WELCOME

Dear Rocky 2017 participant

Welcome to the 15th Rocky Mountain Bioinformatics Conference. We 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 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 fifteen years; to SomaLogic for their financial and administrative support; to Biodesix and PatientsLikeMe for their continued support of this meeting. 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 Campbell, Elizabeth Wethington, and Caitlin Moloney.

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 6, 2017

<table>
<thead>
<tr>
<th>TIME</th>
<th>SESSION TYPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>4:00 PM – 6:00 PM</td>
<td>REGISTRATION</td>
</tr>
</tbody>
</table>

THURSDAY – DECEMBER 7, 2017

<table>
<thead>
<tr>
<th>TIME</th>
<th>SESSION TYPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:00 AM – 6:00 PM</td>
<td>REGISTRATION</td>
</tr>
<tr>
<td>8:00 AM – 9:00 AM</td>
<td>BREAKFAST</td>
</tr>
<tr>
<td>9:00 AM – 9:45 AM</td>
<td>KEYNOTE 1: MARICEL KANN, PhD</td>
</tr>
<tr>
<td></td>
<td>Associate Professor, Department of Biological Sciences, University of Maryland, United States</td>
</tr>
<tr>
<td></td>
<td>A Protein-Domain Approach for the Analysis of Disease Mutations</td>
</tr>
<tr>
<td>9:45 AM – 10:25 AM</td>
<td>OP 01 – OP 04</td>
</tr>
<tr>
<td>10:25 AM – 10:45 AM</td>
<td>BREAK</td>
</tr>
<tr>
<td>10:45 AM – 11:15 AM</td>
<td>KEYNOTE 2: PATRICK ZHAO, PhD</td>
</tr>
<tr>
<td></td>
<td>Professor and Head, Computational Biology &amp; Bioinformatics Lab, Noble Research Institute, LLC, Oklahoma, United States</td>
</tr>
<tr>
<td></td>
<td>Genotype-phenotype Association Discovery through Accurately Dissecting Phenotype-associated Genetic Causal Effects</td>
</tr>
<tr>
<td>11:15 AM – 11:55 AM</td>
<td>OP 05 – OP 08</td>
</tr>
<tr>
<td>12:00 PM – 04:00 PM</td>
<td>SKI BREAK</td>
</tr>
<tr>
<td>4:00 PM – 4:30 PM</td>
<td>KEYNOTE 3: KRISTA MEYER, PhD</td>
</tr>
<tr>
<td></td>
<td>Director of Proteomic Analysis, R&amp;D, Biodesix: Making Medicine Personal®, Colorado, United States</td>
</tr>
<tr>
<td></td>
<td>Development of Clinically Relevant Tests from Human Serum Samples: A Look at the Circulating Proteome</td>
</tr>
<tr>
<td>4:30 PM – 5:10 PM</td>
<td>OP 09 – OP 12</td>
</tr>
<tr>
<td>5:10 PM – 5:30 PM</td>
<td>BREAK</td>
</tr>
<tr>
<td>5:30 PM – 6:00 PM</td>
<td>OP 13 – OP 15</td>
</tr>
<tr>
<td>6:30 PM – 9:30 PM</td>
<td>DINNER – IL POGGIO</td>
</tr>
</tbody>
</table>
All sessions will take place at Snowmass Convention Center CC

<table>
<thead>
<tr>
<th>TIME</th>
<th>SESSION TYPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:00 AM – 6:00 PM</td>
<td>REGISTRATION</td>
</tr>
<tr>
<td>8:00 AM – 9:00 AM</td>
<td>BREAKFAST</td>
</tr>
</tbody>
</table>
| 9:00 AM – 9:45 AM | KEYNOTE 4: BJOERN PETERS, PhD  
Associate Professor, La Jolla Institute for Allergy & Immunology, California, United States  
Open Biomedical Ontology Services |
| 9:45 AM – 10:25 AM | OP 16 – OP 19 |
| 10:25 AM – 10:45 AM | BREAK       |
| 10:45 AM – 11:15 AM | KEYNOTE 5: RENÉE DEEHAN KENNEY, PhD  
Head of Biology and Bioinformatics, PatientsLikeMe, Massachusetts, United States  
Digital Me: Developing a Learning Health System through the Integration of Longitudinal Patient-generated Health Data with Multi-omic Molecular Measurements |
| 11:15 AM – 11:55 AM | OP 20 - OP 23 |
| 12:00 PM – 4:00 PM | SKI BREAK    |
| 4:00 PM – 4:30 PM | KEYNOTE 6: TOM HRAHA, PhD  
Research Scientist, Bioinformatics, SomaLogic, Inc., Boulder, Colorado, United States  
SOMAscan™ Proteomic Insights Contrasted with Genetic Insights in Down Syndrome Patients: The Need for a Molecular Phenotype Approach |
| 4:30 PM – 5:10 PM | OP 24 – OP 27 |
| 5:10 PM – 5:30 PM | BREAK        |
| 5:30 PM – 6:00 PM | OP 28 – OP 30 |
| 6:30 PM – 8:30 PM | POSTER SESSION |
All sessions will take place at Snowmass Convention Center CC

<table>
<thead>
<tr>
<th>TIME</th>
<th>SESSION TYPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:00 AM – 11:00 AM</td>
<td>REGISTRATION</td>
</tr>
<tr>
<td>8:00 AM – 9:00 AM</td>
<td>BREAKFAST</td>
</tr>
</tbody>
</table>
| 9:00 AM – 9:30 AM  | KEYNOTE 7: OLIVIER LICHTARGE, PhD & STEPHEN WILSON  
Baylor College of Medicine, Texas, United States  
MeTeOR: a Literature Network for Hypothesis Generation and Precision Medicine |
| 9:30 AM – 10:20 AM | OP 31 – OP 35                 |
| 10:20 AM – 10:40 AM | BREAK                       |
| 10:40 AM – 11:30 AM | OP 36 – OP 40               |
| 11:30 AM – 12:00 PM | KEYNOTE 8: KIRK E. JORDAN, PhD  
IBM Distinguished Engineer, Data Centric Solutions, IBM  
T.J. Watson Research & Chief Science Officer, IBM Research UK  
Data Centric Systems: Algorithm Exploitation & Evolving Cognitive Examples |
| 12:00 PM – 12:15 PM | CLOSING RAFFLE AND AWARDS     |
**DETAILED AGENDA**

*All sessions will take place at Snowmass Convention Center CC*

### WEDNESDAY – DECEMBER 6, 2017

<table>
<thead>
<tr>
<th>TIME</th>
<th>SESSION TYPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>4:00 PM – 6:00 PM</td>
<td>REGISTRATION</td>
</tr>
</tbody>
</table>

### THURSDAY – DECEMBER 7, 2017

<table>
<thead>
<tr>
<th>TIME</th>
<th>SESSION TYPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:00 AM – 6:00 PM</td>
<td>REGISTRATION</td>
</tr>
<tr>
<td>8:00 AM – 9:00 AM</td>
<td>BREAKFAST</td>
</tr>
</tbody>
</table>
| 9:00 AM – 9:45 AM | KEYNOTE 1: MARICEL KANN, PhD  
Associate Professor, Department of Biological Sciences,  
University of Maryland, United States  
A Protein-Domain Approach for the Analysis of Disease Mutations |
| 9:45 AM – 9:55 AM | OP01: Comparative Genomics and Alliance for Genome Resources  
Judith Blake, Jackson Laboratory |
| 9:55 AM – 10:05 AM | OP02: RSEQREP: RNA-Seq Reports, an open-source cloud-enabled framework for reproducible RNA-Seq data processing, analysis, and result reporting  
Travis Jensen, The Emmes Corporation |
| 10:05 AM – 10:15 AM | OP03: Confounds in biomedical natural language processing  
Kevin Cohen, University of Colorado School of Medicine |
| 10:15 AM – 10:25 AM | OP04: Integrated molecular and clinical analysis for understanding human disease relationships  
Winston Haynes, Stanford University |
| 10:25 AM – 10:45 AM | BREAK                                                                       |
| 10:45 AM – 11:15 AM | KEYNOTE 2: PATRICK ZHAO, PhD  
Professor and Head, Computational Biology and Bioinformatics Lab, Noble Research Institute, LLC, Oklahoma, United States  
Genotype-phenotype Association Discovery through Accurately Dissecting Phenotype-associated Genetic Causal Effects |
| 11:15 AM – 11:25 AM | OP05: Continuous integration for ensuring interoperability of the Open Biomedical Ontologies  
William Baumgartner, University of Colorado Anschutz Medical Campus |
<table>
<thead>
<tr>
<th>TIME</th>
<th>SESSION TYPE</th>
</tr>
</thead>
</table>
| 11:25 AM – 11:35 AM | OP06: Classification of glioblastoma subtypes by integrating genomic and histopathological image features with probabilistic graphical models  
Dimitris Manatakis, University of Pittsburgh |
| 11:35 AM – 11:45 AM | OP07: MPEG-G the emerging standard for genomic data compression  
Mikel Hernaez, University of Illinois |
| 11:45 AM – 11:55 AM | OP08: Physical aspects of the -helix and their implications towards amyloidal tendency  
Simcha Srebnik, Technion - Israel Institute of Technology |
| 12:00 PM – 4:00 PM | SKI BREAK                                                                   |
| 4:00 PM – 4:30 PM | KEYNOTE 3: KRISTA MEYER, PhD  
Director of Proteomic Analysis, R&D, Biodesix: Making Medicine Personal®, Colorado, United States  
Development of Clinically Relevant Tests from Human Serum Samples: A Look at the Circulating Proteome |
| 4:30 PM – 4:40 PM | OP09: A Novel Systematic Analysis of ALDH Isozyme Specificity in Head and Neck Squamous Cell Carcinoma  
Brian Jackson, University of Colorado Anschutz Medical Campus |
| 4:40 PM – 4:50 PM | OP10: Building a Systems-level model of the immune response to Salmonella infection  
Marta Andres-Terre, Stanford University |
| 4:50 PM – 5:00 PM | OP11: Development of a statistical model of the genetics of maternal-fetal dyads in neonatal abstinence syndrome  
James Denvir, Marshall University |
| 5:00 PM – 5:10 PM | OP12: Tracing the Innate Genetic Evolution and Spatial Heterogeneity in Treatment Naïve Lung Cancer Lesions  
Jihye Kim, University of Colorado Anschutz Medical Campus |
| 5:10 PM – 5:30 PM | BREAK                                                                       |
| 5:30 PM – 5:40 PM | OP13: Locating sites of ribonucleotide incorporation in RNase H2-deficient cells  
Alli Gombolay, Georgia Institute of Technology |
| 5:40 PM – 5:50 PM | OP14: Increasing flexibility of Laboratory Information Management System (LIMS) through extension of relational database model with NoSQL  
Marcin J. Domagalski, University of Virginia |
| 5:50 PM – 6:00 PM | OP15: Correcting for non-random fragmentation allows for more accurate estimation of relative abundance in NGS metagenomic data  
Elmar Pruesse, University of Colorado Denver |
<p>| 6:30 PM – 9:30 PM | DINNER - IL POGGIO                                                           |</p>
<table>
<thead>
<tr>
<th>TIME</th>
<th>SESSION TYPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:00 AM – 6:00 PM</td>
<td>REGISTRATION</td>
</tr>
<tr>
<td>8:00 AM – 9:00 AM</td>
<td>BREAKFAST</td>
</tr>
</tbody>
</table>
| 9:00 AM – 9:45 AM| KEYNOTE 4: BJOERN PETERS, PhD  
Associate Professor, La Jolla Institute for Allergy & Immunology,  
California, United States  
Open Biomedical Ontology Services |
| 9:45 AM – 9:55 AM| OP16: Assembly-free metagenomic analyses on distributed computational platforms  
Anna Paola Carrieri, IBM Research UK |
| 9:55 AM – 10:05 AM| OP17: Discovering the Contribution of the Gut Microbiome to the Plasma Metabolome  
Michael Shaffer, University of Colorado Denver Anschutz Medical Campus |
| 10:05 AM – 10:15 AM| OP18: Statistical Integration and Feature Selection for Candidate Biomarker Discovery  
Bobbie-Jo Webb-Robertson, Pacific Northwest National Laboratory |
Natasa Miskov-Zivanov, University of Pittsburgh |
| 10:25 AM – 10:45 AM| BREAK                                                                     |
| 10:45 AM – 11:15 AM| KEYNOTE 5: RENÉE DEEHAN KENNEY, PhD  
Head of Biology and Bioinformatics, PatientsLikeMe,  
Massachusetts, United States  
Digital Me: Developing a Learning Health System through the Integration of Longitudinal Patient-generated Health Data with Multi-omic Molecular Measurements |
| 11:15 AM – 11:25 AM| OP20: A systems biology approach using cell-to-cell heterogeneity to map cell fate decisions in developmental processes  
Jens Preussner, Max Planck Institute for Heart and Lung Research |
| 11:25 AM – 11:35 AM| OP21: Enabling Deep Learning on Structural Biological Data  
Tom Corcoran, Lawrence Berkeley National Laboratory |
| 11:35 AM – 11:45 AM| OP22: Unsupervised discovery of phenotype specific heterogeneous gene networks  
Jenny Shi, University of Colorado Anschutz Medical Campus |
| 11:45 AM – 11:55 AM| OP23: Detecting mutated driver genes from pan-cancer data through a novel matrix factorization based model by learning similarities from pairwise constraints  
Jianing Xi, University of Science and Technology of China |
<table>
<thead>
<tr>
<th>TIME</th>
<th>SESSION TYPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>12:00 PM – 4:00 PM</td>
<td>SKI BREAK</td>
</tr>
</tbody>
</table>
| 4:00 PM – 4:30 PM | KEYNOTE 6: TOM HRAHA, PhD  
*Research Scientist, Bioinformatics, SomaLogic, Inc., Boulder, Colorado, United States*  
*SOMAscan™ Proteomic Insights Contrasted with Genetic Insights in Down Syndrome Patients: The Need for a Molecular Phenotype Approach* |
| 4:30 PM – 4:40 PM | OP24: Automated Biomedical Text Classification with Research Domain Criteria  
*Mohammad Anani, Montana State University* |
| 4:40 PM – 4:50 PM | OP25: Unsupervised methods for lexical ambiguity in concept normalization  
*Negacy Hailu, University of Colorado* |
| 4:50 PM – 5:00 PM | OP26: A Comparison of Viral Identification Techniques in Longitudinal Metagenomic Datasets  
*Cody Glickman, University of Colorado Anschutz Medical Campus* |
| 5:00 PM – 5:10 PM | OP27: A Pilot Systematic Genomic Comparison of Recurrence Risks of Hepatitis B Virus-associated Hepatocellular Carcinoma with low and high degree of liver fibrosis  
*Seungyeul Yoo, Icahn School of Medicine at Mount Sinai* |
| 5:10 PM – 5:30 PM | BREAK                                                                       |
| 5:30 PM – 5:40 PM | OP28: Identifying and Annotating Uninvestigated Preeclampsia-Related Genes Using Linked Open Biomedical Data  
*Tiffany Callahan, University of Colorado Denver Anschutz Medical Campus* |
| 5:40 PM – 5:50 PM | OP29: Novel Approaches for Accuracy Assessment of 3D Chromatin Configuration Reconstructions  
*Mark Segal, University of California San Francisco* |
| 5:50 PM – 6:00 PM | OP30: Codon Aversion: An alignment-free method to recover phylogenies  
*Justin Miller, Brigham Young University* |
<p>| 6:30 PM – 8:30 PM | POSTER SESSION                                                             |</p>
<table>
<thead>
<tr>
<th>TIME</th>
<th>SESSION TYPE</th>
<th>SPEAKER</th>
<th>INSTITUTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:00 AM – 11:00 AM</td>
<td>REGISTRATION</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8:00 AM – 9:00 AM</td>
<td>BREAKFAST</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9:00 AM – 9:30 AM</td>
<td>KEYNOTE 7: OLIVIER LICHTARGE, PhD &amp; STEPHEN WILSON, PhD, Baylor College of Medicine, Texas, United States</td>
<td><em>MeTeOR: a Literature Network for Hypothesis Generation and Precision Medicine</em></td>
<td></td>
</tr>
<tr>
<td>9:30 AM – 9:40 AM</td>
<td>OP31: Isoform level differential network analysis: A proffer alternative for identifying cancer markers</td>
<td>Manoj Kandpal, Northwestern University</td>
<td></td>
</tr>
<tr>
<td>9:40 AM – 9:50 AM</td>
<td>OP32: Non-B DNA Predictor: A Web Application for Identifying Potential Intramolecular Non-B DNAs</td>
<td>Hannah Ajoge, Western University</td>
<td></td>
</tr>
<tr>
<td>9:50 AM – 10:00 AM</td>
<td>OP33: IMPACT Web Portal: An Oncology Database Integrating Molecular Profiles with Actionable Therapeutics from Next Generation Sequencing Data</td>
<td>Jennifer Hintzsche, University of Colorado Anschutz Medical Campus</td>
<td></td>
</tr>
<tr>
<td>10:00 AM – 10:10 AM</td>
<td>OP34: Exploring Frequented Regions in Pan-Genomic Graphs</td>
<td>Brendan Mumey, Montana State University</td>
<td></td>
</tr>
<tr>
<td>10:10 AM – 10:20 AM</td>
<td>OP35: TOBIAS — Transcription factor Occupancy prediction By Investigation of ATAC-seq Signals</td>
<td>Mario Looso, Max Planck Institute for Heart and Lung Research</td>
<td></td>
</tr>
<tr>
<td>10:20 AM – 10:40 AM</td>
<td>BREAK</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10:40 AM – 10:50 AM</td>
<td>OP36: Integrating Protein Localization Information in Signaling Pathway Reconstructions</td>
<td>Anna Ritz, Reed College</td>
<td></td>
</tr>
<tr>
<td>10:50 AM – 11:00 AM</td>
<td>OP37: Reducing the footprint of mass spectrometry data</td>
<td>Idoia Ochoa, University of Illinois at Urbana-Champaign</td>
<td></td>
</tr>
<tr>
<td>11:00 AM – 11:10 AM</td>
<td>OP38: Meeting today’s challenging computational biology demands by integrating a focused IT team into the research process</td>
<td>Dan Timmons, University of Colorado Boulder/BioFrontiers Institute</td>
<td></td>
</tr>
<tr>
<td>11:10 AM – 11:20 AM</td>
<td>OP39: Predicting Adverse Events Associated with Kinase Inhibitors from Clinical Drug Trials</td>
<td>Ilyssa Summer, University of Colorado Anschutz Medical Campus</td>
<td></td>
</tr>
<tr>
<td>11:20 AM – 11:30 AM</td>
<td>OP40: Characterization of the mechanism of drug-drug interactions from PubMed</td>
<td>Presenting Author: Feng Cheng, University of South Florida</td>
<td></td>
</tr>
<tr>
<td>11:30 AM – 12:00 PM</td>
<td>KEYNOTE 8: KIRK E. JORDAN, PhD</td>
<td>IBM Distinguished Engineer, Data Centric Solutions, IBM T.J. Watson Research &amp; Chief Science Officer, IBM Research UK</td>
<td><em>Data Centric Systems: Algorithm Exploitation &amp; Evolving Cognitive Examples</em></td>
</tr>
<tr>
<td>12:00 PM – 12:15 PM</td>
<td>CLOSING RAFFLE AND AWARDS</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Tom Hraha, PhD

