FOURTEENTH ROCKY MOUNTAIN BIOINFORMATICS CONFERENCE

DECEMBER 8 TO 10, 2016
SNOWMASS/ASPEN COLORADO

Conference Chair
Lawrence Hunter, PhD
University of Colorado Denver
School of Medicine
Welcome

Dear Rocky 2016 participant

Welcome to the 14th Rocky Mountain Bioinformatics Conference. The organizers hope that you enjoy the program, and find the meeting a productive opportunity to meet researchers, students and industrial users of bioinformatics technology.

We are grateful for your continued interest in the meeting. We are also grateful for the support of our sponsors. We want to thank and acknowledge the ongoing support of IBM who has provided significant sponsorship funds to this conference for the past fourteen years; to SomaLogic for their financial and administrative support as our new administrative host for Rocky through the GoldLab in Boulder, Colorado; to Biodesix for their continued support of this meeting and to our newest sponsor, PatientsLikeMe. We are so fortunate to have such support and we hope to have them support this conference for many years to come. It is only with the help of these sponsors that we can make this meeting as affordable as it is. Please seek out attendees from the sponsoring organizations, and let them know that their participation is important to you!

Finally, the meeting would simply not be possible without organizational help from Stephanie Hagstrom, Kathy Thomas, Morgan Williams, Elizabeth Wetherington and Heather McNurney. We hope you enjoy the science, the company, the hotel and the spectacular scenery of the Rocky Mountains.

Welcome!

Larry Hunter
Conference Chair
# AGENDA AT-A-GLANCE

All sessions will take place at Snowmass Convention Center CC

## WEDNESDAY – DECEMBER 7, 2016

<table>
<thead>
<tr>
<th>Time</th>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>4:00 PM–6:00 PM</td>
<td>REGISTRATION</td>
</tr>
</tbody>
</table>

## THURSDAY – DECEMBER 8, 2016

<table>
<thead>
<tr>
<th>Time</th>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:00 AM</td>
<td>BREAK</td>
</tr>
<tr>
<td>8:00 AM–6:00 PM</td>
<td>REGISTRATION</td>
</tr>
<tr>
<td>9:00 AM</td>
<td>KEYNOTE 1: Joshua M. Stuart, UC Santa Cruz</td>
</tr>
<tr>
<td></td>
<td>The Blue-prints of Tumors Uncovered Through Network Integration</td>
</tr>
<tr>
<td>9:45 AM</td>
<td>OP01 to OP04</td>
</tr>
<tr>
<td>10:25 AM</td>
<td>BREAK</td>
</tr>
<tr>
<td>10:45 AM</td>
<td>Keynote 2: Renee Deehan Kenney, PatientsLikeMe</td>
</tr>
<tr>
<td></td>
<td>Measuring Disease Through Real World Evidence and Biological State: Impact For Patients (and Animal Models, too)</td>
</tr>
<tr>
<td>11:15 AM</td>
<td>OP05 to OP08</td>
</tr>
<tr>
<td>12:00 PM</td>
<td>SKI BREAK</td>
</tr>
<tr>
<td>4:00 PM</td>
<td>KEYNOTE 3: Marco Masseroli, Politecnico di Milano</td>
</tr>
<tr>
<td></td>
<td>Next Generation Genomic Computing</td>
</tr>
<tr>
<td>4:30 PM</td>
<td>OP09 to OP12</td>
</tr>
<tr>
<td>5:10 PM</td>
<td>BREAK</td>
</tr>
<tr>
<td>5:30 PM</td>
<td>OP13 to OP15</td>
</tr>
<tr>
<td>6:00 PM</td>
<td>BREAK</td>
</tr>
<tr>
<td>6:30 PM</td>
<td>BANQUET: Il Poggio Restaurant, Snowmass Village</td>
</tr>
</tbody>
</table>
## Friday – December 9, 2016

<table>
<thead>
<tr>
<th>Time</th>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:00 AM–6:00 PM</td>
<td>REGISTRATION</td>
</tr>
<tr>
<td>8:30 AM</td>
<td>BREAK</td>
</tr>
</tbody>
</table>
| 9:00 AM | KEYNOTE 4: Laura K. Wiley, University of Colorado Anschutz Medical Campus  
Precision Medicine and the Learning Healthcare System: Leveraging Informatics to Improve Care |
| 9:45 AM | OP16 to OP19                                                            |
| 10:25 AM | BREAK                                                                   |
| 10:45 AM | Keynote 5: Alan Williams, SomaLogic                                     
Data Analytics and the SOMAscan™ Proteomic Platform |
| 11:15 AM | OP20 to OP23                                                            |
| 12:00 PM | SKI BREAK                                                              |
| 4:00 PM | KEYNOTE 6: Temple F. Smith, Boston University                          
Current Insights into the Evolution of the Genetic Translation System and the Genetic Code Itself |
| 4:30 PM | OP24 to OP27                                                            |
| 5:10 PM | BREAK                                                                   |
| 5:30 PM | OP28 to OP30                                                            |
| 6:00 PM | BREAK                                                                   |
| 6:30 PM | POSTER SESSION                                                         |

## Saturday – December 10, 2016

<table>
<thead>
<tr>
<th>Time</th>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:00 AM–11:00 AM</td>
<td>REGISTRATION</td>
</tr>
<tr>
<td>8:00 AM</td>
<td>BREAK</td>
</tr>
</tbody>
</table>
| 9:00 AM | KEYNOTE 7: Krista Meyer, Biodesix: Making Medicine Personal®           
Leveraging Biology to Guide Feature Enrichment for Designing Multivariate Classifiers for Clinical Tests |
| 9:30 AM | OP31 to OP35                                                            |
| 10:20 AM | BREAK                                                                   |
| 10:40 AM | OP36 to OP40                                                            |
| 11:30 AM | Keynote 8: Kirk E. Jordan, IBM Research UK                             
Data Centric Cognitive Computing: IBM’s Direction, Workflow Challenges and Opportunities |
| 11:50 AM | CLOSING RAFFLE AND AWARDS                                              |
# Detailed Agenda

All sessions will take place at Snowmass Convention Center CC

## Wednesday – December 7, 2016

<table>
<thead>
<tr>
<th>Time</th>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>4:00 PM–6:00 PM</td>
<td><strong>Registration</strong></td>
</tr>
</tbody>
</table>

## Thursday – December 8, 2016

<table>
<thead>
<tr>
<th>Time</th>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:00 AM–6:00 PM</td>
<td><strong>Registration</strong></td>
</tr>
<tr>
<td>8:00 AM</td>
<td><strong>Breakfast</strong></td>
</tr>
</tbody>
</table>
| 09:00 AM | **KEYNOTE 1**: Joshua M. Stuart, UC Santa Cruz  
*The Blue-prints of Tumors Uncovered Through Network Integration*** |
| 09:45 AM | **OP01**: IndeCut: A Cut-norm Based Method for Evaluating Independent and Uniform Sampling in Network Motif Discovery Algorithms  
*Mitra Ansariola, Oregon State University, Corvallis*** |
| 09:55 AM | **OP02**: Reproducible Computational Workflows with Continuous Analysis  
*Brett Beaulieu-Jones, University of Pennsylvania*** |
| 10:05 AM | **OP03**: Application Ontologies Supporting Phenotyping from Clinical Text  
*Wendy Chapman, University of Utah*** |
| 10:15 AM | **OP04**: SPARQLer: Making Knowledge Functional  
*Daniel McShan, University of Colorado School of Medicine*** |
| 10:25 AM | **Break**                                                                 |
| 10:45 AM | **KEYNOTE 2**: Renee Deehan Kenney, PatientsLikeMe  
*Measuring Disease Through Real World Evidence and Biological State: Impact For Patients (and Animal Models, too)*** |
| 11:15 AM | **OP05**: Improved Network Ontology Analysis by Segmentation  
*Ananda Mondal, Claflin University*** |
| 11:25 AM | **OP06**: An Image Phenotyping Environment Based on Open-Source Tools  
*Brian Chapman, University of Utah*** |
| 11:35 AM | **OP07**: InterViewer, a new Cytoscape-based viewer that displays interactions between selected sets of proteins  
*Marek Tutaj, Medical College of Wisconsin*** |
| 11:45 AM | **OP08**: CAMSA: a Tool for Comparative Analysis and Merging of Scaffold Assemblies  
*Max Alekseyev, George Washington University*** |
| 12:00 PM | **Ski Break**                                                             |
| 4:00 PM  | **KEYNOTE 3**: Marco Masseroli, Politecnico di Milano  
*Next Generation Genomic Computing*** |
<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>4:30 PM</td>
<td>OP09: Analysis of Tobacco Users Admitted to Intensive Care Units</td>
</tr>
<tr>
<td></td>
<td>Andrey Soares, University of Colorado School of Medicine</td>
</tr>
<tr>
<td>4:40 PM</td>
<td>OP10: A new molecular signature approach for prediction of driver cancer</td>
</tr>
<tr>
<td></td>
<td>pathways from transcriptional data</td>
</tr>
<tr>
<td></td>
<td>Boris Reva, Icahn School of Medicine at Mount Sinai</td>
</tr>
<tr>
<td>4:50 PM</td>
<td>OP11: Computational analysis of breakome reveals replication fork</td>
</tr>
<tr>
<td></td>
<td>movement and elucidates mechanisms of DNA double-stranded break</td>
</tr>
<tr>
<td></td>
<td>formation</td>
</tr>
<tr>
<td></td>
<td>Maga Rowicka, University of Texas Medical Branch at Galveston</td>
</tr>
<tr>
<td>5:00 PM</td>
<td>OP12: HRC3 – A new class of motifs involved in chromatin organization</td>
</tr>
<tr>
<td></td>
<td>and development</td>
</tr>
<tr>
<td></td>
<td>Andrzej Kudlicki, University of Texas Medical Branch</td>
</tr>
<tr>
<td>5:10 PM</td>
<td>BREAK</td>
</tr>
<tr>
<td>5:30 PM</td>
<td>OP13: Network Inference and the Knowledge Base of Biomedicine</td>
</tr>
<tr>
<td></td>
<td>Tiffany Callahan, University of Colorado Denver Anschutz Medical Campus</td>
</tr>
<tr>
<td>5:40 PM</td>
<td>OP14: ShinyLearner: Enabling biologists to perform robust</td>
</tr>
<tr>
<td></td>
<td>machine-learning classification</td>
</tr>
<tr>
<td></td>
<td>Stephen Piccolo, Brigham Young University</td>
</tr>
<tr>
<td>5:50 PM</td>
<td>OP15: Stratification of prostate cancer patients based on</td>
</tr>
<tr>
<td></td>
<td>molecular interaction profiles</td>
</tr>
<tr>
<td></td>
<td>Roland Mathis, IBM Research</td>
</tr>
<tr>
<td>6:00 PM</td>
<td>BREAK</td>
</tr>
<tr>
<td>6:30 PM</td>
<td>DINNER - IL POGGIO</td>
</tr>
</tbody>
</table>

**FRIDAY – DECEMBER 9, 2016**

8:00 AM–6:00 PM REGISTRATION

8:30 AM BREAKFAST

9:00 AM KEYNOTE 4: Laura K. Wiley, University of Colorado Anschutz Medical Campus

Precision Medicine and the Learning Healthcare System: Leveraging Informatics to Improve Care

9:45 AM OP16: Medication Data Mining of Electronic Medical Records Reveal Race-Specific Prescription Patterns

Benjamin Glicksberg, Icahn School of Medicine at Mount Sinai

9:55 AM OP17: Comparative analysis of the expression patterns and regulation of histone variant genes reveals a novel epigenetic pathway related to cancer

Michael Tolstorukov, Massachusetts General Hospital and Harvard Medical School

10:05 AM OP18: The Cognoma Collaborative creates a webapp to predict cancer mutations from gene expression

Daniel Himmelstein, University of Pennsylvania

10:15 AM OP19: Functionally prioritizing candidate genes from genome-wide association studies

Kelsey Anderson, University of Colorado School of Medicine
<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>10:25 AM</td>
<td>BREAK</td>
</tr>
</tbody>
</table>
| 10:45 AM  | **KEYNOTE 5**: Alan Williams, SomaLogic  
Data Analytics and the SOMAscan™ Proteomic Platform |
| 11:15 AM  | **OP20**: Deriving Population-Scale Therapeutic Trajectories to Enable Precision Pharmacology  
Kipp Johnson, Icahn School of Medicine at Mount Sinai |
| 11:25 AM  | **OP21**: Comparison of Relief-F Nucleotide Differences for GWAS Data with Application to Bipolar Disorder  
Marziyeh Arabnejad Khanouki, University of Tulsa |
| 11:35 AM  | **OP22**: ModEvo: A Web-Based Tool for Modeling Evolutionary Dynamics  
Filip Jagodzinski, Western Washington University |
| 11:45 AM  | **OP23**: Predicting Neural Fluctuations in the Primary Visual Cortex  
William Kindel, University of Colorado School of Medicine |
| 12:00 PM  | **SKI BREAK**                                                                                   |
| 4:00 PM   | **KEYNOTE 6**: Temple F. Smith, Boston University  
Current Insights into the Evolution of the Genetic Translation System and the Genetic Code Itself |
| 4:30 PM   | **OP24**: De novo protein structure prediction by big data and deep learning  
Sheng Wang, Toyota Technological Institute at Chicago |
| 4:40 PM   | **OP25**: Identifying the mechanism for the metastatic spread of breast cancer through integration of gene expression, whole genome sequencing and functional screens  
Eran Andrechek, Michigan State University |
| 4:50 PM   | **OP26**: Allelic Maps of Cancer  
Anelia Horvath, George Washington University |
| 5:00 PM   | **OP27**: Identifying non-specific effects of small molecule treatment through GSEA meta-analysis  
Rani Powers, University of Colorado Anschutz Medical Campus |
| 5:10 PM   | **BREAK**                                                                                       |
| 5:30 PM   | **OP28**: Insights into Bathyarcheota Ecology and Co-occurrence Patterns as Revealed by Public Metagenome Sequencing Data  
David Banks-Richardson, University of Colorado-Denver |
| 5:40 PM   | **OP29**: The SNPPhenA Corpus: An annotated research abstract corpus for extracting ranked association of single-nucleotide polymorphisms and phenotypes  
Hamidreza Chitsaz, Colorado State University |
| 5:50 PM   | **OP30**: Toward a Metric Learning Model for Protein Fold Recognition Using a Novel Feature Extraction Technique Based on the Mixture of Evolutionary and Secondary Structural Information  
Pooya Zakeri, 1)KU Leuven. 2)iMinds |
| 6:00 PM   | **BREAK**                                                                                       |
| 6:30 PM   | **POSTER SESSION**                                                                               |
### Detailed Agenda

**Saturday – December 10, 2016**

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:00 AM</td>
<td>Registration</td>
</tr>
<tr>
<td>8:00 AM</td>
<td>Breakfast</td>
</tr>
<tr>
<td>09:00 AM</td>
<td><strong>KEYNOTE 7:</strong> Krista Meyer, Biodesix: Making Medicine Personal®</td>
</tr>
<tr>
<td></td>
<td>Leveraging Biology to Guide Feature Enrichment for Designing Multivariate Classifiers for Clinical Tests</td>
</tr>
<tr>
<td>09:30 AM</td>
<td><strong>OP31:</strong> Development of a diagnostic to profile eukaryotic microbes of the human microbiome</td>
</tr>
<tr>
<td></td>
<td>Ana Popovic, Hospital for Sick Children, University of Toronto</td>
</tr>
<tr>
<td>09:40 AM</td>
<td><strong>OP32:</strong> G4 quadruplexes in and near regulatory elements of maize genes predict tissue type and altered transcriptional and translational response to submergence and heat stress</td>
</tr>
<tr>
<td></td>
<td>Mingze He, Iowa State University</td>
</tr>
<tr>
<td>09:50 AM</td>
<td><strong>OP33:</strong> Modeling heterogeneous cell populations using Boolean networks</td>
</tr>
<tr>
<td></td>
<td>Brian Ross, University of Colorado</td>
</tr>
<tr>
<td>10:00 AM</td>
<td><strong>OP34:</strong> Enhancer Reprogramming in Mammalian Genomes</td>
</tr>
<tr>
<td></td>
<td>Mario Flores, NIH</td>
</tr>
<tr>
<td>10:10 AM</td>
<td><strong>OP35:</strong> The Finite State Projection based Fisher Information Matrix for the Design of Single-Cell Experiments</td>
</tr>
<tr>
<td></td>
<td>Zachary Fox, Colorado State University</td>
</tr>
<tr>
<td>10:20 AM</td>
<td>BREAK</td>
</tr>
<tr>
<td>10:40 AM</td>
<td><strong>OP36:</strong> 2-Scale KNN Classifications</td>
</tr>
<tr>
<td></td>
<td>Destiny Anyaiwe, Oakland University</td>
</tr>
<tr>
<td>10:50 AM</td>
<td><strong>OP37:</strong> Best practices for reproducible and robust data analysis in a bioinformatics core facility</td>
</tr>
<tr>
<td></td>
<td>James Denvir, Marshall University</td>
</tr>
<tr>
<td>11:00 AM</td>
<td><strong>OP38:</strong> The Affinity Data Bank for biophysical analysis of regulatory sequences</td>
</tr>
<tr>
<td></td>
<td>Todd Riley, University of Massachusetts, Boston</td>
</tr>
<tr>
<td>11:10 AM</td>
<td><strong>OP39:</strong> Pattern-based estimation of the extent of explicit contradiction in the scientific literature</td>
</tr>
<tr>
<td></td>
<td>Elizabeth White, University of Colorado Denver, Anschutz</td>
</tr>
<tr>
<td>11:20 AM</td>
<td><strong>OP40:</strong> Towards Highly Accurate Mapping of Protein Glycosylation Sites in the Human Proteome</td>
</tr>
<tr>
<td></td>
<td>Chen Li, Monash University</td>
</tr>
<tr>
<td>11:30 AM</td>
<td><strong>KEYNOTE 8:</strong> Kirk E. Jordan, IBM Research UK</td>
</tr>
<tr>
<td></td>
<td>Data Centric Cognitive Computing: IBM’s Direction, Workflow Challenges and Opportunities</td>
</tr>
<tr>
<td>11:50 AM</td>
<td>CLOSING RAFFLE AND AWARDS</td>
</tr>
</tbody>
</table>
KEYNOTE SPEAKERS

KIRK E. JORDAN
IBM Research UK, United Kingdom

Data Centric Cognitive Computing: IBM’s Direction, Workflow Challenges and Opportunities

The volume, variety, velocity and veracity data is pushing how we think about computer systems. In this talk, I will describe the IBM Research’s Data Centric Solutions directions to develop systems that handle large data sets shortening time to solution. I will give an overview of our motivation, describe some of our current design thinking for these systems and some of our current work on these systems. I will describe some of the challenges and opportunities these systems may have for workflows for application solutions and some of our thinking how workflows might be changed given such data centric systems. This includes a description of some of our work incorporating cognitive enhanced simulations that are changing our approaches to large scale simulation.

RENEE DEEHAN KENNEY
PatientsLikeMe, Massachusetts, United States

Measuring Disease Through Real World Evidence and Biological State: Impact For Patients (and Animal Models, too)

Individuals living with a medical condition (and clinicians/researchers) are interested in learning from others who have exhibited a similar disease trajectory as their own to identify the medical and lifestyle interventions that optimize outcomes. PatientsLikeMe has developed a science-based social platform where individuals with medical conditions can share information about their symptoms, treatments, health outcomes and experiences living with disease, and engage with others for information and support, all to gain new insights and improve outcomes.

The term “big data” is no longer synonymous only with platforms such as this, or electronic health record systems, but also includes “big biological data” from which hundreds of thousands of measurements can be taken form a single biosample. The combination of patient reported health data and biological measurements will enable detailed tracking of disease and lifestyle measurements together over time in a new way, and at a scale never before achieved. Ultimately, individuals will benefit from the conclusions made available from the
direct combination of these data streams and the utility and efficiency of collecting and analyzing relevant data outside of the standard methods of costly, often proprietary, and time-consuming clinical trials. Additionally, parallel systems-level evaluation of animal models will enable academics doing basic research, and industry scientists doing pre-clinical discovery, to more quickly evaluate the true translational capabilities of their work.

M ARCO M ASSEROLI
Politecnico di Milano, Milano, Italy

Next Generation Genomic Computing

Next-generation sequencing (NGS) technologies and data processing pipelines are rapidly and inexpensively providing increasingly numerous sequencing data and associated (epi)genomic features of many individual genomes in multiple biological and clinical conditions, generally made publicly available within well-curated repositories. Answers to fundamental biomedical problems are hidden in these data; yet, their efficient management and integrative processing is becoming the biggest and most important “big data” problem of mankind. Multi-sample processing of heterogeneous information can support data-driven discoveries and biomolecular sense making, such as discovering how heterogeneous genomic, transcriptomic and epigenomic features cooperate to characterize biomolecular functions; yet, it requires state-of-the-art “big data” computing strategies, with abstractions beyond commonly used tool capabilities.

We recently proposed a new paradigm in NGS data management and processing by introducing an essential Genomic Data Model (GDM) using few general abstractions for genomic region data and associated experimental, biological and clinical metadata that guarantee interoperability between existing data formats. Leveraging on GDM, we developed a next-generation, high-level, declarative GenoMetric Query Language (GMQL) for genomics data; here, we demonstrate its usefulness, flexibility and simplicity of use through several biological query examples. GMQL operates downstream of raw data preprocessing pipelines and supports queries over thousands of heterogeneous samples; computational efficiency and high scalability are achieved by using parallel computing on clusters or public clouds. GDM and GMQL can be exploited to provide integrated access to curated data, made available by large consortia such as ENCODE, Epigenomics Roadmap, or TCGA, through user-friendly search services.
Biodesix uses a hypothesis-independent approach to building clinically relevant tests allowing the creation of multivariate classifiers that reflect the complexity of biological interactions without any bias from expectations about their mechanisms. We use mass spectral data collected from patient serum samples in combination with the Diagnostic CortexTM robust data analytics platform to design classifiers with clinical relevance. Once the classifier is created, it is of interest to understand the biological underpinnings of its performance. We applied ideas similar to GSEA (Gene Set Enrichment Analysis) to mass spectral data (ProteinSEA). This approach allowed us to find correlations between classification and sets of proteins associated with known biological functions, such as acute response, wound healing, and complement system. With this in hand, we wondered if the biological insight gained from our protein sets could be leveraged to guide test development. We developed a method to enrich the pools of mass spectral features for signals related to biological processes thought to be relevant to the clinical problem of interest. The approach allows us to probe several biological pathways and is modular such that the results of each protein set can capture information from influential and more subtle biological processes in independent steps. While the Diagnostic CortexTM itself does not require biological input to develop clinically relevant tests, the ability to tune the feature space using biological insight can be a useful tool to tease apart the pathways driving disease.
Current Insights into the Evolution of the Genetic Translation System and the Genetic Code Itself

Given the massive amount of determined structure of the translational system components we appear to be in position to respond to Crick's famous 1968 statement:

“It is almost impossible to discuss the origin or the Genetic Code without discussing the origin of the actual biochemical mechanisms of protein synthesis. This is very difficulty for two reasons; it is complex and many of its details are not yet understood.”

Along with the full ribosomal structure and its associated proteins, perhaps it is the aminoacyl tRNA synthetases that provide the most interesting information. These along with the structures of the related nucleotide transferases and the tRNAs themselves have provided a number of new additional insights. And while, there is no final answers to the evolution of the full genetic translational system, there are a number of very suggestive ideas now being discussed. Many presented at a recent Cold Spring Harbor Laboratory small symposium.

The Blue-prints of Tumors Uncovered Through Network Integration

DNA sequencing of cancerous tissue has revealed a complex landscape of mutations. Many altered genes represent the “usual suspects” known to drive the disease. Still, the majority of patients have alterations in genes of unknown significance or in non-coding parts of the genome. Gene expression levels reflected in the RNA integrate information about how genomic alterations affect the circuitry of tumor cells. I’ll discuss methods that find clues about tumors from DNA, RNA and protein data. Our approach identifies treatment options for an individual patient by matching their tumor to the closest “neighbors” in an encyclopedia of other tumors. It builds a logical wiring diagram that represent the misfirings inside and among the cells in the tumor and its microenvironment. We resolved diagrams for six patients with metastatic prostate cancer and found several new angles of attack. Finally, I’ll show how this work is informing clinicians for pediatric cancer cases as part of the ongoing California Kids Cancer Consortium.
LAURA K. WILEY

University of Colorado Anschutz Medical Campus, Colorado, United States

PRECISION MEDICINE AND THE LEARNING HEALTHCARE SYSTEM: LEVERAGING INFORMATICS TO IMPROVE CARE

Precision medicine and the learning healthcare system are predicated on the assumption that by turning data (be they molecular, clinical, or social) into knowledge we can improve health. Informatics is a key enabler of this vision. Using warfarin pharmacogenomics as an example, we will examine the role of informatics for data discovery, translation, implementation and evaluation.

