]Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, Indiana 46202

The Global Evaluation of SARS-Cov-2/nCoV-19 Sequences (GESS) v2 (GESS v2, https://shiny.ph.iu.edu/GESS_v2/) is an updated version of GESS, which has offered a handy query platform for results based on our comprehensive analysis on over one million high-coverage and high-quality SARS-CoV-2 complete genomes from GISAID. As of July 30th, 2021, the GESS v2 has analyzed 2,333,322 hCOV-19 viral genome sequences. In the first published version, the focus of GESS database allows users to search SNVs at any individual or multiple SARS-CoV-2 genome positions, or within a chosen genomic region, or in a certain country/area of interest. It reveals the information of the geographical relations of SNVs around the world and within the US or other countries. In addition to these functions in the first version, the updated GESS v2 extends search with more useful options, e.g., sequence search given nucleotides, amino acids, or proteins. The new version explores the emerging SNVs in a given month to monitor new dominant SNVs in time during the viral transmission. The functions of “Lineage Analysis” on GESS v2 can identify specific lineages or SNVs which are over-represented in certain countries of users’ interests. GESS v2 also provides another handy tool called “Lineage Migration” which can intuitively show how the selected lineage migrates from one country to others on a world map. Overall, with these significant updates, the GESS v2 will be much more powerful to serve researchers as a great platform to find critical information about SARS-CoV-2 study and prevalence prediction, public health policy making, and vaccine designs.

P24

Investigating Inheritability of Alternative Splicing in Alcohol Use Disorders

Rudong Li[1,2], Andy B. Chen[1,2], Steven X. Chen[3], Dongbing Lai[2], Yunlong Liu[1,2]

[1]Center for Computational Biology and Bioinformatics, Indiana University School of Medicine

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

[3]Department of Medicine, Indiana University School of Medicine

Alcohol use disorder (AUD) is a serious and prevalent psychiatric disorder characterized by excessive alcohol consumption. Despite environmental factors, early studies demonstrated that the disorder is genetically heritable. Yet much about the genetic basis of AUD have remained to be revealed. Majority of research efforts on AUD focused on identification of risk loci or genes through genome-wide association studies (GWAS). Nevertheless, less was known for regulatory events at the RNA level (e.g. mRNA alternative splicing). In the present work, we have built prediction models for 6284 RNA alternative splicing events transcriptome-wide using the prefrontal cortex data from Common Mind Consortium (CMC) via Elastic net (EN) regularization. Using these models, we assessed genetically inheritable associations of alternative splicing to alcoholic outcomes, including phenotypes DSM-4 and symptom counts (SXCT), across 8038 individuals in the Collaborative Studies on Genetics of Alcoholism (COGA) via Generalized Estimation Equation (GEE). We have found 27 events showing significant associations of genetic inheritability, by an FDR threshold of 0.05. To replicate the results, we constructed an independent dataset of 2856 individuals using the Australian twin-family study of alcohol use disorder (OZ-ALC) by excluding overlapping individuals from COGA. Six of the 27 have been successfully replicated under FDR 0.05. The 6 events were among the top of significance in the 27 candidates and showed consistency in the effect sizes estimated in COGA and OZ-ALC, respectively. Host genes for these events included a lncRNA and 5 protein-coding genes. To further investigate the biological functions of the events, we explored their potential targets genes via differential expression analyses by stratifying CMC samples upon high and low splicing rates, for each of the events respectively. Gene Ontology (GO) enrichment and Gene Set Enrichment Analysis (GSEA) were performed on the differential genes identified for each event, and we found most of them were enriched for regulation of immune response or neurologic process. In conclusion, the present work investigated the genetic basis of AUD from the RNA splicing prospective. In additional to the novel splicing events, genes and variants identified, the result also consolidated the relevance of neurodegenerative disorder and immune response to AUD. This research will fill in the blank for the knowledge of RNA splicing in genetic regulation of AUD.

P25

The Alternative Splicing Landscape In Multiple Myeloma Is Determined by *IGH* Translocations and Mutations of RNA Processing Genes

Enze Liu, Sabine Wenzel, Brian A. Walker

Melvin and Bren Simon Comprehensive Cancer Center, Division of Hematology Oncology, School of Medicine, Indiana University, Indianapolis, IN, USA

Background: Alternative Splicing (AS) plays a key role in regulating numerous cellular processes in both normal and malignant cells. Previous studies have revealed mutations in the spliceosome complex, such as *SF3B1,* can cause increased AS frequencies in multiple myeloma (MM) patients, and patients with increased levels of AS are associated with a poor prognosis. Other frequently mutated genes involved in RNA processing include *DIS3* and *FAM46C*, thus, systematically investigating other causes of AS abnormalities and pathologies in MM patients is highly necessary.