Research Scientist, Bioinformatics, SomaLogic, Inc., Boulder, Colorado, United States

SOMAscan™ Proteomic Insights Contrasted with Genetic Insights in Down Syndrome Patients: The Need for a Molecular Phenotype Approach

Your genome is essentially the same as it was the day you were born (excepting cancer and some immune cells), but your body is likely not. Consulting the genome sequence to understand your health is akin to unrolling the blueprint of a 50-year-old building to understand the leaky pipes and sagging walls that exist today. The blueprint might suggest potential future health issues, but not to a high degree of certainty.

So what should we measure in addition to the genome? To understand the current health status of the individual (and changes in that status over time), we have to go beyond genotype to a deep measure of phenotype, which has always been the essential focus of medicine to diagnose and treat the individual. Tools are only now emerging that allow us to see individual phenotypes broadly at the molecular level.

One such tool is SOMAscan®, a high-throughput, multiplexed proteomic assay that is 15 years in the making. Using modified aptamer binding reagents (SOMAmer® reagents) SOMAscan® measures >5,000 unique proteins – or approximately 25% of the proteome. This has allowed us to develop “molecular phenotype” approaches with the goal of providing actionable health insights that go far beyond the intrinsic limits of genomic analyses.

The example presented will be Down Syndrome (trisomy 21), where an additional chromosome 21 does not lead to a proportional increase in circulating proteins encoded on that chromosome compared to the other 22. Therefore, the etiology of this disease must be understood from a phenotypic level.

Kirk E. Jordan, PhD

IBM Distinguished Engineer, Data Centric Solutions, IBM T.J. Watson Research & Chief Science Officer, IBM Research UK

Data Centric Systems: Algorithm Exploitation & Evolving Cognitive Examples

As has been stated many times, the volume, variety, velocity and veracity of data is pushing how we think about computer systems. IBM Research’s Data Centric Solutions organization has been developing systems that handle large data
sets shortening time to solution. As various features to improve data handling now exist in these systems, we need to begin to rethink the algorithms and their implementations to exploit these features. As bioinformatics often deals with large data sets, I will briefly describe some of the hardware features, and give an example of our early work to exploit these features. In addition, cognitive computing using machine learning techniques is all the rage. I will show how we are continuing to evolve and apply cognitive computing to address a variety of problems in the biological sciences areas.

Maricel Kann, PhD
Associate Professor, Department of Biological Sciences, University of Maryland, United States

A Protein-Domain Approach for the Analysis of Disease Mutations

Identifying the functional context for key molecular disruptions in complex diseases is a major goal of modern medicine that will lead to earlier diagnosis and more effective personalized therapies. Most available resources for visualization and analysis of disease mutations center on gene analysis and do not leverage information about the functional context of the mutation. In addition, these gene-centric approaches are confounded by the fact that gene products (proteins) may share some functional sub-units or protein domains but not others. I will describe a resource for domain mapping of disease mutations, DMDM, a protein domain database developed by our group in which each disease mutation is aggregated and displayed by its protein domain location. We have also developed a methodology using domain significance scores (DS-Scores) to detect statistically significant disease mutation clusters at the protein domain level. When we applied the DS-Scores to human data, we identified domain hotspots in oncogenes, tumor suppressors, as well as in genes associated with Mendelian diseases. In addition, I will describe recent work on analyzing cancer somatic mutations from individual cancer patient genomes. We found that incorporating information about classification of proteins and protein sites leads to new hypotheses regarding the role of tumor somatic mutations in cancer. Our analysis confirms that the domain-centric approach creates a framework for leveraging structural genomics and evolution into the analysis of disease mutations.
Renee Deehan Kenney, PhD
Head of Biology and Bioinformatics, PatientsLikeMe, Massachusetts, United States

Digital Me: Developing a Learning Health System through the Integration of Longitudinal Patient-generated Health Data with Multi-omic Molecular Measurements

PatientsLikeMe has over a decade of experience collecting patient-generated health data (PGHD) with the purpose of helping individuals track information about their health to improve their outcomes. This year, we launched Digital Me, a program whereby individuals can also donate longitudinally collected biosamples for multi-omic analysis. The goal is to develop a learning system whereby individuals benefit from insights gained about their health state through the analysis of both PGHD and multi-omics data. PatientsLikeMe has collected over 1000 blood draws from healthy controls and patients with neurological, immunological, mental health and pain and fatigue disorders, and are running DNA and RNAseq, proteomics, metabolomics, lipidomics, immunosignature and methylomics measurements on these samples. Health questionnaires are administered at the time of biosample collection and in between biosample collection, and particular emphasis is taken to collect biosamples during changes in health state; e.g., starting or stopping a new medication, the initiation or resolution of a flare. Here, we will share our development of an analysis platform to extract scientific insights from these data, as well as early results from one of the studies.

Olivier Lichtarge, PhD & Stephen Wilson, PhD
Baylor College of Medicine, Texas, United States • Joint Presentation

MeTeOR: a Literature Network for Hypothesis Generation and Precision Medicine

The paradigm of hypothesis-driven discovery depends on an understanding of the literature to guide experiments. However, as the number of publications grows exponentially, the ability to read the literature increasingly falls short. To relieve this bottleneck, we introduce MeTeOR, a network for automated knowledge summarization and hypothesis generation, which aggregates PubMed articles by connecting MeSH indexing terms curated by the National Library of Medicine. MeTeOR tallies more novel associations among genes, drugs and diseases than other databases and is more reliable than existing natural language processing algorithms. When combined with automated hypothesis generation, MeTeOR analyses of past literature predicted associations discovered afterwards. In a prospective example, immunoprecipitation
mass spectrometry supported both known and novel epidermal growth factor receptor associations predicted by MeTeOR. We conclude that MeTeOR generates valuable integrative hypotheses through a uniquely broad and accurate summarization of PubMed knowledge.

Krista Meyer, PhD

Director of Proteomic Analysis, R&D, Biodesix: Making Medicine Personal®, Colorado, United States

Co-authors: Julia Grigorieva. Biodesix, Inc, Research, US; Carlos Oliveira, Biodesix, Inc, Research, PT; Heinrich Roder, Biodesix, Inc, Research, DE; Joanna Roder, Biodesix, Inc, Research, UK

Development of Clinically Relevant Tests from Human Serum Samples: A Look at the Circulating Proteome

Therapeutics for the treatment of cancer patients are changing with the introduction of immunotherapies. Instead of targeting the tumor itself, the mechanism of action for immunotherapies relies on reactivation of the immune system such that the host can beat the cancer using the appreciated complex system evolutionarily developed to heal human disease. In oncology, clinical trials have proven that immunotherapies are effective at reducing tumor burden and extending survival in cancer patients across many indications. However, not all patients benefit from all immunotherapies and some patients may have primary immunotherapy resistance. A test identifying patients, prior to treatment that are unlikely to benefit from particular immunotherapies would be useful for guiding therapeutic decision making. Biodesix uses a hypothesis-free 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. Mass spectral data collected from human serum samples are analyzed by the Diagnostic CortexTM robust data analytics platform to design classifiers with clinical relevance. Using this approach, we have developed a test to identify patients with melanoma that have particular poor outcomes on anti-PD1 immunotherapy and therefore may be unsuitable candidates for treatment with drugs targeting this immune suppressing signal. While oncology has been the primary focus of our development work to date, understanding circulating proteome with our analytics platform could be extended to make new discoveries in a broad spectrum of human diseases.
Open Biomedical Ontology Services

The Open Biomedical Ontologies (OBO) Foundry is a community-organized project that aims to facilitate the discovery, development, application, harmonization, interoperability and sharing of ontologies. Its member ontologies include among many others the Gene Ontology (GO), Relations Ontology (RO), Disease Ontology (DO) and Ontology for Biomedical Investigations (OBI). For over a decade, OBO has provided a variety of services to the community, including the OBO registry, which collects standardized metadata for each member ontology, configurable Persistent URLs (PURLs), which provide an unchanging address for ontologies and ontology classes, OBO principles intended to enhance quality and interoperability, and an outreach and coordination effort to mediate between different member ontologies with the goal to coordinate and reduce overlap. Here we present our plans to increase robustness of existing services, and introduce new capabilities into the OBO infrastructure. We are working on a suite of automated quality controls that can be applied to any ontology project. These will be extended from single ontologies to groups of ontologies, testing their interdependencies. We believe that formulating such a computational test framework will enable productive discussions in the broader ontology developer and user community what standards and associated tests are useful, while also facilitating their implementation and eliminating the need for manual ontology review.