ALAN WILLIAMS

SomaLogic, Colorado, United States

Data Analytics and the SOMAscan™ Proteomic Platform

ABSTRACT: Unavailable
ORAL PRESENTATION LIST

OP01: IndeCut: A CUT-NORM BASED METHOD FOR EVALUATING INDEPENDENT AND UNIFORM SAMPLING IN NETWORK MOTIF DISCOVERY ALGORITHMS
Presenting Author: Mitra Ansariola, Oregon State University

OP02: REPRODUCIBLE COMPUTATIONAL WORKFLOWS WITH CONTINUOUS ANALYSIS
Presenting Author: Brett Beaulieu-Jones, University of Pennsylvania

OP03: APPLICATION ONTOLOGIES SUPPORTING PHENOTYPING FROM CLINICAL TEXT
Presenting Author: Wendy Chapman, University of Utah

OP04: SPARQLer: MAKING KNOWLEDGE FUNCTIONAL
Presenting Author: Daniel McShan, University of Colorado School of Medicine

OP05: IMPROVED NETWORK ONTOLOGY ANALYSIS BY SEGMENTATION
Presenting Author: Ananda Mondal, Claflin University

OP06: AN IMAGE PHENOTYPING ENVIRONMENT BASED ON OPEN-SOURCE TOOLS
Presenting Author: Brian Chapman, University of Utah

OP07: InterViewer, A NEW CYTOSCAPE-BASED VIEWER THAT DISPLAYS INTERACTIONS BETWEEN SELECTED SETS OF PROTEINS
Presenting Author: Marek Tutaj, Medical College of Wisconsin

OP08: CAMSA: A TOOL FOR COMPARATIVE ANALYSIS AND MERGING OF SCAFFOLD ASSEMBLIES
Presenting Author: Max Alekseyev, George Washington University

OP09: ANALYSIS OF TOBACCO USERS ADMITTED TO INTENSIVE CARE UNITS
Presenting Author: Andrey Soares, University of Colorado School of Medicine

OP10: A NEW MOLECULAR SIGNATURE APPROACH FOR PREDICTION OF DRIVER CANCER PATHWAYS FROM TRANSCRIPTIONAL DATA
Presenting Author: Boris Reva, Icahn School of Medicine at Mount Sinai

OP11: COMPUTATIONAL ANALYSIS OF BREAKOME REVEALS REPLICATION FORK MOVEMENT AND ELUCIDATES MECHANISMS OF DNA DOUBLE-STRANDED BREAK FORMATION
Presenting Author: Maga Rowicka, University of Texas Medical Branch at Galveston

OP12: HRC3 – A NEW CLASS OF MOTIFS INVOLVED IN CHROMATIN ORGANIZATION AND DEVELOPMENT.
Presenting Author: Andrzej Kudlicki, University of Texas Medical Branch

OP13: NETWORK INERENCE AND THE KNOWLEDGE BASE OF BIOMEDICINE
Presenting Author: Tiffany Callahan, University of Colorado Denver Anschutz Medical Campus

OP14: ShinyLearner: ENABLING BIOLOGISTS TO PERFORM ROBUST MACHINE-LEARNING CLASSIFICATION
Presenting Author: Stephen Piccolo, Brigham Young University

OP15: STRATIFICATION OF PROSTATE CANCER PATIENTS BASED ON MOLECULAR INTERACTION PROFILES
Presenting Author: Roland Mathis, IBM Research

OP16: MEDICATION DATA MINING OF ELECTRONIC MEDICAL RECORDS REVEAL RACE-SPECIFIC PRESCRIPTION PATTERNS
Presenting Author: Benjamin Glicksberg, Icahn School of Medicine at Mount Sinai
OP17: COMPARATIVE ANALYSIS OF THE EXPRESSION PATTERNS AND REGULATION OF HISTONE VARIANT GENES REVEALS A NOVEL EPIGENETIC PATHWAY RELATED TO CANCER
Presenting Author: Michael Tolstorukov, Massachusetts General Hospital and Harvard Medical School

OP18: THE COGNOMA COLLABORATIVE CREATES A WEBAPP TO PREDICT CANCER MUTATIONS FROM GENE EXPRESSION
Presenting Author: Daniel Himmelstein, University of Pennsylvania

OP19: FUNCTIONALLY PRIORITIZING CANDIDATE GENES FROM GENOME-WIDE ASSOCIATION STUDIES
Presenting Author: Kelsey Anderson, University of Colorado School of Medicine

OP20: DERIVING POPULATION-SCALE THERAPEUTIC TRAJECTORIES TO ENABLE PRECISION PHARMACOLOGY
Presenting Author: Kipp Johnson, Icahn School of Medicine at Mount Sinai

OP21: COMPARISON OF RELIEF-F NUCLEOTIDE DIFFERENCES FOR GWAS DATA WITH APPLICATION TO BIPOLAR DISORDER
Presenting Author: Marziyeh Arabnejad Khanouki, University of Tulsa

OP22: ModEvo: A WEB-BASED TOOL FOR MODELING EVOLUTIONARY DYNAMICS
Presenting Author: Filip Jagodzinski, Western Washington University

OP23: PREDICTING NEURAL FLUCTUATIONS IN THE PRIMARY VISUAL CORTEX
Presenting Author: William Kindel, University of Colorado School of Medicine

OP24: DE NOVO PROTEIN STRUCTURE PREDICTION BY BIG DATA AND DEEP LEARNING
Presenting Author: Sheng Wang, Toyota Technological Institute at Chicago

OP25: IDENTIFYING THE MECHANISM FOR THE METASTATIC SPREAD OF BREAST CANCER THROUGH INTEGRATION OF GENE EXPRESSION, WHOLE GENOME SEQUENCING AND FUNCTIONAL SCREENS.
Presenting Author: Eran Andrechek, Michigan State University

OP26: ALLELIC MAPS OF CANCER
Presenting Author: Anelia Horvath, George Washington University

OP27: IDENTIFYING NON-SPECIFIC EFFECTS OF SMALL MOLECULE TREATMENT THROUGH GSEA META-ANALYSIS
Presenting Author: Rani Powers, University of Colorado Anschutz Medical Campus

OP28: INSIGHTS INTO BATHYARCHEOTA ECOLOGY AND CO-OCCURRENCE PATTERNS AS REVEALED BY PUBLIC METAGENOME SEQUENCING DATA
Presenting Author: David Banks-Richardson, University of Colorado-Denver

OP29: THE SNPPhenA CORPUS: AN ANNOTATED RESEARCH ABSTRACT CORPUS FOR EXTRACTING RANKED ASSOCIATION OF SINGLE-NUCLEOTIDE POLYMORPHISMS AND PHENOTYPES
Presenting Author: Hamidreza Chitsaz, Colorado state university

OP30: TOWARD A METRIC LEARNING MODEL FOR PROTEIN FOLD RECOGNITION USING A NOVEL FEATURE EXTRACTION TECHNIQUE BASED ON THE MIXTURE OF EVOLUTIONARY AND SECONDARY STRUCTURAL INFORMATION
Presenting Author: Pooya Zakeri, 1)KU Leuven. 2)iMinds

OP31: DEVELOPMENT OF A DIAGNOSTIC TO PROFILE EUKARYOTIC MICROBES OF THE HUMAN MICROBIOME
Presenting Author: Ana Popovic, Hospital for Sick Children, University of Toronto
ORAL PRESENTATION LIST

OP32: G4 QUADRUPLEXES IN AND NEAR REGULATORY ELEMENTS OF MAIZE GENES PREDICT TISSUE TYPE AND ALTERED TRANSCRIPTIONAL RESPONSE TO ABIOTIC STRESSES
Presenting Author: Mingze He, Iowa State University

OP33: MODELING HETEROGENEOUS CELL POPULATIONS USING BOOLEAN NETWORKS
Presenting Author: Brian Ross, University of Colorado

OP34: ENHANCER REPROGRAMMING IN MAMMALIAN GENOMES
Presenting Author: Mario Flores, NIH

OP35: THE FINITE STATE PROJECTION BASED FISHER INFORMATION MATRIX FOR THE DESIGN OF SINGLE-CELL EXPERIMENTS.
Presenting Author: Zachary Fox, Colorado State University

OP36: 2-SCALE KNN CLASSIFICATIONS
Presenting Author: Destiny Anyaiwe, Oakland University

OP37: BEST PRACTICES FOR REPRODUCIBLE AND ROBUST DATA ANALYSIS IN A BIOINFORMATICS CORE FACILITY
Presenting Author: James Denvir, Marshall University

OP38: THE AFFINITY DATA BANK FOR BIOPHYSICAL ANALYSIS OF REGULATORY SEQUENCES
Presenting Author: Todd Riley, University of Massachusetts Boston

OP39: PATTERN-BASED ESTIMATION OF THE EXTENT OF EXPLICIT CONTRADICTION IN THE SCIENTIFIC LITERATURE
Presenting Author: Elizabeth White, University of Colorado Denver, Anschutz

OP40: TOWARDS HIGHLY ACCURATE MAPPING OF PROTEIN GLYCOSYLATION SITES IN THE HUMAN PROTEOME
Presenting Author: Chen Li, Monash University
ORAL PRESENTATION ABSTRACTS

OP01: IndeCut: A CUT-NORM BASED METHOD FOR EVALUATING INDEPENDENT AND UNIFORM SAMPLING IN NETWORK MOTIF DISCOVERY ALGORITHMS

Presenting Author: Mitra Ansariola, Oregon State University

Co-Author(s): David Koslicki, Oregon State University; Molly Megrew, Oregon State University

ABSTRACT: Network motif discovery is a well-established general statistical strategy for identifying over-represented sub-network structures within a larger network. In the biosciences, it serves as a prominent conceptual tool that enables scientists to recognize biologically important patterns and generate testable hypotheses within large genetic networks of interest. Network motif discovery algorithms function by comparing the frequency of particular sub-network of interest within a given ‘real-world’ network to its frequency in a large collection of randomized networks. While the method of randomization may differ, all algorithms face the challenge of how to sample uniformly and independently from the set of all possible randomized networks that may be generated. Though several network motif discovery tools with different underlying random sampling strategies are available, scientists who want to apply these tools on their own networks of interest currently do not have any method by which to assess whether any tool will provide an accurate outcome. Most users will not be able to test the correctness of detected motifs in the laboratory due to prohibitive cost, so it is essential to have access to such an evaluation method. In this talk, we present IndeCut, the very first method that numerically determines the degree of sampling uniformity and independence of network motif discovery algorithms. IndeCut is the first and only method to date that allows characterization of network motif finding algorithm performance beyond computational efficiency.

OP02: REPRODUCIBLE COMPUTATIONAL WORKFLOWS WITH CONTINUOUS ANALYSIS

Presenting Author: Brett Beaulieu-Jones, University of Pennsylvania

Co-Author(s): Casey Greene, University of Pennsylvania

ABSTRACT: Reproducing experiments is vital to science. Being able to replicate, validate and extend previous work also speeds new research projects. Reproducing computational biology experiments, which are scripted, should be straightforward. But reproducing such work remains challenging and time consuming. In the ideal world we would be able to quickly and easily rewind to the precise computing environment where results were generated. We would then be able to reproduce the original analysis or perform new analyses. We introduce a process
termed “continuous analysis” which provides inherent reproducibility to computational research at a minimal cost to the researcher. Continuous analysis combines Docker, a container service similar to virtual machines, with continuous integration, a popular software development technique, to automatically re-run computational analysis whenever relevant changes are made to the source code. This allows results to be reproduced quickly, accurately and without needing to contact the original authors. Continuous analysis also provides an audit trail for analyses that use data with sharing restrictions. This allows reviewers, editors, and readers to verify reproducibility without manually downloading and rerunning any code.

**OP03: APPLICATION ONTOLOGIES SUPPORTING PHENOTYPING FROM CLINICAL TEXT**

Presenting Author: *Wendy Chapman, University of Utah*

**ABSTRACT:** Representation of the knowledge described in clinical reports is critical to accurate phenotyping of patients. We have developed two application ontologies for modeling annotations of clinical reports: the schema ontology describes the clinical entities that are described in reports, such as findings, procedures, and medications. The modifier ontology enumerates the allowable modifiers for those entities with three types of modifiers: shared modifiers that apply to all entities: negation, uncertainty, and temporality; semantic modifiers specific to particular entities, such as dose and route for medications; and numeric modifiers for specifying numeric values such as body temperature. A user can create a domain ontology by creating instances of entity-modifier combinations, accommodating rich phenotypic representation for concepts like no family history of colon cancer or severe carotid stenosis in the right internal carotid artery. In addition to modeling the semantic composition of the concepts, the ontologies provide value sets and lexical variants that can be customized and enhanced. Our long-term goal is to create shareable libraries of domain ontologies.

In addition to supporting annotation of concept mentions, swirl rules stored in the ontology support inferencing over mention annotations for classification at the document, encounter, and phenotypic/patient level. The ontologies support rich phenotypic characterizations to go beyond binary phenotypes toward answering questions like “what histologic types of breast cancer are associated with patients that have a substitution mutation on BRCA-1?” and “for patients with a papillary breast carcinoma that underwent neoadjuvant treatment regimen, what number of patients have had a recurrence or metastasis?”
**OP04: SPARQLer: MAKING KNOWLEDGE FUNCTIONAL**

Presenting Author: Daniel McShan, University of Colorado School of Medicine

**ABSTRACT:** Knowledge base triplestores are notoriously challenging to navigate. This talk will cover some approaches for making the stored knowledge more functional. While SPARQL 2.0 is extraordinarily powerful, it can be difficult for someone unfamiliar with the underlying graph structure to quickly construct meaningful queries.

SPARQLer is a simple syntactic sugar layer in Clojure which allows for the construction of functional decomposable queries that are modular and reusable. These functional components can then be easily tested against easily validated examples. SPARQLer is demonstrated as a functional front end to the Hunter lab’s KABOB Knowledge Base of Biology, and various examples will illustrate the modularity and the reuse potential of the approach.

**OP05: IMPROVED NETWORK ONTOLOGY ANALYSIS BY SEGMENTATION**

Presenting Author: Ananda Mondal, Claflin University

Co-Author(s): Charles Schultz, University of Utah; Markea Sheppard, Claflin University; Jasmine Carson, North Carolina A&T State University

**ABSTRACT:** Our recent study in filtering disease proteins, based on subcellular protein locations, from a protein network biomarker for liver cancer, resulted in groups of proteins at different locations with two distinct network structures, namely, clique and bipartite graph. This motivated us developing a Segmentation Algorithm, which can be used as a preprocessing tool to provide a better gene-term enrichment analysis based on network ontology.

The proposed algorithm breaks the source network into component subgraphs using an appropriate metric such as bipartite-like or clique-like subgraphs. The component subgraphs generated above are independently analyzed using the Network Ontology Analysis (NOA) method. The independent results, which contain overlapping ontological components, are integrated to form a single representation for gene-ontology analysis.

We applied the developed technique as a preprocessing tool for NOA analysis of protein network biomarkers for Adherens Junction and Breast Cancer. Results showed that the proposed algorithm produces a more concise and easily interpretable representation of the gene-term relationship compare to the representation produced using NOA only.
**OP06: AN IMAGE PHENOTYPING ENVIRONMENT BASED ON OPEN-SOURCE TOOLS**

Presenting Author: Brian Chapman, University of Utah  
Co-Author(s): John Roberts, University of Utah

**ABSTRACT:** Medical imaging data are an often-overlooked resource for defining patient phenotypes. Because images data are unstructured, in order to extract information from the images requires creating pipelines for identifying relevant studies, segmenting and quantifying features from the images, and linking these features to other data sources (e.g. the EHR). We are building an image phenotyping environment based on open-source deployed using Docker (https://www.docker.com/), allowing us to version-control our environments, which are defined with simple text files. Our phenotyping pipeline is built using three open-source projects: 1) Orthanc (http://www.orthanc-server.com/), a light-weight DICOM server for communicating with the clinical PACS and scrubbing images for research purposes. Orthanc allows for persistent, customized scrubbing processes. 2) Girder (https://girder.readthedocs.io/en/latest/), an open-source, web-based data management system developed by Kitware, Inc. Girder provides user authentication, access control and a framework for linking data and defining meta-data. We have integrated Girder with bioportal so that data uploads are tagged with concepts from relevant ontologies. 3) JupyterHub for providing web-based computational environments. JupyterHub provides Docker containers serving up Jupyter notebooks. Jupyter notebooks allow for programming through the web browser and supports a number of languages including Python and a number of other languages. Jupyter notebooks contain image processing pipelines for extracting features from medical images using SimpleITK and other software packages. Our initial use-cases are drawn from dermatology and radiology and require both 2D and 3D feature extraction tasks.

**OP07: InterViewer, A NEW CYTOSCAPE-BASED VIEWER THAT DISPLAYS INTERACTIONS BETWEEN SELECTED SETS OF PROTEINS**

Presenting Author: Marek Tutaj, Medical College of Wisconsin  
Co-Author(s): Jyothi Thota, Medical College of Wisconsin; Jeff De Pons, Medical College of Wisconsin; Jennifer Smith, Medical College of Wisconsin; Thomas G Hayman, Medical College of Wisconsin; Victoria Petri, Medical College of Wisconsin; Stan Laulederkind, Medical College of Wisconsin; Shur-Jen Wang, Medical College of Wisconsin; Mary Shimoyama, Medical College of Wisconsin

**ABSTRACT:** InterViewer, RGD’s new Cytoscape-based protein-protein interactions viewer, (https://rgd.mcw.edu/rgdweb/cytoscape/query.html), facilitates a detailed visualization of interactions between
proteins. As usual, RGD provides interaction data not only for rat, but also for mouse and human. The tool accepts input in multiple ways: as a list of UniProt IDs, RGD IDs or gene symbols. On the display page, binary interaction data from IMEX are displayed as nodes and edges, which can be zoomed in or out using controls. Clicking on a protein node provides links to UniProtKB and to RGD gene report pages. Detailed information about the protein appears in the upper right. Clicking on an edge shows additional information about that interaction. Also for more complex networks, multiple display filters can be applied. The user can pick a set of interaction types of interest, one or more species or common interactors. In addition, several layout modes common for Cytoscape graphs like ‘cose’ or ‘circle’, are available. A legend details the color-coded interaction types and protein species. The table beneath the display lists downloadable characteristics of each pair of interactors, the complete node list and node/edge statistics. The bird’s-eye view panel facilitates movement of the display. The tool also has options to generate printable reports and graph images for user convenience.

**OP08: CAMSA: A TOOL FOR COMPARATIVE ANALYSIS AND MERGING OF SCAFFOLD ASSEMBLIES**

Presenting Author: Max Alekseyev, George Washington University

Co-Author(s): Sergey Aganezov, George Washington University

**ABSTRACT:** Motivation: Despite the recent progress in genome sequencing and assembly, many of the currently available assembled genomes come in a draft form. Such draft genomes consist of a large number of genomic fragments (scaffolds), whose positions and orientations along the genome are unknown. While there exists a number of methods for reconstruction of the genome from its scaffolds, utilizing various computational and wet-lab techniques, they often can produce only partial error-prone scaffold assemblies. It therefore becomes important to compare and merge scaffold assemblies produced by different methods, thus combining their advantages and highlighting present conflicts for further investigation. These tasks may be labor intensive if performed manually.

Results: We present CAMSA—a tool for comparative analysis and merging of two or more given scaffold assemblies. The tool (i) creates an extensive report with several comparative quality metrics; (ii) constructs a most consistent combined scaffold assembly; and (iii) provides an interactive framework for a visual comparative analysis of the given assemblies.

Availability: CAMSA is available for download from http://cblab.org/camsa/
OP09: ANALYSIS OF TOBACCO USERS ADMITTED TO INTENSIVE CARE UNITS

Presenting Author: Andrey Soares, University of Colorado School of Medicine
Co-Author(s): Sonia Leach, National Jewish Health, University of Colorado School of Medicine; Kevin Cohen, University of Colorado School of Medicine; Joan Davis, Southern Illinois University

ABSTRACT: Smoking is known to cause numerous tobacco-related diseases such as cancer, heart disease, diabetes, respiratory disease, as well as death. The Center for Disease Control and Prevention warns that over 16 million Americans have some disease caused by smoking, with about 480,000 deaths in the United States. Thus, it is critical for healthcare professionals to identify and treat every tobacco user seen in any healthcare facilities. This research seeks to examine if patients, who are current tobacco users, have been correctly identified as smokers, their smoking status and behaviors have been documented, and they have received appropriate treatment recommendations (prescriptions) based on their health conditions. In particular, we will perform text analysis of the chart notes recorded during the patient stay to collect information that can be used to offer tailored treatment recommendations such as the number of cigarettes used per day, and to verify inconsistencies in documenting information about smoking. Preliminary data analysis shows that some tobacco users have not been diagnosed as smokers using the appropriate ICD9 code, leaving the information about smoking to be retrieved from the text notes or inferred from the prescribed tobacco medications. This research will also evaluate the treatment recommendations based on patient health conditions and risks, and will cluster smokers to identify emerging patterns and relationships among characteristics and diagnoses that can support tobacco intervention strategies for patients admitted to intensive care units. We will focus on comorbidities as tobacco use can trigger new diseases or complicate existing ones.

OP10: A NEW MOLECULAR SIGNATURE APPROACH FOR PREDICTION OF DRIVER_CANCER PATHWAYS FROM TRANSCRIPTIONAL DATA

Presenting Author: Boris Reva, Icahn School of Medicine at Mount Sinai
Co-Author(s): Noam Beckmann, Icahn School of Medicine at Mount Sinai; Hui Li, Icahn School of Medicine at Mount Sinai; Andrew Uzilow, Icahn School of Medicine at Mount Sinai; Dmitry Rykunov, Icahn School of Medicine at Mount Sinai; Eric Schadt, Icahn School of Medicine at Mount Sinai

ABSTRACT: Assigning cancer patients to the most effective treatments requires an understanding of the molecular basis of their disease. While DNA-based molecular profiling approaches have flourished over the past several years to transform our understanding of driver pathways across a broad range of tumors, a systematic characterization of key driver pathways based on RNA data has not been undertaken. Here
we introduce a new approach to predict the status of driver cancer pathways based on signature functions we constructed using weighted sums of gene expression levels derived from RNA sequencing data. To identify the driver cancer pathways of interest, we mined DNA variant data from TCGA and nominated driver alterations in seven major cancer pathways in breast, ovarian, and colon cancer tumors. The activation status of these driver pathways was then characterized using RNA sequencing data by constructing signature functions in training datasets and then testing the accuracy of the signatures in test datasets. The signature functions differentiated tumors with nominated active pathways from tumors with no genomic signs of activation very well (average AUC equals to 0.8), and they systematically exceeded the accuracies obtained by ten other known classification methods we employed as a control. A typical pathway signature is composed of ~20 biomarker genes that are unique to a given pathway and cancer type. Our results confirm that driver genomic alterations are distinctively displayed at the transcriptional level and that the transcriptional signatures can generally provide an alternative to DNA sequencing methods in detecting specific driver pathways.

**OP11: COMPUTATIONAL ANALYSIS OF BREAKOME REVEALS REPLICATION FORK MOVEMENT AND ELUCIDATES MECHANISMS OF DNA DOUBLE-STRANDED BREAK FORMATION**

Presenting Author: Maga Rowicka, University of Texas Medical Branch at Galveston

Co-Author(s): Yingjie Zhu, University of Texas Medical Branch at Galveston; Norbert Dojer, University of Texas Medical Branch at Galveston; Anna Biernacka, University of Warsaw; Jules Nde, University of Texas Medical Branch at Galveston; Bernard Fongang, University of Texas Medical Branch at Galveston; Razieyeh Yousefi, University of Texas Medical Branch at Galveston; Abhishek Mitra, University of Texas Medical Branch at Galveston; Ji Li, University of Texas Medical Branch at Galveston; Andrzej Kudlicki, University of Texas Medical Branch at Galveston; Krzysztof Ginalski, University of Warsaw; Philippe Pasero, French National Center for Scientific Research

**ABSTRACT:** DNA double-stranded breaks (DSB) can result from endogenous processes, such as replication stress, or exogenous ones, like chemotherapeutics. We developed the method, termed BLESS, to label DSBs with single-nucleotide resolution and used it to detect them in samples with various levels of replication stress, in yeast and human. DSBs induced by replication stress such as replication fork collapse are asymmetric, which we exploited to infer fork position and reconstruct its movement. DSBs are rare events, therefore, signal-to-noise is typically low in DSB sequencing data. To address this, we built the model of the expected BLESS read pattern around an origin and used Fourier transform based filtering to improve our ability to detect pattern related to replication stress. Thus constructed model allowed us to predict 169 early origins in the budding yeast genome. Our predictions were confirmed to have at least 94% accuracy by BrdU staining. Our approach is applicable to other organisms, such
as human, although accuracy of our predictions is unclear in humans due to lack of high-quality data on origin location. Our analysis also suggests putative displacement of MCM double-hexamers in close vicinity of replication origins. Finally, analysis of the breakome obtained from the BLESS method and alternative Break-Seq technology allowed us to clarify what specific types of DSBs are formed during replication fork collapse and thus infer more precisely than previously possible where in the fork vicinity the breaks take place and which among several proposed mechanisms of break creation is most likely occurring in our experiments.

OP12: HRC3 – A NEW CLASS OF MOTIFS INVOLVED IN CHROMATIN ORGANIZATION AND DEVELOPMENT

Presenting Author: Andrzej Kudlicki, University of Texas Medical Branch

ABSTRACT: Chromatin modifications, such as methylation and acetylation of lysine residues in histone tails, are an important mechanism of epigenetic regulation. It remains unclear how the enzymes responsible for histone modifications are directed to the correct loci, in a manner that is specific to the cell type and outside stimuli.

We report the discovery of a conserved structural signature of DNA fragment that coincides with experimental binding sites of histone-modifying enzymes, such as KDM5B, KDM5A, PHF8, EZH2, RBBP5, SAP30, HDAC1 and HDAC6, also SUZ12, CHD1, SMARCB1 – involved in regulation of chromatin organization and silencing. The signature (“the HRC3 motif”) is approximately 180 base pairs long and is defined by a specific, periodic pattern in the Hydroxyl Radical Cleavage profile of a dsDNA interval. The pattern is present in both non-coding and coding sequences; in coding sequences it is produced by a very specific choice of codons in the region. The HRC3 signature is associated with several thousand genes; functional analysis show highly significant enrichment of genes involved in processes related to development (GO:0009888, GO:0048731, GO:00325020), regulation of gene expression and in DNA binding (GO:0003677). The HRC3 motifs are highly conserved, remaining unchanged from human to Drosophila. The most intriguing property of these motifs is their association with pairs or clusters of developmental transcription factors with a conserved synteny, including Hox genes. We present a model that uses HRC3s to explain the colinearity of HOX clusters in segmented animals. We also discuss their possible role in control of replication initiation.
OP13: NETWORK INFERENCE AND THE KNOWLEDGE BASE OF BIOMEDICINE

Presenting Author: Tiffany Callahan, University of Colorado Denver Anschutz Medical Campus

Co-Author(s): William A. Baumgartner Jr, University of Colorado Denver Anschutz Medical Campus; Marc Daya, University of Colorado Denver Anschutz Medical Campus; Lawrence E. Hunter, University of Colorado Denver Anschutz Medical Campus

ABSTRACT: Structural transformation of biological knowledge represented using Semantic Web standards significantly improves the utility of visualization tools and network analytics. Link prediction algorithms are powerful tools for predicting unobserved connections between nodes in a network. The application of such algorithms to biological networks has lead to the correct prediction of previously unobserved relationships ranging from protein-protein interactions to novel P53 kinases. The use of such algorithms to analyze larger and more complex representations has the potential to generate novel and important hypotheses, and insights into biological mechanisms. Unfortunately, the direct application of these algorithms to biological knowledge is limited by the representational complexity of the web ontology language standard OWL. The Network Information Content Entity (NICE) approach, a novel transformation method, reversibly transforms OWL-compliant biomedical knowledge into a representation better suited for visualization and network inference algorithms. Using several illustrative biomedical queries, the NICE transformation produces simpler network representations that are more visually comprehensible and whose structural properties (e.g. clustering coefficient, modularity, number of shortest paths, number of average neighbors, and diameter and radius) are significantly improved over raw OWL. Furthermore, comparison of the results from the application of several state-of-the-art link prediction algorithms on raw OWL versus NICE networks shows that the NICE transformation results in more accurate and biologically meaningful predictions. For each query and each algorithm, the top-ten predicted links for both OWL and NICE networks were validated via evidence from literature review and domain expert consultation.

OP14: ShinyLearner: ENABLING BIOLOGISTS TO PERFORM ROBUST MACHINE-LEARNING CLASSIFICATION

Presenting Author: Stephen Piccolo, Brigham Young University

Co-Author(s): Terry Lee, Brigham Young University; Shelby Taylor, Brigham Young University

ABSTRACT: Machine-learning classification is an invaluable tool for biologists. In one type of application, biomedical researchers use classification algorithms to predict whether patients will respond to a particular drug or belong to a specific disease subtype. Although the research community has developed many classification algorithms and
corresponding software libraries, considerable barriers exist for non-computational biologists to take advantage of these tools. Different algorithms are written in different programming languages and require different input formats. Software libraries may require dependencies that are difficult to install, and the software may fail if incompatible versions are installed. If a researcher wanted to employ algorithms implemented in multiple software libraries, she/he may need to learn multiple programming languages and be careful to avoid biases as comparisons were made across the algorithms.

We developed ShinyLearner (https://github.com/srp33/ShinyLearner), an open-source software tool that reduces these barriers. ShinyLearner integrates several popular machine-learning libraries (e.g., scikit-learn, mlr, weka) within a Docker container that includes all software dependencies. Accordingly, ShinyLearner can be installed with ease. ShinyLearner supports Monte Carlo and k-fold cross validation and provides an option for feature selection. When multiple classification algorithms are used, ShinyLearner dynamically selects the best algorithm via nested evaluation. A simple Web interface facilitates the process of selecting parameters. Output files are in “tidy” format to enable easier processing with external tools. New algorithms can be integrated into ShinyLearner with a simple GitHub pull request.

Finally, we will describe findings from a comprehensive benchmark comparison across classification algorithms applied to 20+ gene-expression data sets.

**OP15: STRATIFICATION OF PROSTATE CANCER PATIENTS BASED ON MOLECULAR INTERACTION PROFILES**

Presenting Author: Roland Mathis, IBM Research

Co-Author(s): Matteo Manica, IBM Research; Maria Rodriguez Martinez, IBM Research

**ABSTRACT:** Prostate cancer is a leading cause of cancer death amongst men, however the molecular-level understanding of disease onset and progression are largely unknown. Specifically, stratification of intermediate prostate tumor states based on current markers is difficult. The aim of this project is to integrate multi-omics data from individual patients with knowledge from literature and public databases to infer a molecular interaction network specific to prostate cancer. Inspired by the DREAM5 challenge we integrate predictions from multiple inference methods based on information theory, correlation and regression models to build a disease specific interactome. Emphasis is put on combining different data types and systematically integrating prior information using natural language processing and knowledge graphs. From the interactome we identify relevant interaction modules through
graph-theory approaches. For each interaction module we cluster the patients based on molecular states measurements. The patient-specific cluster assignment vectors serve as a personalized interaction signatures and is used to stratify patients.