Method: RNA-seq data from 598 newly diagnosed MM patients from the MMRF CoMMpass study were utilized to generate AS comparisons. They were previously annotated for cytogenetic, copy number, and mutation data (version IA16). RNA-seq data were aligned and AS events are called from aligned reads. Thresholds were determined to find high-quality differentially spliced events and were filtered for those also present in normal PCs (GSE110486). Dysregulated pathways caused by differential splicing or by the interactions between differential expression and differential splicing are identified. Survival analysis was performed on clinical annotations of 598 NDMM patients for evaluating risks of AS.

Results: We compared 16 major cytogenetic subgroups, including Ig translocations (t(4;14), t(14;16), t(11;14)), hyperdiploidy, mutations in *KRAS*, *NRAS*, *BRAF*, *FAM46C*, *SF3B1*, *DIS3* and *TP53*, combined events (t(4;14) plus *DIS3* mutation), as well as those with biallelic abnormalities (*DIS3*, *FAM46C*, and *TP53*). Samples with *SF3B1* hotspot mutations identified the greatest number of significantly and differentially spliced (SDS) events (n=862), and samples with any *SF3B1* mutation had approximately half as many. Translocations and DIS3-related comparisons reflect more number (N>200) of SDS events compared to others. Bi-allelic comparisons *consistently reflected more SDS events than mono-allelic events.* Combinational event comparison reflected significantly more (P<0.01) SDS events than corresponding single event comparisons. *RAS* related comparisons reflect the minimum number of SDS events.

Identified AS events exemplify heterogeneity in all comparisons. Few of them are consistent among translocations (t(11;14), t(4;14) and t(14;16)) and *DIS3*-related comparisons (*DIS3* mutation, Biallelic *DIS3* and t(4;14) plus *DIS3* mutation), indicating their similar roles .

AS heterogeneity also leads to functional heterogeneity. Pathway analysis has been conducted for each comparison and only few overlapped pathways are found, even though some pathways fall into the same categories (e.g. metabolic processes).

High-risk events were identified through survival analysis such as Skipped exon on *AIB1* in both bi-allelic *TP53* (dPSI=0.12, Hazard Ratio (HR) =3.0, P=4.8e-8, OS) and *TP53* mutation (dPSI=0.12, HR=2.5, P=1.5e-7, OS)

P26

Detecting Cell-type Specific Variations within Pathway Gene Interaction in AD using Covariance Regression

Xiaoyu Lu[1], Chi Zhang[1,2], Yi Zhao[2*], Sha Cao[1,2*]

[1]Department of BioHealth, Indiana University–Purdue University Indianapolis, Indianapolis, IN, 46202, USA

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

Alzheimer's disease (AD) and related dementia are among the greatest public health challenges in the United States. A deeper understanding of the disease pathology on a molecular and cellular level is critical to devise treatments to slow down the disease progression and better understand the mechanisms behind AD. We here present a statistical framework for extracting the variability of gene-gene interaction in a disease-, sex- and cell type- specific manner, to identify AD-associated markers and pathways and to articulate disease mechanisms on cellular and molecular levels. Our covariance regression based analytical pipeline represents a statistically solid tool to investigate the abnormalities in gene interaction on pathway-level using single-nucleus RNA sequencing data, and to make inference on key covariates such as sex. Then we applied our method to the Religious Orders Study and Memory and Aging Project (ROS/MAP) single nucleic RNA-Sequencing dataset. In general, majority of the pathways tend to have lower co-expression strength among its genes in AD cells than healthy cells for female subjects in astrocyte, neuron and microglial. This also holds true for neuron cells in male subjects. While not many pathways are differentially co-expressed in male AD vs healthy subjects, neuron and astrocyte are the two major cell types with the largest number of alternatively co-expressed pathways in female subjects, with 3972 and 4218 pathways respectively. These altered pathways belong to known pathway categories, such as synapse, neuronal development, lipid metabolism, calcium transport, etc. Interestingly, cell subtype heterogeneity analysis demonstrated that under AD condition, different subtypes of cells tend to be more homogeneous particularly in female subjects. While the same phenomenon doesn't hold true for male subjects. This indicates that cell subtypes that perform specific functions in normal conditions now tend to homogenize their functions in AD. On an independent brain single cell RNA-Seq data, we validated our model by observing a consistent trend of the pathway connectivity patterns.