Genotype-phenotype Association Discovery through Accurately Dissecting Phenotype-associated Genetic Causal Effects

Phenotypes, often called traits, are controlled by many genes (G), the interactions among genes (GxG) and genes with environment (GxE). The GxG and GxE effects are thought to contribute significantly to the phenotypic variations of the complex traits. Generally, genetic variances (G; GxG) can be partitioned into 1) main effect that represents the cumulative effect of both individual genes/loci (additive effects) and dominant allelic interactions (dominance effects) to a given phenotype and 2) epistasis effect that refers to a trait that can be attributed to the interactions between multiple genetic loci. We developed a trio of genotype-phenotype association analysis tools, namely 1) GWASPRO (bioinfo.noble.org/GWASPRO/), which is designed to analyze main effects in large-scale
genome-wide association studies (GWASs); 2) PEPIS (bioinfo.noble.org/PolyGenic_QTL/), which adopts a full polygenic linear mixed model (LMM) to analyze the additive, dominance and epistatic effects in GWASs and quantitative trait loci mapping; and 3) PATOWAS (bioinfo.noble.org/PATOWAS/), which further extends the LMMs for broader associative ‘omics’ studies, i.e. can be applied to not only GWASs, but also transcriptomics-wide association studies (TWASs), and metabolomics-wide association studies (MWASs). In our case analysis of a publically available Immortalized F2 (IMF2) associative omics study data, we found only about 66% of the total phenotypic variance could be explained in GWAS results, while about 99% of the phenotypic variances were accounted for in the TWAS results, suggesting that the GxG and GxE involved in transcriptomic level. Our case studies demonstrated the high performance of our tools for genetic variances analysis, enabling genotype-phenotype association discovery.
ORAL PRESENTATION LIST

OP01: Comparative Genomics and Alliance for Genome Resources  
Presenting Author: Judith Blake, Jackson Laboratory

OP02: RSEQREP: RNA-Seq Reports, an open-source cloud-enabled framework for reproducible RNA-Seq data processing, analysis, and result reporting  
Presenting Author: Travis Jensen, The Emes Corporation

OP03: Confounds in biomedical natural language processing  
Presenting Author: Kevin Cohen, University of Colorado School of Medicine

OP04: Integrated molecular and clinical analysis for understanding human disease relationships  
Presenting Author: Winston Haynes, Stanford University

OP05: Continuous integration for ensuring interoperability of the Open Biomedical Ontologies  
Presenting Author: William Baumgartner, University of Colorado Anschutz Medical Campus

OP06: Classification of glioblastoma subtypes by integrating genomic and histopathological image features with probabilistic graphical models  
Presenting Author: Dimitris Manatakis, University of Pittsburgh

OP07: MPEG-G the emerging standard for genomic data compression  
Presenting Author: Mikel Hernaez, University of Illinois

OP08: Physical aspects of the α-helix and their implications towards amyloidal tendency  
Presenting Author: Simcha Srebnik, Technion — Israel Institute of Technology

OP09: A Novel Systematic Analysis of ALDH Isozyme Specificity in Head and Neck Squamous Cell Carcinoma  
Presenting Author: Brian Jackson, University of Colorado Anschutz Medical Campus

OP10: Building a Systems-level model of the immune response to Salmonella infection  
Presenting Author: Marta Andres-Terre, Stanford University

OP11: Development of a statistical model of the genetics of maternal-fetal dyads in neonatal abstinence syndrome  
Presenting Author: James Denvir, Marshall University

OP12: Tracing the Innate Genetic Evolution and Spatial Heterogeneity in Treatment Naïve Lung Cancer Lesions  
Presenting Author: Jihye Kim, University of Colorado Anschutz Medical Campus

OP13: Locating sites of ribonucleotide incorporation in RNase H2-deficient cells  
Presenting Author: Alli Gombolay, Georgia Institute of Technology

OP14: Increasing flexibility of Laboratory Information Management System (LIMS) through extension of relational database model with NoSQL  
Presenting Author: Marcin J. Domagalski, University of Virginia

OP15: Correcting for non-random fragmentation allows for more accurate estimation of relative abundance in NGS metagenomic data.  
Presenting Author: Elmar Pruesse, University of Colorado Denver

OP16: Assembly-free metagenomic analyses on distributed computational platforms  
Presenting Author: Anna Paola Carrieri, IBM Research UK
OP17: Discovering the Contribution of the Gut Microbiome to the Plasma Metabolome
Presenting Author: Michael Shaffer, University of Colorado Denver Anschutz Medical Campus

OP18: Statistical Integration and Feature Selection for Candidate Biomarker Discovery
Presenting Author: Bobbie-Jo Webb-Robertson, Pacific Northwest National Laboratory

OP19: DySE – Dynamic System Explanation Framework, Application in Cancer
Presenting Author: Natasa Miskov-Zivanov, University of Pittsburgh

OP20: A systems biology approach using cell-to-cell heterogeneity to map cell fate decisions in developmental processes
Presenting Author: Jens Preussner, Max Planck Institute for Heart and Lung Research

OP21: Enabling Deep Learning on Structural Biological Data
Presenting Author: Tom Corcoran, Lawrence Berkeley National Laboratory

OP22: Unsupervised discovery of phenotype specific heterogeneous gene networks
Presenting Author: Jenny Shi, University of Colorado Anschutz Medical Campus

OP23: Detecting mutated driver genes from pan-cancer data through a novel matrix factorization based model by learning similarities from pairwise constraints
Presenting Author: Jianing Xi, University of Science and Technology of China

OP24: Automated Biomedical Text Classification with Research Domain Criteria
Presenting Author: Mohammad Anani, Montana State University

OP25: Unsupervised methods for lexical ambiguity in concept normalization
Presenting Author: Negacy Hailu, University of Colorado

OP26: A Comparison of Viral Identification Techniques in Longitudinal Metagenomic Datasets
Presenting Author: Cody Glickman, University of Colorado Anschutz Medical Campus

OP27: A Pilot Systematic Genomic Comparison of Recurrence Risks of Hepatitis B Virus-associated Hepatocellular Carcinoma with low and high degree of liver fibrosis
Presenting Author: Seungyeul Yoo, Icahn School of Medicine at Mount Sinai

OP28: Identifying and Annotating Uninvestigated Preeclampsia-Related Genes Using Linked Open Biomedical Data
Presenting Author: Tiffany Callahan, University of Colorado Denver Anschutz Medical Campus

OP29: Novel Approaches for Accuracy Assessment of 3D Chromatin Configuration Reconstructions
Presenting Author: Mark Segal, University of California San Francisco

OP30: Codon Aversion: An alignment-free method to recover phylogenies
Presenting Author: Justin Miller, Brigham Young University

OP31: Isoform level differential network analysis: A proffer alternative for identifying cancer markers
Presenting Author: Manoj Kandpal, Northwestern University

OP32: Non-B DNA Predictor: A Web Application for Identifying Potential Intramolecular Non-B DNAs
Presenting Author: Hannah Ajoge, Western University
OP33: IMPACT Web Portal: An Oncology Database Integrating Molecular Profiles with Actionable Therapeutics from Next Generation Sequencing Data  
Presenting Author: Jennifer Hintzsche, University of Colorado Anschutz Medical Campus

OP34: Exploring Frequented Regions in Pan-Genomic Graphs  
Presenting Author: Brendan Mumey, Montana State University

OP35: TOBIAS — Transcription factor Occupancy prediction By Investigation of ATAC-seq Signals  
Presenting Author: Mario Looso, Max Planck Institute for Heart and Lung Research

OP36: Integrating Protein Localization Information in Signaling Pathway Reconstructions  
Presenting Author: Anna Ritz, Reed College

OP37: Reducing the footprint of mass spectrometry data  
Presenting Author: Idoia Ochoa, University of Illinois at Urbana-Champaign

OP38: Meeting today’s challenging computational biology demands by integrating a focused IT team into the research process  
Presenting Author: Dan Timmons, University of Colorado Boulder/BioFrontiers Institute

OP39: Predicting Adverse Events Associated with Kinase Inhibitors from Clinical Drug Trials.  
Presenting Author: Ilyssa Summer, University of Colorado Anschutz Medical Campus

OP40: Characterization of the mechanism of drug-drug interactions from PubMed  
Presenting Author: Feng Cheng, University of South Florida
OP01: Comparative Genomics and Alliance for Genome Resources

Presenting Author: Judith Blake, Jackson Laboratory

ABSTRACT: I will report on the emergence of the Alliance for Genome Resources (AGR), a unified approach to bringing together data from the major model organisms to facilitate comparative genomics approaches to understanding human biology and disease. The objective of the Alliance initiative is to develop a common framework for acquiring, curating, and accessing genome data from different model organisms, facilitating utilization of different systems for experimentation, while preserving the details of the unique biology of each model organism. The Alliance, formed in the fall of 2016 under NHGRI guidance, combines work from six model organism databases: (Mouse Genome Database, Saccharomyces Genome Database, WormBase, FlyBase, Zebrafish Information Network, Rat Genome Database) and the Gene Ontology Consortium. With initial release of web portal in October, we demonstrate approaches to data unification. I will report on process of integration, development of standards for data exchange and representation, and next steps and future plans for the Alliance community.

OP02: RSEQREP: RNA-Seq Reports, an open-source cloud-enabled framework for reproducible RNA-Seq data processing, analysis, and result reporting

Presenting Author: Travis Jensen, The Emmes Corporation

ABSTRACT: We present RNA-Seq Reports (RSEQREP), a new open-source cloud-enabled framework for reproducible and scalable RNA-Seq analysis. The software allows users to execute start-to-end gene-level RNA-Seq analysis on a pre-configured RSEQREP Amazon Virtual Machine Image (AMI) hosted by AWS or on their own Ubuntu Linux machine. As input, users can specify unstranded, stranded, and paired-end sequence data stored locally, on AWS S3, or at the Sequence Read Archive (SRA). The software automates a series of customizable steps including human reference genome alignments, reference-based compression of BAM files (CRAM), reference alignment QC, data normalization, multivariate data visualization, identification of differentially expressed genes, heatmaps, co-expressed gene clusters, enriched pathways, and a series of custom visualizations. The strength of the RSEQREP software is its start-to-end open-source solution that combines operating system, bioinformatics software, reference data set
download, data processing, analysis, advanced data visualizations, and automatic reporting. In addition to intermediate files, RSEQREP outputs dynamically generated PDF reports using R, knitR, and LaTeX as well as table and figure files that can readily be integrated into manuscripts. All RSEQREP components are built using open-source software, R code, and Bioconductor packages allowing for further customization. To highlight and exemplify RSEQREP functionality, we provide RSEQREP example results for a publicly available RNA-Seq dataset (GEO ID: GSE45764).

**OP03: Confounds in biomedical natural language processing**

*Presenting Author: Kevin Cohen, University of Colorado School of Medicine*

**ABSTRACT:** Typical machine learning papers, in natural language processing or otherwise, compare two algorithmic approaches and label one as better than the other on the basis of some figure of merit. The assumption is that the algorithmic approach under consideration is the causal factor for the observed difference in performance. However, a reanalysis of the literature on natural language processing is consistent with the hypothesis that “minor” differences in preprocessing are plausible explanations for the performance differences that are usually ascribed to algorithmic differences. This work examines that possibly, focussing on research in the biomedical domain. It focusses on the biomedical domain because the results may be explanatory with respect to understanding the ubiquitous phenomenon of failure of general-domain language processing approaches to generalize to the biomedical domain.

**OP04: Integrated molecular and clinical analysis for understanding human disease relationships**

*Presenting Author: Winston Haynes, Stanford University*

Co-Author(s): Rohit Vashisht, Stanford University; Francesco Vallania, Stanford University; Charles Liu, Stanford University; Gregory Gaskin, Stanford University; Erika Bongen, Stanford University; Shane Lofgren, Stanford University; Timothy Sweeney, Stanford University; Paul J. Utz, Stanford University; Nigam Shah, Stanford University; Purvesh Khatri, Stanford University

**ABSTRACT:** A detailed understanding of relationships among diseases will enable a deeper understanding of disease causation as well as offer opportunities to reposition drugs. To identify unbiased cluster of molecularly and clinically related diseases, we performed gene expression meta-analysis of 104 diseases using 600 studies with 41,000 samples and electronic health record analysis of over two
million patients. Based on molecular data, we observed autoimmune diseases clustering with their specific infectious triggers and brain disorders clustering by disease class. In contrast, clinical data clustered diseases based on clinical practice. Our integrated molecular and clinical analysis spanned vastly different scales to identify robust disease clusters. We identified diseases with under-appreciated, therapeutically actionable relationships in our analysis. We highlighted the relationship between myositis and interstitial cystitis to encourage collaboration by connecting the seemingly disparate research communities.

**OP05: Continuous integration for ensuring interoperability of the Open Biomedical Ontologies**

*Presenting Author: William Baumgartner, University of Colorado Anschutz Medical Campus*

**ABSTRACT:** The modern ecosystem of biomedical ontologies coupled with Semantic Web computational technology has the potential to transform knowledge-based interpretation of biomedical research results, however, the distributed nature of biomedical ontology development and the varying degrees to which quality control measures are used within the community may hinder this potential. This talk will present a continuous integration (CI) system to monitor development of the community-driven Open Biomedical Ontologies (OBOs), a family of ontologies guided by the principles of openness, orthogonality, and interoperability. Our recent work shows that, since the OBOs are developed in a distributed and loosely coupled way, unintentional inter-ontology conflicts and logical inconsistencies can arise and persist despite being able to be detected automatically. The potential for inter-ontology conflict has become more prevalent as many OBOs now define their concepts logically, using concepts spanning other ontologies. Such logical conflicts hamper efforts to reason over this base of knowledge as a whole and should alarm the ontology developer and user communities. Further, as the OBOs are under continuous development new conflicts are sure to arise. CI plays an integral quality assurance role when engineering software by monitoring code for changes, executing tests whenever change is detected, and providing feedback for errors induced by change. We will present data supporting the need for CI over the OBOs to ensure their global interoperability, and will introduce to the ontology developer and user communities a soon-to-be deployed system capable of continuously monitoring the OBOs for logical consistency and ontology development best practices.
OP06: Classification of glioblastoma subtypes by integrating genomic and histopathological image features with probabilistic graphical models

Presenting Author: Dimitris Manatakis, University of Pittsburgh

Co-Author(s): Panayiotis Benos, University of Pittsburgh; Akif Tosun, University of Pittsburgh; Chakra Chennubhotla, University of Pittsburgh

ABSTRACT: Integrating multi-modal biomedical data types under the same analytic framework is an important step towards harvesting the existing, fragmented knowledge that is collected in a clinical setting. H&E staining has been used extensively in pathology for diagnosing various diseases or disease subtypes, but it has limitations. Computational pathology, i.e. the processing and analysis of H&E stained images with computational methods augmented by high-throughput data collection from the same patients aims in improving disease diagnosis by combining multiple, complementary sources of information. Currently, the methods for integrating clinical image and –omics features have used mainly second order methods. These methods, however, cannot properly capture the complexity of the direct and indirect relations between these features. In this paper, we present an alternative method for data integration: the use of directed graphs, a form of probabilistic graphical models. Using new methods we have developed to (1) parse tissue heterogeneity information from H&E images and (2) learn directed graphs over mixed data types, we test the hypothesis that there is a relationship between gene expression and spatial information in the tissue organization in glioblastoma.