OP16: MEDICATION DATA MINING OF ELECTRONIC MEDICAL RECORDS REVEAL RACE-SPECIFIC PRESCRIPTION PATTERNS

Presenting Author: Benjamin Glicksberg, Icahn School of Medicine at Mount Sinai

Co-Author(s): Kipp Johnson, Icahn School of Medicine at Mount Sinai; Khader Shameer, Icahn School of Medicine at Mount Sinai; Joel Dudley, Icahn School of Medicine at Mount Sinai

ABSTRACT: Introduction: Disparities in medication availability, tolerability, and effectiveness exist and patient outcomes. We aimed to mine electronic medical records (EMR) and quantify differences in medication counts, prescription-record counts, and drug-class enrichment using the New York Metropolitan area population compiled from Mount Sinai Data Warehouse.

Methods: Self-reported ancestry was abstracted from EMR (n=2.1 million) as Caucasian (EA), African-American (AA), Hispanic/Latino (HL), Asian (A), or Other (O). Medications were normalized with RxNorm and mapped to Anatomical Therapeutic Chemical (ATC) drug-classes using the PharmaFactors software framework.

Results: We found differences in prescription and unique medication count between races (one-way ANOVA, p<5E-16 for both). AA individuals had more prescription instances and unique medications compared to all other racial groups (Tukey HSD, p<10-16, all comparisons). Conversely, HL individuals had the fewest prescription instances and unique medications compared to all other groups (Tukey HSD, p<10-16, all comparisons). Polypharmacy (4+ simultaneous drug prescriptions) varied according to race (χ² p<10-16), EA having the highest rates (0.58) and AA the lowest (0.43). ATC drug-class enrichment varied with race: of 473 level 4 ATC classes, we found 125 and 70 enriched for EA and AA respectively (Fisher’s Exact Q<0.05, OR>1). The most enriched classes per group were EA, joint muscle pain and bowel disorders (OR=8.73 for both); AA, antiseptics (OR=8.38); HL, thiazolidinediones (OR=1.14); and A, Nucleoside/nucleotide reverse transcriptase inhibitors (OR=7.42).

Conclusion: We identified various ancestry-specific prescription data patterns. Further investigation of these patterns may help to develop prescription practices and improve therapeutic outcomes by optimizing drug efficacy and lowering side effects.
OP17: COMPARATIVE ANALYSIS OF THE EXPRESSION PATTERNS AND REGULATION OF HISTONE VARIANT GENES REVEALS A NOVEL EPIGENETIC PATHWAY RELATED TO CANCER

Presenting Author: Michael Tolstorukov, Massachusetts General Hospital and Harvard Medical School

Co-Author(s): Jakub Mieczkowski, Massachusetts General Hospital and Harvard Medical School; Sihem Cheloufi, Massachusetts General Hospital and Harvard Medical School; Konrad Hochedlinger, Massachusetts General Hospital and Harvard Medical School

ABSTRACT: Minor histone variants replace canonical histones in replication-independent manner, altering chromatin structure and thereby affecting gene expression. This constitutes a distinct mechanism of genome regulation, extending the function of nucleosomes beyond ‘simple’ DNA packaging. In an unusual genomic arrangement, two genes with unique sequences (H3F3A and H3F3B), both encode the same protein – a developmentally essential histone variant H3.3. It has been recently discovered that the mutations in each of these genes occur in different cancers, including pediatric brain tumors. To understand the biological role and regulation of each of these genes we performed an integrative analysis of gene expression, chromatin organization and DNA mutability. We show that the H3.3 genes have distinct expression patterns in human cell types. This difference is most pronounced between differentiated and stem-like cells, whose expression profile resembles that of some cancers. Further analysis and experimental tests reveal that the transcription factors, including Oct4/Sox2 and N-Myc, can differentially regulate these genes. We directly demonstrate that Oct4 and Sox2 upregulate H3f3a but not H3f3b in mouse embryonic stem cells. Notably, the increased H3F3A contribution to the total H3.3 pool, i.e. its ‘transcriptional dosage’, correlates with tumor malignancy in humans. We infer that a similar increase in the H3F3A transcriptional dosage in stem-like cells enables the mutations in this gene to impact cell fate determination. Collectively, our findings provide new insights into the interplay between gene expression and DNA mutations, and point to potential therapeutic strategies in the case of the H3.3-related cancers.
OP18: THE COGNOMA COLLABORATIVE CREATES A WEBAPP TO PREDICT CANCER MUTATIONS FROM GENE EXPRESSION

Presenting Author: Daniel Himmelstein, University of Pennsylvania

Co-Author(s): Gregory Way, University of Pennsylvania; Andrew Madonna, DataPhilly; Alan Elkner, DataPhilly; Benjamin Dolly, DataPhilly; Yichuan Liu, DataPhilly; Claire McLeod, DataPhilly; Derek Goss, DataPhilly; Haitao Cai, DataPhilly; Chris Fuller, DataPhilly; Robert Paul Miller, DataPhilly; Nabeel Sarwar, DataPhilly; Branka Jokanovic, DataPhilly; Mans Singh, DataPhilly; Stephen Shank, DataPhilly; Joel Eden, DataPhilly; Karin Wolok, DataPhilly; Casey Greene, University of Pennsylvania

ABSTRACT: Precision oncology requires that we functionally categorize cancers into treatment-relevant subtypes. The predominant approach—characterizing tumors based solely on actionable mutations—struggles to detect complex changes in gene or pathway function. Alternatively, genome-wide expression profiles provide a comprehensive reflection of aberrant cellular states resulting from mutation events. Therefore, we embarked on Project Cognoma to translate between gene expression and mutation in cancer.

Cognoma is an open-source/citizen-science philanthropy being developed as a collaboration between the Greene Lab at Penn and the DataPhilly and Code for Philly meetups. This arrangement leverages the collective fullstack expertise of our diverse contributor base. Hitherto, hundreds of individuals have attended Cognoma meetups, and more than fifty have gotten involved on GitHub (https://github.com/cognoma/cognoma). Our priorities are “everyone learns something new” and “putting machine learning in the hands of cancer biologists.” Our product is cognoma.org, a webapp that makes it easy to build mutation status classifiers from gene expression on 7,306 TCGA samples representing 33 cancer types. The publicly available dataset contains RNA-seq gene expression for 20,530 genes, non-silent mutation calls for 21,940 genes, and sample attributes such as the patient’s disease, age, sex, and survival. Cognoma enables a cancer biologist to assign each sample a mutation status based on one or more selected genes. Next, a disciplined classifier is trained using gene expression and sample attributes as features. As output, the user receives the importance of each feature—offering insight into the molecular effects of their chosen mutation—as well as a mutation scores for samples—which potentially identify hidden responders to targeted pharmacotherapies.
OP19: FUNCTIONALLY PRIORITIZING CANDIDATE GENES FROM GENOME-WIDE ASSOCIATION STUDIES

Presenting Author: Kelsey Anderson, University of Colorado School of Medicine
Co-Author(s): Sonia Leach, National Jewish Health

ABSTRACT: Genome-wide association studies (GWAS) have become the main approach for studying the genetic architecture of common diseases and traits. GWAS correlate variants at genomic loci with the trait under study. Recovery of the important genes from these loci, however, is not always straightforward. Recent evidence suggests the majority of associations found in GWAS do not change the protein-coding region of genes, but instead affect the regulation of gene transcription. Since regulatory regions like enhancers can be hundreds of kilobases away from their target gene’s promoter, a locus from a GWAS may reasonably contain dozens of plausible candidate genes. Methods to computationally select or prioritize these candidates can help save researcher time and/or verify decisions. They can also suggest the underlying biology and inform mechanistic hypotheses. Here we propose a method for functionally prioritizing the candidate genes from GWAS data. We use orthogonal evidence from protein-protein interaction (PPI) networks to score each candidate, under the assumption that the true causal proteins will be functionally related. Unlike other prioritization approaches that search for dense modules in the protein network, we use a Regularized Laplacian graph kernel to measure similarity between proteins in the network. Candidates score highly if they are strongly associated with other candidates, all of whom are similar to each other according to the graph kernel. The method is evaluated against a number of existing network-based prioritization approaches on several complex disorders and traits. In nearly all cases, our method outperforms the competition.

OP20: DERIVING POPULATION-SCALE THERAPEUTIC TRAJECTORIES TO ENABLE PRECISION PHARMACOLOGY

Presenting Author: Kipp Johnson, Icahn School of Medicine at Mount Sinai
Co-Author(s): Benjamin Glicksberg, Icahn School of Medicine at Mount Sinai; Khader Shameer, Icahn School of Medicine at Mount Sinai; Joel Dudley, Icahn School of Medicine at Mount Sinai

ABSTRACT: Introduction: Treatment pathways provide standard guidelines for treating the primary diseases of patients. However, patients present with comorbidities, side effects and comply poorly with treatment adherence. Availability of a precision prescription data analytics platform may help to understand factors driving better therapeutic outcomes and lower side effects.

Methods and Results: The Mount Sinai EMR contains over 18.5 million prescriptions od 1,510 unique medications. Of the entire
hospital population used in this study, 803,157 (38.2%) had at least one prescription (23.25±87.21). Polypharmacy prevalence (co-administration of 4+ prescriptions) increased in an age-dependent manner, from 4% in those 0-10 years old to 62.8% in those >80. 95,373 drug pairs were enriched for co-administration (Exact-test Q<0.01). 23,656 drug-pair sequences (drug 1 followed by drug 2) were detected (Binomial Q<0.01) including the stimulants modafinil to armodafinil (OR=185), antiplatelet therapies aspirin to ticagrelor (OR=139), diabetes drugs liraglutide to canagliflozin (OR=79), antipsychotics olanzapine to haloperidol (OR=63), and drug-antidote pair naloxone and hydromorphone (OR=22). We assembled a directed network of drug trajectories with 838 nodes and 23656 edges (diameter=13) from drug pair trajectories. Greedy clustering partitioned the network into 7 subgraphs. Network hubs were detected and scored with Kleinberg’s method (principal eigenvectors of Adj(M)*t(Adj(M)). Top hub drugs were lisinopril, amlodipine, aspirin, fluticasone/salmetrol, hydrochlorothiazide, simvastatin, ergocalciferol, albuterol, furosemide, and omeprazole.

Conclusion: Systematic mining of prescription data could help to uncover relationships between therapies and outcomes and aid in the implementation of precision prescription workflows.

**OP21: COMPARISON OF RELIEF-F NUCLEOTIDE DIFFERENCES FOR GWAS DATA WITH APPLICATION TO BIPOLAR DISORDER**

Presenting Author: Marziyeh Arabnejad Khanouki, University of Tulsa

Co-Author(s): Brett McKinney, University of Tulsa; Bill White, University of Tulsa

**ABSTRACT:** Genetic studies of the bipolar disorder disease have found an overall strong inheritance pattern for the disease but have not found specific genes with individually strong effects. Thus, there are likely many genes that increase the susceptibility to develop bipolar disorder for many individuals, but additional algorithms are needed to help identify the patterns of genes and other factors that act together to produce the illness or increase risk.

In this work, a feature selection algorithm called ReliefF is used to rank the SNPs for the cases of bipolar disorder and normal controls in two published genome-wide association studies from the National Institute of Mental Health and the Wellcome Trust Case-Control Consortium. The ranking is done through our software called “ReliefSeq”. There are two places in the ReliefSeq implementation of ReliefF where distance is used: ReliefF feature weighting and computing the nearest neighbors. The distance in ReliefSeq computed with three metrics: genotype mismatch(gm), allele mismatch(am), and Transition/Transversion(Ti/Tv). In total, nine combinations of the metrics are implemented in the analysis of 5000 SNPs.
Analysis yielded several SNPs that may have involvement in the pathophysiology of bipolar disorder. After finding the relevant genes and pathways for the high-rank SNPs, it was observed that most of the metrics combinations enriched Neuronal System and Axon-Guidance pathways in both data sets. The combinations of Ti/Tv metric performed relatively better than the other combinations in enriching the Neuronal System pathways in both data sets while most of the combinations with gm were found to be less successful in enriching the Axon-Guidance pathway.

OP22: ModEvo: A WEB-BASED TOOL FOR MODELING EVOLUTIONARY DYNAMICS

Presenting Author: Filip Jagodzinski, Western Washington University
Co-Author(s): Rainier Harvey, Western Washington University; Jesse Sliter, Western Washington University; Elizabeth Brooks, Western Washington University; Ali Scoville, Central Washington University

ABSTRACT: Quantitative genetics is concerned with developing computational models to predict the evolution of traits in response to selection. Most models for analyzing the evolution of multiple traits employ a constant genetic variance co-variance matrix (G-Matrix). However, non-linear interactions between developmental factors underlying the production of traits can drastically affect how they co-vary.

We have developed a code-base, ModEvo, to assist in testing hypothesis about the evolutionary dynamics among multiple phenotypic traits affected by non-linear developmental interactions. Our software implements and extends a novel mathematical framework developed by Sean Rice that synthesizes concepts central to evolutionary developmental biology and quantitative genetics.

We are developing a Graphical User Interface (GUI) and the accompanying back-end infrastructure to permit biologists to interface with ModEvo via a publicly available web server. Users specify input parameters for the quantitative genetics models and invoke the back-end modeling software with a single button click. The evolutionary dynamics output by ModEvo are displayed both graphically and numerically. The front-end, back-end infrastructure uses Google Go as the back-end server and Angular as the front-end model-view controller. Our web tool is easy enough to use by a non-specialist, but also allows an experienced user to specify model parameters for a more detailed analysis.
**OP23: PREDICTING NEURAL FLUCTUATIONS IN THE PRIMARY VISUAL CORTEX**

Presenting Author: *William Kindel, University of Colorado School of Medicine*

**ABSTRACT:** The images we perceive are processed by our brain through evoking neural activity. Understanding how these images are turned into patterns of neural activity is an outstanding question in neuroscience. The solution will enable new therapies such as synthetic vision for those with damaged eyes because technology already exists to excite arrays of neurons— the difficulty is knowing precisely which ones to excite. Moreover, cracking this neural code is complicated because the brain responds uniquely to repeated presentations of the same image. The number of times an individual neuron fires within a window in time fluctuates greatly.

In this presentation, I am focused on understanding and predicting these neuronal fluctuations as part of a translator that predicts the neural activity in the primary visual cortex (V1) in response to seeing any image. Fortunately, the seemingly random neuronal fluctuations contain many correlations between near neurons and over a short time window. Thus over many pairs of correlated neurons, knowing the history of one neuron’s activity can improve the predictions of another neuron’s activity. Synthesizing all of this information to form my fluctuation predictor, I utilize artificial neural networks, which can, in principle, find any relationship between input and output variables. Using data from nonhuman primates, I build and then benchmark the predictor by bounding how much of neuronal noise I can predetermine. Thus, I shed light on the information stored in the neuronal fluctuations.

**OP24: DE NOVO PROTEIN STRUCTURE PREDICTION BY BIG DATA AND DEEP LEARNING**

Presenting Author: *Sheng Wang, Toyota Technological Institute at Chicago*

Co-Author(s): *Jinbo Xu, Toyota Technological Institute at Chicago*

**ABSTRACT:** Recently ab initio protein folding using predicted contacts as restraints has made some progress, but it requires accurate contact prediction, which by existing methods can only be achieved on some large-sized protein families. To deal with small-sized protein families, we employ the powerful deep learning technique from Computer Science, which can learn complex patterns from large datasets and has revolutionized object and speech recognition and the GO game. Our deep learning model for contact prediction is formed by two deep residual neural networks. The first one learns relationship between contacts and sequential features from protein...
databases, while the second one models contact occurring patterns and their relationship with pairwise features such as contact potential, residue co-evolution strength and the output of the first network. Experimental results suggest that our deep learning method greatly improves contact prediction and contact-assisted folding. Tested on 579 proteins dissimilar to training proteins, the average top L (L is sequence length) long-range prediction accuracy of our method, the representative evolutionary coupling method CCMpred and the CASP11 winner MetaPSICOV is 0.47, 0.21 and 0.30, respectively; their average top L/10 long-range accuracy is 0.77, 0.47 and 0.59, respectively. Using our predicted contacts we can correctly fold 203 test proteins, while MetaPSICOV and CCMpred can do so for only 79 and 62 proteins, respectively. In the three weeks of blind test with the weekly benchmark CAMEO (http://www.cameo3d.org/), our method successfully folded three hard targets with a new fold and only 1.5L-2.5L sequence homologs while all template-based methods failed.

**OP25: IDENTIFYING THE MECHANISM FOR THE METASTATIC SPREAD OF BREAST CANCER THROUGH INTEGRATION OF GENE EXPRESSION, WHOLE GENOME SEQUENCING AND FUNCTIONAL SCREENS**

Presenting Author: Eran Andrechek, Michigan State University

**ABSTRACT:** Breast cancer mortality is usually caused by metastasis to distant sites. Using genomic signatures to predict cell signaling pathway activation has allowed us to develop hypotheses about key signaling pathways that are involved in the metastatic progression of breast cancer. To test the hypothesis that the E2F transcription factors are involved in metastasis, we generated a mouse model of breast cancer lacking E2F1 or E2F2. Consistent with our hypothesis, these mice developed breast cancer lacking metastasis. The E2F family of transcription factors is classically known to regulate G1 to S-phase transition in cell cycle, however, other functions have emerged. To address the function of the E2Fs in metastasis, gene expression of tumors from wild type and E2F knockout backgrounds were analyzed. This was integrated with whole genome sequence data from matched tumor samples. Potential genetic mechanisms identified through this approach were validated for human relevance using TCGA data. Patient outcomes were screened for these genes through the application of a predictive gene expression signature. Further, the Achilles project and a drug screening study in patient derived xenograft tumors were two functional screens that were also integrated with this work. The outcome of the integrated study was the identification of an amplification event in breast cancer that correlates with metastasis. Genetic ablation of genes in this amplicon revealed specific roles in metastatic migration. Together this work demonstrates the utility of integrating multiple data platforms to address key biological problems.
OP26: ALLELIC MAPS OF CANCER

Presenting Author: Anelia Horvath, George Washington University

Co-Author(s): Paula Restrepo, GWU; Muzzi Li, Georgetown University; Nawaf Alomran, Georgetown University; Piotr Slawinski, University of Exeter; Mercedeh Movassagh, GWU; Sonali Bahl, GWU; Wesley Waterhouse, GWU; Christian Miller, GWU; Chris Trenkov, GWU; Julian Manchev, GWU; Tatiyana Apanasovich, GWU; Nathan Edwards, Georgetown University; Krasimira Atanasova-Tsaneva, University of Exeter

ABSTRACT: The relationships between genome- and transcriptome dosage have been challenging to study. Allele-specific signals, especially when integrated between DNA and RNA, allow tracking chromosome-of-origin transcripts and may provide insights on the transcriptional dynamics of the aneuploid cell. We present a novel approach for building integrated RNA-DNA maps that depict allelic asymmetries, including those corresponding to aneuploidy, and enable correlation to expression features both at nucleotide resolution and continuously across genes and chromosomes. Using a currently developed in our lab software - RNA2DNAAlign – we counted the reference and variant sequencing reads at every variant position in all matching datasets, and computed the variant allele fraction VAF (VAF = nvar/(nref+nvar)) where, in a pure diploid sample, VAFDNA of 0.5 corresponds to a true allelic ratio of 1:1 for heterozygote sites. To build the Allelic Maps, for all heterozygote sites in the normal DNA, we plotted VAFnDNA along the chromosomes of the matching normal exome, normal transcriptome, tumor exome, and tumor transcriptome. To define regions of aneuploidy, we adopted a model based on widely accepted notion, where aneuploidy is depicted through bimodal VAF distribution. We simulated ideal distributions corresponding to mono-, tri-, tetra-, and penta-ploid status, and tested the in the tumor VAFDNA distribution for highest similarity using the Earth Movers Distance measure (EMD). We demonstrate the application of this approach using sequencing data from human tumor tissues, cell lines, and single cell experiments.

OP27: IDENTIFYING NON-SPECIFIC EFFECTS OF SMALL MOLECULE TREATMENT THROUGH GSEA META-ANALYSIS

Presenting Author: Rani Powers, University of Colorado Anschutz Medical Campus

Co-Author(s): Andrew Goodspeed, University of Colorado Anschutz Medical Campus; James Costello, University of Colorado Anschutz Medical Campus

ABSTRACT: Despite advancements in therapeutic strategies such as antibodies and gene therapy, small molecules remain the gold standard of treatment for numerous diseases, including cancer. Small molecules are low molecular weight compounds that rapidly diffuse across cell membranes to reach their molecular target, which is often a protein or nucleic acid. For example, many small molecule therapies inhibit the activity of a specific kinase. When investigating the effect of a small molecule on cell state or disease, researchers
often compare the genome-wide mRNA levels of drug-treated cells and vehicle-treated control cells. The output of this experiment is a list of differentially expressed genes, which either increase or decrease in expression following drug treatment. This list can then be analyzed with gene set enrichment analysis (GSEA), an algorithm which performs hypergeometric tests with curated gene sets to determine which biological processes are more or less active in the drug-treated cells.

We hypothesized that even a highly-specific small molecule drug may result in non-specific effects in the cell, such as an up-regulation of generic stress response pathways. These non-specific pathways may appear as significant in the GSEA output, potentially overshadowing crucial biological processes specific to the drug under investigation. To address this problem, we aggregated several hundred gene expression experiments where human tissues or primary cells were treated with a small molecule drug. These experiments were annotated before analysis with GSEA. Our results identified pathways that are overrepresented in small molecule drug screens, providing valuable experimental and biological insight into therapeutic drug development.

OP28: INSIGHTS INTO BATHYARCHEOTA ECOLOGY AND CO-OCCURRENCE PATTERNS AS REVEALED BY PUBLIC METAGENOME SEQUENCING DATA

Presenting Author: David Banks-Richardson, University of Colorado-Denver
Co-Author(s): Christopher Miller, University of Colorado-Denver; Adrienne Narrowe, University of Colorado-Denver

ABSTRACT: Members of the archaeal phylum Bathyarchaeota are a major component of aquatic sediment microbial communities. To date, a comprehensive inter-domain assessment of the co-occurrence patterns between this phylum and other organisms has not been done, and surveys of Bathyarchaeota habitat preferences have relied on amplicon-based 16S rRNA studies of limited ecosystems. Our in-silico analyses suggest that commonly used primers in such 16S rRNA amplicon studies may be obscuring large portions of Bathyarchaeota phylogeny. Shotgun metagenomic sequencing has the potential to shed light onto Bathyarchaeota habitat preferences, especially for the portions of the clade that PCR-primer bias may be hiding, but shotgun assembly is often incomplete or lacks the context of existing 16S phylogeny. Here, we employ a targeted gene assembly approach (EMIRGE) to reconstruct 16S rRNA sequences from publically available shotgun metagenomes representing several environmental and host-associated habitats. We quantify the degree to which PCR-primer bias is obscuring the diversity of the Bathyarchaeota, build a cross-domain co-occurrence network between members of Bathyarchaeota and other microorganisms, and identify association patterns between environmental variables and the Bathyarchaeota. Preliminary results suggest that members of this phylum may co-occur with members of
the bacterial phyla Proteobacteria, Planctomycetes, and OP1, and that sub-clades within this group respond differentially to depth gradients in estuarine sediments. This study informs future work seeking to characterize the roles these broadly distributed archaea play in microbial communities across the globe.

**OP29: THE SNPPhenA CORPUS: AN ANNOTATED RESEARCH ABSTRACT CORPUS FOR EXTRACTING RANKED ASSOCIATION OF SINGLE-NUCLEOTIDE POLYMORPHISMS AND PHENOTYPES**

Presenting Author: Hamidreza Chitsaz, Colorado state university

Co-Author(s): Behrouz Bokharaeian, Complutense University of Madrid; Alberto Diaz, Complutense University of Madrid; Ramyar Chavoshinejad, Royan Institute for Reproductive Biomedicine

**ABSTRACT:** Single Nucleotide Polymorphisms (SNPs) are the most comprehensively studied type of genetic variations that influence a number of diseases and phenotypes. Recently, some corpora and methods have been developed for extracting SNPs, diseases, and their associations from scientific text. However, there is no available method and corpus for extracting those SNP-disease associations that have been annotated with linguistic based negation, modality markers, neutral candidates, and the level of confidence of association.

**Method:** In this research, we present different steps for producing the SNPPhenA corpus. They include automatic Named Entity Recognition (NER) followed by the manual annotation of SNP and phenotype names, annotating the SNP-phenotype associations and their level of confidence, as well as modality markers. Moreover, the produced corpus has been annotated with negation scopes and cues as well as neutral candidates that have an important role in dealing with negation and the modality phenomenon in relation extraction tasks.

**Result:** The agreements between annotators were measured by Cohen’s Kappa coefficient and the resulting scores showed reliability of the corpus. The Kappa score was 0.86 for annotating the associations and 0.80 for annotating the degree of confidence of associations. Additionally, basic statistics for extracting ranked SNP-phenotype associations are presented here, with regard to the annotated features of the corpus besides the results of our first experiments. Moreover, we prepared guidelines for using the corpus. The guidelines and the corpus are available at http://nil.fdi.ucm.es/?q=node/639.

**Conclusion:** Estimating confidence of SNP-phenotype associations could help determine phenotypic plasticity and the importance of environmental factors. Moreover, our first experiments with the corpus show that linguistic-based confidence alongside other non-linguistic features can be utilized to estimate strength of the observed SNP-phenotype association. Trial Registration: Not Applicable
OP30: TOWARD A METRIC LEARNING MODEL FOR PROTEIN FOLD RECOGNITION USING A NOVEL FEATURE EXTRACTION TECHNIQUE BASED ON THE MIXTURE OF EVOLUTIONARY AND SECONDARY STRUCTURAL INFORMATION

Presenting Author: Pooya Zakeri, 1)KU Leuven. 2)iMinds
Co-Author(s): Forough Amini, Institute for Research in Fundamental Sciences (IPM); Mehdi Sadeghi, 1) National Institute of Genetic Engineering and Biotechnology, 2)Institute for Research in Fundamental Sciences (IPM); Yves Moreau, 1) KU Leuven, 2)iMinds

ABSTRACT: It has been demonstrated that integrating the evolutionary and secondary structural information could be useful in explicating the relationship between primary and tertiary structure in proteins. In this work, we propose a protein feature extraction technique based on the blending of evolutionary and secondary structural information. A protein feature vector is created by summing up each column of the same predicted secondary structure elements in the position-specific scoring matrix and dividing by the length of the protein domain. Then, the kernel-based kNN is employed to address the prediction of protein folds. Nevertheless, the performance of kNN-based fold predictors depends crucially on the metric used to measure distances between protein sequences. To overcome this limitation, we develop for the first time a protein fold predictor based on the large margin nearest neighbor algorithm, which learns a Mahalanobis distance metric for the kNN algorithm in a supervised fashion to improve its performance.

In order to evaluate our models in a more realistic task setting, we develop a time-stamped benchmark based on the SCOP database. Our models are trained based on the proteins of known fold discovered before a certain time using protein features released prior to that time. Then, we assess our model on the prediction of proteins of known fold reported afterwards. The experimental results on our prospective benchmark which covers about two-hundreds folds demonstrate that our developed model based on our proposed feature extraction technique can effectively improve the accuracy of the state-of-the-art protein fold predictors, such as TAXFOLD and GeoFold.