P27

Needles in a Stack of Needles: Finding Significant Thermal Shifts in Thermal Proteome Profiling with Inflect Statistical Analysis and Bioinformatics

Neil McCracken,* Hao Liu,**,+ Amber Mosley*,#

*Department of Biochemistry and Molecular Biology, **Department of Biostatistics, Indiana University School of Medicine, Indianapolis, IN; +current address: Rutgers School of Public Health, Department of Bioinformatics and Epidemiology, Rutgers, NJ; #Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, IN

Thermal Proteome Profiling (TPP) has been used to discover drug targets and detect changes within the cellular "melt-ome" for nearly a decade.[1-3] The method typically involves treating cells or tissues with a drug or vehicle followed by thermal treatment, tryptic digestion, isobaric labeling and tandem MS. Protein abundance values are then used to determine changes in protein stability (melt shifts) for treatment relative to control. We have developed an R based workflow, available to the scientific community, that has been used to determine statistically significant melt shifts in a TPP experiment. We have also implemented bioinformatics tools that allow for a holistic understanding of experimental outputs within the context of previously defined protein functional groups. Our R based program, Inflect-SP, is an extension of our earlier work.[4] Briefly, the workflow normalizes abundance data, corrects for potential thermal treatment anomalies, fits melt curves, calculates melt shifts, selects proteins based on statistical criteria and reports results. The workflow has been coded to allow for any number of vehicle/control or treatment/condition experiments that are analyzed simultaneously. Analysis uses normalized abundance variability at the melt point to determine the p-value based significance of the shift. Significant shifts are then analyzed further by STRING and PANTHER based tools using available application programming interface (API).

Two recent and publicly available data sets were used to test the functionality of our program.[5, 6] We defined significant protein melt shifts as those with melt shift p-values less that 0.05. In the case of the dataset from Kalxdorf et.al., K562 cells in biological triplicate were treated with Dasatinib, a tyrosine kinase inhibitor and cancer treatment. Several of the proteins identified are tyrosine kinases with some being previously reported targets for Dasatinib. At least one kinase, YES1, was not discussed in the original report by Kalxdorf et.al suggesting increased statistical utility of our tool in dataset interpretation. The work from Sridharan involved treating Jurkat cells in biological replicate with ATP to understand related signaling. An example output using the Sriharan data set is in Figure 1 and shows several select proteins (green boxes) that are nucleotide binding based on Uniprot.[7] These STRING results reinforce the utility of our workflow to detect potential protein-protein interactions in the data sets. Overall, our Inflect-SP is an essential tool for any TPP data analysis pipeline.

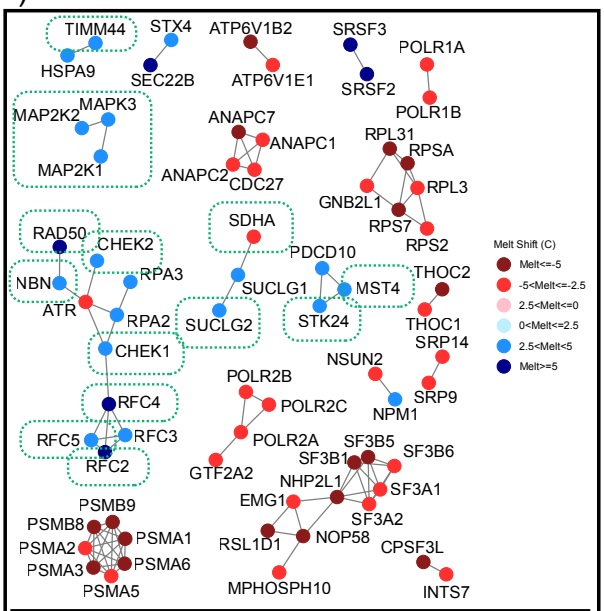


Figure 1. STRING output from Sriharan melt shift analysis using Inflect program. STRING Criteria 0.97, PSM>2, UP>3, p<0.05, Melt Limit:>3˚C. Select ATP utilizing proteins highlighted. Nucleotide binding proteins identified by Uniprot are highlighted.

1. Martinez Molina, D, et. al, Monitoring drug target engagement in cells and tissues using the cellular thermal shift assay. *Science* **2013,** *341* (6141), 84-7.

2. Savitski, M. M, et. al. Tracking cancer drugs in living cells by thermal profiling of the proteome. *Science* **2014,** *346* (6205), 1255784.

3.mJarzab, A, et al Meltome atlas-thermal proteome stability across the tree of life. *Nat Methods* **2020,** *17* (5), 495-503.

4. McCracken, N. A.; Peck Justice, S. A.; Wijeratne, A. B.; Mosley, A. L., Inflect: Optimizing Computational Workflows for Thermal Proteome Profiling Data Analysis. *J Proteome Res* **2021**.

5. Kalxdorf, M, et. al. Cell surface thermal proteome profiling tracks perturbations and drug targets on the plasma membrane. *Nature Methods* **2021,** *18* (1), 84-91.

6. Sridharan, S., et al., Proteome-wide solubility and thermal stability profiling reveals distinct regulatory roles for ATP. *Nature Communications* **2019,** *10* (1), 1155.