OP07: MPEG-G the emerging standard for genomic data compression

Presenting Author: Mikel Hernaez, University of Illinois

Co-Author(s): Claudio Alberti, GenomSys; Marco Mattavelli, EPFL; Idoia Ochoa, University of Illinois

ABSTRACT: The development of Next Generation Sequencing technologies might enable the usage of genomic information as everyday practice in public health, but the large volumes of raw data generated becomes a serious obstacle for its wide diffusion. The lack of an appropriate representation for compressed genomic data is widely recognized as a critical element limiting its application potential. Beside efficient compression, which is an essential element for any usage of genomic information, several other functionality that the current data formats do not support are of paramount importance.

Two ISO committees, TC 276 and JTC 1/SC 29/WG 11 (MPEG), have combined their efforts and are jointly working to develop MPEG-G, a
new compression standard for genomic sequencing data aiming at providing new effective solutions for genomic information processing applications. Efficient compression, selective data access and queries, support for data streaming and data protection are among the new functionality natively provided in the compressed domain.

The different elements composing the MPEG-G standard (transport, compression technology, APIs, conformance and reference SW) intend to provide the standard framework that will enable stakeholders to exploit to its full potential the large body of genomic data that the current sequencing technology is generating.

This paper provides a summary of the components, main innovations, functionality and performance of the current version of the MPEG-G technology currently undergoing the final steps of the ISO standardization process.

OP08: Physical aspects of the α-helix and their implications towards amyloidal tendency

Presenting Author: Simcha Srebnik, Technion – Israel Institute of Technology

ABSTRACT: α-Helices are the most abundant structures found within proteins and play an important role in the determination of the global structure of proteins and their function. It is common to describe protein structure using Ramachandran (φ,ψ) dihedrals, which reveal a diagonally aligned ellipsoidal region of the α-helices on the Ramachandran Map. We show that an alternative orthogonal coordinates system can be used to describe the helical conformation in terms of physical parameters, i.e., the number of residues per turn (ρ) and the angle (ϑ) between backbone carbonyls relative to the helix direction through a linear transformation between the two coordinates system (φ,ψ and ρ,ϑ). When described in this way, we observe a direct correlation between the physical interpretation of the α-helical structure and its tendency for amyloid formation. We conclude with qualitative scenarios that may affect the helical structure and induce modifications to ρ and ϑ.

OP09: A Novel Systematic Analysis of ALDH Isozyme Specificity in Head and Neck Squamous Cell Carcinoma

Presenting Author: Brian Jackson, University of Colorado Anschutz Medical Campus

Co-Author(s): AC Tan, University of Colorado Anschutz Medical Campus; Antonio Jimeno, University of Colorado Anschutz Medical Campus

ABSTRACT: Head and Neck Squamous Cell Carcinoma (HNSCC) is among the top 10 diagnosed cancers in the United States, and is one of the few cancer that has increased in incidence over the last 10
years. It has recently been elucidated that Aldehyde Dehydrogenase (ALDH) superfamily members play important roles in cancer, and in particular ALDH1 family members are implicated as markers and effectors of stemness in HNSCC. The ALDH superfamily consists of 19 members in humans that catalyze the conversion of an aldehyde to its corresponding carboxylic acid. This enzyme superfamily is responsible for the detoxification of a large number of endogenous and exogenous substrates. Interest in inhibition of ALDH members has increased in recent years, but ALDH isozymes have overlapping but often distinct substrates and roles in human physiology. We attempt to address these problems by 1) building a comprehensive database of ALDH substrates and inhibitors and 2) systematic computational modeling of ALDH isozymes to better predict the action and specificity of novel ALDH substrates and inhibitors. With these tools, future efforts at specific ALDH isozyme inhibition will be more efficiently realized.

**OP10: Building a Systems-level model of the immune response to Salmonella infection**

*Presenting Author: Marta Andres-Terre, Stanford University*

*Co-Author(s): Adityia Rao, Stanford University; Michele Donato, Stanford University; Purvesh Khatri, Stanford University*

**ABSTRACT:** Infectious diseases are the result of a molecular warfare between the host immune system and the pathogen. Their treatment and eradication are complicated by the heterogeneous nature of these interactions, which remains poorly understood. Here, we have identified a set of genetic and molecular determinants that characterize the host immune response to bacterial infection. First, we conducted an integrated multi-cohort analysis of publicly available gene expression data and identified a common host gene signature across different bacterial infections. This meta-bacterial signature can 1) distinguish bacterial from viral and fungal infections, and 2) predict symptom onset and disease outcome in infected individuals. We then identified a Salmonella-specific host-response gene signature, which can be used as a prognostic marker for typhoid fever, as well as for understanding the biology of Salmonella infections. Second, we applied cell mixture deconvolution to the same datasets we used to obtain the gene expression signature and estimated the cellular populations driving the response to Salmonella infection. We are currently working on building a quantitative model of bacterial infections in which we take into account both the cellular and gene expression signatures identified using heterogeneous data sources. Defining the metrics that characterize the immune response to bacterial infection will go beyond the concept of identifying biomarkers, as this model could potentially be used as a platform to identify and understand novel mechanisms underlying host-pathogen interactions.
OP11: Development of a statistical model of the genetics of maternal-fetal dyads in neonatal abstinence syndrome

Presenting Author: James Denvir, Marshall University

Co-Author(s): Richard Egleton, Marshall University; Don Primerano, Marshall University; Jun Fan, Marshall University; Vincent Setola, West Virginia University; Laura Lander, West Virginia University

ABSTRACT: As a consequence of the dramatic increase in opioid use in the US in the last decade, a large increase in the number of infants born with Neonatal Abstinence Syndrome (NAS) has been observed. Opioids entering the mother’s bloodstream during pregnancy are metabolized at three primary sites before having the opportunity to influence brain function in the developing fetus: the maternal liver, the placenta, and the fetal liver. We are in the process of performing whole exome sequencing on mother-infant dyads in order to identify genetic variants in either the mother or infant that may be predictors of NAS severity or prognosis, and may inform treatment plans for the neonate. Here, we present the development of a genetic model that regards the mother-infant dyad as a single, quasi-tetraploid, entity that may be used for hypothesis testing in this context.

OP12: Tracing the Innate Genetic Evolution and Spatial Heterogeneity in Treatment Naïve Lung Cancer Lesions

Presenting Author: Jihye Kim, University of Colorado Anschutz Medical Campus

Co-Author(s): Kenichi Suda, Kindai University Faculty of Medicine; Isao Murakami, Higashi-Hiroshima Medical Center; Leslie Rozeboom, University of Colorado Anschutz Medical Campus; Christopher J. Rivard, University of Colorado Anschutz Medical Campus; Tetsuya Mitsudomi, Kindai University Faculty of Medicine; Fred R. Hirsch, University of Colorado Anschutz Medical Campus; Aik-Choon Tan, University of Colorado Anschutz Medical Campus

ABSTRACT: Extensive intratumor heterogeneity (ITH) have been observed in individual patient tumors by large-scale sequencing analyses. ITH can contribute to drug resistance and cancer metastasis. Distinct microenvironments provide selection advantages to sub-population of cells for cancer metastasis. We hypothesized that ITH can contribute to metastasis by disseminating different sub-populations of cancer cells to distant sites. We collected tumor specimens and non-cancer tissues from treatment naïve autopsied patients to study the innate genetic evolution and spatial heterogeneity. Our cohort consists of four (two adenocarcinoma and two squamous cell carcinoma) NSCLC patients and one SCLC patient. Each patient had 5–9 primary and metastatic lesions. Comprehensive data analyses were performed on the RNA-seq that include gene expression and pathway analyses, fusion detection and somatic variants detection. Global unsupervised
clustering of expression data reveals the NSCLC patients clustered together from the SCLC, and the two adenocarcinoma and squamous cell carcinoma patients formed two clusters. Within each patient, metastatic lesions clustered according to the distant metastatic sites. Pathway analysis and somatic mutation analysis in individual patients revealed that, in general, the primary lesion is distinct from metastatic lesions in NSCLCs. For the SCLC patient, distant metastases and lymph node metastases clustered according to different parts of the primary tumors. We also identified KIF5B-RET fusion as a founder mutation in all tumor specimens obtained from a never-smoking adenocarcinoma patient. This study provides evidence that ITH contributes to distant metastasis sites based on the similarity and the heterogeneity between primary and metastatic lesions in lung cancer patients.

OP13: Locating sites of ribonucleotide incorporation in RNase H2-deficient cells

Presenting Author: Alli Gombolay, Georgia Institute of Technology

Co-Author(s): Francesca Storici, Georgia Institute of Technology; Fredrik Vannberg, Georgia Institute of Technology

ABSTRACT: Ribonucleoside monophosphates (ribonucleotides or rNMPs) that are inadvertently incorporated into DNA can wreak havoc on genome stability. Under normal cellular conditions, the RNase H2 enzyme initiates the removal of rNMPs by efficiently cleaving these toxic nucleotides. However, when left unrepaired, rNMPs can cause breaks in the DNA strand, replication stress, and spontaneous mutagenesis. To understand the biological consequences of rNMPs and their role in the pathogenesis of disease, we must profile the distribution of rNMPs in the genome. Determining where rNMPs are differentially incorporated will allow us to identify how rNMPs cause genome instability. Recent advances in laboratory techniques and computational methods provide the unique opportunity to capture these non-standard nucleotides and map their locations in the genome. One of these techniques is ribose-seq (Koh et al. Nature Methods 2015). In contrast to other techniques, ribose-seq directly captures rNMPs embedded in DNA and may be applied to any cell type at any stage of the cell cycle. Achieving the full potential of ribose-seq is dependent upon computational methods tailored to analyzing this type of data. The Ribose-Map toolkit is a novel collection of user-friendly, open-source, and well-documented scripts developed to profile the incorporation of rNMPs captured via ribose-seq. Ribose-Map allows the user to determine the genomic coordinates of rNMPs, calculate nucleotide frequencies, locate rNMP genomic hotspots, and create publication-ready figures. By exploring the location and distribution of rNMPs in RNase H2-deficient cells, we may begin to understand the role rNMPs play in genome instability and, ultimately, disease.
OP14: Increasing flexibility of Laboratory Information Management System (LIMS) through extension of relational database model with NoSQL

Presenting Author: Marcin J. Domagalski, University of Virginia

Co-Author(s): Przemyslaw Porebski, University of Virginia; David Cooper, University of Virginia; Marcin Cymborowski, University of Virginia; Heping Zheng, University of Virginia; Marek Grabowski, University of Virginia; Wladek Minor, University of Virginia

**ABSTRACT:** LabDB is a modular, specialized super-LIMS, originally developed to track the macromolecular structure determination pipeline from cloning to structure solution (Zimmerman et al. 2014, Methods Mol. Biol. 1140). The system has modules to automatically or semi-automatically import data from a variety of different types of laboratory equipment, including chromatography and electrophoresis systems, crystallization observation robots, isothermal titration calorimetry, and others. Here we present a novel abstract data model for a new generation of LabDB LIMS, designed to address the extreme complexity of biology workflows. A purely relational database model was reduced to minimum number of database tables capable of storing experiment specific data structures and their predicate schemas (definitions) in JavaScript Object Notation (JSON) form. The model is suitable for extension with programmatic plugins as well as for manual definition of samples and workflows through user interface. Its “generic” design limits the structural requirements for stored data and provides better usability and higher performance. The data is accessible using a representational state transfer (REST) API, which provides easy interoperability with other software systems. In conclusion, LabDB flexible schema can integrate instruments and manage various experimental samples and workflows from end-to-end, ensuring traceable results, reproducibility of experiments, and last but not least improving overall lab efficiency.

OP15: Correcting for non-random fragmentation allows for more accurate estimation of relative abundance in NGS metagenomic data.

Presenting Author: Elmar Pruesse, University of Colorado Denver

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

**ABSTRACT:** Accurately estimating the relative abundances of genes, contigs or entire genomes is essential to metagenomic studies. Variation in contig abundance between samples is used for de-novo in-silico organism isolation (binning); per genome or gene abundances allow comparative microbial community studies to establish associations of genes or microbes with disease states or environmental factors. However, current methods are surprisingly naive. Using simple
averaging, most tools assume that reads are randomly sampled from the source DNA material.

This assumption, while acceptable in the past, is broken by modern NGS sequencing methods. Tagmerase based library preparation protocols such as Nextera exhibit distinct biases in fragmentation point distribution. Sonication based fragmentation protocols exhibit less, but still noticeable, bias. These biases result in read depth variation at gene-scale windows far greater than to be expected from a Poisson distribution. This inhibits the binning of small contigs which often comprise the majority of an assembly. Moreover, the association of bias with DNA patterns potentially distorts relative abundances within gene families, which can lead to false positives or false negatives in comparative analyses.

We further present an algorithm for abundance estimation in the presence of non-random fragmentation. By learning fragmentation point preferences of the employed library preparation protocols from all mapped reads, the read depths can be adjusted for the positional likelihood of observing a read. Positions at which no fragmentation was observed are accounted for by incorporating the ratio of expected and observed zero observation positions. Overall, our Cython implementation demonstrates far lower variance in predicted abundance at small (1kb) window sizes than naive approaches, compensating for the effects of non-random fragmentation nearly completely.