OP31: DEVELOPMENT OF A DIAGNOSTIC TO PROFILE EUKARYOTIC MICROBES OF THE HUMAN MICROBIOME

Presenting Author: Ana Popovic, Hospital for Sick Children, University of Toronto
Co-Author(s): John Parkinson, Hospital for Sick Children, University of Toronto; Michael Grigg, National Institutes of Health

ABSTRACT: Human microbiome studies have implicated the composition of gut bacteria in function of the immune system,
obesity, drug metabolism, even human behaviour. While much has been learned about the contribution of bacteria to human health and disease, few studies have addressed the role of the eukaryotic members of the microbiome. This represents a considerable gap in knowledge, as single celled eukaryotes such as Giardia, Cryptosporidium and Entamoeba infect hundreds of millions of people worldwide, and are responsible for a significant burden of gastrointestinal illnesses. In addition to pathogenic eukaryotes, several studies have identified particular species of Blastocystis and Entamoeba as residents of the healthy gut, suggesting that eukaryotic microbes play a larger role than previously appreciated in human health. A key challenge in establishing the contribution of the eukaryotic microbiome to health and disease is the lack of accurate diagnostic technology. Here, we will present our efforts to develop a new multi DNA biomarker technology, based on several hypervariable regions in the small and large ribosomal subunit genes, to accurately profile eukaryotic microbes in the human gut.

**OP32: G4 QUADRUPLEXES IN AND NEAR REGULATORY ELEMENTS OF MAIZE GENES PREDICT TISSUE TYPE AND ALTERED TRANSCRIPTIONAL RESPONSE TO ABIOTIC STRESSES**

Presenting Author: Mingze He, Iowa State University

Co-Author(s): Angélica Sanclemente, University of Florida; Carson Andorf, USDA; Hank W. Bass, Florida State University; Harkamal Walia, University of Nebraska-Lincoln; Justin W. Walley, Iowa State University; Karen Koch, University of Florida; Peng Liu, Iowa State University; Carolyn J. Lawrence-Dill, Iowa State University

**ABSTRACT:** In maize shoot tissues genes with G4-quadruplexes in or near regulatory regions respond strongly to diverse stress conditions including submergence, cold, heat UV, salt, and cold stress. GO enrichment studies indicate that differentially expressed G4-containing genes are likely to be involved in developmental processes, suggesting that altered growth rates may be a specific component of the stress response. To further investigate the function of these G4 genes, we carried out transcriptomic and proteomic analyses across 55 tissues and developmental stages in non-stress conditions. We found G4 could be applied as a marker to predict transcription rate and specific tissue type in normal tissues. In addition, co-expression network analysis between maize atlas and stressed tissues revealed G4 motifs strongly associated with transcription factors activation in response to stresses. Our results provide novel evidence to the association of G4 with emergent energy status in maize. Our findings suggest a new component in maize stress response mechanism.
**OP33: MODELING HETEROGENEOUS CELL POPULATIONS USING BOOLEAN NETWORKS**

Presenting Author: Brian Ross, University of Colorado  
Co-Author(s): James Costello, University of Colorado

**ABSTRACT:** Cellular processes can be simulated using Monte Carlo (random sampling) methods, but these have difficulty capturing rare outcomes, particularly when the state space is huge. Yet in many cases (such as cancer) these infrequent outcomes are the ones with the most impact. Here we present a Boolean network method for modeling mixed cell populations using a single simulation, which captures these very rare subpopulations. Our method works by treating the dynamics as a system of linear equations which allow superposition of different cell populations, in a basis rotated from the state space so that the equations tend to close with relatively few variables. For cases when the variable space is still too large, we show how to efficiently remove degeneracies in our linear system as it is being built, thereby capturing the later-time evolution with a reduced system of equations. Our method generalizes to probabilistic Boolean networks, and works for both discrete and continuous time-evolution.

We evaluate our method using a >50-gene network modeling prostate cancer. Our method reproduces the results of Monte Carlo while capturing rare events that Monte Carlo cannot find. As a proof of concept, we simulate the dynamics of a mixed population spanning $>10^{15}$ different cell states with all possible combinations of loss-of-function mutations. Finally, we use these simulations to find the likely mutational trajectories of an evolving tumor in our prostate-network model. Our method can thus identify the extraordinary, as well as the typical, fates of cells.

**OP34: ENHANCER REPROGRAMMING IN MAMMALIAN GENOMES**

Presenting Author: Mario Flores, NIH

**ABSTRACT:** It has been shown that changes in regulatory regions (enhancers) have supported evolution in mammals. However there is still a lack of knowledge about the distinct types of enhancers, their identification in more tissues/cell types and the mechanisms that act to modify these regulatory regions during evolution. Here we study a type of enhancers that we have named reprogrammed enhancers. Enhancer reprogramming establish that changes in the transcription factor binding sites of noncoding regulatory DNA sequences could potentially change their regulatory function. In this context, TFBSSs loss, gain and reshuffling within an enhancer can change its function (spatial and/or temporal regulatory activity). We have identified reprogrammed enhancers in 11 tissues/cell types in human and mouse. We estimate
that in average 30% of the total number of enhancers in a gene locus had been reprogrammed in the course of evolution. Furthermore, the analysis of DNA sequence changes underlying enhancer reprogramming shows a change in the transcription factor binding site (TFBS) composition that significantly overlaps with the TFBS composition of tissue specific enhancers. Our observations provide evidence that reprogrammed enhancers are important contributors of the shaping of the regulatory landscape during evolution.

This research is supported by the Intramural Research Program of the NIH, National Library of Medicine

OP35: THE FINITE STATE PROJECTION BASED FISHER INFORMATION MATRIX FOR THE DESIGN OF SINGLE-CELL EXPERIMENTS

Presenting Author: Zachary Fox, Colorado State University
Co-Author(s): Brian Munsky, Colorado State University

ABSTRACT: Measuring and understanding gene expression fluctuations is key to predicting and controlling gene regulation dynamics. Rapidly advancing experiments enable precise quantification of RNA and protein in single cells. However, to keep pace with expanding experimental capabilities, computational and theoretical approaches must also improve. If tightly coupled with experiments, computational analyses can extract improved insight from previous measurements and enhance the effectiveness of future experiments. The Fisher Information Matrix (FIM) is a tool that is often used to aid experiment design for engineering applications, but common FIM approaches focus on deterministic models and cannot capture the full information contained in stochastic single-cell distributions. Such distributions are known to be well captured by the chemical master equation (CME). However, the CME is frequently too difficult or impossible to solve, which precludes rigorous computation of the FIM. The finite-state projection (FSP) approach systematically reduces the CME to a finite, solvable set of ordinary differential equations. In this study, we extend the FSP to compute the FSP-FIM and estimate the expected information for potential single-cell experiments. In contrast to existing experiment design strategies, our FSP-FIM approach makes no assumptions about the underlying distributions. We demonstrate the advantage of the FSP-FIM approach on several common models of stochastic gene expression, for which previous approaches and assumptions of normal distributions are not justified. Our results allow for the computational exploration of many potential experiments, and can promote iterative and efficient integration of modeling and experimentation to understand, predict and control gene expression.
OP36: 2-SCALE KNN CLASSIFICATIONS

Presenting Author: Destiny Anyaiwe, Oakland University
Co-Author(s): George D. Wilson, William Beaumont Hospital, Royal Oak, MI; Timothy J. Geddes, William Beaumont Hospital, Royal Oak, MI; Gautam B. Singh, Oakland University

ABSTRACT: Diverse algorithms and methods are needed to answer the ever increasing need of adequately harnessing Mass Spectrometer generated data. The unique nature and structure of mass spectra data usually, requires a high level of expertise and rigorous algorithms. This study’s methodology discusses feature selection based on direct and simple mathematical observations of variables and their inter-relationships, Jackknife technique for data re-sampling, matrix to vector decomposition and successfully classifies Alzheimer’s disease patients into three disease stages; age-matched controls without any evidence of dementia, patients with mild cognitive impairment and patients with clinical symptoms of Alzheimer’s disease (AD), using a 2-scale principle of K-nearest neighbor (KNN) algorithm on SELDI data and without collaborating clinical records. Hitherto, there exists no clinical diagnostic tool for AD, in lieu of this, patient cognitive abilities are clinically followed-up over a period of time (may be months) to make a diagnosis. This practice usually leads to inconclusive diagnosis and results obtained from it are not generalizable. Our model provides an immediate classification and correctly classifies test data sets with 82% confidence. It can also identify traces of positive/negative change within and across data sets in regards to severity of the disease over time.

OP37: BEST PRACTICES FOR REPRODUCIBLE AND ROBUST DATA ANALYSIS IN A BIOINFORMATICS CORE FACILITY

Presenting Author: James Denvir, Marshall University
Co-Author(s): Don Primerano, Marshall University; Jun Fan, Marshall University; Swanthana Rekulapally, Marshall University

ABSTRACT: With the publication of standards for Minimum Information About a Microarray Experiment in 2001, and the subsequent establishment of global repositories for gene expression and sequencing data, the research community has substantial achievements in making research data associated with published, peer reviewed manuscripts available for reuse and evaluation. However, there are currently no standards for the amount of detail of the analysis performed that should be provided in a publication. Consequently, it is rare to find publications for which the data analysis pipeline has sufficiently detailed description for the analysis to be reproduced, or in some cases critically evaluated.

We adopted simple practices used in software engineering, including version control management, self-documenting code, and convention
over configuration techniques into the data analysis pipelines used in a small genomics and bioinformatics core facility. Adoption of these techniques both improved the ability of our facility to create reproducible pipelines, and enhanced operational efficiency.

OP38: THE AFFINITY DATA BANK FOR BIOPHYSICAL ANALYSIS OF REGULATORY SEQUENCES

Presenting Author: Todd Riley, University of Massachusetts Boston

Co-Author(s): Cary Calaneri, UMass Boston; Aadish Shah, UMass Boston; Brandon Phan, UMass Boston; Pritesh Patel, UMass Boston; Zazil Villanueva, UMass Boston; Devesh Bhimsaria, University of Wisconsin-Madison

ABSTRACT: We present The Affinity Data Bank (ADB), a suite of tools that provides biologists with novel aids to deeply investigate the sequence-specific binding properties of a transcription factor (TF) or an RNA-binding protein (RBP), and to study subtle differences in specificity between homologous nucleic acid-binding proteins. Also, integrated with Pfam, the PDB, and the UCSC database, The ADB allows for simultaneous interrogation of protein-DNA and protein-RNA specificity and structure in order to find the biochemical basis for differences in specificity across protein families. The ADB also includes a biophysical genome browser for quantitative annotation of in vivo binding – using free protein concentrations to model the non-linear saturation effect that relates binding occupancy with binding affinity. Importantly, the in vivo TF and RBP protein concentrations can be inferred from transcriptome or proteome data – including RNA-seq data. The biophysical browser also integrates dbSNP and other polymorphism data in order to depict changes in affinity due to genetic polymorphisms – which can aid in finding both functional SNPs and functional binding sites. Lastly, the biophysical browser also supports biophysical positional priors to allow for quantitative designation of the in vivo, locus-specific accessibility that a protein has to the DNA. With the inclusion of these biophysical occupancy-based and affinity-based positional priors, the ADB can properly model in vivo protein-DNA binding by integrating the effects of chromatin accessibility and epigenetic marks.

OP39: PATTERN-BASED ESTIMATION OF THE EXTENT OF EXPLICIT CONTRADICTION IN THE SCIENTIFIC LITERATURE

Presenting Author: Elizabeth White, University of Colorado Denver, Anschutz

Co-Author(s): K. Brettonnel Cohen, University of Colorado Denver, Anschutz; Jennifer Panzo, De La Salle University

ABSTRACT: Enormous amounts of effort are put into manually extracting findings from the scientific literature and associating them with entries in model organism databases, and as natural language
processing techniques improve, it may soon be possible to accelerate that effort considerably. But, what is the reliability of statements in that literature? One way to estimate that is to look for evidence of explicit contradictions in scientific journals. We used search engines and a set of patterns that can indicate a deliberate claim of contradiction of another paper’s findings to gauge the number of contradictions in a variety of scientific fields. The findings are consistent with the conclusion that there is a large amount of contradiction in the scientific literature, and that the amount of contradictions varies considerably from one sub-field to the next.

**OP40: TOWARDS HIGHLY ACCURATE MAPPING OF PROTEIN GLYCOSYLATION SITES IN THE HUMAN PROTEOME**

Presenting Author: Chen Li, Monash University  
Co-Author(s): Fuyi Li, Monash University; Jerico Revote, Monash University; Yang Zhang, Northwest A&F University; Geoffrey Webb, Monash University; Jian Li, Monash University; Jiangning Song, Monash University; Trevor Lithgow, Monash University

**ABSTRACT:** Glycosylation is a crucial and ubiquitous type of protein post-translational modification and plays an important role in cell-cell adhesion, ligand-binding and subcellular recognition. To facilitate high-throughput prediction of protein glycosylation sites, we proposed GlycoMine, a comprehensive tool for in silico identification of C-, N-, and O-linked glycosylation sites in the human proteome. Heterogeneous sequences and functional features were derived from various sources, and subjected to further two-step feature selection to characterize a condensed subset of optimal features that contributed most to the type-specific prediction of glycosylation sites. Experimental studies showed that GlycoMine significantly improved the prediction performance compared with existing prediction tools. Given the fact that little work has been done to systematically assess the importance of structural features to glycosylation prediction, we then proposed an updated version of GlycoMine, GlycoMine_struct, for improved prediction of human N- and O-linked glycosylation sites by combining sequence and structural features in an integrated computational framework with a two-step feature-selection strategy. Experiments indicated that GlycoMine_struct outperformed currently existing predictor incorporating both sequence and structure features, achieving AUC values of 0.941 and 0.922 for N- and O-linked glycosylation, respectively, on an independent test dataset. Both GlycoMine and GlycoMine_struct have been used to screen the human proteome to obtain high-confidence predictions for glycosylation sites. We anticipate that GlycoMine and GlycoMine_struct can be used as powerful computational approaches to expedite the discovery of glycosylation events and substrates to facilitate hypothesis-driven experimental studies.
POSTERS

LOCATION
Snowmass Convention Center Ballroom B

POSTER SESSION HOURS
Posters will only be available for viewing Friday. The Poster Session with authors present will be on Friday evening (see schedule below).
Poster Presenters must be available for presentation during the scheduled poster session.

POSTER NUMBER ASSIGNMENTS
Posters have been assigned a number. There will also be lists at the conference. Please put your poster on the poster board corresponding to the number assigned:

SCHEDULE

| FRIDAY, DECEMBER 9 | 9:00 AM – 12:00 PM | SET UP POSTERS |
|                   | 12:00 PM – 6:00 PM | POSTER VIEWING |
|                   | 6:30 PM – 7:30 PM  | (no authors present) |
|                   | 7:30 PM – 8:30 PM  | POSTER SESSION WITH AUTHORS |
|                   |                     | Even Number Posters |

* Authors please remove posters from boards at end of this session.*

POSTER LIST

P01: CAMSA: A TOOL FOR COMPARATIVE ANALYSIS AND MERGING OF SCAFFOLD ASSEMBLIES
Presenting Author: Max Alekseyev, George Washington University

P02: IDENTIFYING THE MECHANISM FOR THE METASTATIC SPREAD OF BREAST CANCER THROUGH INTEGRATION OF GENE EXPRESSION, WHOLE GENOME SEQUENCING AND FUNCTIONAL SCREENS.
Presenting Author: Eran Andrechek, Michigan State University

P03: IndeCut: A CUT-NORM BASED METHOD FOR EVALUATING INDEPENDENT AND UNIFORM SAMPLING IN NETWORK MOTIF DISCOVERY ALGORITHMS
Presenting Author: Mitra Ansariola, Oregon State University

P04: 2-SCALE KNN CLASSIFICATIONS
Presenting Author: Destiny Anyaiwe, Oakland University

P05: WITHDRAWN

P06: INSIGHTS INTO BATHYARCHEOTA ECOLOGY AND CO-OCCURRENCE PATTERNS AS REVEALED BY PUBLIC METAGENOME SEQUENCING DATA
Presenting Author: David Banks-Richardson, University of Colorado-Denver

P07: INTER-ANNOTATOR AGREEMENT AND THE UPPER BOUND ON SYSTEM PERFORMANCE IN BIOMEDICAL AND GENERAL-DOMAIN NATURAL LANGUAGE PROCESSING
Presenting Author: Mayla Boguslav, University of Colorado School of Medicine

P08: NETWORK INFERENCE AND THE KNOWLEDGE BASE OF BIOMEDICINE
Presenting Author: Tiffany Callahan, University of Colorado Denver Anschutz Medical Campus
P09: ANTIBACTERIAL POTENTIAL OF TWO PEPTIDES DERIVED FROM A RIBOSOMAL PROTEIN FROM PYROBACULUM AERO PHILUM
Presenting Author: Elizabete Cândido, Universidade Católica de Brasília

P10: A POLY ALANINE PEPTIDE DERIVED FROM POLAR FISH WITH ANTI-INFECTIOUS ACTIVITIES
Presenting Author: Marlon Cardoso, Universidade de Brasília

P11: APPLICATION ONTOLOGIES SUPPORTING PHENOTYPING FROM CLINICAL TEXT
Presenting Author: Wendy Chapman, University of Utah

P12: AN IMAGE PHENOTYPING ENVIRONMENT BASED ON OPEN-SOURCE TOOLS
Presenting Author: Brian Chapman, University of Utah

P13: THE SNNPhenA CORPUS: AN ANNOTATED RESEARCH ABSTRACT CORPUS FOR EXTRACTING RANKED ASSOCIATION OF SINGLE-NUCLEOTIDE POLYMORPHISMS AND PHENOTYPES
Presenting Author: Hamidreza Chitsaz, Colorado State University

P14: COMPUTATIONAL DRUG DISCOVERY: AN IN SILICO AND IN VITRO EXPLORATION INTO COMBINING ESTABLISHED THERAPIES FOR TREATMENT-RESISTANT MELANOMA
Presenting Author: Brian Cicale, Stockton University

P15: VISUALIZING THE ROLE OF HORIZONTAL GENE TRANSFER WITHIN PSEUDOMONAS AERUGINOSA
Presenting Author: Evan Cudone, Loyola University, Chicago

P16: IMPROVING USER EXPERIENCE AND TOOL INTEROPERABILITY AT THE RAT GENOME DATABASE
Presenting Author: Jeff De Pons, Medical College of Wisconsin

P17: TOWARDS EFFICIENT PATIENT CARE MANAGEMENT SYSTEM FOR TERMINALLY ILL PATIENTS
Presenting Author: Avinda De Silva, Corona del Sol High School

P18: BEST PRACTICES FOR REPRODUCIBLE AND ROBUST DATA ANALYSIS IN A BIOINFORMATICS CORE FACILITY
Presenting Author: James Denvir, Marshall University

P19: PyoFuel - USING PYTHON AND PATHWAY TOOLS TO ENGINEER SYNTHETIC BIOFUEL
Presenting Author: Ashley D’Souza, Westwood High School, Austin, Texas

P20: ENHANCER REPROGRAMMING IN MAMMALIAN GENOMES
Presenting Author: Mario Flores, NIH

P21: THE FINITE STATE PROJECTION BASED FISHER INFORMATION MATRIX FOR THE DESIGN OF SINGLE-CELL EXPERIMENTS.
Presenting Author: Zachary Fox, Colorado State University

P22: UNBIASED SEQUENCE IDENTIFICATION USING MULTIPLE K-MERS
Presenting Author: Cody Glickman, University of Colorado Denver

P23: MEDICATION DATA MINING OF ELECTRONIC MEDICAL RECORDS REVEAL RACE-SPECIFIC PRESCRIPTION PATTERNS
Presenting Author: Benjamin Glicksberg, Icahn School of Medicine at Mount Sinai

P24: REPRODUCIBLE COMPUTATIONAL WORKFLOWS WITH CONTINUOUS ANALYSIS
Presenting Author: Brett Beaulieu-Jones, University of Pennsylvania
P25: INTEGRATIVE GENOMIC ANALYSIS OF CANDIDATE LONG NON-CODING RNAS ASSOCIATED WITH AUTISM
Presenting Author: Brian Gudenas, Clemson University

P26: ModEvo: A WEB-BASED TOOL FOR MODELING EVOLUTIONARY DYNAMICS
Presenting Author: Filip Jagodzinski, Western Washington University

P27: G4 QUADRUPLEXES IN AND NEAR REGULATORY ELEMENTS OF MAIZE GENES PREDICT TISSUE TYPE AND ALTERED TRANSCRIPTIONAL RESPONSE TO ABIOTIC STRESSES
Presenting Author: Mingze He, Iowa State University

P28: POPULATION-SPECIFIC DIAGNOSTIC ANALYSIS FOR IMPROVING DETECTION OF DISEASE-ASSOCIATED GENES IN TYPE 2 DIABETES
Presenting Author: Michael Hinterberg, University of Colorado

P29: MULTIMETHOD COMPUTATIONAL MODELING ANALYSIS OF SPONTANEOUS AND XENOBIOTIC-MODULATED MITOCHONDRIAL DYSFUNCTION UNDERLYING DEGENERATIVE SENESCENCE
Presenting Author: Timothy Hoffman, Colorado State University

P30: PREDICTION OF PROKARYOTIC OPTIMUM GROWTH TEMPERATURE BASED ON GENOMIC AND PROTEOMIC FEATURES
Presenting Author: Mallika Iyer, University of Colorado Denver

P31: DERIVING POPULATION-SCALE THERAPEUTIC TRAJECTORIES TO ENABLE PRECISION PHARMACOLOGY
Presenting Author: Kipp Johnson, Icahn School of Medicine at Mount Sinai

P32: KScope: A FAST MACHINE LEARNING COMPOSITION-BASED GENOMIC READ CLASSIFICATION TOOL
Presenting Author: Laurynas Kalesinskas, Loyola University Chicago

P33: A SPATIOTEMPORAL MODEL TO SIMULATE CHEMOTHERAPY REGIMENS FOR HETEROGENEOUS BLADDER CANCER METASTASES TO THE LUNG
Presenting Author: Kimberly Kanigel Winner, University of Colorado School of Medicine

P34: ScanGEO - MINING HIGH-THROUGHPUT FUNCTIONAL GENOMICS DATA
Presenting Author: Katja Koeppen, Geisel School of Medicine at Dartmouth

P35: INEXPENSIVE MOBILE DIAGNOSIS OF DIABETIC RETINOPATHY USING DEEP LEARNING
Presenting Author: Kavya Kopparapu, Thomas Jefferson High school

P36: HRC3 – A NEW CLASS OF MOTIFS INVOLVED IN CHROMATIN ORGANIZATION AND DEVELOPMENT.
Presenting Author: Andrzej Kudlicki, University of Texas Medical Branch

P37: RECONSTRUCTING PROTEIN AND GENE PHYLOGENIES BY EXTENDING THE FRAMEWORK OF RECONCILIATION
Presenting Author: Esaie Kuitche, Université de Sherbrooke

P38: A NEW ALGORITHM FOR BIOMEDICAL ARTICLE RANKING
Presenting Author: Ying Liu, St. John’s University

P39: STRATIFICATION OF PROSTATE CANCER PATIENTS BASED ON MOLECULAR INTERACTION PROFILES
Presenting Author: Roland Mathis, IBM Research

P40: PROTEOMIC ANALYSIS OF HUMAN SERUM SAMPLES TO REVEAL NEW BIOMARKERS AND MECHANISMS OF NSAID-INDUCED CARDIOVASCULAR TOXICITY
Presenting Author: Jane Mitchell, Imperial College London
P41: WITHDRAWN

P42: USING SEGMENTAL DUPLICATIONS TO ANALYZE THE ACCURACY OF TE CLASSIFICATION AND THE FREQUENCY OF GENE CONVERSION BETWEEN TE REMNANTS
Presenting Author: Gilia Patterson, University of Montana

P43: ShinyLearner: ENABLING BIOLOGISTS TO PERFORM ROBUST MACHINE-LEARNING CLASSIFICATION
Presenting Author: Stephen Piccolo, Brigham Young University

P44: DEVELOPMENT OF A DIAGNOSTIC TO PROFILE EUKARYOTIC MICROBES OF THE HUMAN MICROBIOME
Presenting Author: Ana Popovic, Hospital for Sick Children, University of Toronto

P45: IDENTIFYING NON-SPECIFIC EFFECTS OF SMALL MOLECULE TREATMENT THROUGH GSEA META-ANALYSIS
Presenting Author: Rani Powers, University of Colorado Anschutz Medical Campus

P46: WHOLE GENOME SEQUENCING AND DE NOVO ASSEMBLY FROM A CRITICALLY ENDANGERED MAMMAL, THE SUMATRAIN RHINOCEROS (DICERORHINUS SUMATRENSIS)
Presenting Author: Swanthana Rekulapally, Marshall University

P47: A NEW MOLECULAR SIGNATURE APPROACH FOR PREDICTION OF DRIVER CANCER PATHWAYS FROM TRANSCRIPTIONAL DATA
Presenting Author: Boris Reva, Icahn School of Medicine at Mount Sinai

P48: THE AFFINITY DATA BANK FOR BIOPHYSICAL ANALYSIS OF REGULATORY SEQUENCES
Presenting Author: Todd Riley, University of Massachusetts Boston

P49: MODELING HETEROGENEOUS CELL POPULATIONS USING BOOLEAN NETWORKS
Presenting Author: Brian Ross, University of Colorado

P50: A NEW APPROACH FOR PREDICTION OF MOLECULAR SIGNATURES OF OUTCOME IN CANCER
Presenting Author: Dmitry Rykunov, Icahn School of Medicine at Mount Sinai

P51: SCNIC: FINDING AND SUMMARIZING MODULES OF CORRELATED OBSERVATIONS
Presenting Author: Michael Shaffer, University of Colorado - Denver

P52: AUTOMATED OPTIMAL DESIGN OF VOLTAGE CLAMP PROTOCOLS TO STUDY SODIUM CHANNEL KINETICS USING A MINIMAL CARDIAC ION CHANNEL MODEL
Presenting Author: Matthew Shotwell, Vanderbilt University Medical Center

P53: INTELLIGENT 3D Cryo-EM IMAGE ANALYSIS FOR NEXT GENERATION BIOMEDICINE
Presenting Author: Dong Si, University of Washington Bothell

P54: NETWORKS IN SYSTEMS IMMUNOLOGY
Presenting Author: Janet Siebert, University of Colorado Denver

P55: DIFFERENTIATING BETWEEN AUTHENTIC AND CRYPTIC 5’ SPLICING SITES
Presenting Author: Kiruthika Sivaraman, San Jose State University

P56: THE RGD PhenoMiner DATABASE AND TOOL
Presenting Author: Jennifer Smith, Medical College of Wisconsin
<table>
<thead>
<tr>
<th>Poster No.</th>
<th>Title</th>
<th>Presenting Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>P57</td>
<td>WHO WANTS TO QUIT: CHARACTERISTICS AND PREDICTION OF SMOKERS INTERESTED IN QUITTING TOBACCO USE</td>
<td>Andrey Soares, University of Colorado</td>
</tr>
<tr>
<td>P58</td>
<td>ANALYSIS OF TOBACCO USERS ADMITTED TO INTENSIVE CARE UNITS</td>
<td>Andrey Soares, University of Colorado School of Medicine</td>
</tr>
<tr>
<td>P59</td>
<td>NETWORK BASED ANALYTICS FOR DOWN STREAM GENOMICS</td>
<td>Nahil Sobh, Carle R. Woess Institute for Genomic Biology</td>
</tr>
<tr>
<td>P60</td>
<td>GPU-ACCELERATED IDENTIFICATION OF DIFFERENTIAL GENETIC DEPENDENCY</td>
<td>Gil Speyer, The Translational Genomics Research Institute</td>
</tr>
<tr>
<td>P61</td>
<td>KNOWLEDGE-BASED ANALYSIS AND INTERPRETATION OF GENOME WIDE ASSOCIATION STUDIES</td>
<td>Laura Stevens, University of Colorado, Anschutz Medical Campus</td>
</tr>
<tr>
<td>P62</td>
<td>CROSS-PLATFORM NORMALIZATION ENABLES MACHINE LEARNING MODEL TRAINING ON MICROARRAY AND RNA-SEQ DATA SIMULTANEOUSLY</td>
<td>Jaclyn N Taroni, University of Pennsylvania Perelman School of Medicine</td>
</tr>
<tr>
<td>P63</td>
<td>USING KaBOB TO FIND NOVEL ADVERSE DRUG-DRUG INTERACTIONS</td>
<td>Ignacio Tripodi, University of Colorado, Boulder</td>
</tr>
<tr>
<td>P64</td>
<td>InterViewer, A NEW CYTOSCAPE-BASED VIEWER THAT DISPLAYS INTERACTIONS BETWEEN SELECTED SETS OF PROTEINS</td>
<td>Marek Tutaj, Medical College of Wisconsin</td>
</tr>
<tr>
<td>P65</td>
<td>ESTIMATING LOCAL AND REGIONAL EFFECTS ON SUBSTITUTION RATES</td>
<td>Aaron Wacholder, University of Colorado Anschutz Medical Campus</td>
</tr>
<tr>
<td>P66</td>
<td>METHODS FOR INFERRING CONSENSUS ACROSS TUMOR PHYLOGENETIC HISTORIES</td>
<td>Allie Warren, Carleton College</td>
</tr>
<tr>
<td>P67</td>
<td>INTEGRATION OF PROTEIN FAMILIES, LOCALIZATIONS, AND MODIFICATIONS IN A BIOLOGICAL KNOWLEDGE BASE</td>
<td>Elizabeth White, University of Colorado Denver, Anschutz</td>
</tr>
<tr>
<td>P68</td>
<td>DE NOVO PROTEIN STRUCTURE PREDICTION BY BIG DATA AND DEEP LEARNING</td>
<td>Sheng Wang, Toyota Technological Institute at Chicago</td>
</tr>
<tr>
<td>P69</td>
<td>RANK AGGREGATION FOR FEATURE SCORING AND SELECTION</td>
<td>Tara Yankee, University of Connecticut</td>
</tr>
<tr>
<td>P70</td>
<td>BOOTSTRAPPING ESTIMATES OF STABILITY FOR CLUSTERS, OBSERVATIONS AND MODEL SELECTION</td>
<td>Han Yu, University at Buffalo</td>
</tr>
<tr>
<td>P71</td>
<td>THE HETNET AWAKENS AT <a href="HTTPS://NEO4J.HET.IO">HTTPS://NEO4J.HET.IO</a></td>
<td>Daniel Himmelstein, University of Pennsylvania</td>
</tr>
</tbody>
</table>
P01: CAMSA: A TOOL FOR COMPARATIVE ANALYSIS AND MERGING OF SCAFFOLD ASSEMBLIES

Subject: Graphics and user interfaces

Presenting Author: Max Alekseyev, George Washington University

Co-Author(s): Sergey Aganezov, George Washington University, United States

ABSTRACT: Motivation: Despite the recent progress in genome sequencing and assembly, many of the currently available assembled genomes come in a draft form. Such draft genomes consist of a large number of genomic fragments (scaffolds), whose positions and orientations along the genome are unknown. While there exists a number of methods for reconstruction of the genome from its scaffolds, utilizing various computational and wet-lab techniques, they often can produce only partial error-prone scaffold assemblies. It therefore becomes important to compare and merge scaffold assemblies produced by different methods, thus combining their advantages and highlighting present conflicts for further investigation. These tasks may be labor intensive if performed manually.