7. Consortium, T. U., UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research* **2020,** *49* (D1), D480-D489.

P28

Identification of Splice Isoforms in T helper (Th) Cells using Nanopore-based cDNA Sequencing

Quoseena Mir[1], Rajneesh Srivastava[1], Benjamin Ulrich[2,3], Mark H. Kaplan[2,3,4], Sarath Chandra Janga[1,4,5] *

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

[2]Department of Microbiology and Immunology, Indiana University School of Medicine, Indianapolis, Indiana

[3]Wells Center for Pediatric Research, Indiana University School of Medicine, Indianapolis, Indiana

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

[5]Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Health Information and Translational Sciences (HITS), Indianapolis, Indiana

Differentiation of naïve T cells to effector subsets is regulated through a complex gene regulatory network resulting in transcriptomic alterations. Our understanding of how this process is controlled transcriptionally is limited. Nanopore-based RNA sequencing technology enables the identification of full-length transcripts and alternatively spliced isoforms. Here, we present a single molecule transcriptome map of differentiated Th1, Treg and naïve T cells, obtained from 8 weeks mouse spleen, to characterize the full-length transcripts and determine the splicing potential of gene to produce spliced isoforms. We obtained 6.47M, 9.5M and 10.85 M reads from cDNA sequencing of Treg, Th1 and naïve T cells using Nanopore with mean length of 876.4 bp, 706.7bp and 704.6bp respectively. Quality filtered reads were aligned to the mouse genome (mm10) and resulted in ~5.7 million mapped reads (>85% alignment rate with 15% error allowance) in respective cell type. We calculated the splicing efficiency (SE) of each gene by dividing the summation of read depth per base of exons in a gene with the sum of read depth per base for respective gene. On comparison of the genes with their SE scores, falling in the interquartile range (i.e. 0.25-0.75) along with several other parameters (such as #exons>2, #mapped reads≥10), we obtained 245 genes (including 26 RBPs) in Th1 and 486 genes (including 84 RBPs) in Treg cells, exhibiting ±1.5 fold difference in SE with respect to naïve T cells. Function annotation analysis of these genes revealed significantly enriched ($p<0.01$) biological processes such as protein transport, ubiquitin mediated proteolysis, cell cycle, autophagy in Th1 cells and RNA splicing & transport, ribosome biogenesis & translational regulation, phosphate metabolism, chromosome segregation etc. in Treg cells along with; cellular response to DNA damage stimulus, covalent chromatin modification and DNA repair, enriched in both the cell types. Hence, this study decodes the transcriptional signatures of differentiating T cell subtypes at single molecule resolution.

P29

Thermal Proteome Profiling of Triple Negative Breast Cancer Cells Treated with IB-DNQ and Rucaparib

Avery Runnebohm[1], Sagara Wijeratne[1], Neil McCracken[1], Aruna Wijeratne[1,2], Gitanjali Roy[1], Edward Motea[1,2], David Boothman[1,2], Amber Mosley[1,2,3]

[1]Department of Biochemistry and Molecular Biology, Indiana University School of Medicine, Indianapolis, IN; [2]IU Simon Comprehensive Cancer Center, Indianapolis, IN; [3]Center for Computational Biology and Bioinformatics, Indianapolis, IN

Triple negative breast cancer (TNBC) accounts for 15-20% of all breast cancer cases, with TNBC patients having a higher risk of relapse and lower survival rates. TNBC is defined by the lack of the estrogen, progesterone, and epidermal growth factor 2 receptors. Due to the lack of these receptors, treating the tumors with commonly used therapies, including hormonal therapies, is ineffective. Thus, more research is required to find TNBC-specific drug targets for better therapeutic development. One such potential target is NAD(P)H:quinone oxidoreductase 1 (NQO1), which is highly expressed in TNBC tissue and lowly expressed in normal tissues (1). The NQO1 bio-activatable drug isobutyl-deoxynyboquinone (IB-DNQ) acts as a futile cycling substrate for NQO1, which leads to an accumulation of reactive oxygen species that results in DNA damage (2). Upon DNA damage, PARP1 is hyperactivated to initiate DNA repair. However, inhibition of PARP1 using Rucaparib, in combination with IB-DNQ treatment, is proposed to synergistically induce cell death. To further understand the mechanism of how Rucaparib and IB-DNQ induce death of TNBC cells, MDA-MB-231 TNBC cells expressing endogenous NQO1 or the poorly expressed NQO1*2 variant were subject to IB-DNQ, Rucaparib, and a combination treatment followed by thermal proteome profiling (TPP). TPP is a mass spectrometry based approach used to evaluate protein thermal stability. Protein melt temperature shifts report the stability of proteins under the different conditions, which can help identify cellular pathways that may be effected by the treatments.