**OP16: Assembly-free metagenomic analyses on distributed computational platforms**

*Presenting Author: Anna Paola Carrieri, IBM Research UK*

*Co-Author(s): Philippe Gambron, Science Technology Facilities Council; Martyn Winn, Science and Technology Facilities Council; Neil Venables, Cancer Research UK Manchester Institute; Vipin Sachdeva, Silicon Therapeutics; Will Rowe, Science and Technology Facilities Council; Kirk Jordan, IBM Research*

**ABSTRACT:** Recent advances in high-throughput sequencing technologies have enabled the characterization and comparison of microbial communities in very diverse environments, such as the human gut microbiome. Characterizing microbial function and composition across different individuals and diseases is important to understand the role of microbiota in disease development. Consider comparing metagenomes of healthy and disease subjects, or pre-and-post antibiotic treatment microbial communities. Current taxonomic classification methods focus on sequencing of specific marker genes, such as 16S rRNA, and rely on existing microbial reference databases, often incomplete. A more informative method is whole-metagenome shotgun sequencing which generates huge collections of short reads. The use of larger reference databases and/or the need to assemble reads makes whole-metagenome analysis both data and computation intensive. We present
a high performance computing (HPC) tool which generates distributed k-mer spectra from metagenomic sequences and facilitates assembly free metagenomics analyses. K-mer spectra are compressed representations of metagenomic reads that allow comparisons of metagenomes including the entire data volume into analysis. The HPC k-mer library rapidly estimates pairwise dissimilarities between metagenomes and accelerates the metagenomic analyses exploiting distributed computational platforms. The computed dissimilarity measures can be applied for cluster analysis and classification, for differentiating between phenotype groups, for inferring features of metagenomic composition and for phylogenetic reconstruction from metagenomes. The latter problems are computationally expensive therefore the use of high performance computing can overcome limitations in time and space.

**OP17: Discovering the Contribution of the Gut Microbiome to the Plasma Metabolome**

*Presenting Author: Michael Shaffer, University of Colorado Denver Anschutz Medical Campus*

*Co-Author(s): Catherine Lozupone, University of Colorado Denver Anschutz Medical Campus; Nichole Reisdorph, University of Colorado Denver Anschutz Medical Campus*

**ABSTRACT:** The human body contains roughly the same number of bacterial cells as human cells and these bacteria encode for 150 times more genes. These microbes primarily live in the gut, produce metabolites that are transported all over the human body and have potential to influence disease. While untargeted metabolomics can be used to investigate influence of microbial metabolites, determining whether disease-associated metabolites come from microbes, the host or the environment can be challenging. We have developed methods that use the KEGG and PICRUSt databases to predict the origin of metabolites and applied them to a set of 54 paired plasma metabolome and stool 16S microbiome samples. Using human KEGG genes we predicted 1376 unique compounds could be produced by the host and using PICRUSt and KEGG we predicted 1321 compounds produced by the detected gut microbial community. Of 1018 KEGG annotated compounds in the plasma metabolome, 155 were predicted human metabolites and 135 were bacterial. This cohort contains individuals with HIV and with HIV and lipodystrophy (a metabolic comorbidity); of compounds that differed with lipodystrophy 2 were predicted to only be produced by bacteria, 11 only by humans and 11 by both. However, the majority of metabolites in the plasma metabolome could not be assigned to KEGG IDs, limiting these techniques to only a small subset of metabolites. Our results highlight both the promise and challenges of using metabolic networks to predict bacterial origin of metabolites.
OP18: Statistical Integration and Feature Selection for Candidate Biomarker Discovery

Presenting Author: Bobbie-Jo Webb-Robertson, Pacific Northwest National Laboratory

Co-Author(s): Lisa Bramer, Pacific Northwest National Laboratory; Sarah Reehl, Pacific Northwest National Laboratory; Brigette Frohnert, University of Colorado; Marian Rewers, University of Colorado

ABSTRACT: High-throughput technologies currently have the capability to capture information at both global and targeted scales for the transcriptome, proteome, metabolome, and lipidome as well as determining functional aspects of these biomolecules. The promise of data integration is that by utilizing these disparate data streams, in combination with low-throughput clinical information, a more complete or accurate estimate of system behavior can be obtained. In the case of biomarker discovery to better diagnose and predict outcomes of disease one goal is to identify the best subset of molecules that can separate specific phenotypes of interest. However, in a space of tens of thousands of variables (e.g., genes, proteins), feature selection approaches often yield over-trained models with poor predictive power. Moreover, feature selection algorithms are typically focused on single sources of information and do not evaluate the effect on downstream statistical integration models. We present an ensemble-based feature selection approach that optimizes the outcome of interest in the context of the integrated posterior probability. We demonstrate that this approach improves sensitivity and specificity over simple selection routines based on individual datasets and present the application of the approach on a juvenile diabetes cohort.

OP19: DySE – Dynamic System Explanation Framework, Application in Cancer

Presenting Author: Natasa Miskov-Zivanov, University of Pittsburgh

ABSTRACT: Biomedical research results are being published at a high rate, and with existing search engines, the vast amount of published work is usually easily accessible. To accurately reuse this voluminous knowledge that is fragmented and sometimes inconsistent, one can extract and assemble published information into models. However, the creation of models often relies on intense human effort: model developers have to read hundreds of published papers, and conduct discussions with domain experts. This laborious process results in slow development of models, as it includes steps such as model validation with experimental data and model extension with new available information. To automate the process of explaining biological observations and answering biological questions, we have built a Dynamic System Explanation (DySE) framework. DySE automatically assembles models, extends them with the interactions extracted by
automated reading engines, analyzes the models for various conditions and scenarios, and iteratively improves assembled models. The framework includes techniques such as stochastic simulation, statistical model checking, static and dynamic sensitivity analysis, and hypothesis generation. With the automated process of reading, assembly, and reasoning, our framework allows for rapidly designing and conducting thousands of in silico experiments, and thus, can speed up discoveries from the order of decades to the order of hours or days. We have applied DySE to studying cancer microenvironment, as well as on explaining mechanisms of several melanoma drugs. The techniques and the modeling approach incorporated within the framework are not disease specific, and therefore, DySE can be used for explaining systems in many other domains.

**OP20: A systems biology approach using cell-to-cell heterogeneity to map cell fate decisions in developmental processes**

*Presenting Author: Jens Preussner, Max Planck Institute for Heart and Lung Research*

*Co-Author(s): Guangshuai Jia, Max Planck Institute for Heart and Lung Research; Stefan Guenther, Max Planck Institute for Heart and Lung Research; Xuejun Yuan, Max Planck Institute for Heart and Lung Research; Michael Yekelchyk, Max Planck Institute for Heart and Lung Research; Carsten Kuenne, Max Planck Institute for Heart and Lung Research; Mario Looso, Max Planck Institute for Heart and Lung Research; Yonggang Zhou, Max Planck Institute for Heart and Lung Research; Thomas Braun, Max Planck Institute for Heart and Lung Research*

**ABSTRACT:** Challenges for understanding developmental processes are the characterization of cell fate transitions, specification and hierarchical lineage descendants of progenitor cells. Differentiating cell populations typically manifest in profound cell-to-cell heterogeneity, justified by gene expression changes and gradual remodeling of underlying gene regulatory networks (GRNs). RNA sequencing at the single cell level allows the comprehensive characterization and analysis of cell population heterogeneity, especially if large ensembles of identical systems are profiled. Here we present computational approaches to harness statistical properties of heterogeneity with the objective of in-silico reconstruction of lineage trajectories and cell fate decision mapping in ~1500 differentiating cardiac progenitor cells. We define suitable measures of cell-to-cell heterogeneity and employ self-organizing maps and a density-based, hierarchical clustering to identify heterogeneous sub-populations. Reconstruction of the developmental trajectory revealed a bifurcation event when cells segregated into their terminal fates. Next, we show how a correlation-based analysis of cells in a transiently unstable state can be used to identify key regulatory elements in the underlying gene regulatory network that drive differentiation. We additionally show that those determinants progressively overcome epigenetic barriers to achieve open chromatin states associated with elevated expression in differentiated cell
populations. Our approach not only comprehensively exploits heterogeneity and emerging correlations among large numbers of cells and genes to study early cardiogenesis at single-cell resolution, but also sets a general systems biology framework of transcriptional and epigenetic regulations in cell fate decisions.

**OP21: Enabling Deep Learning on Structural Biological Data**

*Presenting Author: Tom Corcoran, Lawrence Berkeley National Laboratory*

*Co-Author(s): Rafael Zamora, Lawrence Berkeley National Laboratory*

**ABSTRACT:** Convolutional Neural Network (CNN)-based machine learning has made noticeable breakthroughs in feature extraction tasks, but its applications in protein research are constrained by limited training data availability, as well as by the disparity between the 2D-oriented functionality of mainstream CNN technology and the inherently 3D nature of protein structures. We present a mapping algorithm that converts 3D structures to 2D data grids by first traversing the 3D space with a space-filling curve, encoding the 3D structural information into a 1D vector, and then projecting that vector into 2D via a reverse process with a complementary curve. For comparison against state-of-the-art CNN-based classification methods, we explore the performance of 2D CNNs trained on data encoded with our method versus comparable volumetric CNNs operating upon raw 3D data. Our results indicate that our mapping process preserves sufficient locality information across the transformation to be useful for training 2D CNNs on classification tasks between proteins and other structural models. We show that 2D CNNs trained on data generated using our method perform equivalently to state-of-the-art volumetric methods in terms of accuracy and generalizability while offering the potential for decreased training time cost compared to their 3D counterparts. We discuss several experiments that show the effectiveness of our approach, including classifying between everyday 3D object models from the ModelNet10 benchmarking dataset and between protein models sourced from the Protein Data Bank such as KRas and HRas. An implementation of our encoding process and neural network architectures is available for download at GitHub.

**OP22: Unsupervised discovery of phenotype specific heterogeneous gene networks**

*Presenting Author: Jenny Shi, University of Colorado Anschutz Medical Campus*

*Co-Author(s): Pamela Russell, University of Colorado Anschutz Medical Campus; Pratyaydipta Rudra, University of Colorado Anschutz Medical Campus; Brian Vestal, National Jewish Health; Brian Hobbs, Brigham and Women’s Hospital; Craig Hersh, Brigham and Women’s Hospital; Laura Saba, University of Colorado Anschutz Medical Campus; Katerina Kechris, University of Colorado Anschutz Medical Campus*
ABSTRACT: Complex diseases often have a wide spectrum of symptoms. Better understanding the biological mechanism behind each symptom (a.k.a. phenotype) promotes targeted and effective treatment plans. We propose to utilize a machine learning technique, sparse canonical correlation analysis (SCCA), to integrate messenger RNA (mRNA) and microRNA (miRNA) expression data, taking a phenotype of interest into account. With canonical weights, mRNA-miRNA (i.e. heterogeneous) subnetworks that are specific to the phenotype can be constructed. Unlike traditional pairwise target prediction, the SCCA approach allows identification of associations that can be missed based on marginal correlations. We applied the method to a recombinant inbred mouse panel with endophenotypes that are associated with alcohol use disorders in human, and constructed heterogeneous subnetworks that are specific to drinking behavior. The leading subnetwork identified included three mRNA and four miRNAs, including two miRNAs from the same family. Most of the SCCA detected network connections were not predicted using pairwise methods, yielding novel associations. The strong associations discovered will be validated through biological experiments, such as knock-out. We also applied the method to a chronic obstructive pulmonary disease pilot study. The preliminary results revealed three subnetworks, which contain candidate features for more focused studies. The proposed SCCA method is not limited to expression data. It can be easily generalized to other data types, such as copy number variation, etc. It can also be applied to integrating more than three data types. The versatility of the approach will be useful in many other applications.

OP23: Detecting mutated driver genes from pan-cancer data through a novel matrix factorization based model by learning similarities from pairwise constraints

Presenting Author: Jianing Xi, University of Science and Technology of China
Co-Author(s): Ao Li, University of Science and Technology of China; Minghui Wang, University of Science and Technology of China

ABSTRACT: Identifying cancer-causing mutated driver genes from passenger mutations is crucial to enhance the development of cancer diagnostics and therapeutics, and many previous efforts have been undertaken to identify driver genes from mutation data of specific types of cancers. However, many driver genes are underestimated when the mutation data of only specific cancers are investigated, which complicates the understanding of tumorigenesis. According to recent studies, cancers of disparate organs have many shared genomic mutations, and some driver genes that are not highly frequently mutated
in patients of one cancer type may display considerable mutation frequencies across patients of multiple cancer types. By taking into account both the similarities of mutation profiles of different cancer types and the information of gene interaction network, we detect driver genes from pan-cancer data through a novel unsupervised learning model based on matrix tri-factorization by learning the similarities from pairwise constraints. In the evaluation of known benchmarking genes, the detection performance of our model is largely increased when compared with existing methods. Furthermore, our model achieve better performance than those of the existing matrix factorization based methods which do not consider the pairwise similarites between cancers. Moreover, our model discovers some novel driver gene candidates that have been reported in recent published studies, which show great potential for further wet experimental verification.

**OP24: Automated Biomedical Text Classification with Research Domain Criteria**

*Presenting Author: Mohammad Anani, Montana State University*

**ABSTRACT:** Research Domain Criteria (RDoC) is a recently introduced framework for accurate diagnosis of mental illness. This framework contains five domains of interest, where each domain contains a number of constructs that define a specific behavior. Developing a method to automate the process of labeling biomedical articles with RDoC constructs is highly useful to advance research efforts in the area of mental illness. Therefore, this study focuses on exploring the feasibility of developing a tool for this purpose. Using a gold-standard dataset of about 40,000 Medline abstracts tagged with 26 RDoC constructs, we model this both as binary and multilabel classification problem, to perform document classification using several supervised learning algorithms. We use a simple Bag-of-Words model and apply standard preprocessing steps such as stemming and stop words removal. According to AUROC (Area Under Receiver Operating Characteristic curve) values obtained through 5-fold cross validation, we observe that overall, Artificial Neural Networks and Support Vector Machines perform best on the multilabel problem providing 96% average AUROC across all the constructs. Interestingly, all the binary classifiers showed the same level of performance. However, the cohort of binary classifiers took significantly longer time to train compared to their multilabel counterparts, showing the utility of modeling this as a multi-label problem. We also note that the articles labeled with more specific constructs were predicted better than the rest of the constructs. To the best of our knowledge, this is the first study on automated prediction of RDoC constructs for biomedical literature.
OP25: Unsupervised methods for lexical ambiguity in concept normalization

Presenting Author: Negacy Hailu, University of Colorado

Co-Author(s): Mike Bada, University of Colorado; William A. Baumgartner Jr, University of Colorado; Kevin Bretonnel Cohen, University of Colorado; Lawrence E. Hunter, University of Colorado

ABSTRACT: Background • Previous studies have argued that ConceptMapper, which is a dictionary-based look up concept normalization tool, is the best performing system. The big limitation of dictionary-based look up systems is that they do not correctly map lexically ambiguous terms. In this work, we investigate various methods that will tackle lexical ambiguity in the task of concept normalization. We specifically apply part-of-speech (POS) tag information, Word Sense Disambiguation (WSD) and word embeddings to remove false positive annotations of ConceptMapper. We used the CRAFT corpus that is manually annotated for eight biomedical ontologies for evaluation. The ontologies are Cell Ontology (CL), Gene Ontology (biological process (GOBP), cellular component (GOCC), and molecular function (GOMF)), Sequence Ontology, Chemical Entities of Biological Interest (CHEBI), NCBI Taxonomy and Protein Ontology (PR).