Results: We present CAMSA—a tool for comparative analysis and merging of two or more given scaffold assemblies. The tool (i) creates an extensive report with several comparative quality metrics; (ii) constructs a most consistent combined scaffold assembly; and (iii) provides an interactive framework for a visual comparative analysis of the given assemblies.

Availability: CAMSA is available for download from http://cblab.org/camsa/

P02: IDENTIFYING THE MECHANISM FOR THE METASTATIC SPREAD OF BREAST CANCER THROUGH INTEGRATION OF GENE EXPRESSION, WHOLE GENOME SEQUENCING AND FUNCTIONAL SCREENS

Subject: Metagenomics

Presenting Author: Eran Andrechek, Michigan State University

ABSTRACT: Breast cancer mortality is usually caused by metastasis to distant sites. Using genomic signatures to predict cell signaling pathway activation has allowed us to develop hypotheses about key signaling pathways that are involved in the metastatic progression of breast cancer. To test the hypothesis that the E2F transcription factors are involved in metastasis, we generated a mouse model of breast cancer lacking E2F1 or E2F2. Consistent with our hypothesis, these mice developed breast cancer lacking metastasis. The E2F family of transcription factors is classically known to regulate G1 to S-phase
transition in cell cycle, however, other functions have emerged. To address the function of the E2Fs in metastasis, gene expression of tumors from wild type and E2F knockout backgrounds were analyzed. This was integrated with whole genome sequence data from matched tumor samples. Potential genetic mechanisms identified through this approach were validated for human relevance using TCGA data. Patient outcomes were screened for these genes through the application of a predictive gene expression signature. Further, the Achilles project and a drug screening study in patient derived xenograft tumors were two functional screens that were also integrated with this work. The outcome of the integrated study was the identification of an amplification event in breast cancer that correlates with metastasis. Genetic ablation of genes in this amplicon revealed specific roles in metastatic migration. Together this work demonstrates the utility of integrating multiple data platforms to address key biological problems.

**P03: IndeCut: A CUT-NORM BASED METHOD FOR EVALUATING INDEPENDENT AND UNIFORM SAMPLING IN NETWORK MOTIF DISCOVERY ALGORITHMS**

**Subject:** Graph Theory

**Presenting Author:** Mitra Ansariola, Oregon State University

**Co-Author(s):** David Koslicki, Oregon State University, United States; Molly Megrew, Oregon State University, United States

**ABSTRACT:** Network motif discovery is a well-established general statistical strategy for identifying over-represented sub-network structures within a larger network. In the biosciences, it serves as a prominent conceptual tool that enables scientists to recognize biologically important patterns and generate testable hypotheses within large genetic networks of interest. Network motif discovery algorithms function by comparing the frequency of particular sub-network of interest within a given ‘real-world’ network to its frequency in a large collection of randomized networks. While the method of randomization may differ, all algorithms face the challenge of how to sample uniformly and independently from the set of all possible randomized networks that may be generated. Though several network motif discovery tools with different underlying random sampling strategies are available, scientists who want to apply these tools on their own networks of interest currently do not have any method by which to assess whether any tool will provide an accurate outcome. Most users will not be able to test the correctness of detected motifs in the laboratory due to prohibitive cost, so it is essential to have access to such an evaluation method. In this talk, we present IndeCut, the very first method that numerically determines the degree of sampling uniformity and independence of network motif discovery algorithms. IndeCut is the first and only method to date that allows characterization of network motif finding algorithm performance beyond computational efficiency.
**ABSTRACT**: Diverse algorithms and methods are needed to answer the ever increasing need of adequately harnessing Mass Spectrometer generated data. The unique nature and structure of mass spectra data usually, requires a high level of expertise and rigorous algorithms. This study’s methodology discusses feature selection based on direct and simple mathematical observations of variables and their inter-relationships, Jackknife technique for data re-sampling, matrix to vector decomposition and successfully classifies Alzheimer’s disease patients into three disease stages; age-matched controls without any evidence of dementia, patients with mild cognitive impairment and patients with clinical symptoms of Alzheimer’s disease (AD), using a 2-scale principle of K-nearest neighbor (KNN) algorithm on SELDI data and without collaborating clinical records. Hitherto, there exists no clinical diagnostic tool for AD, in lieu of this, patient cognitive abilities are clinically followed-up over a period of time (may be months) to make a diagnosis. This practice usually leads to inconclusive diagnosis and results obtained from it are not generalizable. Our model provides an immediate classification and correctly classifies test data sets with 82% confidence. It can also identify traces of positive/negative change within and across data sets in regards to severity of the disease over time.
amplicon studies may be obscuring large portions of Bathyarchaeota phylogeny. Shotgun metagenomic sequencing has the potential to shed light onto Bathyarchaeota habitat preferences, especially for the portions of the clade that PCR-primer bias may be hiding, but shotgun assembly is often incomplete or lacks the context of existing 16S phylogeny. Here, we employ a targeted gene assembly approach (EMIRGE) to reconstruct 16S rRNA sequences from publically available shotgun metagenomes representing several environmental and host-associated habitats. We quantify the degree to which PCR-primer bias is obscuring the diversity of the Bathyarchaeota, build a cross-domain co-occurrence network between members of Bathyarchaeota and other microorganisms, and identify association patterns between environmental variables and the Bathyarchaeota. Preliminary results suggest that members of this phylum may co-occur with members of the bacterial phyla Proteobacteria, Planctomycetes, and OP1, and that sub-clades within this group respond differentially to depth gradients in estuarine sediments. This study informs future work seeking to characterize the roles these broadly distributed archaea play in microbial communities across the globe.

P07: INTER-ANNOTATOR AGREEMENT AND THE UPPER BOUND ON SYSTEM PERFORMANCE IN BIOMEDICAL AND GENERAL-DOMAIN NATURAL LANGUAGE PROCESSING

Subject: Text Mining

Presenting Author: Mayla Boguslav, University of Colorado School of Medicine

Co-Author(s): Kevin Cohen, University of Colorado School of Medicine, United States

ABSTRACT: In natural language processing in general and machine learning in particular, we often use data that has been labelled by humans (annotators) with the correct answers. For various reasons, we often compute the agreement between annotators — if two annotators look at the same texts, how often do they agree about its classification? It’s often thought the agreement between annotators is the upper limit on how well a system can perform: if humans can’t agree with each other about the classification more than some percentage of the time, then it’s not reasonable to expect a computer to do any better. We trace the logical positivist roots of the motivation for measuring inter-annotator agreement, show what happens when we try to trace the origins of the widely-held belief about the relationship between inter-annotator agreement and system performance, and then present data that suggests that inter-annotator agreement is not in fact an upper bound on system performance in natural language processing, with evidence from both the biomedical and the general domains.
Subject: Machine learning, inference and pattern discovery

Presenting Author: Tiffany Callahan, University of Colorado Denver Anschutz Medical Campus

Co-Author(s): William A. Baumgartner Jr, University of Colorado Denver Anschutz Medical Campus, United States; Marc Daya, University of Colorado Denver Anschutz Medical Campus, United States; Lawrence E. Hunter, University of Colorado Denver Anschutz Medical Campus, United States

ABSTRACT: Structural transformation of biological knowledge represented using Semantic Web standards significantly improves the utility of visualization tools and network analytics. Link prediction algorithms are powerful tools for predicting unobserved connections between nodes in a network. The application of such algorithms to biological networks has lead to the correct prediction of previously unobserved relationships ranging from protein-protein interactions to novel P53 kinases. The use of such algorithms to analyze larger and more complex representations has the potential to generate novel and important hypotheses, and insights into biological mechanisms. Unfortunately, the direct application of these algorithms to biological knowledge is limited by the representational complexity of the web ontology language standard OWL. The Network Information Content Entity (NICE) approach, a novel transformation method, reversibly transforms OWL-compliant biomedical knowledge into a representation better suited for visualization and network inference algorithms. Using several illustrative biomedical queries, the NICE transformation produces simpler network representations that are more visually comprehensible and whose structural properties (e.g. clustering coefficient, modularity, number of shortest paths, number of average neighbors, and diameter and radius) are significantly improved over raw OWL. Furthermore, comparison of the results from the application of several state-of-the-art link prediction algorithms on raw OWL versus NICE networks shows that the NICE transformation results in more accurate and biologically meaningful predictions. For each query and each algorithm, the top-ten predicted links for both OWL and NICE networks were validated via evidence from literature review and domain expert consultation.
**P09: ANTIBACTERIAL POTENTIAL OF TWO PEPTIDES DERIVED FROM A RIBOSOMAL PROTEIN FROM PYROBACULUM AEROPHILUM**

Subject: Qualitative modeling and simulation

Presenting Author: Elizabete Cândido, Universidade Católica de Brasília

Co-Author(s): Marlon Henrique Cardoso, Universidade de Brasília, Brazil; Karen Oshiro, Universidade Católica Dom Bosco, Brazil; Suzana Ribeiro, Universidade Católica Dom Bosco, Brazil; Diego Nolasco, Universidade Católica de Brasília, Brazil; William Porto, Universidade Católica de Brasília, Brazil; Octávio Luiz Franco, Universidade Católica de Brasília, Brazil

**ABSTRACT:** Antimicrobial peptides have emerged as promising antimicrobial molecules, being prospected by several methods including the screening for potential antimicrobial sequences within proteins already described. Here we focused on the functional/structural characterization of two novel peptides derived from a Pyrobaculum aerophilum bacterial ribosomal protein. Protein sequences from the non-redundant database were submitted to antimicrobial predictors, where we could identify an 18-amino acid residue fragment from P. aerophilum. This fragment was used as template for the generation of nine analogues by using the JOKER algorithm based on a pattern of P(K)2LA. Among the generated analogues, the third one, PaAMP1R3, was the most active against Pseudomonas aeruginosa. Furthermore, a sliding window of 10 amino acid residues was applied to PaAMP1R3, where the tenth peptide (PaAMP1R3F10) was selected for further analysis due to its higher antibacterial potential. Both peptides were synthetized by Fmoc and analyzed on MALDI-ToF, revealing two ions of 2296.4 and 1244.9 Da for PaAMP1R3 and PaAMP1R3F10, respectively. Antibacterial activities were accessed, where PaAMP1R3 showed minimum inhibitory concentrations (MICs) ranging from 4 to 32 μg.mL-1 against resistant/susceptible Escherichia coli strains and susceptible Klebsiella pneumoniae, Enterococcus faecalis and P. aeruginosa. PaAMP1R3F10 revealed MICs between 8 to 64 μg.mL-1 against the same strains. In addition, both peptides could completely inhibit methicillin-resistant Staphylococcus aureus strains at 64 μg.mL-1. Molecular dynamics simulations were performed during 200 ns in water, using the CHARMM 27 force field. PaAMP1R3 model presented a stable α-helical structure in hydrophilic environments, while a random coil was observed for PaAMP1R3F10.
P10: A POLYALANINE PEPTIDE DERIVED FROM POLAR FISH WITH ANTI-INFECTIOUS ACTIVITIES

Subject: Qualitative modeling and simulation

Presenting Author: Marlon Cardoso, Universidade de Brasília

Co-Author(s): Suzana Ribeiro, Universidade Católica Dom Bosco, Brazil; Diego Nolasco, Universidade Católica de Brasília, Brazil; César de la Fuente-Núñez, Massachusetts Institute of Technology, United States; Mário Felício, Universidade de Lisboa, Portugal; Sónia Gonçalves, Universidade de Lisboa, Portugal; Carolina Matos, Universidade Federal de Goiás, Brazil; Luciano Liao, Universidade Federal de Goiás, Brazil; Nuno Santos, Universidade de Lisboa, Portugal; Robert Hancock, University of British Columbia, Canada; Ludovico Migliolo, Universidade Católica Dom Bosco, Brazil; Octávio Franco, Universidade Católica de Brasília, Brazil

ABSTRACT: Due to a growing concern about bacterial infections, increasing supports has been given to drug discovery. As a promising alternative, the antimicrobial peptides (AMP) have appeared, being mother molecules for rational design strategies. Here we focused on the structural and functional characterization of Pa-MAP 1.9, an AMP rationally designed based on a peptide derived from Pleuronectes americanus. Pa-MAP 1.9 was synthetized by Fmoc and further analyzed by MALDI-ToF, revealing a 2668.0 Da peptide. Antibacterial and anti-biofilm assays showed that Pa-MAP 1.9 could inhibit Enterococcus faecalis, Escherichia coli and Klebsiella pneumoniae growth from 6 to 96 μM. In addition, Pa-MAP 1.9 could also prevent E. coli and K. pneumoniae biofilm formation, as well as eradicate them at 3.0 and 1.1 μM, respectively. Atomic force microscopy (AFM) was also conducted, revealing that Pa-MAP 1.9 did not cause morphological damages on E. coli. Otherwise, at 50-fold higher doses it could be observed membrane destabilization. None cytotoxic (RAW 267.4) and hemolytic (human erythrocytes) activities were reported at 115 μM. Leakage assays and molecular docking simulations showed that Pa-MAP 1.9 could interact with higher specificity to anionic membranes and vesicles mimicking Gram-negative bacteria. Circular dichroism (CD) and computational simulations allowed characterizing the secondary structure of this peptide in hydrophobic (TFE 50% v:v), hydrophilic (water) and anionic (SDS) environments. As result, CD spectra and molecular dynamics, using the GROMOS96 43a1 force field, revealed that Pa-MAP 1.9 is a linear, amphipathic peptide presenting high helical contents when in hydrophobic and anionic environments characteristics from Gram-negative bacteria.

P11: APPLICATION ONTOLOGIES SUPPORTING PHENOTYPING FROM CLINICAL TEXT

Subject: Data management methods and systems

Presenting Author: Wendy Chapman, University of Utah

ABSTRACT: Representation of the knowledge described in clinical reports is critical to accurate phenotyping of patients. We have developed two
application ontologies for modeling annotations of clinical reports: the schema ontology describes the clinical entities that are described in reports, such as findings, procedures, and medications. The modifier ontology enumerates the allowable modifiers for those entities with three types of modifiers: shared modifiers that apply to all entities: negation, uncertainty, and temporality; semantic modifiers specific to particular entities, such as dose and route for medications; and numeric modifiers for specifying numeric values such as body temperature. A user can create a domain ontology by creating instances of entity-modifier combinations, accommodating rich phenotypic representation for concepts like no family history of colon cancer or severe carotid stenosis in the right internal carotid artery. In addition to modeling the semantic composition of the concepts, the ontologies provide value sets and lexical variants that can be customized and enhanced. Our long-term goal is to create shareable libraries of domain ontologies.

In addition to supporting annotation of concept mentions, swirl rules stored in the ontology support inferencing over mention annotations for classification at the document, encounter, and phenotypic/patient level. The ontologies support rich phenotypic characterizations to go beyond binary phenotypes toward answering questions like “what histologic types of breast cancer are associated with patients that have a substitution mutation on BRCA-1?” and “for patients with a papillary breast carcinoma that underwent neoadjuvant treatment regimen, what number of patients have had a recurrence or metastasis?”

**P12: AN IMAGE PHENOTYPING ENVIRONMENT BASED ON OPEN-SOURCE TOOLS**

Subject: *Data management methods and systems*

Presenting Author: *Brian Chapman, University of Utah*

Co-Author(s): *John Roberts, University of Utah, United States*

**ABSTRACT:** Medical imaging data are an often-overlooked resource for defining patient phenotypes. Because images data are unstructured, in order to extract information from the images requires creating pipelines for identifying relevant studies, segmenting and quantifying features from the images, and linking these features to other data sources (e.g. the EHR). We are building an image phenotyping environment based on open-source deployed using Docker (https://www.docker.com/), allowing us to version-control our environments, which are defined with simple text files. Our phenotyping pipeline is built using three open-source projects: 1) Orthanc (http://www.orthanc-server.com/), a light-weight DICOM server for communicating with the clinical PACS and scrubbing images for research purposes. Orthanc allows for persistent, customized scrubbing processes. 2) Girder (https://girder.readthedocs.io/en/latest/), an open-source, web-based data management system developed by Kitware, Inc. Girder provides user...
authentication, access control and a framework for linking data and defining meta-data. We have integrated Girder with bioportal so that data uploads are tagged with concepts from relevant ontologies. 3) JupyterHub for providing web-based computational environments. JupyterHub provides Docker containers serving up Jupyter notebooks. Jupyter notebooks allow for programming through the web browser and supports a number of languages including Python and a number of other languages. Jupyter notebooks contain image processing pipelines for extracting features from medical images using SimpleITK and other software packages. Our initial use-cases are drawn from dermatology and radiology and require both 2D and 3D feature extraction tasks.

**P13: THE SNPPhenA CORPUS: AN ANNOTATED RESEARCH ABSTRACT CORPUS FOR EXTRACTING RANKED ASSOCIATION OF SINGLE-NUCLEOTIDE POLYMORPHISMS AND PHENOTYPES**

Subject:

Presenting Author: Hamidreza Chitsaz, Colorado state university

Co-Author(s): Behrouz Bokharaeian, Complutense University of Madrid, Spain; Alberto Diaz, Complutense University of Madrid, Spain; Ramyar Chavoshinejad, Royan Institute for Reproductive Biomedicine, Iran

**ABSTRACT:** Single Nucleotide Polymorphisms (SNPs) are the most comprehensively studied type of genetic variations that influence a number of diseases and phenotypes. Recently, some corpora and methods have been developed for extracting SNPs, diseases, and their associations from scientific text. However, there is no available method and corpus for extracting those SNP-disease associations that have been annotated with linguistic based negation, modality markers, neutral candidates, and the level of confidence of association.

Method: In this research, we present different steps for producing the SNPPhenA corpus. They include automatic Named Entity Recognition (NER) followed by the manual annotation of SNP and phenotype names, annotating the SNP-phenotype associations and their level of confidence, as well as modality markers. Moreover, the produced corpus has been annotated with negation scopes and cues as well as neutral candidates that have an important role in dealing with negation and the modality phenomenon in relation extraction tasks.

Result: The agreements between annotators were measured by Cohen’s Kappa coefficient and the resulting scores showed reliability of the corpus. The Kappa score was 0.86 for annotating the associations and 0.80 for annotating the degree of confidence of associations. Additionally, basic statistics for extracting ranked SNP-phenotype associations are presented here, with regard to the annotated features.
of the corpus besides the results of our first experiments. Moreover, we prepared guidelines for using the corpus. The guidelines and the corpus are available at http://nil.fdi.ucm.es/?q=node/639.

Conclusion: Estimating confidence of SNP-phenotype associations could help determine phenotypic plasticity and the importance of environmental factors. Moreover, our first experiments with the corpus show that linguistic-based confidence alongside other non-linguistic features can be utilized to estimate strength of the observed SNP-phenotype association. Trial Registration: Not Applicable

P14: COMPUTATIONAL DRUG DISCOVERY: AN IN SILICO AND IN VITRO EXPLORATION INTO COMBINING ESTABLISHED THERAPIES FOR TREATMENT-RESISTANT MELANOMA

Subject: Simulation and numeric computing

Presenting Author: Brian Cicali, Stockton University

Co-Author(s): Robert Olsen, Stockton University, United States

ABSTRACT: Melanoma is the most deadly form of skin cancer, killing more than 10,000 people annually in the United States. Although multiple FDA-approved therapies exist, melanoma is still a very serious form of cancer. This project centers on the computational modeling for potential combination melanoma therapies. The therapies examined in this project are drugs that target proteins within the mitogen-activated protein kinase (MAPK) pathway. The MAPK pathway is involved in many cellular functions, and mutations in this pathway are associated with melanoma as well as other forms of cancer such as lung and breast cancer. This particular project focuses on a computational examination of the potential synergistic effects of combining melanoma therapies that target a portion of the MAPK pathway. To perform this work, a model of the RAS/B-RAF/MEK/ERK portion of the MAPK signaling pathway was constructed using PySB, a Python-based software for modeling networks of biochemical reactions. Inhibitory drug pathways were added that represent therapies centered on four cancer drugs, dabrafenib, vemurafenib, trametinib, and binimetinib. Results of simulations of the model, with the drug pathways both deactivated and activated, were analyzed for significance. To validate the results of the computational model, a cellular study was performed to measure the effect these drugs have on MEK phosphorylation in melanoma cells. These results give a deeper look into the efficacy of combining melanoma therapies, as well as demonstrate the application of computational modeling to the field of drug discovery.
P15: VISUALIZING THE ROLE OF HORIZONTAL GENE TRANSFER WITHIN PSEUDOMONAS AERUGINOSA

Subject: Metagenomics
Presenting Author: Evan Cudone, Loyola University, Chicago

ABSTRACT: Metagenomics has uncovered the complexities of microbial communities in various environmental niches on Earth. While horizontal gene transfer (HGT), the exchange of genes between unrelated organisms, can vary between environments, it has nevertheless been shown to play a significant role in the dynamics of complex communities in addition to an important role in prokaryotic evolution. Thus, the ability to readily identify horizontally acquired elements as well as their putative sources can profoundly advance our understanding of both the interactions between the host and pathogen, as well as interactions between microbes within the host microbiota. Furthermore, identifying horizontally acquired genes and genome rearrangements is of particular importance for clinical isolates, as HGT of antibiotic-resistance and virulence genes is of critical concern. We recently developed a tool, SPlot2.0, for the expedient analysis of genomic sequences (partial or complete). The tool creates an interactive, two-dimensional heat map capturing the similarities and dissimilarities in nucleotide usage at various levels both within and between genomic sequences. Exogenous sequences acquired via HGT can thus be easily identified and further examined for their source(s). Using this tool, we have conducted an extensive analysis of Pseudomonas aeruginosa genomes. Prior evidence has shown that HGT is a key factor in the genetic diversity of this medically important organism. Through our analysis of all publicly available P. aeruginosa genomes, we can visualize and capture the evolution of this species, both strains from the environment and clinical isolates.

P16: IMPROVING USER EXPERIENCE AND TOOL INTEROPERABILITY AT THE RAT GENOME DATABASE

Subject: Graphics and user interfaces
Presenting Author: Jeff De Pons, Medical College of Wisconsin

Co-Author(s): Jennifer Smith, Medical College of Wisconsin, United States; Stan Laulederkind, Medical College of Wisconsin, United States; G Thomas Hayman, Medical College of Wisconsin, United States; Victoria Petri, Medical College of Wisconsin, United States; Shur-Jen Wang, Medical College of Wisconsin, United States; Jyothi Thota, Medical College of Wisconsin, United States; Marek Tutaj, Medical College of Wisconsin, United States; Melinda Dwinell, Medical College of Wisconsin, United States; Mary Shimoyama, Medical College of Wisconsin, United States

ABSTRACT: The Rat Genome Database (RGD, http://rgd.mcw.edu), the premier online resource for rat genetic, genomic and phenotypic data, offers a large body of cross-species functional, phenotype and disease data and multiple innovative software tools to assist in analysis. As the
number of analysis tools and datasets at RGD increases, it has become more important to find ways to educate users as to what is available. In addition, tool interoperability and a consistent and recognizable interface across disparate sections of the site are important to enhance user experience. To address this, RGD has implemented a context sensitive dynamic interface that allows for interoperability between tools and gene lists. After finding a gene list of interest, selecting the tools icon renders a navigation window listing of all analysis options available for the gene set. Selecting a tool then submits your gene list to the analysis tool selected. The interface is consistent, recognizable, and allows users to navigate seamlessly amongst the many tools available at RGD. Available analysis options include protein-protein interactions, functional annotation for selected species and orthologs, annotation distribution and comparison, strain variation for sequenced strains, variant damage predictions, OLGA (Object List Generator and Analyzer) integration, and the ability to download gene lists.

P17: TOWARDS EFFICIENT PATIENT CARE MANAGEMENT SYSTEM FOR TERMINALLY ILL PATIENTS

Subject: Networking, web services, remote applications
Presenting Author: Avinda De Silva, Corona del Sol High School

ABSTRACT: Terminally ill patients experience various conditions such as nausea, depressions, and pain. Addressing these issues as fast as possible will help patients to obtain some relief faster and reduce number of emergency care visits and hospitalizations. In this work, we are proposing a mobile app based system to enhance patient, doctor communication and improve medical treatments of these patients. The system consists of a patient profile where patients or the caregiver for the patient can enter patient’s condition to the doctor daily basis or the interval determined by the doctor. Once data is entered, the data will be analyzed and the condition of the patient will be notified to the doctor. Doctor’s profile enables doctors to receive notifications and take appropriate medical decisions based on the patient’s condition. Usually the use of the pain scales is relative to individual patient. Therefore, a baseline must be made for each individual person. In this work, we use extended Edmonton pain scale. This comparative pain scale, will help doctors to determine pain relativity of patients and give a more accurate pain tolerance.