We hypothesized that in the presence of NQO1, IB-DNQ and Rucaparib treatment would result in changes in protein thermal stability, illuminating how the different treatments altered the proteome. Following TPP analysis, protein abundance values were used to produce melt curves using the R package “Inflect” (3). Inflect also uses STRING and Panther to identify protein interactions and cellular pathways for significantly stabilized or destabilized proteins. Our analysis found that several cell cycle regulators were stabilized when cells were treated with IB-DNQ, suggesting IB-DNQ induced changes in cell cycle. Additionally, several ribosomal subunits were destabilized following combination treatment, which may correspond to a downregulation of protein synthesis. Currently, further analysis of this data set and comparison to phosphoproteomics data is being performed to better characterize the effects of Rucaparib and IB-DNQ on the proteome.

In summary, advanced mass spectrometry-based methodologies along with innovative curve-fitting software have allowed for a more exploratory analysis of the proteome and protein-protein interactions within TNBC cells following drug treatment.

References:

1. Huang X, et al. Leveraging an NQO1 Bioactivatable Drug for Tumor-Selective Use of Poly(ADP-ribose) Polymerase Inhibitors. Cancer Cell. 2016 Dec 12; 30(6):940-952.
2. Bair JS, Palchaudhuri R, Hergenrother PJ. Chemistry and biology of deoxynyboquinone, a potent inducer of cancer cell death. J Am Chem Soc. 2010 Apr 21; 132(15):5469-78.
3. McCracken NA, Peck Justice SA, Wijeratne AB, Mosley AL. Inflect: Optimizing Computational Workflows for Thermal Proteome Profiling Data Analysis. J Proteome Res. 2021 Apr 2; 20(4):1874-1888.

P30

Protein Occupancy Profile-sequencing (POP-seq) for Global Mapping of Protein RNA Interactions Across Human Cell Lines

Mansi Srivastava[1], Neel Sangani[1], Gayathri Panangipalli[1], Rajneesh Srivastava[1], Quoseena Mir[1] Hunter Gill[1], Mark H. Kaplan[4], Sarath Chandra Janga[1,2,3*]

[1]Department of Biohealth Informatics, School of Informatics and Computing, Indiana University Purdue University Indianapolis; [2]Center for Computational Biology and Bioinformatics, [3]Department of Medical and Molecular Genetics, [4]Department of Microbiology and Immunology, Indiana University School of Medicine, Indianapolis, IN

Interaction between RNA-binding proteins (RBPs) and RNA is critical for post-transcriptional regulatory processes[1]. Existing high throughput methods based on crosslinking of the protein-RNA are reported to contribute to biases in the resulting protein occupancy profiles[2–6]. We present Protein Occupancy Profile-Sequencing (POP-seq), a phase separation-based method that does not require crosslinking thus providing unbiased protein occupancy profiles on whole cell transcriptomes. In order to compare the robustness of identified protein occupied sites, POP-seq was implemented in two phases: the first phase comprised of UV crosslinking and no-crosslinking approaches on K562 and HepG2 cells, resulting ~200,000 peaks detected across the protocols. The results were further cross-validated with the publicly available ENCODE eCLIP profile for scores of RBPs in the two cell lines. Our analysis reveals that majority of detected genes (>70%) overlap between the two approaches indicating the reproducible nature of the generated interaction maps in the two cell lines. We observed an abundance of binding sites on the intronic region of the genomic location (~40%) for both the cell lines with maximum number of POP-seq peaks detected on protein-coding genes (>85%). Cross-validation with the ENCODE eCLIP profile of RBPs demonstrated relatively higher recall for UV-crosslinking compared to no-crosslinking approach in K562 and HepG2 cells respectively, indicating the robustness of both approaches. In the second phase, we expanded POP-seq on multiple cell lines (MCF7, A549, Jurkat and HEK293) using the no-crosslinking approach resulting in ~60,000 reproducible peaks between replicates. POP-seq detected differential peaks (~2000-6000) between pair of cell lines corresponding to ~1000-2000 genes that were enriched for vital cellular functions. Motif enrichment analysis revealed the presence of high confident RBP motifs on POP-seq identified sites including motifs for crucial transcription regulators SRSFs and HNRNPs. We also detected somatic and clinical SNPs on the motifs associated to POP-seq profiles, that could potentially contribute to dysregulated post- transcriptional functions. We used CRISPR Cas 9 system to evaluate the functionality of POP-seq sites and observed that identified interactions could potentially impact the expression of nearby exons. Altogether, our data supports POP-seq as a robust and cost-effective method enabling comprehensive mapping and understanding of post-transcriptional regulatory networks with a potential to be applied to primary tissues.