Results • Word embeddings improved performance of GOBP, SO, CHEBI, and PR; and POS information improved performance of SO, CHEBI and NCBI Taxonomy on F1 measure. But, WSD didn’t bring any improvement on F1 measure. All the three methods achieved improvement on precision for all the ontologies. Our evaluations are on a five fold cross validation. All our improvements are statistically significant.

Conclusions • We propose three unsupervised methods to solve the problem of lexical ambiguity on concept normalization. The performance of the proposed methods varies depending on the ontology. Also, it can be tuned depending the need (whether high precision or high F1 measure). In general word embeddings achieved best results.

OP26: A Comparison of Viral Identification Techniques in Longitudinal Metagenomic Datasets

Presenting Author: Cody Glickman, University of Colorado Anschutz Medical Campus

ABSTRACT: Viruses that influence bacterial community compositions, also known as bacteriophages, are commonly small genomes that lack a universally conserved genetic marker. As a result, bacteriophages make up a small proportion of reads in traditional DNA shotgun metagenomic experiments. Biological methods to enrich for virus-like particles (VLPs) suffer from the biases of filtering for free floating elements, thus limiting the recovery of endogenous viral elements. The importance of endogenous viral elements in the life cycles and
adaptation of pathogenic bacteria is well known (Schuch et al. 2009-Aug-12; van et al. 2017-Sep-01). To capture novel endogenous viral elements in DNA shotgun metagenomic data, I propose a novel methodology to extract and pool viral reads across longitudinal patient samples. I propose that this methodology will produce longer viral read assemblies and more accurate taxonomic assignments than assembly of individual samples and full cross-sample assembly.

To test the performance of viral read assembly across time, I performed a simulation study using bacteria with endogenous viral elements. The study tests the lengths of viral read assemblies and the accuracy of the methodologies to recapitulate the elements in the synthetic dataset. In addition, I performed a synthetic spike-in on a real longitudinal metagenomic dataset with rare viral species to measure the sensitivity of the methodologies within noisy data. These simulations are a benchmark for mining viral elements from longitudinal datasets in the publicly available databases. Mining viral reads from metagenomic experiments allows researchers to study endogenous viral elements not typically found in viral enrichment studies.

**OP27: A Pilot Systematic Genomic Comparison of Recurrence Risks of Hepatitis B Virus-associated Hepatocellular Carcinoma with low and high degree of liver fibrosis**

*Presenting Author: Seungyeul Yoo, Icahn School of Medicine at Mount Sinai*

*Co-Author(s): Wenhui Wang, Icahn School of Medicine at Mount Sinai; Qin Wang, Icahn School of Medicine at Mount Sinai; Maria Fiel, Icahn School of Medicine at Mount Sinai; Eunjee Lee, Icahn School of Medicine at Mount Sinai; Spiros Hiotis, Icahn School of Medicine at Mount Sinai; Jun Zhu, Icahn School of Medicine at Mount Sinai*

*ABSTRACT: Chronic Hepatitis B virus (HBV) infection leads to liver fibrosis which is a major risk factor in Hepatocellular carcinoma (HCC). HBV genome can be inserted into human genome, and chronic inflammation may trigger somatic mutations. However, how HBV integration and other genomic changes contribute to the risk of tumor recurrence with regard to different degree of liver fibrosis is not clearly understood. We performed comprehensive genomic analyses of our RNAseq data from 21 HBV-HCC patients treated at Mount Sinai Medical Center, and public available HBV-HCC sequencing data. Using a robust pipeline we developed, consistently more HBV integrations were identified in non-neoplastic liver than in tumor tissues. HBV host genes identified in non-neoplastic liver tissues significantly overlapped with known tumor suppressor genes. More significant enrichment of tumor suppressor genes was observed among HBV host genes identified from patients with tumor recurrence, indicating potential risk of tumor recurrence driven by HBV integration in non-neoplastic liver tissues. Pathogenic SNP loads in non-neoplastic liver were consistently higher.
than ones in normal liver tissues. And HBV host genes identified in non-neoplastic liver tissues significantly overlapped with pathogenic somatic mutations, suggesting HBV integration and somatic mutations targeting the same set of genes important to tumorigenesis. HBV integrations and pathogenic mutations showed distinct patterns between low and high liver fibrosis patients with regard to tumor recurrence. The results suggest that HBV integrations and pathogenic SNPs in non-neoplastic tissues are important for tumorigenesis and different recurrence risk models are needed for patients of low and high liver fibrosis.

**OP28: Identifying and Annotating Uninvestigated Preeclampsia-Related Genes Using Linked Open Biomedical Data**

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

*Co-Author(s): Adrienne Stefanski, University of Colorado Denver Anschutz Medical Campus; William Baumgartner Jr, University of Colorado Denver Anschutz Medical Campus; Jin-Dong Kim, Database Center for Life Science; Toyofumi Fujiwara, Database Center for Life Science; Ann Cirincione, University of Maryland; Maricel Kann, University of Maryland; Lawrence Hunter, University of Colorado Denver Anschutz Medical Campus*

**ABSTRACT:** Preeclampsia is a leading cause of maternal and fetal morbidity and mortality. Currently, there is no cure for preeclampsia except delivery of the placenta, which is central to preeclampsia pathogenesis. Transcriptional profiling of human placenta from pregnancies complicated by preeclampsia and control has been extensively performed to identify differentially expressed genes (DEGs). DEGs are identified using unbiased assays, however the decisions to investigate DEGs experimentally are biased by many factors (e.g., investigator interests, available reagents, knowledge of gene function) causing many DEGs to remain uninvestigated. To address this shortcoming, we utilized existing linked open biomedical resources and publicly available high-throughput transcriptional profiling data to identify and annotate the function of currently uninvestigated preeclampsia-associated DEGs. Using the keyword “preeclampsia”, we identified and reviewed 68 publicly available human gene expression experiments deposited in the Gene Expression Omnibus. Meta-analysis of the 13 experiments meeting our inclusion criteria generated a list of 273 DEGs. We annotated these genes using a knowledge graph constructed with Semantic Web Technologies that contained several Open Biomedical Ontologies and publicly available datasets. The relative complement of the annotation-derived and meta-analysis derived genes were identified as the uninvestigated preeclampsia-associated genes. Experimentally investigated DEGs were then identified from published literature based on semantic and syntactic annotations to PubMed abstracts by PubTator and PubAnnotation.
Finally, novel biological relationships between experimentally investigated and uninvestigated preeclampsia-associated genes were identified by learning neuro-symbolic logic embeddings to predict missing edges in the knowledge graph. Detailed documentation and source code can be found on GitHub (https://github.com/callahantiff/ignorenet/wiki).

**OP29: Novel Approaches for Accuracy Assessment of 3D Chromatin Configuration Reconstructions**

*Presenting Author: Mark Segal, University of California San Francisco*

**ABSTRACT:** Assessing accuracy of 3D chromatin configurations inferred from HiC data is crucial. Such evaluation is challenging given the absence of gold standards and the many competing algorithms. Using recent multiplexed FISH and genomic architecture mapping assays we devise novel approaches to accuracy assessment, with rigorous attendant inferential methods.

**OP30: Codon Aversion: An alignment-free method to recover phylogenies**

*Presenting Author: Justin Miller, Brigham Young University*

*Co-Author(s): Perry Ridge, Brigham Young University; Michael Whiting, Brigham Young University; Lauren McKinnon, Brigham Young University*

**ABSTRACT:** Codon bias refers to the non-random usage of synonymous codons, and differs between organisms, between genes, and even within a gene. We previously identified a strong phylogenetic signal, based on codon usage preferences, in 72 tetrapod species, focusing on stop codon usage preferences. Here we report the expansion of our previous work into >20,000 species across all kingdoms of life, and the development of tools to streamline phylogenetic inference based on codon usage preferences, and here specifically codon non-usage (or codon aversion). For each organism, we constructed a set of tuples, where each tuple contains a list of unused codons for a given gene. We define the pairwise distance between two species, A and B, as the ratio of total possible overlaps to direct overlaps. Total possible overlaps is the number of tuples in the set, for A or B, containing the fewest tuples, and direct overlaps is the intersection of tuples in the two sets. This approach allows us to calculate pairwise distances, even though there are substantial differences in the number of genes for each species. Finally, we use neighbor-joining to recover phylogenies. Using the Open Tree of Life and NCBI Taxonomy Database as expected phylogenies, our approach compares well, recovering phylogenies that largely match expected trees.
**OP31: Isoform level differential network analysis: A proffer alternative for identifying cancer markers**

*Presenting Author: Manoj Kandpal, Northwestern University*

*Co-Author(s): Ramana Davuluri, Northwestern University*

**Abstract:** Given that the majority of multi-exon genes generate diverse functional products, it is important to study expression at the isoform level. Furthermore, the interaction networks may have distinct connections among participating entities. In the proposed two-level analysis, our first objective is to observe the capability of differential network analysis (DifNA) between normal and diseased state in detecting the important markers for cancer. We then, through a comparative study, propose isoform based differential network as a better alternative than its gene counter in identifying cancer biomarkers. To test the proposed methodology, two comparative studies have been done using data from Gene Expression Omnibus (GEO) and The Cancer Genome Atlas (TCGA). Raw data is collected from publically available GEO repository in form of 160 Affymetrix exon-array mammalian cell line data comprising of 87 normal and 73 tumor samples or varied origin. We used Partial Least Square based Differential Network Analysis algorithm on top variable genes/isoforms. Differential network analysis provides each gene/isoform in the network with a connectivity score. Permutation tests based on these scores determined if the connectivity of a gene differs between the two networks. Top differentially connected genes/isoforms with p-value less than 0.05 were selected and Ingenuity Pathway Analysis (IPA) is performed to find out the involvement of selected genes/isoforms in various metabolic pathways and diseases. The analysis is repeated for 105 GBM samples from TCGA. DiffNA performed great in finding markers genes and isoforms. Further, Isoform level analysis showed better performance than gene level.

**OP32: Non-B DNA Predictor: A Web Application for Identifying Potential Intramolecular Non-B DNAs**

*Presenting Author: Hannah Ajoge, Western University*

*Co-Author(s): Hinissan Kohio, Western University; Henry He, Western University; Stephen Barr, Western University*

**Abstract:** Genome biology is not limited to the confines of the canonical right-handed double-helix DNA (a.k.a. B DNA), but includes other secondary structures that are collectively termed non-B DNA (NBD) structures. NBDs play an important role in biological processes such as DNA replication, telomere maintenance, gene expression, viral replication, immune evasion, neurologic disorders and cancers.

We developed a tool called ‘Non-B DNA Predictor’ (NBDP) to enable researchers with little computational skills to identify potential
intramolecular NBDs from nucleic acid sequences. NBDP is an interactive web application developed in R with Shiny with the back-end based on R package GQUAD, which is the only standalone tool available for identifying multiple NBDs. NBDP implements knowledge-based algorithms that are highly sensitive and accurate and out-performs all other tools with human genome-wide data containing known NBDs. It accurately identifies millions of NBD motifs in the human genome at a sensitivity level of >96%.

NBDP is presently the only NBD prediction application capable of: (1) identifying multiple NBDs, (2) accepting single/multiple sequence inputs, (3) accepting data in three possible forms (raw sequences, fasta and accession numbers), (4) irrespective of source (e.g. species) and (5) online/offline availability. Motifs that are identified (with/without overlaps) are A-phased, G-quadruplexes, H-DNA, slipped, short tandem repeats, triplex forming oligonucleotides and Z-DNA. NBDP has been used to show the difference in NBD density between human female and male genome and that NBDs are highly enriched in oncogenes. NBDP is hosted on http://ext0574.mni.fmd.uwo.ca/Non-BDNAP/.

OP33: IMPACT Web Portal: An Oncology Database Integrating Molecular Profiles with Actionable Therapeutics from Next Generation Sequencing Data

Presenting Author: Jennifer Hintzsche, University of Colorado Anschutz Medical Campus

Co-Author(s): Minjae Yoo, University of Colorado; Jihye Kim, University of Colorado; Carol Amato, University of Colorado; William Robinson, University of Colorado; Aik Choon Tan, University of Colorado

ABSTRACT: Next generation sequencing (NGS) technology allows researchers to identify important variants and structural changes in DNA and RNA in cancer patient samples. Using this information, we can now correlate specific variants and/or structural changes with known inhibitory actionable therapeutics. We introduce the creation of the IMPACT Web Portal, a new online resource that links molecular profiles from NGS of tumors to approved drugs, investigational therapeutics, and pharmacogenetics associated drugs. IMPACT Web Portal contains a total of 776 drugs connected to 1,326 target genes and 435 target variants, fusion, and copy number alterations. The online IMPACT Web Portal allows users to search for genetic alterations and connects them to three levels of actionable therapeutics. The results are categorized into 3 levels: Level 1 contains approved drugs separated into two groups; Level 1A contains approved drugs with variant specific information while Level 1B contains approved drugs with gene level information. Level 2 contains drugs currently in oncology clinical trials. Level 3 provides pharmacogenetic associations between approved drugs and genes. The results for each level are ranked by a p-value calculated from a hypergeometric test of all overlapping gene targets.
This allows users to understand the specificity of each actionable therapeutic. Each drug also links to another web page containing external information and additional gene targets, allowing further investigation of each actionable therapeutic. This resource is a valuable database for personalized medicine and drug repurposing translational oncology studies. IMPACT Web Portal is freely available for non-commercial use at http://tanlab.ucdenver.edu/IMPACT/.