Our mobile app is kind of a “click-and-choose system” where user data entry is streamlined. The app has been designed and implemented to make it more usable. Aesthetically speaking, everything should be big, clear and straightforward.

*This project was initially proposed by Dr. Lipinski at Mayo clinic. Also, Jarrett M. Wilkes, David Ganey, and Lelan Dao at Arizona State University worked on different aspects of this project."
P18: BEST PRACTICES FOR REPRODUCIBLE AND ROBUST DATA ANALYSIS IN A BIOINFORMATICS CORE FACILITY

Subject: Data management methods and systems
Presenting Author: James Denvir, Marshall University
Co-Author(s): Don Primerano, Marshall University, United States; Jun Fan, Marshall University, United States; Swanthana Rekulapally, Marshall University, United States

Abstract: With the publication of standards for Minimum Information About a Microarray Experiment in 2001, and the subsequent establishment of global repositories for gene expression and sequencing data, the research community has substantial achievements in making research data associated with published, peer reviewed manuscripts available for reuse and evaluation. However, there are currently no standards for the amount of detail of the analysis performed that should be provided in a publication. Consequently, it is rare to find publications for which the data analysis pipeline has sufficiently detailed description for the analysis to be reproduced, or in some cases critically evaluated.

We adopted simple practices used in software engineering, including version control management, self-documenting code, and convention over configuration techniques into the data analysis pipelines used in a small genomics and bioinformatics core facility. Adoption of these techniques both improved the ability of our facility to create reproducible pipelines, and enhanced operational efficiency.

P19: PyoFuel — USING PYTHON AND PATHWAY TOOLS TO ENGINEER SYNTHETIC BIOFUEL

Subject: Simulation and numeric computing
Presenting Author: Ashley D’Souza, Westwood High School, Austin, Texas

Abstract: Pathway Tools is a collection of biological modeling tools with databases of organism models, metabolic flux analysis, and query and visualization tools. Pythoncyc is a Python programming interface to Pathway Tools. In two earlier projects I had experimented with flux balance analysis on models of bacteria that had been modified with pathways to synthesize biofuel, and with wet-lab recombineering of the DHX35 gene using E.coli. The former was quick, easy, and fun; the latter was slow, painful, and fun. So I wanted to use Python to script Pathway Tools, to help find candidate biofuel pathways across organisms, identify the corresponding gene-edits to engineer biofuel-friendly E.coli, and evaluate how effective each engineered organism might be -- all as a precursor to either more detailed modeling or wet-lab work.

PyoFuel is the resulting project. It is ongoing work, and my poster will report on the following using flowcharts, relevant Pythoncyc API calls, PyoFuel code snippets, and Pathway visualizations:

• Find candidate biofuel metabolites in MetaCyc, a multi-organism database
• Identify the pathways that produce those metabolites
• Generate a modified organism database to evaluate via FBA
• Run MetaFlux on the modified organism with suitable objectives and constraints
• Filter out those organisms if key flux numbers are poor
• Identify enzymes and corresponding genes for the modified pathways

I am currently a senior in high school. If accepted, I plan to open-source the current Jupyter notebook and pgdb databases. My info is at http://ashdza.github.io/.

**P20: ENHANCER REPROGRAMMING IN MAMMALIAN GENOMES**

Subject: *Simulation and numeric computing*

Presenting Author: *Mario Flores, NIH*

**ABSTRACT:** It has been shown that changes in regulatory regions (enhancers) have supported evolution in mammals. However there is still a lack of knowledge about the distinct types of enhancers, their identification in more tissues/cell types and the mechanisms that act to modify these regulatory regions during evolution. Here we study a type of enhancers that we have named reprogrammed enhancers. Enhancer reprogramming establish that changes in the transcription factor binding sites of noncoding regulatory DNA sequences could potentially change their regulatory function. In this context, TFBSs loss, gain and reshuffling within an enhancer can change its function (spatial and/or temporal regulatory activity). We have identified reprogrammed enhancers in 11 tissues/cell types in human and mouse. We estimate that in average 30% of the total number of enhancers in a gene locus had been reprogrammed in the course of evolution. Furthermore the analysis of DNA sequence changes underlying enhancer reprogramming shows a change in the transcription factor binding site (TFBS) composition that significantly overlaps with the TFBS composition of tissue specific enhancers. Our observations provide evidence that reprogrammed enhancers are important contributors of the shaping of the regulatory landscape during evolution.

This research is supported by the Intramural Research Program of the NIH, National Library of Medicine
P21: THE FINITE STATE PROJECTION BASED FISHER INFORMATION MATRIX FOR THE DESIGN OF SINGLE-CELL EXPERIMENTS

Subject: Simulation and numeric computing
Presenting Author: Zachary Fox, Colorado State University
Co-Author(s): Brian Munsky, Colorado State University, United States

ABSTRACT: Measuring and understanding gene expression fluctuations is key to predicting and controlling gene regulation dynamics. Rapidly advancing experiments enable precise quantification of RNA and protein in single cells. However, to keep pace with expanding experimental capabilities, computational and theoretical approaches must also improve. If tightly coupled with experiments, computational analyses can extract improved insight from previous measurements and enhance the effectiveness of future experiments. The Fisher Information Matrix (FIM) is a tool that is often used to aid experiment design for engineering applications, but common FIM approaches focus on deterministic models and cannot capture the full information contained in stochastic single-cell distributions. Such distributions are known to be well captured by the chemical master equation (CME). However, the CME is frequently too difficult or impossible to solve, which precludes rigorous computation of the FIM. The finite-state projection (FSP) approach systematically reduces the CME to a finite, solvable set of ordinary differential equations. In this study, we extend the FSP to compute the FSP-FIM and estimate the expected information for potential single-cell experiments. In contrast to existing experiment design strategies, our FSP-FIM approach makes no assumptions about the underlying distributions. We demonstrate the advantage of the FSP-FIM approach on several common models of stochastic gene expression, for which previous approaches and assumptions of normal distributions are not justified. Our results allow for the computational exploration of many potential experiments, and can promote iterative and efficient integration of modeling and experimentation to understand, predict and control gene expression.

P22: UNBIASED SEQUENCE IDENTIFICATION USING MULTIPLE K-MERS

Subject: Metagenomics
Presenting Author: Cody Glickman, University of Colorado Denver

ABSTRACT: Metagenomic sequencing has transformed the understanding of the role a microbial community plays in human health. The crux of metagenomic studies is proper identification of microbial organisms or functional genes in a sample. The accuracy of taxonomic and functional annotation is correlated with the length of the sequence. Current sequencing technology produces short reads,
which are commonly assembled to form longer contiguous sequences or contigs. The assembly of contiguous sequences can produce misassemblies known as chimeras. One way to reduce the formation of chimeras and increase the accuracy of calls against a database is to perform sequence filtering to remove contamination. Sequence filtration methods include mapping reads to known genomes and referencing sequences against a genetic database. The issue with both processes is the reliance on the completeness of extant databases to retain or discard reads. Here, we propose an unbiased metagenomic sequence identification model using a multiple k-mer approach. Our approach explores the feasibility of directly using the k-mer composition of metagenomic reads to classify the sequence origin as that of bacterial or viral. By utilizing information stored within the reads themselves, we avoid relying on the completeness of extant databases to perform filtering. We test our model against sheared sequences from extant databases and against a randomly generated null sequence model. The unbiased filtration methodology presented is capable of expansion into areas such as bacterial or viral functional metagenomics, where the presence of one conflates the functional observations of the other.

P23: MEDICATION DATA MINING OF ELECTRONIC MEDICAL RECORDS REVEAL RACE-SPECIFIC PRESCRIPTION PATTERNS

Subject: Machine learning, inference and pattern discovery

Presenting Author: Benjamin Glicksberg, Icahn School of Medicine at Mount Sinai

Co-Author(s): Kipp Johnson, Icahn School of Medicine at Mount Sinai, United States; Khader Shameer, Icahn School of Medicine at Mount Sinai, United States; Joel Dudley, Icahn School of Medicine at Mount Sinai, United States

ABSTRACT: Introduction: Disparities in medication availability, tolerability, and effectiveness exist and patient outcomes. We aimed to mine electronic medical records (EMR) and quantify differences in medication counts, prescription-record counts, and drug-class enrichment using the New York Metropolitan area population compiled from Mount Sinai Data Warehouse.

Methods: Self-reported ancestry was abstracted from EMR (n=2.1 million) as Caucasian (EA), African-American (AA), Hispanic/Latino (HL), Asian (A), or Other (O). Medications were normalized with RxNorm and mapped to Anatomical Therapeutic Chemical (ATC) drug-classes using the PharmaFactors software framework.

Results: We found differences in prescription and unique medication count between races (one-way ANOVA, p<5E-16 for both). AA individuals had more prescription instances and unique medications compared to all other racial groups (Tukey HSD, p<10-16, all comparisons). Conversely, HL individuals had the fewest prescription instances and unique medications compared to all other groups (Tukey
HSD, p<10-16, all comparisons). Polypharmacy (4+ simultaneous drug prescriptions) varied according to race (\(\chi^2\) p<10-16), EA having the highest rates (0.58) and AA the lowest (0.43). ATC drug-class enrichment varied with race: of 473 level 4 ATC classes, we found 125 and 70 enriched for EA and AA respectively (Fisher’s Exact Q<0.05, OR>1). The most enriched classes per group were EA, joint muscle pain and bowel disorders (OR=8.73 for both); AA, antiseptics (OR=8.38); HL, thiazolidinediones (OR=1.14); and A, Nucleoside/nucleotide reverse transcriptase inhibitors (OR=7.42).

Conclusion: We identified various ancestry-specific prescription data patterns. Further investigation of these patterns may help to develop prescription practices and improve therapeutic outcomes by optimizing drug efficacy and lowering side effects.

**P24: REPRODUCIBLE COMPUTATIONAL WORKFLOWS WITH CONTINUOUS ANALYSIS**

Subject: *Data management methods and systems*

Presenting Author: Brett Beaulieu-Jones, University of Pennsylvania

Co-Author(s): Casey Greene, University of Pennsylvania, United States

**ABSTRACT:** Reproducing experiments is vital to science. Being able to replicate, validate and extend previous work also speeds new research projects. Reproducing computational biology experiments, which are scripted, should be straightforward. But reproducing such work remains challenging and time consuming. In the ideal world we would be able to quickly and easily rewind to the precise computing environment where results were generated. We would then be able to reproduce the original analysis or perform new analyses. We introduce a process termed “continuous analysis” which provides inherent reproducibility to computational research at a minimal cost to the researcher.

Continuous analysis combines Docker, a container service similar to virtual machines, with continuous integration, a popular software development technique, to automatically re-run computational analysis whenever relevant changes are made to the source code. This allows results to be reproduced quickly, accurately and without needing to contact the original authors. Continuous analysis also provides an audit trail for analyses that use data with sharing restrictions. This allows reviewers, editors, and readers to verify reproducibility without manually downloading and rerunning any code.
P25: INTEGRATIVE GENOMIC ANALYSIS OF CANDIDATE LONG NON-CODING RNAS ASSOCIATED WITH AUTISM

Subject: Machine learning, inference and pattern discovery
Presenting Author: Brian Gudenas, Clemson University
Co-Author(s): Liangjiang Wang, Clemson University, United States; Anand Srivastava, Greenwood Genetic Center, United States

ABSTRACT: Genetic studies have identified many risk loci for autism spectrum disorder (ASD) although causal factors in the majority of cases are still unknown. Currently, known ASD risk genes are all protein-coding genes; however, the vast majority of transcripts in humans are non-coding RNAs (ncRNAs) which do not encode proteins. Recently, long non-coding RNAs (lncRNAs) were shown to be enriched in the human brain and be crucial for normal brain development. A major functional theme of lncRNAs is to regulate the gene expression of other genes through transcriptional, post-transcriptional and epigenetic mechanisms. LncRNAs affected by mutations could cause abnormal lncRNA expression and/or function causing downstream regulatory effects disrupting regulatory pathways during brain development.

To identify lncRNAs associated with ASD, we integrated differential gene expression patterns with gene co-expression networks. We analyzed RNA-seq data from the cortical tissue of brains from ASD cases and controls to identify lncRNAs differentially expressed in ASD. We derived a gene co-expression network from an independent human brain developmental transcriptome and detected a convergence of the differentially expressed lncRNAs and known ASD risk genes into a gene co-expression module. Co-expression network analysis facilitates the discovery of associations between uncharacterized lncRNAs with known ASD risk genes, affected molecular pathways and at-risk developmental periods. Utilizing an integrative approach comprised of differential expression analysis in affected tissues and connectivity metrics from a developmental co-expression network, we prioritized a set of candidate ASD-associated lncRNAs. The identification of lncRNAs as novel ASD susceptibility genes could help explain the genetic pathogenesis of ASD.

P26: ModEvo: A WEB-BASED TOOL FOR MODELING EVOLUTIONARY DYNAMICS

Subject: Simulation and numeric computing
Presenting Author: Filip Jagodzinski, Western Washington University
Co-Author(s): Rainier Harvey, Western Washington University, United States; Jesse Sliter, Western Washington University, United States; Elizabeth Brooks, Western Washington University, United States; Ali Scoville, Central Washington University, United States

ABSTRACT: Quantitative genetics is concerned with developing computational models to predict the evolution of traits in response to selection. Most models for analyzing the evolution of multiple traits
employ a constant genetic variance co-variance matrix (G-Matrix). However, non-linear interactions between developmental factors underlying the production of traits can drastically affect how they co-vary.

We have developed a code-base, ModEvo, to assist in testing hypothesis about the evolutionary dynamics among multiple phenotypic traits affected by non-linear developmental interactions. Our software implements and extends a novel mathematical framework developed by Sean Rice that synthesizes concepts central to evolutionary developmental biology and quantitative genetics.

We are developing a Graphical User Interface (GUI) and the accompanying back-end infrastructure to permit biologists to interface with ModEvo via a publicly available web server. Users specify input parameters for the quantitative genetics models and invoke the back-end modeling software with a single button click. The evolutionary dynamics output by ModEvo are displayed both graphically and numerically. The front-end, back-end infrastructure uses Google Go as the back-end server and Angular as the front-end model-view controller. Our web tool is easy enough to use by a non-specialist, but also allows an experienced user to specify model parameters for a more detailed analysis.

P27: G4 QUADRUPLEXES IN AND NEAR REGULATORY ELEMENTS OF MAIZE GENES PREDICT TISSUE TYPE AND ALTERED TRANSCRIPTIONAL RESPONSE TO ABIOTIC STRESSES

Subject: Qualitative modeling and simulation

Presenting Author: Mingze He, Iowa State University

Co-Author(s): Angélica Sanclemente, University of Florida, United States; Carson Andorf, USDA, United States; Hank W. Bass, Florida State University, United States; Harkamal Walia, University of Nebraska-Lincoln, United States; Justin W. Walley, Iowa State University, United States; Karen Koch, University of Florida, United States; Peng Liu, Iowa State University, United States; Carolyn J. Lawrence-Dill, Iowa State University, United States

ABSTRACT: In maize shoot tissues genes with G4-quadruplexes in or near regulatory regions respond strongly to diverse stress conditions including submergence, cold, heat UV, salt, and cold stress. GO enrichment studies indicate that differentially expressed G4-containing genes are likely to be involved in developmental processes, suggesting that altered growth rates may be a specific component of the stress response. To further investigate the function of these G4 genes, we carried out transcriptomic and proteomic analyses across 55 tissues and developmental stages in non-stress conditions. We found G4 could be applied as a marker to predict transcription rate and specific tissue type in normal tissues. In addition, co-expression network analysis between maize atlas and stressed tissues revealed G4 motifs strongly associated with transcription factors activation in response to stresses. Our results provide novel evidence to the association of G4 with emergent energy
status in maize. Our findings suggest a new component in maize stress response mechanism.

P28: POPULATION-SPECIFIC DIAGNOSTIC ANALYSIS FOR IMPROVING DETECTION OF DISEASE-ASSOCIATED GENES IN TYPE 2 DIABETES

Subject: Machine learning, inference and pattern discovery

Presenting Author: Michael Hinterberg, University of Colorado

Co-Author(s): David Kao, University of Colorado, United States; Judy Regensteiner, University of Colorado, United States; Jane Reusch, University of Colorado, United States; Carsten Goerg, University of Colorado, United States

ABSTRACT: Diagnostic measurements serve as surrogate endpoints for health and disease status. Usually, the threshold used to distinguish healthy from diseased individuals is based upon population parameters or outcomes related to the disease. However, in smaller subsampled populations, such as those found in clinical trials, this clinical diagnostic cutoff for disease status may not be optimal for a particular group of subjects. Intrinsic factors of the trial population, such as age, biological sex, comorbid conditions and other potential confounding variables, can bias the subject distribution. In this work, we demonstrate that applying different diagnostic thresholds in Type 2 diabetes reveals different gene expression associations within a specific sample population. To accomplish this, we used visually-interactive algorithms and representations for rapid reconfiguration of phenotype definitions for hypothesis testing. Just as the choice of diagnostic cutoff influences sensitivity and specificity of disease detection, it also affects the sensitivity and specificity of gene-association hypotheses. Using publicly-available gene expression data from pancreatic and skeletal muscle tissue, we show that stratification by biological sex suggests different diagnostic thresholds for genes associated with glucose control within a specific trial population. Furthermore, we describe distinct patterns of association of different genes along the continuum of clinical diagnostic cutoffs. Our results suggest that population-specific phenotype definitions may be important to detect robust associations between disease phenotype and gene expression.

P29: MULTIMETHOD COMPUTATIONAL MODELING ANALYSIS OF SPONTANEOUS AND XENOBIOTIC-MODULATED MITOCHONDRIAL DYSFUNCTION UNDERLYING DEGENERATIVE SENESCENCE

Subject: Simulation and numeric computing

Presenting Author: Timothy Hoffman, Colorado State University

Co-Author(s): William Hanneman, Colorado State University, United States

ABSTRACT: The past two decades have proven fruitful for the field of biogerontology, but much of the research has focused solely on
snapshots of the relationship between mitochondrial dysfunction and biological aging. In particular, the mitochondrial-free-radical-theory-of-aging (MFRTA) has focused on reactive oxygen species (ROS) produced by the electron transport chain (ETC) and the resulting aberrations that persist primarily within the mitochondrial genome. However, this theory has lost momentum in the wake of recent studies that have shown minimal ROS as not outright deleterious in nature, but in fact beneficial under the appropriate circumstances. Additional dimensions that may account for these observations are the mitochondrial unfolded protein response (UPRmt) and the process of selective mitophagy. A multi-level hybrid-modeling paradigm, containing agent-based elements among probabilistic system-dynamics environments of logically-derived ODEs, is utilized here to simulate aging mitochondrial phenotypes within a population of cells, equipped with specific characteristics intended to mimic neuronal behavior. The model is based upon an integrated network of known cellular mechanisms pertaining to the biology of Caenorhabditis elegans, and also draws upon conserved biochemical characteristics of other eukaryotic cell types. The integration of such processes provides a deeper understanding of age-related mechanisms, as the in silico experiments performed here account for the spontaneous quantitative decline in mitochondrial function and the subsequent onset of cell death. Additionally, the simulation was virtually probed with xenobiotics in a variety of dosing schemes to enhance or inhibit specific mechanistic targets, providing insights into chemical agents that may shorten or improve neurological health-span.

P30: PREDICTION OF PROKARYOTIC OPTIMUM GROWTH TEMPERATURE BASED ON GENOMIC AND PROTEOMIC FEATURES

Subject: Other

Presenting Author: Mallika Iyer, University of Colorado Denver

Co-Author(s): Christopher Miller, University of Colorado Denver, United States

ABSTRACT: Prokaryotes are known to grow at a wide range of temperatures. Many studies have been conducted to determine what genomic and proteomic features are responsible for growth at different temperatures. For example, it has been found that the GC content of RNA stems, and the fraction of IVYWREL amino acids in a proteome separately correlate with prokaryotic optimum growth temperature (OGT). However, many of these studies were performed over 5-10 years ago, when genomic databases were more limited and phylogenetically biased. Modern sequencing technology has resulted in exponential growth in the number of genomes added to public databases. This calls for validation of these studies on an updated dataset of genomes. Furthermore, a combination of these features could produce a highly accurate predictor of OGT. We have collected
~3000 genomes annotated with optimum growth temperatures to investigate these correlations. The calculation of many of these proteomic features requires protein structures, thus we are also modeling the structures of conserved proteins across all prokaryotes. Our initial results show, for example, that the fraction of IVYWREL in the proteomes correlates strongly with OGT \((r=0.760)\) in our expanded set of prokaryotes. In general, our preliminary results confirm the utility of many of the metrics used to predict OGT, but highlight the need to integrate multiple metrics in order to achieve accuracy across the full spectrum of phylogeny and temperatures.

**P31: DERIVING POPULATION-SCALE THERAPEUTIC TRAJECTORIES TO ENABLE PRECISION PHARMACOLOGY**

Subject: Machine learning, inference and pattern discovery

Presenting Author: Kipp Johnson, Icahn School of Medicine at Mount Sinai

Co-Author(s): Benjamin Glicksberg, Icahn School of Medicine at Mount Sinai, United States; Khader Shameer, Icahn School of Medicine at Mount Sinai, United States; Joel Dudley, Icahn School of Medicine at Mount Sinai, United States

**ABSTRACT:** Introduction: Treatment pathways provide standard guidelines for treating the primary diseases of patients. However, patients present with comorbidities, side effects and comply poorly with treatment adherence. Availability of a precision prescription data analytics platform may help to understand factors driving better therapeutic outcomes and lower side effects.

Methods and Results: The Mount Sinai EMR contains over 18.5 million prescriptions od 1,510 unique medications. Of the entire hospital population used in this study, 803,157 (38.2%) had at least one prescription \((23.25\pm87.21)\). Polypharmacy prevalence (co-administration of 4+ prescriptions) increased in an age-dependent manner, from 4% in those 0-10 years old to 62.8% in those >80. 95,373 drug pairs were enriched for co-administration \((\text{Exact-test } Q<0.01)\). 23,656 drug-pair sequences (drug 1 followed by drug 2) were detected \((\text{Binomial } Q<0.01)\) including the stimulants modafinil to armodafinil \((\text{OR}=185)\), antiplatelet therapies aspirin to ticagrelor \((\text{OR}=139)\), diabetes drugs liraglutide to canagliflozin \((\text{OR}=79)\), antipsychotics olanzapine to haloperidol \((\text{OR}=63)\), and drug-antidote pair naloxone and hydromorphone \((\text{OR}=22)\). We assembled a directed network of drug trajectories with 838 nodes and 23656 edges \((\text{diameter}=13)\) from drug pair trajectories. Greedy clustering partitioned the network into 7 subgraphs. Network hubs were detected and scored with Kleinberg’s method \((\text{principal eigenvectors of } \text{Adj(M)}^*\text{t(Adj(M))})\). Top hub drugs were lisinopril, amlodipine, aspirin, fluticasone/salmeterol, hydrochlorothiazide, simvastatin, ergocalciferol, albuterol, furosemide, and omeprazole.
Conclusion: Systematic mining of prescription data could help to uncover relationships between therapies and outcomes and aid in the implementation of precision prescription workflows.

**P32: KScope: A FAST MACHINE LEARNING COMPOSITION-BASED GENOMIC READ CLASSIFICATION TOOL**

Subject: Metagenomics

Presenting Author: Laurynas Kalesinskas, Loyola University Chicago

Co-Author(s): Maxwell Kelly, Rose Hulman Institute of Technology, United States; Catherine Putonti, Loyola University Chicago, United States

**ABSTRACT:** With the onset of contemporary high-throughput sequencing technologies, we are able to generate massive amounts of reads in a very short period. However, assigning taxonomic classifications to these reads remains a rate-limiting step and is computationally expensive. While alignment-based classifiers, such as those founded on BLAST searches, are the most sensitive and precise, they require substantial CPU time and necessitate that the organism(s) under investigation are represented within existing databases. In the case of viruses, the latter is not true: despite being the most ubiquitous biological entities on earth, there is a dearth of viral genome sequences. Herein, we introduce KScope, a machine learning k-mer-composition-based read classification tool. KScope uses a modified, hash-based, k-nearest neighbor algorithm and SQL databases to speed and reduce the computational expense of classifying sequencing reads. KScope examines reads based upon the frequency of occurrence of short k-mers, and conducts these analyses for multiple values of k. As such, KScope is capable of readily classifying species based upon underlying phylogenetic signals, e.g. codon usage, tetranucleotide usage, etc.

**P33: A SPATIOTEMPORAL MODEL TO SIMULATE CHEMOTHERAPY REGIMENS FOR HETEROGENEOUS BLADDER CANCER METASTASES TO THE LUNG**

Subject: Qualitative modeling and simulation

Presenting Author: Kimberly Kanigel Winner, University of Colorado School of Medicine

Co-Author(s): James Costello, University of Colorado School of Medicine, United States

**ABSTRACT:** Tumors are composed of heterogeneous populations of cells. Somatic genetic aberrations are one form of heterogeneity that allows clonal cells to adapt to chemotherapeutic stress, thus providing a path for resistance to arise. In silico tumor modeling provides a platform for rapid, quantitative experiments to inexpensively study how compositional heterogeneity contributes to drug resistance. Accordingly, we have built a spatiotemporal model of a lung metastasis originating
from a primary bladder tumor, incorporating in vivo drug concentrations of first-line chemotherapy, vascular density of lung metastases, and increased resistance in cells that survive chemotherapy. In metastatic bladder cancer, a first-line drug regimen includes six 21-day cycles, with gemcitabine plus cisplatin (GC) delivered simultaneously on day 1, and gemcitabine on day 8. After simulated treatment, post-regimen tumor cell populations are mixtures of originally resistant clones and/or new clones that have gained resistance to cisplatin, gemcitabine, or both drugs. The emergence of a tumor with increased resistance is qualitatively consistent with the five-year survival of 6.8% for patients with metastatic transitional cell carcinoma of the urinary bladder treated with a GC or MVAC regimen. We have also explored the effects of the synergistic interaction between gemcitabine and cisplatin, and of the disbursement of cellular drug damage between daughter cells. The model can be adapted to other cancers, and can be further used to explore the parameter space for clinically relevant variables, including drug delivery timing, increased dosage within toxicity limits, and patient-specific data such as rates of resistance gain, disease progression, and molecular profiles.

P34: ScanGEO - MINING HIGH-THROUGHPUT FUNCTIONAL GENOMICS DATA

Subject: Networking, web services, remote applications

Presenting Author: Katja Koeppen, Geisel School of Medicine at Dartmouth

Co-Author(s): Thomas Hampton, Geisel School of Medicine at Dartmouth, United States; René Zelaya, Perelman School of Medicine at the University of Pennsylvania, United States; Casey Greene, Perelman School of Medicine at the University of Pennsylvania, United States; Bruce Stanton, Geisel School of Medicine at Dartmouth, United States

ABSTRACT: The NCBI gene expression omnibus (GEO) is a repository of high-throughput data containing millions of significant results, less than 1% of which have been reported in the literature. ScanGEO is a user-friendly open source web application designed to facilitate efficient mining of this under-utilized resource.

While the NCBI GEO web portal is limited to looking at differential gene expression one study or one gene at a time, ScanGEO allows users to rapidly identify differentially expressed genes of interest across all relevant GEO data sets and visualize the results.

The application is written in R and implemented as a Shiny App to allow access to users without knowledge of R. A ScanGEO search can be limited to a particular organism and/or keyword of interest and uses a custom list of relevant genes to be tested for differential gene expression using ANOVA. Users can either input genes or use any public or, with login, private geneset available in the Tribe webserver for reproducible geneset-based analyses.
Outputs of the application include a summary table of all GEO data sets with the selected characteristics and PDF files with box plots of significantly differentially expressed genes.