References:

1. Hentze, M. W., Castello, A., Schwarzl, T. & Preiss, T. A brave new world of RNA-binding proteins. *Nat. Rev. Mol. Cell Biol.* **19**, 327–341 (2018).
2. Baltz, A. G. *et al.* The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. *Mol. Cell* **46**, 674–690 (2012).
3. Schueler, M. *et al.* Differential protein occupancy profiling of the mRNA transcriptome. *Genome Biol.* **15**, R15 (2014).
4. Silverman, I. M. *et al.* RNase-mediated protein footprint sequencing reveals protein-binding sites throughout the human transcriptome. *Genome Biol.* **15**, R3 (2014).
5. Queiroz, R. M. L. *et al.* Comprehensive identification of RNA-protein interactions in any organism using orthogonal organic phase separation (OOPS). *Nat. Biotechnol.* **37**, 169–178 (2019).
6. Trendel, J. *et al.* The Human RNA-Binding Proteome and Its Dynamics during Translational Arrest. *Cell* **176**, 391-403.e19 (2019). Schueler, M. et al. Differential protein occupancy profiling of the mRNA transcriptome. Genome Biol. 15,R15 (2014).
7. Silverman, I. M. et al. RNase-mediated protein footprint sequencing reveals protein-binding sites throughout the human transcriptome. Genome Biol. 15, R3 (2014).
8. Queiroz, R. M. L. et al. Comprehensive identification of RNA-protein interactions in any organism using orthogonal organic phase separation (OOPS). Nat. Biotechnol. 37, 169–178 (2019).
9. Trendel, J. et al. The Human RNA-Binding Proteome and Its Dynamics during Translational Arrest. Cell 176, 391-403.e19 (2019).

P31

Integrative -omics for Discovery of Network-level Disease Biomarkers for Alzheimer's Disease

Linhui Xie[1], Bing He[2], Pradeep Varathan[2], Kwangsik Nho[3], Shannon Risacher[3], Andrew J. Saykin[3], Paul Salama[1], Jingwen Yan[2,3]

[1]Purdue School of Engineering and Technology, Indiana University Purdue University Indianapolis, Indianapolis, Indianapolis, IN, USA; [2]Department of BioHealth Information, Indiana University Purdue University Indianapolis, Indianapolis, IN, USA; [3]Department of Radiology and Imaging Sciences, Indiana University School of Medicine, Indianapolis, IN, USA. Contact e-mail: linhxie@iupui.edu

Background: A large number of genetic variations have been identified to be associated with Alzheimer's disease (AD) and related traits. But the downstream biology through which they exert effect on the development of AD remains unknown. Leveraging genomic, transcriptomic and proteomic data, and various biological networks, we proposed a modularity-constrained logistic regression model (M-logistic) to identify a set of functionally connected SNPs, genes and proteins related to AD. **Methods:** The GWAS genotype data, RNA-Seq gene expression data and Protein expression data were obtained from the ROS/MAP Project. Detailed data processing steps can be found in AMP-AD (https://www.synapse.org). In total, 179 subjects with full set of data were included (Table. 1). Using peptides as seed, we prefiltered the data and included 186 peptides, 743 unique genes and 822 Single-nucleotide polymorphisms (SNPs) from upstream of their connected genes (Fig. 1). Functional interactions from REACTOME database [1] and the SNP-gene mapping relationship were used to construct the prior network. These functional connected -omics features were used to classify AD patients from non-symptomatic group, including cognitive normal and mild cognitive impairment. **Results:** Across 10-fold cross-validation, we observed that M-logistic largely outperforms other state-of-the-art models (Table. 2). For feature selection, the proposed model identified 305 features predictive of disease status, which appeared in more than 6 folds. When mapped back to the prior network, these features formed three big sub-networks with > 10 nodes (Fig. 2). For 20 genes in the largest sub-network, their brain-wide expression profile in the Allen Human Brain Atlas were found to be mostly correlated with the activation map of brain function term "vision" in Neurosynth [2] (Fig. 3). **Conclusion:** M-Logistic identified a set of SNPs, genes and proteins that are not only predictive of AD but also densely connected in the prior network. The trans-omic paths in the selected sub-networks indicates a potential pathway underlying the development of AD from SNPs to gene expression, protein expression and ultimately brain functional and structural changes. Future efforts are warranted to enable integrative analysis of multi-omics data from decoupled subjects. [1] Fabregat et al., *Nucleic acids research*, 2018. [2] Yarkoni et al., *Nature Method*, 2011.

Figure 1. Major steps taken to pre-filter SNPs, genes and proteins. (a) 186 peptides in the proteomic layer were used as seeds. (b) 186 peptides were mapped to 126 unique genes (gene set A). (c) 954 genes were found to interact with gene set A in REACTOME functional protein interaction database. (d) For gene set A and B, we extract SNPs that are located within upstream 5K boundary and have potential effect on transcription factor binding activity according to SNP2TFBS database.