**OP34: Exploring Frequented Regions in Pan-Genomic Graphs**

*Presenting Author:* Brendan Mumey, Montana State University  
*Co-Author(s):* Alan Cleary, Montana State University; Thiruvarangan Ramaraj, National Center for Genome Resources; Indika Kahanda, Montana State University; Joann Mudge, National Center for Genome Resources; Shubhang Kulkarni, Purdue University

**ABSTRACT:** We consider the problem of identifying regions within a pan-genome de Bruijn graph that are traversed by many sequence paths. Such regions and the subpaths that traverse them are denoted frequented regions (FRs). We have recently developed efficient algorithms for finding FRs. FRs have proved useful in pan-genomic applications such as providing distinguishing features for classifiers and for visualizing the pan-genomic graph. In initial work, we corroborated the biological relevance of FRs to identify introgressions in yeast that aid in alcohol tolerance. This talk will concentrate on (1) recent results to scale the approach to larger pan-genomic data sets such as the plant Medicago truncatula, with a 500 Mb genome, (2) developing visualization approaches that provides a simplified representation of pan-genomic space.

**OP35: TOBIAS — Transcription factor Occupancy prediction By Investigation of ATAC-seq Signals**

*Presenting Author:* Mario Looso, Max Planck Institute for Heart and Lung Research  
*Co-Author(s):* Mette Bentsen, Max Planck Institute for Heart and Lung Research; Annika Fust, Max Planck Institute for Heart and Lung Research; Jens Preussner, Max Planck Institute for Heart and Lung Research; Carsten Kuenne, Max Planck Institute for Heart and Lung Research

**ABSTRACT:** Chromatin accessibility and transcription factor (TF) binding are important aspects for mechanisms of gene regulation. Sequencing data from high-throughput chromatin accessibility assays such as ATAC-seq provides the possibility to identify opened chromatin regions as well as binding locations of a wide range of TFs within a single experiment. The exact location of DNA bound TFs can be identified by the intrinsic effect of the enzymes used by these assays to only act on accessible DNA locations, generating a coverage drop on occupied regions with a single base resolution, referred to as footprints. However,
an advanced method for footprint detection and quantification in ATAC-seq data is critically missing. Here we introduce our open source bioinformatics tool TOBIAS. It supports enzyme cutting bias normalization and applies a density-based clustering algorithm and footprint model to ATAC-seq data. In a first step, TOBIAS identifies potential TF binding sites in a global genome scale. Next, it combines these in silico predicted footprints with PWM scored motifs from hundreds of TFs. It provides individual binding locations, scores and probabilities for each TF as well as summarizing profile plots of bound and unbound TF binding sites. Ongoing algorithm development targets the quantification of footprints between biological conditions and downstream analysis of TF specific gene sets.

**OP36: Integrating Protein Localization Information in Signaling Pathway Reconstructions**

*Presenting Author: Anna Ritz, Reed College*

*Co-Author(s): Ibrahim Youssef, Reed College*

**ABSTRACT:** Understanding the components of intracellular signaling pathways is an important task in systems biology. Computational methods have been developed to automatically reconstruct signaling pathways from large networks of protein-protein interactions (interactomes). These pathway reconstruction methods can accelerate the time-consuming manual curation of pathway databases and generate hypothesis that aid in the discovery of novel signaling components. PathLinker is a pathway reconstruction method that computes many short paths within an interactome that connect known receptors to known transcriptional regulators specific to a particular pathway (e.g., Wnt or EGFR). PathLinker has previously been shown to outperform other state-of-the-art methods (e.g., Steiner trees, network flow, and random walk approaches), and RNAi experiments have confirmed a PathLinker prediction of CFTR’s involvement in canonical Wnt signaling. Despite the improved performance over other methods, PathLinker reconstructions still contain erroneous signaling interactions. Here, we utilize information about protein localization within the cell to improve pathway reconstructions. By adding information about cellular compartments, we preserve the spatial hierarchy of signaling flow by explicitly finding paths that begin at the membrane, terminate at the nucleus, and pass through intermediate compartments such as the cytosol. The new pathway reconstructions contain fewer false positives than the original PathLinker reconstructions, based on a benchmark dataset of pathways associated with cancer and immune signaling from the NetPath database. This work illustrates the utility of using additional biological information about protein interactions to improve predictions about signaling pathway interactions.
OP37: Reducing the footprint of mass spectrometry data

Presenting Author: Idoia Ochoa, University of Illinois at Urbana-Champaign

ABSTRACT: High-resolution mass spectrometry (MS) is a powerful technique used to identify and quantify molecules in simple and complex mixtures by separating molecular ions on the basis of their mass and charge. MS has become invaluable across a broad range of fields and applications, including biology, chemistry and physics, but also clinical medicine and even space exploration. As a result, there has been a rapid growth in the volume of MS data being generated. It is therefore crucial to design better compression algorithms to lessen the pressure of data storage and transmission.

The MS data is mainly composed of the mass-to-charge ratio (m/z)-intensity pairs, which can take several GB of space per experiment. We present a lossless compressor for these data that utilizes an adaptive scheme for compression. Specifically, the proposed algorithm compresses the floating-point pairs efficiently by calculating the hexadecimal difference of consecutive m/z values, and by searching for parts of the intensity values that match previous ones. The proposed algorithm delivers an average compression ratio improvement of 40% when compared to the lossless compressors GZIP, the current de-facto compressor for MS data, and FPC, a state-of-the-art compressor for floating point data. In particular, the proposed algorithm reduces the MS file sizes by 56% on average.

OP38: Meeting today’s challenging computational biology demands by integrating a focused IT team into the research process

Presenting Author: Dan Timmons, University of Colorado Boulder/BioFrontiers Institute

Co-Authors: Matt Hynes-Grace, University of Colorado Boulder; Jon Demasi, University of Colorado Boulder; Robin Dowell, University of Colorado Boulder; Mary Allen, University of Colorado Boulder; Cassidy Thompson, University of Colorado Boulder

ABSTRACT: BioFrontiers IT (BIT) team is taking steps towards meeting today’s rapidly changing HPC needs to support life science and health informatics research. This initiative includes team members immersing themselves in researchers’ work through one-on-one support and consultation, workshops to introduce tools to aid in reproducibility and data integrity, and the facilitation of hackathons to bring together researchers and programmers working through innovative, interdisciplinary projects. Through these personal interactions, BioFrontiers IT members are able to ascertain practical knowledge
from researchers that helps to engineer large-scale data and project management-oriented tools to support the community at large. To further this goal, and recognizing that computational research in academia benefits from collaboration with industry partners, BIT strives to create environments to foster mutually beneficial connections. These efforts have led to an increased integration between BIT and supported researchers which allows a more complete involvement in experimental design and execution, thus allowing BIT to step beyond a traditional troubleshooting role to provide more value with direct partnership. *

**OP39: Predicting Adverse Events Associated with Kinase Inhibitors from Clinical Drug Trials.**

*Presenting Author: Ilyssa Summer, University of Colorado Anschutz Medical Campus*

*Co-Author(s): Minjae Yoo, University of Colorado Anschutz Medical Campus; Jimin Shin, University of Colorado Anschutz Medical Campus; Aik Choon Tan, University of Colorado Anschutz Medical Campus*

**ABSTRACT:** Drug adverse events (AEs) are a major health threat to patients seeking medical treatment and a significant barrier in drug discovery and development. We performed systematic analysis of kinase inhibitors and their associated adverse events by querying clinical trial results reported in Clinicaltrials.gov. We obtained 368 kinase inhibitors from the Drug Repurposing Hub, and queried against our Kinase Inhibition Experiments Omnibus (KIEO) database to obtain kinase inhibition profiles. In total, we collected kinase inhibition and clinical trials data for 83 kinase inhibitors (35 approved and 48 investigational therapeutics) for this study. We extracted >1500 serious adverse events from 758 clinical trials covering 224,200 patients tested with kinase inhibitors. We also collected > 300 kinase inhibition data for these 83 kinase inhibitors. To determine the kinase inhibition – adverse events, we performed various computational and statistical analyses. In particular, we developed a computational method that integrates proportional reporting frequency of adverse events and kinase inhibition data for identifying kinase inhibition-adverse event (KI-AE) relationships. To validate the co-occurrence of kinase inhibition-adverse event detected in the clinical drug trials data, we queried the co-occurrence of KI-AE in PubMed. We computed the co-occurrence of the KI-AE correlation pairs using Fisher’s exact test, corrected by multiple comparisons. We also compared the KI-AE associations against SIDER web resources. Finally, we developed an interaction network for predicting new associations between kinase inhibition and adverse events. We envision that the KI-AE network will assist future drug discovery and development in reducing drug adverse events.
OP40: Characterization of the mechanism of drug-drug interactions from PubMed

Presenting Author: Feng Cheng, University of South Florida

ABSTRACT: Identifying drug-drug interaction (DDI) is an important topic for the development of safe pharmaceutical drugs and for the optimization of multidrug regimens for complex diseases such as cancer and HIV. There have been about 150,000 publications on DDIs in PubMed, which is a great resource for DDI studies. In this abstract, we introduced an automatic computational method for the systematic analysis of the mechanism of DDIs using MeSH (Medical Subject Headings) terms from PubMed literature. MeSH term is a controlled vocabulary thesaurus developed by the National Library of Medicine for indexing and annotating articles. Our method can effectively identify DDI-relevant MeSH terms such as drugs, proteins and phenomena with high accuracy. The connections among these MeSH terms were investigated by using co-occurrence heatmaps and social network analysis. Our approach can be used to visualize relationships of DDI terms, which has the potential to help users better understand DDIs. As the volume of PubMed records increases, our method for automatic analysis of DDIs from the PubMed database will become more accurate.
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. Poster Presenters must be available for presentation during the scheduled poster session.

POSTER NUMBER ASSIGNMENTS
Posters will be assigned even and odd numbers for presentation times. Please put your poster on the poster board corresponding to the number assigned.

SCHEDULE

<table>
<thead>
<tr>
<th>Schedule</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>FRIDAY, DECEMBER 8</td>
<td>9:00 AM – 12:00 PM</td>
</tr>
<tr>
<td></td>
<td>12:00 PM – 6:00 PM</td>
</tr>
<tr>
<td></td>
<td>6:30 PM – 7:30 PM</td>
</tr>
<tr>
<td></td>
<td>7:30 PM – 8:30 PM</td>
</tr>
</tbody>
</table>

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

POSTER LIST

P01: DendroShiny: A Dynamic Visualization Tool for the Analysis of Genome Wide Gene Expression Data
Presenting Author: Natalia Acevedo Luna, Iowa State University

P02: Automated Biomedical Text Classification with Research Domain Criteria
Presenting Author: Mohammad Anani, Montana State University

P03: Building a Systems-level model of the immune response to Salmonella infection
Presenting Author: Marta Andres-Terre, Stanford University

P04: The first de novo genome assembly of the hemiclonal live bearing Poeciliid fish, Poeciliopsis monacha
Presenting Author: Talon Arbuckle, Northwest Indian College

P05: GEOcurate: Enabling biologists to easily curate annotations from GEO datasets
Presenting Author: Avery Bell, Brigham Young University

P06: Using Cross-Species Analysis to investigate mechanisms of chemotherapy agents
Presenting Author: Judith Blake, Jackson Laboratory

P07: Estimating the lower bound of the known unknowns in the scientific literature
Presenting Author: Mayla Boguslav, University of Colorado Anschutz Medical Campus

P08: Multivariate equations increase the accuracy for outcome prediction of children and adolescents with pulmonary hypertension.
Presenting Author: Shawna Burgett, University of Colorado Anschutz Medical Campus
P09: Identifying and Annotating Uninvestigated Preeclampsia-Related Genes Using Linked Open Biomedical Data  
Presenting Author: Tiffany Callahan, University of Colorado Denver Anschutz Medical Campus

P10: Host phenotype prediction from differentially abundant microbes  
Presenting Author: Anna Paola Carrieri, IBM Research UK

P11: Withdrawn

P12: Characterization of the mechanism of drug-drug interactions from PubMed  
Presenting Author: Feng Cheng, University of South Florida

P13: Pathway Networks Generated from Human Disease Phenome  
Presenting Author: Ann Cirincione, University of Maryland, Baltimore County

P14: Development of a statistical model of the genetics of maternal-fetal dyads in neonatal abstinence syndrome  
Presenting Author: James Denvir, Marshall University

P15: Increasing flexibility of Laboratory Information Management System (LIMS) through extension of relational database model with NoSQL  
Presenting Author: Marcin Domagalski, University of Virginia

P16: Dissimilarity Matrix Based Clustering for Phylogenetic Analysis  
Presenting Author: Tara Eicher, The Ohio State University

P17: A Comparison of Viral Identification Techniques in Longitudinal Metagenomic Datasets  
Presenting Author: Cody Glickman, University of Colorado Anschutz Medical Campus

P18: Locating sites of ribonucleotide incorporation in RNase H2-deficient cells  
Presenting Author: Ali Gombolay, Georgia Institute of Technology

P19: Integrated molecular and clinical analysis for understanding human disease relationships  
Presenting Author: Winston Haynes, Stanford University

P20: Modular genomic variant calling workflow in Swift/T  
Presenting Author: Jacob Heldenbrand, University of Illinois at Urbana-Champaign

P21: The regulatory power of lincRNAs  
Presenting Author: Mikel Hernaez, University of Illinois at Urbana-Champaign

P22: IMPACT Web Portal: An Oncology Database Integrating Molecular Profiles with Actionable Therapeutics from Next Generation Sequencing Data  
Presenting Author: Jennifer Hintzsche, University of Colorado Anschutz Medical Campus

P23: A method of reproducing sample clustering by gene-expression with a panel of few genes  
Presenting Author: Christina Horr, University of Notre Dame

P24: Antimicrobial Food Additives Influence the Composition  
Presenting Author: Tomas Hrncir, Czech Academy of Sciences

P25: Simulation study of two-class classification algorithms for focused metabolomics.  
Presenting Author: Akira Imaizumi, Ajinomoto, CO., Inc.
P26: m-NGSG: A modified n-gram and skip-gram-based feature generation model from primary protein sequence.
Presenting Author: S M Ashiqul Islam, Baylor University

P27: Efficiently generating unique vector permutations in R
Presenting Author: Samantha Jensen, Brigham Young University

P28: Tracing the Innate Genetic Evolution and Spatial Heterogeneity in Treatment Naïve Lung Cancer Lesions
Presenting Author: Jihye Kim, University of Colorado Anschutz Medical Campus