In summary, ScanGEO is a new online resource that accelerates the analysis of publicly available high-throughput data for hypothesis generation and validation of experimental data.

**P35: INEXPENSIVE MOBILE DIAGNOSIS OF DIABETIC RETINOPATHY USING DEEP LEARNING**

**Subject:** Machine learning, inference and pattern discovery

**Presenting Author:** Kavya Kopparapu, Thomas Jefferson High school

**ABSTRACT:** Diabetic retinopathy (DR) is the leading cause of blindness among working-age adults and affects over 10 million people worldwide. Many adults, particularly in developing countries, remain undiagnosed due to limited access to the expensive tools needed for diagnosis. Smartphone technology, notably, is cheap, readily available nearly everywhere, and has potential to aid in diagnostics. We developed the Eyeagnosis system, which utilizes machine learning techniques and a smartphone camera for the automatic screening of DR. Specifically, we designed a neural network architecture that uses residual neural networks with cyclic pooling to automatically diagnose DR from retinal images. We were able to obtain an accuracy of 78.9%, sensitivity of 0.675, specificity of 0.812, and area under the receiver operating characteristic curve (ROC) of 0.752. These results are statistically comparable to the results of a group of 74 optometrists.

Additionally, we created a smartphone application which was able to take photos, send them to a server, and display the server’s diagnosis. With a custom-designed 3D-printed lens attachment, Eyeagnosis was able to take focused retinal images, as shown through testing on dilated eyes. These results demonstrate that Eyeagnosis is capable of assisting doctors in diagnosing DR in the field.

**P36: HRC3 – A NEW CLASS OF MOTIFS INVOLVED IN CHROMATIN ORGANIZATION AND DEVELOPMENT.**

**Subject:** Machine learning, inference and pattern discovery

**Presenting Author:** Andrzej Kudlicki, University of Texas Medical Branch

**ABSTRACT:** Chromatin modifications, such as methylation and acetylation of lysine residues in histone tails, are an important mechanism of epigenetic regulation. It remains unclear how the enzymes responsible for histone modifications are directed to the correct loci, in a manner that is specific to the cell type and outside stimuli.

We report the discovery of a conserved structural signature of DNA fragment that coincides with experimental binding sites of histone-
modifying enzymes, such as KDM5B, KDM5A, PHF8, EZH2, RBBP5, SAP30, HDAC1 and HDAC6, also SUZ12, CHD1, SMARCB1 – involved in regulation of chromatin organization and silencing. The signature (“the HRC3 motif”) is approximately 180 base pairs long and is defined by a specific, periodic pattern in the Hydroxyl Radical Cleavage profile of a dsDNA interval. The pattern is present in both non-coding and coding sequences; in coding sequences it is produced by a very specific choice of codons in the region. The HRC3 signature is associated with several thousand genes; functional analysis show highly significant enrichment of genes involved in processes related to development (GO:0009888, GO:0048731, GO:00325020), regulation of gene expression and in DNA binding (GO:0003677). The HRC3 motifs are highly conserved, remaining unchanged from human to Drosophila. The most intriguing property of these motifs is their association with pairs or clusters of developmental transcription factors with a conserved synteny, including Hox genes. We present a model that uses HRC3s to explain the colinearity of HOX clusters in segmented animals. We also discuss their possible role in control of replication initiation.

P37: RECONSTRUCTING PROTEIN AND GENE PHYLOGENIES BY EXTENDING THE FRAMEWORK OF RECONCILIATION

Subject: Optimization and search

Presenting Author: Esaie Kuitche, Université de Sherbrooke

ABSTRACT: Recent genome analyses have revealed the ability of eukaryotic genes to produce several transcripts and proteins. This mechanism plays a major role in the functional diversification of genes. Still, current reconstructions of gene phylogenies are based on a single reference protein per gene, thus neglecting all other alternative products of genes. A first approach for reconstructing gene product phylogenies along gene trees was recently introduced in the literature. It consists in models and algorithms for transcript phylogeny reconstruction, given the gene phylogeny and the gene exon structures. A prerequisite of this approach is to have a correct gene phylogeny, while currently reconstructed gene phylogenies contain errors. Here, we explore a different approach for the joint reconstruction of protein and gene phylogenies using reconciliation. We present an extension of the framework of reconciliation in order to reconstruct conjointly the gene tree and the phylogeny of all the proteins produced by a gene family, given the species tree. We propose a model of protein evolution involving two types of evolutionary event called “protein creation” and “protein loss”, in addition to the classical evolutionary events of speciation, gene duplication and gene loss. We introduce new reconciliation problems derived from the protein evolutionary model. Some preliminary algorithmic results and a method for the joint reconstruction of gene trees and proteins trees are presented. The
applications of the method show that the new framework allows to reconstruct more accurate gene trees than currently available methods. It also allow to reconstruct well-accepted protein phylogenies.

**P38: A NEW ALGORITHM FOR BIOMEDICAL ARTICLE RANKING**

**Subject:** Text Mining  
**Presenting Author:** Ying Liu, St. John’s University

**ABSTRACT:** How to present information retrieval results is one main problem that needs to tackle in biomedical information retrieval. A single query may retrieve a large number of results and advanced ranking algorithms are necessary to rank the results so that most relevant result is shown on the top of the list. In this paper, we explored to rank MEDLINE citations using HITS (Hyperlink-Induced Topic Search) algorithm. HITS uses web links from one page to another to rank web pages. It has proven to be successful in web search engines. We further extended HITS to supervised HITS to rank citations. Our results showed that supervised HITS algorithm significantly outperforms HITS algorithm (p<0.01). Compared with HITS, supervised HITS can improve citation ranking from 15% to more than 59% in almost all the cases we tested. Furthermore, MeSH terms outperforms text words in ranking citations, especially when HITS was applied (p<0.01).

**P39: STRATIFICATION OF PROSTATE CANCER PATIENTS BASED ON MOLECULAR INTERACTION PROFILES**

**Subject:** Machine learning, inference and pattern discovery  
**Presenting Author:** Roland Mathis, IBM Research  
**Co-Author(s):** Matteo Manica, IBM Research, Switzerland; Maria Rodriguez Martinez, IBM Research, Switzerland

**ABSTRACT:** Prostate cancer is a leading cause of cancer death amongst men, however the molecular-level understanding of disease onset and progression are largely unknown. Specifically, stratification of intermediate prostate tumor states based on current markers is difficult. The aim of this project is to integrate multi-omics data from individual patients with knowledge from literature and public databases to infer a molecular interaction network specific to prostate cancer. Inspired by the DREAM5 challenge we integrate predictions from multiple inference methods based on information theory, correlation and regression models to build a disease specific interactome. Emphasis is put on combining different data types and systematically integrating prior information using natural language processing and knowledge graphs. From the interactome we identify relevant interaction modules through
graph-theory approaches. For each interaction module we cluster the patients based on molecular states measurements. The patient-specific cluster assignment vectors serve as a personalized interaction signatures and is used to stratify patients.

**P40: PROTEOMIC ANALYSIS OF HUMAN SERUM SAMPLES TO REVEAL NEW BIOMARKERS AND MECHANISMS OF NSAID-INDUCED CARDIOVASCULAR TOXICITY**

Subject: Other

Presenting Author: Jane Mitchell, Imperial College London

Co-Author(s): Sarah Mazi, Imperial College London, United Kingdom

**ABSTRACT:** Nonsteroidal anti-inflammatory drugs (NSAIDs), which work by inhibiting cyclooxygenase-2, are amongst the most commonly used medications world-wide with an estimated 70 million prescriptions and 30 billion doses consumed annually in the US. However, NSAIDs have serious side effects with the risk of cardiovascular events dominating concern. The anxiety caused by the fear of having a heart attack or stroke whilst taking these drugs, has created a very real unmet clinical need to find biomarkers to predict and mechanisms to explain these side effects. Furthermore, it has prevented the development of NSAIDs as anti-cancer drugs where they have proven potential. Here we have used an unbiased proteomic approach to identify novel biomarkers and mechanistic insights to predict and understand NSAID-induced cardiovascular toxicity.

Proteomic analysis was performed on serum collected from healthy volunteers before and after taking an NSAID (celecoxib) at standard doses for 7 days. Serum was analysed by UPLC-MS/MS using a Thermo QExactive mass spectrometer. Data was processed using Progenesis and MASCOT software, which identified ≈460 proteins across all samples, with ≈30 proteins being altered at p<0.05. Of particular note was increases of 2 and 3 fold respectively in LPS binding protein (LBP) and vascular cell adhesion protein 1 (VCAM-1). Pathway analysis revealed altered proteins map to changes in acute inflammatory response and acute-phase response networks.

These data, whilst preliminary, identify molecules and pathways that may help us predict and understand NSAID-induced cardiovascular toxicity and demonstrate the potential power of a systems biology approach to addressing this research question.

**P41: WITHDRAWN**
P42: USING SEGMENTAL DUPLICATIONS TO ANALYZE THE ACCURACY OF TE CLASSIFICATION AND THE FREQUENCY OF GENE CONVERSION BETWEEN TE REMNANTS

Subject: Other

Presenting Author: Gilia Patterson, University of Montana

Co-Author(s): Travis Wheeler, University of Montana, United States

ABSTRACT: Most of the human genome is derived from the remnants of transposable elements (TEs), sequences of DNA that can move and insert copies of themselves throughout the genome. TEs are annotated and classified into subfamilies based on their DNA sequences. A subfamily is meant to represent all the copies generated in a burst of replication by a few closely related TEs. Different subfamilies within some families, such as Alus, have very similar sequences, so gene conversion can occur between TEs. We use a database of segmental duplications to analyze the accuracy of subfamily classifications and to determine the rate of gene conversion between Alus. When a segment of genome containing a TE remnant is duplicated, the TE remnants in each copy are replicates and so should be in the same subfamily unless one TE is misclassified or has undergone gene conversion. We identified the location and subfamily of all TEs in known segmental duplications and found that many are assigned to different subfamilies. In many cases, these appear to be the result of gene conversion; even in the absence of gene conversion, the rate of TE subfamily misclassification is concerning.

P43: ShinyLearner: ENABLING BIOLOGISTS TO PERFORM ROBUST MACHINE-LEARNING CLASSIFICATION

Subject: Machine learning, inference and pattern discovery

Presenting Author: Stephen Piccolo, Brigham Young University

Co-Author(s): Terry Lee, Brigham Young University, United States; Shelby Taylor, Brigham Young University, United States

ABSTRACT: Machine-learning classification is an invaluable tool for biologists. In one type of application, biomedical researchers use classification algorithms to predict whether patients will respond to a particular drug or belong to a specific disease subtype. Although the research community has developed many classification algorithms and corresponding software libraries, considerable barriers exist for non-computational biologists to take advantage of these tools. Different algorithms are written in different programming languages and require different input formats. Software libraries may require dependencies that are difficult to install, and the software may fail if incompatible versions are installed. If a researcher wanted to employ algorithms implemented in multiple software libraries, she/he may need to learn
multiple programming languages and be careful to avoid biases as comparisons were made across the algorithms.

We developed ShinyLearner (https://github.com/srp33/ShinyLearner), an open-source software tool that reduces these barriers. ShinyLearner integrates several popular machine-learning libraries (e.g., scikit-learn, mlr, weka) within a Docker container that includes all software dependencies. Accordingly, ShinyLearner can be installed with ease. ShinyLearner supports Monte Carlo and k-fold cross validation and provides an option for feature selection. When multiple classification algorithms are used, ShinyLearner dynamically selects the best algorithm via nested evaluation. A simple Web interface facilitates the process of selecting parameters. Output files are in “tidy” format to enable easier processing with external tools. New algorithms can be integrated into ShinyLearner with a simple GitHub pull request.

Finally, we will describe findings from a comprehensive benchmark comparison across classification algorithms applied to 20+ gene-expression data sets.

## P44: DEVELOPMENT OF A DIAGNOSTIC TO PROFILE EUKARYOTIC MICROBES OF THE HUMAN MICROBIOME

**Subject:** Metagenomics

**Presenting Author:** Ana Popovic, Hospital for Sick Children, University of Toronto

**Co-Author(s):** John Parkinson, Hospital for Sick Children, University of Toronto, Canada; Michael Grigg, National Institutes of Health, United States

**ABSTRACT:** Human microbiome studies have implicated the composition of gut bacteria in function of the immune system, obesity, drug metabolism, even human behaviour. While much has been learned about the contribution of bacteria to human health and disease, few studies have addressed the role of the eukaryotic members of the microbiome. This represents a considerable gap in knowledge, as single celled eukaryotes such as Giardia, Cryptosporidium and Entamoeba infect hundreds of millions of people worldwide, and are responsible for a significant burden of gastrointestinal illnesses. In addition to pathogenic eukaryotes, several studies have identified particular species of Blastocystis and Entamoeba as residents of the healthy gut, suggesting that eukaryotic microbes play a larger role than previously appreciated in human health. A key challenge in establishing the contribution of the eukaryotic microbiome to health and disease is the lack of accurate diagnostic technology. Here, we will present our efforts to develop a new multi DNA biomarker technology, based on several hypervariable regions in the small and large ribosomal subunit genes, to accurately profile eukaryotic microbes in the human gut.
ABSTRACT: Despite advancements in therapeutic strategies such as antibodies and gene therapy, small molecules remain the gold standard of treatment for numerous diseases, including cancer. Small molecules are low molecular weight compounds that rapidly diffuse across cell membranes to reach their molecular target, which is often a protein or nucleic acid. For example, many small molecule therapies inhibit the activity of a specific kinase. When investigating the effect of a small molecule on cell state or disease, researchers often compare the genome-wide mRNA levels of drug-treated cells and vehicle-treated control cells. The output of this experiment is a list of differentially expressed genes, which either increase or decrease in expression following drug treatment. This list can then be analyzed with gene set enrichment analysis (GSEA), an algorithm which performs hypergeometric tests with curated gene sets to determine which biological processes are more or less active in the drug-treated cells.

We hypothesized that even a highly-specific small molecule drug may result in non-specific effects in the cell, such as an up-regulation of generic stress response pathways. These non-specific pathways may appear as significant in the GSEA output, potentially overshadowing crucial biological processes specific to the drug under investigation. To address this problem, we aggregated several hundred gene expression experiments where human tissues or primary cells were treated with a small molecule drug. These experiments were annotated before analysis with GSEA. Our results identified pathways that are overrepresented in small molecule drug screens, providing valuable experimental and biological insight into therapeutic drug development.
P46: WHOLE GENOME SEQUENCING AND DE NOVO ASSEMBLY FROM A CRITICALLY ENDANGERED MAMMAL, THE SUMATRAN RHINOCEROS (DICERORHINUS SUMATRENSIS)

Subject: Other

Presenting Author: Swanthana Rekulapally, Marshall University

Co-Author(s): Herman L. Mays Jr, Marshall University, United States; Chih-Ming Hung, Biodiversity Research Center, Taiwan; Terri Roth, Cincinnati Zoo and Botanical Garden, United States; David A. Oehler, Bronx Zoo, United States; Alexander Lange, Cincinnati Children’s Hospital, United States; Jeffery A. Whitsett, Cincinnati Children’s Hospital, United States; James Denvir, Marshall University, United States; Donald A. Primerano, Marshall University, United States; Jun Fan, Marshall University, United States; Megan Justice, Marshall University, United States

ABSTRACT: The Sumatran Rhinoceros (Dicerorhinus sumatrensis) is among the most imperiled mammalian species on earth. Genomics analyses may inform our understanding of the evolution and demographic history of this species in ways that may direct conservation strategies. We assembled a draft genome sequence for D. sumatrensis using a DISCOVAR de novo/SOAPdenovo2 pipeline with data from three 2x250 paired-end and eight mate pair Illumina sequencing libraries. The resulting 1.1 million scaffolds, 4,588 of which were greater than 100Kbases, spanned a total of 2.96Gbases, with an N50 of 0.6Mbases. This genome assembly is currently being used in a pairwise sequential Markovian coalescent (PSMC) approach to assess the demographic history of this critically endangered species. We additionally aligned single-end RNA-Seq reads to the resulting scaffolds using HISAT2 and CUFFLINKS from which a total of 148464 exons and their corresponding 68635 transcripts were identified. Comparison of transcripts to domestic horse (Equus caballus) will aid in identification of homologous genes and gene mutations that may contribute to the unique morphological traits of the Sumatran Rhinoceros.

P47: A NEW MOLECULAR SIGNATURE APPROACH FOR PREDICTION OF DRIVER_CANcER PATHWAYS FROM TRANSCRIPTIONAL DATA

Subject: Machine learning, inference and pattern discovery

Presenting Author: Boris Reva, Icahn School of Medicine at Mount Sinai

Co-Author(s): Noam Beckmann, Icahn School of Medicine at Mount Sinai, United States; Hui Li, Icahn School of Medicine at Mount Sinai, United States; Andrew Uzilow, Icahn School of Medicine at Mount Sinai, United States; Dmitry Rykunov, Icahn School of Medicine at Mount Sinai, United States; Eric Schadt, Icahn School of Medicine at Mount Sinai, United States

ABSTRACT: Assigning cancer patients to the most effective treatments requires an understanding of the molecular basis of their disease. While DNA-based molecular profiling approaches have flourished over the past several years to transform our understanding of driver pathways across a broad range of tumors, a systematic characterization of key
driver pathways based on RNA data has not been undertaken. Here we introduce a new approach to predict the status of driver cancer pathways based on signature functions we constructed using weighted sums of gene expression levels derived from RNA sequencing data. To identify the driver cancer pathways of interest, we mined DNA variant data from TCGA and nominated driver alterations in seven major cancer pathways in breast, ovarian, and colon cancer tumors. The activation status of these driver pathways was then characterized using RNA sequencing data by constructing signature functions in training datasets and then testing the accuracy of the signatures in test datasets. The signature functions differentiated tumors with nominated active pathways from tumors with no genomic signs of activation very well (average AUC equals to 0.8), and they systematically exceeded the accuracies obtained by ten other known classification methods we employed as a control. A typical pathway signature is composed of ~20 biomarker genes that are unique to a given pathway and cancer type. Our results confirm that driver genomic alterations are distinctively displayed at the transcriptional level and that the transcriptional signatures can generally provide an alternative to DNA sequencing methods in detecting specific driver pathways.

P48: THE AFFINITY DATA BANK FOR BIOPHYSICAL ANALYSIS OF REGULATORY SEQUENCES

Subject: System integration
Presenting Author: Todd Riley, University of Massachusetts Boston
Co-Author(s): Cory Colaneri, UMass Boston, United States; Aadish Shah, UMass Boston, United States; Brandon Phan, UMass Boston, United States; Pritesh Patel, UMass Boston, United States; Zazil Villanueva, UMass Boston, United States; Devesh Bhimsaria, University of Wisconsin-Madison, United States

ABSTRACT: We present The Affinity Data Bank (ADB), a suite of tools that provides biologists with novel aids to deeply investigate the sequence-specific binding properties of a transcription factor (TF) or an RNA-binding protein (RBP), and to study subtle differences in specificity between homologous nucleic acid-binding proteins. Also, integrated with Pfam, the PDB, and the UCSC database, The ADB allows for simultaneous interrogation of protein-DNA and protein-RNA specificity and structure in order to find the biochemical basis for differences in specificity across protein families. The ADB also includes a biophysical genome browser for quantitative annotation of in vivo binding – using free protein concentrations to model the non-linear saturation effect that relates binding occupancy with binding affinity. Importantly, the in vivo TF and RBP protein concentrations can be
inferred from transcriptome or proteome data – including RNA-seq data. The biophysical browser also integrates dbSNP and other polymorphism data in order to depict changes in affinity due to genetic polymorphisms – which can aid in finding both functional SNPs and functional binding sites. Lastly, the biophysical browser also supports biophysical positional priors to allow for quantitative designation of the in vivo, locus-specific accessibility that a protein has to the DNA. With the inclusion of these biophysical occupancy-based and affinity-based positional priors, the ADB can properly model in vivo protein-DNA binding by integrating the effects of chromatin accessibility and epigenetic marks.

P49: MODELING HETEROGENEOUS CELL POPULATIONS USING BOOLEAN NETWORKS

Subject: Simulation and numeric computing

Presenting Author: Brian Ross, University of Colorado

Co-Author(s): James Costello, University of Colorado, United States

ABSTRACT: Cellular processes can be simulated using Monte Carlo (random sampling) methods, but these have difficulty capturing rare outcomes, particularly when the state space is huge. Yet in many cases (such as cancer) these infrequent outcomes are the ones with the most impact. Here we present a Boolean network method for modeling mixed cell populations using a single simulation, which captures these very rare subpopulations. Our method works by treating the dynamics as a system of linear equations which allow superposition of different cell populations, in a basis rotated from the state space so that the equations tend to close with relatively few variables. For cases when the variable space is still too large, we show how to efficiently remove degeneracies in our linear system as it is being built, thereby capturing the later-time evolution with a reduced system of equations. Our method generalizes to probabilistic Boolean networks, and works for both discrete and continuous time-evolution.

We evaluate our method using a >50-gene network modeling prostate cancer. Our method reproduces the results of Monte Carlo while capturing rare events that Monte Carlo cannot find. As a proof of concept, we simulate the dynamics of a mixed population spanning >10^15 different cell states with all possible combinations of loss-of-function mutations. Finally, we use these simulations to find the likely mutational trajectories of an evolving tumor in our prostate-network model. Our method can thus identify the extraordinary, as well as the typical, fates of cells.
**P50: A NEW APPROACH FOR PREDICTION OF MOLECULAR SIGNATURES OF OUTCOME IN CANCER**

**Subject:** Machine learning, inference and pattern discovery

**Presenting Author:** Dmitry Rykunov, Icahn School of Medicine at Mount Sinai

**Co-Author(s):** Eric Schadt, Icahn School of Medicine at Mount Sinai, United States; Boris Reva, Icahn School of Medicine at Mount Sinai, United States

**ABSTRACT:** Stratification of cancer patients into different risk groups is one of the key tasks in the development of personalized therapy of cancer. Driven by the hypothesis that the aggressiveness of cancer (and disease outcome) is associated with distinct genomic and transcriptional features, we developed a molecular signature approach for prediction of the disease outcome given a transcriptional or genomic profile of a tumor. The signature of outcome – a weighted sum of gene expression levels - was derived from a training dataset of RNA sequencing profiles of TCGA with available survival information and then tested in a test dataset. In constructing the signature function, we assumed that the expression level of each of the individual gene-biomarkers can be used to differentiate the more aggressive and less aggressive forms of disease. Under this assumption, the biomarker weights in the signature function can be computed analytically. We applied the signature approach to RNAseq profiles of seven cancers and obtained very distinct separation of tumors into poor and better survival classes. The P-values of the survival difference obtained for the combined signatures are substantially lower than any of the P-values obtained for individual genes. This illustrates the power of the general approach to combine individual biomarkers into a consistent signature of outcome.

**P51: SCNIC: FINDING AND SUMMARIZING MODULES OF CORRELATED OBSERVATIONS**

**Subject:** Metagenomics

**Presenting Author:** Michael Shaffer, University of Colorado - Denver

**Co-Author(s):** Catherine Lozupone, University of Colorado Denver, United States

**ABSTRACT:** Microbiome studies are commonly limited by a lack of statistical power. Studies typically have small sample sizes and large numbers of observations finding significant correlations and associations is difficult. By finding modules of autocorrelated observations and summarizing them, the number of observations can be reduced and statistical power increased. The tool WGCNA is the standard for cooccurrence analysis, module formation and feature reduction but uses parametric tests and an assumption of scale-free network topology. We find 16S sequencing data, and metabolomics data, do not meet these assumptions. To remedy this, we developed a tool for correlation network analysis with sparse, compositional data.
To generate the correlation network SparCC is used to avoid the pitfalls of Pearson and Spearman correlations with sparse, compositional data. The clique percolation method is used to find modules in the network. Modules are summarized and a new, smaller table, as well as a network file for import into Cytoscape, is outputted. When applied to a 16S study of the HIV gut microbiome significant differences in module abundances were found when comparing HIV+ untreated individuals to healthy controls. The direction of association with HIV for modules was the same as was found for individual module members. We found a 10% reduction in features yielding increased power for further statistical analysis. The discovery and summarization of modules in 16S sequencing data provides a strong and convenient method for increasing statistical power. Module composition can be associated with disease and indicate potential interactions between groups of microbes.

**P52: AUTOMATED OPTIMAL DESIGN OF VOLTAGE CLAMP PROTOCOLS TO STUDY SODIUM CHANNEL KINETICS USING A MINIMAL CARDIAC ION CHANNEL MODEL**

**Subject:** Simulation and numeric computing

**Presenting Author:** Matthew Shotwell, Vanderbilt University Medical Center

**Co-Author(s):** Richard Gray, Food and Drug Administration, United States

**ABSTRACT:** A “minimal” Hodgkin-Huxley formalism is used to model the behavior of Sodium (Na) channels during the cardiac action potential. This type of model is used to study arrhythmias and cardiac interventions. Conventionally, the model parameters have been estimated in a piecewise fashion using the results of voltage clamp experiments. The design of voltage clamp protocols is a manual and laborious task, and is focused on isolating the time- and voltage-dependence of ion channel conductances. We present an automated optimal design method for selecting voltage clamp protocols from a broad class of protocols such that the associated experimental findings are most informative about the model parameters. We demonstrate the utility of this approach using a series of simulated optimal voltage clamp experiments.

**P53: INTELLIGENT 3D CRYO-EM IMAGE ANALYSIS FOR NEXT GENERATION BIOMEDICINE**

**Subject:** Machine learning, inference and pattern discovery

**Presenting Author:** Dong Si, University of Washington Bothell

**ABSTRACT:** Life ultimately depends on the interactions of large biological molecules, such as proteins. The nature of these interactions depends on the three dimensional (3D) shape and structure of these molecules. Electron cryo-microscopy (cryo-EM) as a cutting edge
technology has carved a niche for itself in the study of large-scale protein complex. Although the protein backbone of complexes cannot be derived directly from the medium resolution (5–10 Å) of amino acids from three-dimensional (3D) density images, secondary structure elements (SSEs) such as alpha-helices and beta-sheets can still be detected. The accuracy of SSE detection from the volumetric protein density images is critical for ab initio backbone structure derivation in cryo-EM. So far it is challenging to detect the SSEs automatically and accurately from the density images at these resolutions. We have combined image processing, machine learning and geometric modeling techniques and developed SSETracer, SSELearner, StrandTwister, StrandRoller along with recent developed deep learning framework to allow for the automatic and accurate detection and prediction of secondary structures from experimental derived cryo-EM images of protein complexes.

P54: NETWORKS IN SYSTEMS IMMUNOLOGY

Subject: Machine learning, inference and pattern discovery

Presenting Author: Janet Siebert, University of Colorado Denver

Co-Author(s): Holden Maecker, Stanford University, United States; Julie Yabu, Stanford University, United States

ABSTRACT: A systems immunology study generates data from a various assays and multiple timepoints. Assays might include gene expression, CyTOF immunophenotyping, and phosphoepitope flow cytometry. These assays interrogate different biological compartments. One question of interest is whether or not analytes are correlated across these compartments. To address this question using data from a study of kidney transplant patients treated with a desensitization therapy, we computed linear regressions between all possible pairs of analytes, where each member of the pair was drawn from a different assay (n=32,772). Then we filtered the results to include only those pairs in which there was both a strong relationship between the analytes, and a credible difference in that relationship between responders and non-responders (n=93). We identified 7 analytes that appeared in at least of 6 these pairs and built a network that included these analytes and their neighbors. An arc diagram of this network illustrated the relationships in the system. Next we characterized the network by the sum of the degrees of the 7 most connected vertices. We used randomly generated graphs of the same degree and size to show that the concentration in our graph was highly unlikely to occur by chance (p < 0.0001). This approach suggests that there are correlations across different compartments that differ for responders and non-responders, and that there are some analytes that may be highly influential. These results might provide insights into the biological mechanisms of responsiveness to therapy.
P55: DIFFERENTIATING BETWEEN AUTHENTIC AND CRYPTIC 5’ SPlice SITES

Subject: Machine learning, inference and pattern discovery
Presenting Author: Kiruthika Sivaraman, San Jose State University
Co-Author(s): Remya Mohanan, San Jose State University, United States; Pratikshya Mishra, San Jose State University, United States; Sami Khuri, San Jose State University, United States

ABSTRACT: The accurate splicing at the 5’ and 3’ splice sites of the pre-mRNA is an extremely important step in the gene expression pathway in eukaryotes. Mis-splicing by the spliceosome at other sites, known as cryptic splice sites, often lead to devastating results. It is now estimated that up to 50% of disease-causing mutations disrupt splicing. Consequently, it is of crucial importance to understand the reasons behind the cryptic splice site selection by the spliceosome. The central question we study is: Can we learn from known cryptic splice sites to predict and detect putative cryptic splice sites in other genes in the human genome?