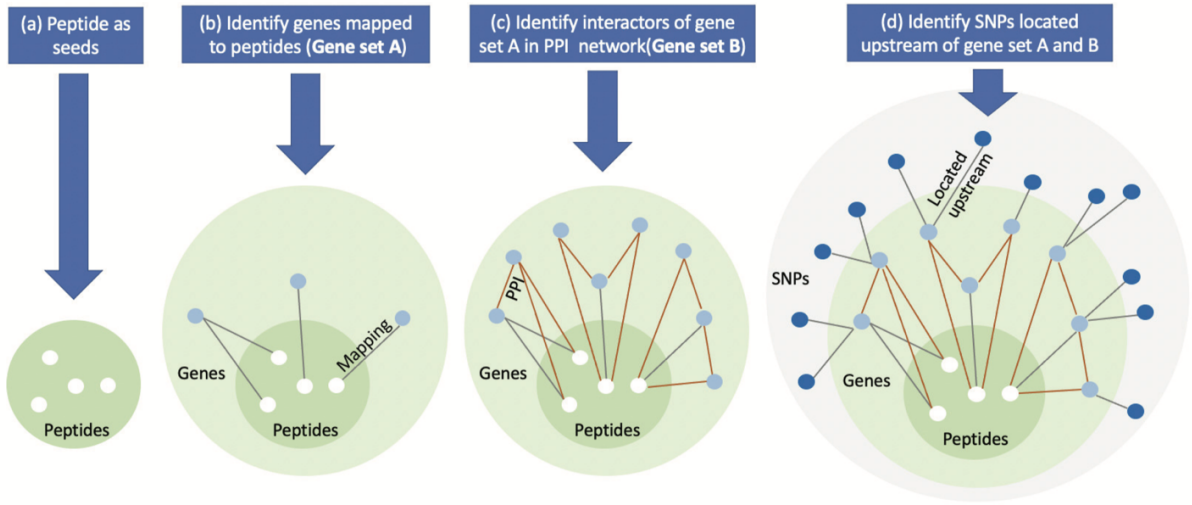


Figure 2. Features selected by M-Logistic were mapped back to prior network. Three sub-networks were found with more than 10 nodes.

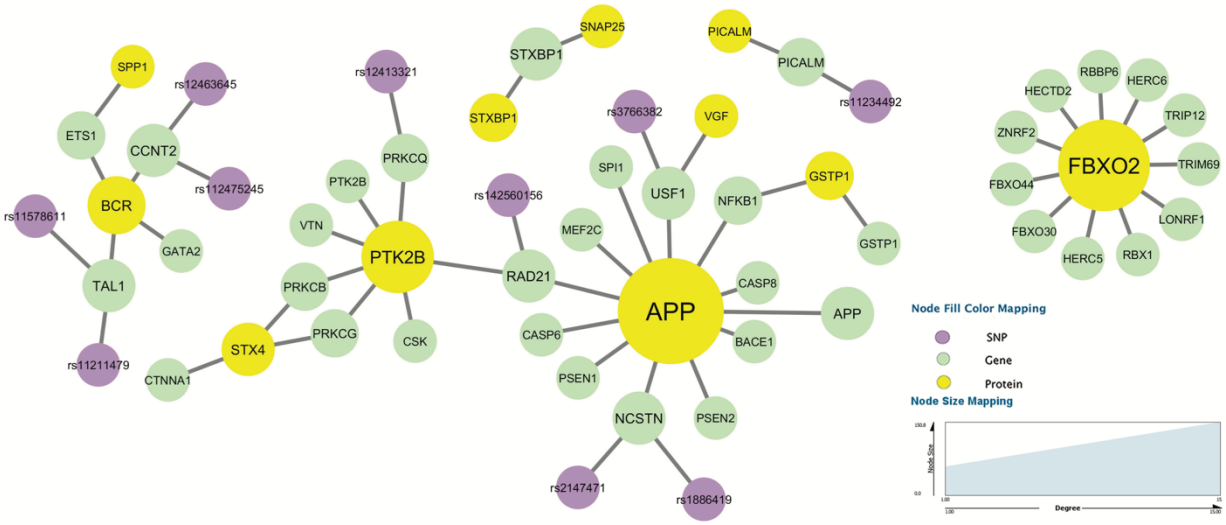
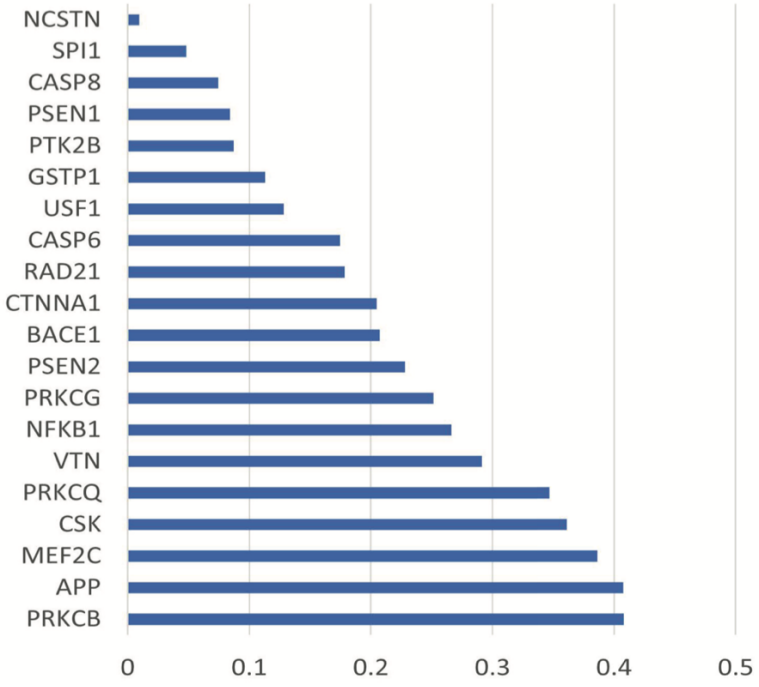