To better understand the mechanics behind the spliceosome’s selection of cryptic splice sites, three data sets, consisting of authentic, cryptic, and random 5’ splice sites were built. The data sets comprise of 9-mers: sequences that are 9 bases long. Nucleotides in positions 1-3 lie at the end of the exon while nucleotides in positions 4-9 lie in the beginning of the intron. Positions 4 and 5 comprise of the invariant GT dinucleotide; this is characteristic of all 5’ splice sites. We then implemented and built a decision tree from the authentic splice sites and scored all three types of sequences. We also built a decision tree from the cryptic splice sites and scored the same three data sets. By comparing the results obtained, one can see if there is an inherent difference between authentic and cryptic splice sites.

P56: THE RGD PhenoMiner DATABASE AND TOOL

Subject: Other
Presenting Author: Jennifer Smith, Medical College of Wisconsin
Co-Author(s): Stanley Laulerderkind, Medical College of Wisconsin, United States; G Thomas Hayman, Medical College of Wisconsin, United States; Victoria Petri, Medical College of Wisconsin, United States; Shur-Jen Wang, Medical College of Wisconsin, United States; Monika Tutaj, Medical College of Wisconsin, United States; Jyothi Thota, Medical College of Wisconsin, United States; Yiqing Zhao, Medical College of Wisconsin, United States; Omid Ghiasvand, Medical College of Wisconsin, United States; Marek Tutaj, Medical College of Wisconsin, United States; Jeffrey De Pons, Medical College of Wisconsin, United States; Melinda Dwinell, Medical College of Wisconsin, United States; Mary Shimoyama, Medical College of Wisconsin, United States

ABSTRACT: Phenotype is defined as a trait which contributes to the physical, biochemical, and physiological makeup of an individual as determined by both genetics and environmental influences. As such, the information needed to fully describe a trait/phenotype
measurement should include information about both the genetics of the organism and environmental influences that might affect the measurement. The Rat Genome Database (http://rgd.mcw.edu) has developed a system to standardize quantitative phenotype measurements using ontologies to capture data for each experiment related to sample, measurements taken, methods used and applicable experimental conditions. Quantitative phenotype records include information on the trait assessed (Vertebrate Trait Ontology), the exact measurement that was made (Clinical Measurement Ontology), the method used (Measurement Method Ontology), the condition(s) under which the measurement was made (Experimental Condition Ontology), and the sample measured—and by extension, the genotype—including information on rat strain (Rat Strain Ontology), number of individuals, sex and age. This has provided the framework to integrate more than 60,000 phenotype records from numerous experiments. PhenoMiner’s query wizard (http://rgd.mcw.edu/phenotypes/) allows researchers to retrieve and view data from multiple studies, and to compare experimental values across multiple strains, methods and/or conditions, allowing them to choose appropriate disease models and controls among the available strains. We are currently working to utilize these data to statistically determine expected ranges for standard measurements in commonly used rat strains. Future development will extend the model to cellular and molecular phenotypes and provide tools with which users can compare their own data to expected ranges.

**P57: WHO WANTS TO QUIT: CHARACTERISTICS AND PREDICTION OF SMOKERS INTERESTED IN QUITTING TOBACCO USE**

**Subject:** Machine learning, inference and pattern discovery

**Presenting Author:** Andrey Soares, University of Colorado

**Co-Author(s):** Sonia Leach, National Jewish Health and University of Colorado School of Medicine, United States

**ABSTRACT:** Smokers that received a brief intervention from healthcare providers have a higher rate of success quitting tobacco use than smokers that try to quit on their own. With 70% of smokers visiting a healthcare provider each year (i.e., doctors, nurses, dentists, and others), there is great potential for change and direct impact helping smokers to quit. However, we need approaches that can support providers during very brief tobacco cessation interventions (less than 3 minutes), including analysis of patient health and treatment recommendations tailored to patient characteristics and health conditions. In addition, short interventions (3 to 10 minutes) may become a disruption for healthcare practices, and may not qualify them to receive reimbursement for smoking cessation counseling.
Lack of time and specialized training are usually reported as issues for providing brief interventions and addressing the complexity of nicotine dependence. This research aims to identify important features that can help predict patients interested in quitting tobacco use so that interventions can start even before a healthcare provider sees a patient. Such features could support predictions by enhancing the questions asked in the patient history form, which are typically completed while patients are in the waiting room of a healthcare facility. This research will build a prediction model to identify who wants to quit, and will perform external validation on multiple cohorts to support generalizability of the model.

P58: ANALYSIS OF TOBACCO USERS ADMITTED TO INTENSIVE CARE UNITS

Subject: Machine learning, inference and pattern discovery
Presenting Author: Andrey Soares, University of Colorado School of Medicine
Co-Author(s): Sonia Leach, National Jewish Health, University of Colorado School of Medicine, United States; Kevin Cohen, University of Colorado School of Medicine, United States; Joan Davis, Southern Illinois University, United States

ABSTRACT: Smoking is known to cause numerous tobacco-related diseases such as cancer, heart disease, diabetes, respiratory disease, as well as death. The Center for Disease Control and Prevention warns that over 16 million Americans have some disease caused by smoking, with about 480,000 deaths in the United States. Thus, it is critical for healthcare professionals to identify and treat every tobacco user seen in any healthcare facilities. This research seeks to examine if patients, who are current tobacco users, have been correctly identified as smokers, their smoking status and behaviors have been documented, and they have received appropriate treatment recommendations (prescriptions) based on their health conditions. In particular, we will perform text analysis of the chart notes recorded during the patient stay to collect information that can be used to offer tailored treatment recommendations such as the number of cigarettes used per day, and to verify inconsistencies in documenting information about smoking. Preliminary data analysis shows that some tobacco users have not been diagnosed as smokers using the appropriate ICD9 code, leaving the information about smoking to be retrieved from the text notes or inferred from the prescribed tobacco medications. This research will also evaluate the treatment recommendations based on patient health conditions and risks, and will cluster smokers to identify emerging patterns and relationships among characteristics and diagnoses that can support tobacco intervention strategies for patients admitted to intensive care units. We will focus on comorbidities as tobacco use can trigger new diseases or complicate existing ones.
ABSTRACT: Advances in technology have resulted in a dramatic decrease of DNA sequencing costs with unprecedented availability of data which has the potential to revolutionize genomics based medicine. In this presentation we report on our experience in harnessing this data growth and deriving personal health and genomics association insights. In particular, we report on our recent implementation of pipelines related to network based clustering (clustering of patients’ omics profiles while utilizing known relationships among genes) and gene set characterization (reporting of annotations such as biological process or pathways significantly enriched in a given gene set). Cloud based solutions to address scalability and sustainability are also discussed. Our future plan calls for implementing other pipelines related to classification and regression.

ABSTRACT: The Evaluation of Differential Dependency (EDDY) detects differential dependencies between two classes (conditions) for a group of genes by computing the probability distribution of dependency networks generated by resampled RNA expression data. EDDY finds the differential dependency between two conditions by calculating the divergence between the condition-specific network distributions for the genes within each annotated pathway and assessing its significance via permutation test. High sensitivity has been one benefit of this statistical rigor yet at a considerable computational cost. As a result, extremely large datasets such as TCGA pan-cancer study were out of reach for EDDY analysis. However, the ample and regular compute coupled with a small memory footprint positioned EDDY as an ideal candidate for GPGPU implementation. Now complete, GPU-EDDY exhibits two orders
of magnitude in performance enhancement and has been applied to pan-cancer datasets involving thousands of samples.

We will present application of GPU-EDDY run across the TCGA pan-cancer dataset, identifying differential pathways between PIK3CA mutation versus wild-type. One discovery involved the TGF-Beta signaling in EMT epithelial to mesenchymal transition pathway, which appears to favor more coherent altered signaling for mutation samples. In the SHC-related events pathway, RAF1 signaling occurs exclusively in wild type samples, pointing to an alternative oncogenic signaling network in wild type. These results will be presented via an interactive network interface made available through our web portal. In addition, we will share insights we’ve gleaned through scaling our application to larger datasets.

P61: KNOWLEDGE-BASED ANALYSIS AND INTERPRETATION OF GENOME WIDE ASSOCIATION STUDIES

Subject: Other

Presenting Author: Laura Stevens, University of Colorado, Anschutz Medical Campus

Co-Author(s): David Kao, University of Colorado, Anschutz Medical Campus, United States; Carsten Gorg, University of Colorado, Anschutz Medical Campus, United States

ABSTRACT: The analysis and interpretation of genome wide association studies (GWAS) is a challenging and significant problem in clinical and biomedical research. Current tools for analyzing these datasets are often based on a linear modeling framework that considers only one single nucleotide polymorphism (SNP) at a time and overlooks the environmental genetic aspect. We propose to employ visual analysis approaches to support the interpretation of results from genome-scale experiments in the context of existing biomedical knowledge. We synthesized a wide range of SNP-specific information from multiple data sources and created two types of networks: gene-centric networks, in which SNPs are mapped to genes (whenever possible) and links represent known relationships between the genes, and SNP-centric networks, in which we directly show known relationships between SNPs. We analyzed these networks with RenoDoI, a visual analysis tool in Cytoscape, which visualizes the knowledge networks and annotations, and allows users to obtain the most useful subnetworks through degree-of-interest functions, statistical analyses, and interactive techniques. Using data from the Framingham Heart Study we performed a case study in collaboration with a cardiologist; the domain expert analyzed phenotypic, genetic, structural, mechanistic, and heritable connections between SNPs related to heart failure with the goal to identify relationships between clusters of comorbid risk factors and incident heart failure.
P62: CROSS-PLATFORM NORMALIZATION ENABLES MACHINE LEARNING MODEL TRAINING ON MICROARRAY AND RNA-SEQ DATA SIMULTANEOUSLY

Subject: Machine learning, inference and pattern discovery

Presenting Author: Jaclyn N Taroni, University of Pennsylvania Perelman School of Medicine

Co-Author(s): Casey S Greene, University of Pennsylvania Perelman School of Medicine, United States

ABSTRACT: Large compendia of gene expression data have proven valuable for the extraction of cell type-specific expression patterns and for the discovery of novel biological relationships. As of August 2016, 1.7 million RNA assays are available from ArrayExpress. The majority of these assays are run on microarray, while RNA-sequencing (RNA-seq) is becoming the platform of choice for new experiments. The data structure and distributions between the two platforms differ, making it challenging to combine them for machine learning applications. Combining both platforms could allow models to take advantage of the additional information captured in some RNA-seq experiments while benefiting from the substantially greater abundance of microarray data. Here, we compare normalization methods when a training set must comprise of samples run on both platforms. We use the Cancer Genome Atlas breast cancer dataset as a test case because many matched samples have been assayed on both platforms. We compare the following normalization methods: log transformation, quantile normalization (QN), nonparanormal normalization (NPN), Training Distribution Matching (TDM), and z transformation, and their use with multiple supervised and unsupervised machine learning algorithms. To test the effect of different proportions of RNA-seq data on performance, we ‘titrate’ a proportion of RNA-seq samples into the training set in 10% intervals (0-100%). We find that QN and TDM perform well on both microarray and RNA-seq test sets when training sets are comprised of moderate amounts of RNA-seq data. This work demonstrates that it is possible to perform model training on microarray and RNA-seq data simultaneously.

P63: USING KaBOB TO FIND NOVEL ADVERSE DRUG-DRUG INTERACTIONS

Subject: Machine learning, inference and pattern discovery

Presenting Author: Ignacio Tripodi, University of Colorado, Boulder

ABSTRACT: A significant number of drugs have adverse interactions, for example due to similarities in their metabolic pathway. Many of these interactions have been studied in depth, but not every possible pair of drugs has been assayed. Past assays [1,2] have used network-based models to attempt solving this problem, even social media [3]. We propose a novel, semantic-reasoning-based approach to
Look for potential drug-drug adverse interactions by using KaBOB, a knowledgebase of biomedical public ontologies and datasets in a complex graph representation. KaBOB makes it possible to find relations between different biological entities like drugs, proteins and biological processes, and perform inferences on those relations. Finding nodes that represent drugs in this graph, and intersecting pathways between these nodes (for example using Reactome data), could yield to novel drug-drug interactions.

1. Li P., et al. “Large-scale exploration and analysis of drug combinations” DOI: 10.1093/bioinformatics/btv080

**P64: InterViewer, A NEW CYTOSCAPE-BASED VIEWER THAT DISPLAYS INTERACTIONS BETWEEN SELECTED SETS OF PROTEINS**

**Subject:** Graphics and user interfaces

**Presenting Author:** Marek Tutaj, Medical College of Wisconsin

**Co-Author(s):** Jyothi Thota, Medical College of Wisconsin, United States; Jeff De Pons, Medical College of Wisconsin, United States; Jennifer Smith, Medical College of Wisconsin, United States; Thomas G Hayman, Medical College of Wisconsin, United States; Victoria Petri, Medical College of Wisconsin, United States; Stan Laulederkind, Medical College of Wisconsin, United States; Shur-Jen Wang, Medical College of Wisconsin, United States; Mary Shimoyama, Medical College of Wisconsin, United States

**ABSTRACT:** InterViewer, RGD’s new Cytoscape-based protein-protein interactions viewer, (https://rgd.mcw.edu/rgdweb/cytoscape/query.html), facilitates a detailed visualization of interactions between proteins. As usual, RGD provides interaction data not only for rat, but also for mouse and human. The tool accepts input in multiple ways: as a list of UniProt IDs, RGD IDs or gene symbols. On the display page, binary interaction data from IMEX are displayed as nodes and edges, which can be zoomed in or out using controls. Clicking on a protein node provides links to UniProtKB and to RGD gene report pages. Detailed information about the protein appears in the upper right. Clicking on an edge shows additional information about that interaction. Also for more complex networks, multiple display filters can be applied. The user can pick a set of interaction types of interest, one or more species or common interactors. In addition, several layout modes common for Cytoscape graphs like ‘cose’ or ‘circle’, are available. A legend details the color-coded interaction types and protein species. The table beneath the display lists downloadable characteristics of each pair of interactors, the complete node list and node/edge statistics. The bird’s-eye view panel facilitates movement of the display. The tool also has options to generate printable reports and graph images for user convenience.
ABSTRACT: Though it is often assumed in evolutionary analysis that the substitution rate is constant across the genome, there is a large body of evidence that this is not so. The substitution rate varies considerably through the genome depending on factors operating at many different scales. At the smallest scale, the substitution rate at a site depends strongly on the identity of nucleotides one to three bases away. At the megabase scale, factors such as replication timing and recombination rate influence substitution rates.

Generally, local and regional influences on substitution rates have been studied individually. This creates a problem, however, because these two sets of factors are not independent. Factors operating at the megabase scale affect the frequency of k-mers in a region, and each k-mer can exert its own influence on neighboring substitution rates at much smaller scales.

To study this local-regional interaction, we developed a Bayesian Markov chain Monte Carlo approach to simultaneously estimate parameters describing the regional contribution to substitution rate for thousands of megabase-sized regions in the human genome, and an additional set of parameters describing the effect of each possible sequences of neighboring nucleotides in the six closest bases to a position. We find that the long-term history of substitution rates in a region strongly influences the frequency of various k-mers in the region, which in turn exerts its own influence on substitution rates. This secondary effect of megabase-scale substitution rate factors appears to be a major contributor to substitution rate variation through the genome.

ABSTRACT: Tumors develop through an evolutionary process where mutations arising over time create distinct subpopulations of cells within a single tumor. Identification of these heterogeneous subpopulations and the evolutionary relationships between them is necessary for better understanding cancer progression. Current computational research has produced many methods to infer the composition and phylogenetic history of tumors, but modeling this
complexity is an uncertain process. Some methods produce multiple possible tumor phylogenies, rather than a single one. Furthermore, different computational approaches can produce contradicting phylogenies for the same dataset—making it difficult to determine the real evolutionary history. Combining information across multiple tumor phylogenies may allow us to identify a phylogeny that better represents the true evolutionary history of the tumor. We formalize the problem of inferring a single phylogeny from a collection of input phylogenies as the Tumor Phylogeny Consensus Problem and propose two approaches to solve this problem. The first approach creates a phylogeny consisting of the ancestral and clustering relations found in the majority of input trees. This approach is also informed by information about the frequency of substructures in the input trees and mutational frequency data from the input trees. The second approach uses a Markov Chain Monte Carlo method to explore the space of possible tumor phylogenies and identifies the phylogeny with minimal distance to the input trees. In tests with low variability simulated data we find that both consensus methods better approximate the true tree, in terms of topology and clustering, than the majority of input trees.

P67: INTEGRATION OF PROTEIN FAMILIES, LOCALIZATIONS, AND MODIFICATIONS IN A BIOLOGICAL KNOWLEDGE BASE

Subject: Qualitative modeling and simulation

Presenting Author: Elizabeth White, University of Colorado Denver, Anschutz

Co-Author(s): Elizabeth White, University of Colorado Denver, Anschutz, United States

ABSTRACT: As scientists accumulate more finely grained knowledge about biology, we still struggle with how to integrate this new information in ways that let us build hypotheses and frame alternative explanations. Our system, the Knowledge Base of Biomedical Ontologies (KaBOB), integrates many data sources into a coherent biological representation using Open Biological Ontologies and OWL semantics. This allows users to explore biological molecules in different stages of processing, in different cellular compartments, and in partnership with other molecules, to predict their involvement in various biological processes and pathways.

Recent work to enrich KaBOB has focused on representing the protein families in the Protein Ontology, along with their homologies, isoforms, variants, and interactions. Integrating entities from this ontology into the taxon-level data already in KaBOB provides significant challenges, including the need to recognize existing entities in the knowledge base and to posit new ones. Incorporating this information allows us to investigate how mutations influence the modification, trafficking and localization of proteins across species; disruptions in these processes are key factors in many diseases. Mutant proteins may be trafficked
and/or modified incorrectly, gain or lose function, coerce partners into pathological behavior, and thus cause varying degrees of havoc in the cell. This talk will demonstrate how KaBOB can be extended to predict variant protein forms, as well as their localizations, modifications, and effects.

**P68: DE NOVO PROTEIN STRUCTURE PREDICTION BY BIG DATA AND DEEP LEARNING**

Subject: Machine learning, inference and pattern discovery

Presenting Author: Sheng Wang, Toyota Technological Institute at Chicago

Co-Author(s): Jinbo Xu, Toyota Technological Institute at Chicago, United States

**ABSTRACT:** Recently ab initio protein folding using predicted contacts as restraints has made some progress, but it requires accurate contact prediction, which by existing methods can only be achieved on some large-sized protein families. To deal with small-sized protein families, we employ the powerful deep learning technique from Computer Science, which can learn complex patterns from large datasets and has revolutionized object and speech recognition and the GO game. Our deep learning model for contact prediction is formed by two deep residual neural networks. The first one learns relationship between contacts and sequential features from protein databases, while the second one models contact occurring patterns and their relationship with pairwise features such as contact potential, residue co-evolution strength and the output of the first network. Experimental results suggest that our deep learning method greatly improves contact prediction and contact-assisted folding. Tested on 579 proteins dissimilar to training proteins, the average top L (L is sequence length) long-range prediction accuracy of our method, the representative evolutionary coupling method CCMpred and the CASP11 winner MetaPSICOV is 0.47, 0.21 and 0.30, respectively; their average top L/10 long-range accuracy is 0.77, 0.47 and 0.59, respectively. Using our predicted contacts we can correctly fold 203 test proteins, while MetaPSICOV and CCMpred can do so for only 79 and 62 proteins, respectively. In the three weeks of blind test with the weekly benchmark CAMEO (http://www.cameo3d.org/), our method successfully folded three hard targets with a new fold and only 1.5L-2.5L sequence homologs while all template-based methods failed.
**P69: RANK AGGREGATION FOR FEATURE SCORING AND SELECTION**

**Subject:** Machine learning, inference and pattern discovery  
**Presenting Author:** Tara Yankee, University of Connecticut  
**Co-Author(s):** Kevin Brown, University of Connecticut, United States

**ABSTRACT:** The collection of all transcribed cellular mRNA for a given set of conditions, the transcriptome, contain prodigious information regarding cell-to-cell interactions and an individual cell’s response to its environment. In the past the limiting factor in genome sequencing was the substantial amount of amplification needed for transcript detection. Amplification results in bias and error and so it was considered too costly and error prone to collect transcriptomic data from single cells. More recently, next-generation sequencing technology (RNA-Seq) has produced robust single-cell transcriptomic data. The dimensionality of these data (50,000 transcripts and of order 1,000 samples) make machine learning approaches crucial in gaining biological understanding from these new data. Unsupervised clustering is used to attempt to group samples into stereotypical cell “types” based on their expression patterns. However, when the ratio of genes to samples may be 50 or more, feature scoring and selection are essential tools to reduce the dimensionality of the problem. A variety of scoring algorithms (linear predictability, laplacian score and related, etc.) have been proposed which are based on dramatically different ideas about feature “importance”. We use a rank aggregation method to combine estimates of feature importance from multiple sources, followed by forward selection using that ordering to obtain optimal feature subsets for subsequent unsupervised clustering. We demonstrate the performance of our algorithm on several real-world datasets, and compare it to naive forward selection (suboptimal but computationally efficient) and simulated annealing (optimal but extremely costly).

**P70: BOOTSTRAPPING ESTIMATES OF STABILITY FOR CLUSTERS, OBSERVATIONS AND MODEL SELECTION**

**Subject:** Machine learning, inference and pattern discovery  
**Presenting Author:** Han Yu, University at Buffalo  
**Co-Author(s):** Brian Chapman, University of Utah, United States; Arianna DiFlorio, Cardiff University School of Medicine, United Kingdom; Ellen Eishen, University of Oregon, United States; David Gotz, University of North Carolina at Chapel Hill, United States; Matthews Jacob, University of Iowa, United States

**ABSTRACT:** Clustering is a challenging problem in unsupervised learning. In lieu of a gold standard, stability is a valuable surrogate to performance and robustness. In this work, we propose a non-parametric bootstrapping approach to estimating the stability of a clustering method, which also captures stability of the individual
clusters and observations. This flexible framework enables different types of comparisons between clusterings that are naive or based on the Jaccard coefficient, can be used in connection with two possible bootstrap approaches for stability. The first approach, scheme 1, can be used to assess confidence (stability) around clustering from the original dataset based on bootstrap replications. A second approach, scheme 2, searches over the bootstrap clusterings for an optimally stable partitioning of the data. The two schemes accommodate different model assumptions, which can be motivated by an investigator’s trust (or lack thereof) in the original data. We propose a hierarchical visualization extrapolated from the stability profiles that give insights into the separation of groups, and projected visualizations for the inspection of individual stability. Our approaches show good performance in simulation and on real data.

**P71: THE HETNET AWAKENS AT HTTPS://NEO4J.HET.IO**

Subject: Graph Theory

Presenting Author: Daniel Himmelstein, University of Pennsylvania

Co-Author(s): Pouya Khankhanian, University of Pennsylvania, United States; Antoine Lizée, UCSF, United States; Leo Brueggeman, University of Iowa, United States; Sabrina Chen, Johns Hopkins University, United States; Dexter Hadley, UCSF, United States; Christine Hessler, UCSF, United States; Ari Green, UCSF, United States; Sergio Baranzini, UCSF, United States

**ABSTRACT:** Hetionet is a hetnet — a network with multiple node and relationship types — which encodes biological information. Version 1.0 contains 47,031 nodes of 11 types and 2,250,197 relationships of 24 types. Data was integrated from 29 public resources to connect compounds, diseases, genes, anatomies, pathways, biological processes, molecular functions, cellular components, perturbations, pharmacologic classes, drug side effects, and disease symptoms. Hetionet is available online as a public Neo4j database instance (https://neo4j.het.io). Hetionet was designed for Project Rephetio, which aims to systematically identify why drugs work and predict new therapies for drugs. Project Rephetio is an open notebook project available at https://thinklab.com/p/rephetio. 209,168 predictions of whether a compound treats a disease are available at http://het.io/repurpose/. 
IBM’s Technical Computing organization is the high performance computing organization in IBM Systems and Technology Group. This group is responsible for the strategy, marketing and identification of areas that can benefit from IBM’s high end technology. The life sciences is such an area, and IBM is and will continue to bring valued solutions to life sciences.

IBM’s Research Division is a partner with IBM’s Technical Computing organization, developing the next generation of high performance computers. In addition, the Research Division has many groups investigating numerous application areas in collaboration with IBM’s customers and partners. This includes IBM’s Computational Biology Center with IBM’s new Computational Science Center.

The Gold Lab at the University of Colorado was established in 1971 by Dr. Larry Gold upon his arrival at the Department of Molecular, Cellular, and Developmental Biology. Starting with basic research on bacteria and bacteriophage, the lab shifted its focus to human disease following the invention of the SELEX process in 1989. While at the university, Dr. Gold served as the chairman of the department from 1988 to 1992. Dr. Gold is one of a handful of people associated with the beginnings of biotechnology in Colorado. He cofounded an early biotech company, Synergen, and then founded NeXstar, a company that discovered Macugen, a drug to treat age-related blindness. Dr. Gold is a member of the National Academy of Sciences. Currently Dr. Gold serves as Chairman of the Board and Chief Executive Officer of SomaLogic, a proteom.

PatientsLikeMe is a free website where people with chronic health conditions get together and share their experiences living with disease. Where newly diagnosed patients can improve their outcomes by connecting with and learning from others who’ve gone before them. Where researchers learn more about what’s working, what’s not, and where the gaps are, so that they can develop new and better treatments.
SomaLogic® was founded in 2000 by Larry Gold, with the goal of improving the well-being and quality of life of every individual by transforming how diseases were detected and diagnosed. Building on the previous decade of aptamer research, SomaLogic scientists have developed a new proteomics technology that overcomes the significant challenges of current technologies, and which has multiple applications across the biological and medical sciences. Our mission is to leverage our proprietary technology to discover, develop and commercialize revolutionary new life science research tools and breakthrough clinical diagnostic products that will transform healthcare. See more at: www.somalogic.com/About-Us

SILVER

Biodesix® is a molecular diagnostics company advancing the development of innovative blood-based tests in oncology to enable precision medicine. At the forefront of precision medicine, Biodesix is developing new blood-based tests to identify patients who may benefit from immunotherapies. Biodesix discovers, develops and commercializes multivariate protein and genomic diagnostic blood tests, including the GeneStrat® and VeriStrat® tests, that deliver results within 72 hours. In addition to developing novel diagnostics independently, the company partners with biotechnology and pharmaceutical companies to develop companion diagnostics for use with therapeutic agents. The company is changing the standard of care by providing physicians with diagnostic tests for better therapeutic guidance, more accurate prognosis and enhanced disease monitoring to improve patient outcomes. For more information, please visit www.biodesix.com/technology/.
Mark your Calendar

December 7–9, 2017
Westin Hotel
Snowmass, CO

www.iscb.org/rocky2017