Figure 3. The correlation between brain activation map of “vision” and the brain wide expression profile of 20 genes in the largest sub-network



P32

Computational Modeling of Metabolic Flux of Branched Chain Amino Acids in Cancer and Alzheimer's Disease

Shaoyang Huang[1,2+], Kevin Hu[1,2+], Noah Meroueh[1,2+], Jonathan Yang[1,2+], Grace Yang[1,2+], Wennan Chang[1,3], Sha Cao[1,4*], Chi Zhang[1,5*]

[1]Center for Computational Biology and Bioinformatics, [4]Department of Biostatistics, [5]Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, IN, USA; [2]Carmel High School, Carmel, IN, USA; [3]Department of Electric Computing Engineering, Purdue University, West Lafayette, IN, USA. Contact e-mail: Chi Zhang, czhang87@iu.edu
+: Equal contributions, *: corresponding authors

Both cancer and Alzheimer's disease have altered branched-chain amino acids (BCAA) metabolisms. Recent studies identified the ratio of leucine, isoleucine, and valine metabolism determines the disease characteristics. In this study, we reconstructed the BCAA metabolic modules to enable the flux estimation of BCAA modules. Our analysis characterized the flux distribution of BCAA metabolism including synthesis of branched-chain fatty acids (BCFA) and production of acetyl-CoA, succinyl-CoA and propanoyl-CoA. We identified cancer trends to have a lower-level isoleucine metabolism to succinyl-CoA and propanoyl-CoA and save most isoleucine for BCFA biosynthesis. In AD, most of the BCAA metabolisms happen in astrocyte, neuron, and oligodendrocyte progenitor cells with different characteristics

P33

Computational Modeling of Cell Type Specific Metabolic Rate of Glucose Flow and Glutaminolysis in Cancer Microenvironment

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

[1]Center for Computational Biology and Bioinformatics, [4]Department of Biostatistics, [5]Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, IN, USA; [2]Carmel High School, Carmel, IN, USA; [3]Department of Electric Computing Engineering, Indiana University Purdue University Indianapolis, Indianapolis, Indianapolis, IN, USA, IN, USA. Contact e-mail: Chi Zhang, czhang87@iu.edu
+: Equal contributions, *: corresponding authors

Glutaminolysis has been considered as a major carbon and energy source in cancer. Recent studies reported immune cells tend to use more glucose while cancer cells use more glutaminolysis for ATP production. However, we suspect the observations only made on cell lines or limited mouse orthotopic model do not reflect the general metabolic changes in real cancer tissue. In this study, we comprehensively estimate the metabolic flux and distribution of glycolysis, pentose phosphate, DNA synthesis, lactate production, TCA cycle, and glutaminolysis in multiple cancer types and normal tissue, by using tissue and single-cell transcriptomics data. Our analysis confirms the increased glucose uptake and upper part of glycolysis and decreased upper part of TCA cycle in all cancer types. However, increased lactate production and the second half of TCA cycle were only seen in certain cancer types. More interestingly, we did not see cancer tissues have shifted glutaminolysis comparing to normal tissue samples. Even cancer cells have the highest glutaminolysis rate compared to stromal and immune cells in tumor microenvironment, we did not observe stromal and immune cells have higher glucose consumption than cancer cells. Further analysis illustrates immune cells may have higher potential in glucose uptake (by glucose transporters), while their downstream consumption of glucose is much lower than cancer cells.