P29: SALSA: Systematic Alternative Splicing Analysis Pipeline for Detecting Cryptic 3’ Splice Site Usage in SF3B1 Mutant Cancer RNA-seq
Presenting Author: Hyunmin Kim, University of Colorado Denver School of Medicine

P30: Not so noisy after all – large meta-analysis of annotated human microarray data reveals strong evidence for the alternative hypothesis in gene expression studies.
Presenting Author: Katja Koeppen, Geisel School of Medicine at Dartmouth

P31: Application of Machine Learning based classification approaches to Identify HAT Inhibitors as Potent Anticancer Agents
Presenting Author: Shagun Krishna, CSIR-Central Drug Research Institute

P32: DeSigN: Exploiting Gene Expression Signatures to Identify Potential Therapeutic Agents
Presenting Author: Bernard Lee, Cancer Research Malaysia

P33: Linking Genes to Cell Function
Presenting Author: Marc Maldaver, Michigan State University

P34: Classification of glioblastoma subtypes by integrating genomic and histopathological image features with probabilistic graphical models
Presenting Author: Dimitris Manatakis, University of Pittsburgh

P35: Optimal equation selection for graphlet counting
Presenting Author: Ine Melckenbeeck, Ghent University

P36: Codon Aversion: An alignment-free method to recover phylogenies
Presenting Author: Justin Miller, Brigham Young University

P37: DySE – Dynamic System Explanation Framework, Application in Cancer
Presenting Author: Natasa Miskov-Zivanov, University of Pittsburgh

P38: An Interactome Analyses to Reveal Chaperone Proteins as Effective Drug Targets against Salmonellosis
Presenting Author: Shama Mujawar, Sunway University

P39: A Bioinformatic Framework to Assess the Transcriptomic Response of Species to Environmental Toxins
Presenting Author: Siavash Nazari, University of Guelph

P40: Reducing the footprint of mass spectrometry data
Presenting Author: Idoia Ochoa, University of Illinois at Urbana-Champaign

P41: Identifying the molecular phenotypes of urothelial carcinoma of the bladder
Presenting Author: Oluwole Olaku, Morgan State University

P42: The use of signature genes and motifs to improve annotations of viruses
Presenting Author: Dnyanada Pande, San Diego State University
P43: National Center for Genome Analysis Support (NCGAS) as a platform for metagenomic analysis
Presenting Author: Bhavya Papudeshi, Indiana University

P44: Using functional interaction networks as prior knowledge in expression epistasis network centrality feature selection improves classification and enrichment of biologically relevant pathway for major depressive disorder
Presenting Author: Saeid Parvandeh, University of Tulsa

P45: Knowtator: A graphical text annotation plugin for Protegé
Presenting Author: Harrison Pielke-Lombardo, University of Colorado Anschutz Medical Campus

P46: Quality Assurance of Bioinformatics Software
Presenting Author: Morteza Pourreza Shahri, Montana State University

P47: Evaluating the Effectiveness of Using Literature Features for Automated Protein Phenotype Prediction using PHENOstruct
Presenting Author: Morteza Pourreza Shahri, Montana State University

P48: A rank-based hypergeometric test to extend GSEA for discerning specific effects of small molecule treatment
Presenting Author: Rani Powers, University of Colorado Anschutz Medical Campus

P49: A systems biology approach using cell-to-cell heterogeneity to map cell fate decisions in developmental processes
Presenting Author: Jens Preussner, Max Planck Institute for Heart and Lung Research

P50: Correcting for non-random fragmentation allows for more accurate estimation of relative abundance in NGS metagenomic data.
Presenting Author: Elmar Pruesse, University of Colorado Denver

P51: Integrating Protein Localization Information in Signaling Pathway Reconstructions
Presenting Author: Anna Ritz, Reed College

P52: Design and evaluation of peptides that recognize the Tn antigen.
Presenting Author: Daniel Armando Romero Frenchy, National University of Colombia

P53: Measuring chromosome conformation by fluorescence microscopy
Presenting Author: Brian Ross, University of Colorado Anschutz Medical Campus

P54: Discovering the Contribution of the Gut Microbiome to the Plasma Metabolome
Presenting Author: Michael Shaffer, University of Colorado Denver Anschutz Medical Campus

P55: Unsupervised discovery of phenotype specific heterogeneous gene networks
Presenting Author: Jenny Shi, University of Colorado Anschutz Medical Campus

P56: Visual analysis of disease-associated multi-omics relationships in the human gut
Presenting Author: Janet Siebert, University of Colorado Denver Anschutz Medical Campus

P57: Knowledge Network-guided analysis of genomics data on the Cloud
Presenting Author: Nahil Sobh, University of Illinois

P58: Harmonization and Preparation of Clinical Data using Shiny
Presenting Author: Laura Stevens, University of Colorado, Anschutz Medical Campus
P59: Predicting Adverse Events Associated with Kinase Inhibitors from Clinical Drug Trials
Presenting Author: Ilyssa Summer, University of Colorado Anschutz Medical Campus

P60: Improved prediction of functionally important residues through phylogenetic analysis
Presenting Author: Haiming Tang, Baylor College of Medicine

P61: “Parent” of duplication gene pairs tends to be more conserved than “daughter” copy, and is more likely to be further duplicated
Presenting Author: Haiming Tang, University of Southern California

P62: Meeting today’s challenging computational biology demands by integrating a focused IT team into the research process
Presenting Author: Dan Timmons, University of Colorado Boulder/BioFrontiers Institute

P63: Knowledge-based chemical relation extraction
Presenting Author: Ignacio Tripodi, University of Colorado, Boulder

P64: Statistical Integration and Feature Selection for Candidate Biomarker Discovery
Presenting Author: Bobbie-Jo Webb-Robertson, Pacific Northwest National Laboratory

P65: Pattern Recognition in Full-Text Biomedical Articles
Presenting Author: Elizabeth White, University of Colorado-Denver

P66: MeTeOR: a Literature Network for Hypothesis Generation and Precision Medicine
Presenting Author: Stephen Wilson, Baylor College of Medicine

P67: A comparison of error correction algorithms in T-cell receptor sequencing experiments
Presenting Author: Nicolle Witte, University of Colorado Anschutz Medical Campus

P68: Abstract Withdrawn

P69: A Pilot Systematic Genomic Comparison of Recurrence Risks of Hepatitis B Virus-associated Hepatocellular Carcinoma with low and high degree of liver fibrosis
Presenting Author: Seungyeul Yoo, Icahn School of Medicine at Mount Sinai

P70: Inference of protein networks and protein function in nontuberculous mycobacterial pathogens
Presenting Author: Diana Zajac, University of Colorado, Denver

P71: Enabling Deep Learning on Structural Biological Data
Presenting Author: Rafael Zamora, Lawrence Berkeley National Laboratory

P72: Genotype-phenotype Association Discovery through Accurately Dissecting Phenotype-associated Genetic Causal Effects
Presenting Author: Patrick Xuechun Zhao, Noble Research Institute

P73: Utilizing DNA methylation for Genome Annotation through Deep Learning
Presenting Author: Gregory Zynda, Texas Advanced Computing Center

P74: Transcriptional regulatory network inference from gene expression and chromatin accessibility measurements
Presenting Author: Ronald Taylor, Cincinnati Children’s Hospital Medical Center
P01: DendroShiny: A Dynamic Visualization Tool for the Analysis of Genome Wide Gene Expression Data

Subject: Graphics and user interfaces

Presenting Author: Natalia Acevedo Luna, Iowa State University

Author(s): Heike Hofmann, Iowa State University, United States; Stephan Q. Schneider, Iowa State University, United States

ABSTRACT: Modern genome-wide gene expression analysis approaches routinely yield prohibitively large sets of potential genes of interest which require further post-processing in order to destil useful information from the data. One such approach attempts to assign potential functions to poorly characterized genes by clustering annotated genes with unannotated candidates based on their expression patterns.

The success of these methods is highly dependent on a sensible choice of numerous clustering parameters. Making an informed decision regarding these parameters and interpreting the clustering results however is challenged by a current lack of user friendly, graphical, and interactive visualization tools.

To close this gap, we have developed DendroShiny, a tool to interactively examine gene expression clusters as the clustering parameters are adjusted. DendroShiny (1) clusters the genes based on a user-defined distance measure, (2) displays an interactive gene tree representing the similarity among the gene expression patterns, (3) allows the user to define the final number of clusters and visualizes the resulting gene sets, (4) interactively displays the expression profiles of genes in each cluster, and (5) computes the cumulative probability of finding known genes in each cluster.

We exemplify the use of DendroShiny through the analysis of RNA-seq data from the early embryonic development of Platynereis dumerilli, an annelid animal model. Here, DendroShiny allows for the identification of candidate ciliary genes despite the lack of an annotated genome. Overall, DendroShiny allows the user to intuitively explore clustered gene expression data and has the potential to facilitate the downstream analysis of transcriptomic data sets.
P02: Automated Biomedical Text Classification with Research Domain Criteria

Subject: Machine learning

Presenting Author: Mohammad Anani, Montana State University

ABSTRACT: Research Domain Criteria (RDoC) is a recently introduced framework for accurate diagnosis of mental illness. This framework contains five domains of interest, where each domain contains a number of constructs that define a specific behavior. Developing a method to automate the process of labeling biomedical articles with RDoC constructs is highly useful to advance research efforts in the area of mental illness. Therefore, this study focuses on exploring the feasibility of developing a tool for this purpose. Using a gold-standard dataset of about 40,000 Medline abstracts tagged with 26 RDoC constructs, we model this both as binary and multilabel classification problem, to perform document classification using several supervised learning algorithms. We use a simple Bag-of-Words model and apply standard preprocessing steps such as stemming and stop words removal. According to AUROC (Area Under Receiver Operating Characteristic curve) values obtained through 5-fold cross validation, we observe that overall, Artificial Neural Networks and Support Vector Machines perform best on the multilabel problem providing 96% average AUROC across all the constructs. Interestingly, all the binary classifiers showed the same level of performance. However, the cohort of binary classifiers took significantly longer time to train compared to their multilabel counterparts, showing the utility of modeling this as a multi-label problem. We also note that the articles labeled with more specific constructs were predicted better than the rest of the constructs. To the best of our knowledge, this is the first study on automated prediction of RDoC constructs for biomedical literature.

P03: Building a Systems-level model of the immune response to Salmonella infection

Subject: other

Presenting Author: Marta Andres-Terre, Stanford University

Author(s): Aditya Rao, Stanford University, United States; Michele Donato, Stanford University, United States; Purvesh Khatri, Stanford University, United States

ABSTRACT: Infectious diseases are the result of a molecular warfare between the host immune system and the pathogen. Their treatment and eradication are complicated by the heterogeneous nature of these interactions, which remains poorly understood. Here, we have identified a set of genetic and molecular determinants that characterize the host immune response to bacterial infection. First,
we conducted an integrated multi-cohort analysis of publicly available
gene expression data and identified a common host gene signature
across different bacterial infections. This meta-bacterial signature can
1) distinguish bacterial from viral and fungal infections, and 2) predict
symptom onset and disease outcome in infected individuals. We then
identified a Salmonella-specific host-response gene signature, which
can be used as a prognostic marker for typhoid fever, as well as for
understanding the biology of Salmonella infections. Second, we applied
cell mixture deconvolution to the same datasets we used to obtain
the gene expression signature and estimated the cellular populations
driving the response to Salmonella infection. We are currently working
on building a quantitative model of bacterial infections in which we
take into account both the cellular and gene expression signatures
identified using heterogeneous data sources. Defining the metrics that
characterize the immune response to bacterial infection will go beyond
the concept of identifying biomarkers, as this model could potentially
be used as a platform to identify and understand novel mechanisms
underlying host-pathogen interactions

P04: The first de novo genome assembly of the
hemiclonal live bearing Poeciliid fish, Poeciliopsis
monacha

Subject: Metagenomics

Presenting Author: Talon Arbuckle, Northwest Indian College

Author(s): Robert C. Vrijenhoek, Monterey Bay Aquarium Research Institute, United States; Shannon Johnson, Monterey Bay Aquarium Research Institute, United States; Nathaniel Jue, California State University Monterey Bay, United States

ABSTRACT: Poeciliopsis monacha is a fish endemic to Northwestern
Mexico that the International Union for Conservation of Nature (IUCN)
lists as data deficient. This species is biologically unique; certain
subpopulations are female-only and hemiclonal wherein females mate
with males from two other species in the Poeciliidae family but only
pass on the maternal genome to their offspring. To better understand
the genetic mechanisms underlying hemiclonality within this species
we generate the first reference genome sequence for this taxon.
De novo construction of the genome of P. monacha was attempted
via the Platanus assembler program, due to its reported accuracy in
assembling heterozygous genomes. Because of the nature of novel
assembly, model validation metrics were used to inquire about the
accuracy and completeness of the genome via N50 (329,723 and
analysis of single copy orthologs, respectively. In addition to assembly
of the sequence and model validation, we will also provide annotations
of repetitive elements and gene structures. The results of the work will
serve as the basis for development of a larger research program on the
evolution of hemiclonality. *
P05: GEOcurate: Enabling biologists to easily curate annotations from GEO datasets

Subject: Data management methods and systems
Presenting Author: Avery Bell, Brigham Young University
Author(s): Stephen Piccolo, Brigham Young University, United States

ABSTRACT: Gene Expression Omnibus (GEO) contains publicly available data for millions of biological samples. Annotations are available for many of these samples, but these metadata are formatted in a way that is difficult for many researchers to extract and analyze. We created GEOcurate, a software tool that walks users through the steps of parsing these annotations, specifying which variables to extract, and performing custom data transformations. GEOcurate uses a command-line interface—consisting of bash, Python, and R scripts, encapsulated in a Docker container—to step the user through the process of curating a GEO dataset. It prompts the user for a GEO series identifier, downloads the corresponding dataset, and asks the user, column by column, which columns to keep and how to transform the data in these columns. In many cases, key-value pairs are stored across multiple, generically-labeled columns. In other cases, multiple key-value pairs are stored in the same column, separated by arbitrary delimiters. GEOcurate enables the user to merge or split such columns in a custom manner. Additionally, the user can specify which values to treat as missing and which values to exclude, discretize, or rename. GEOcurate requires no programming expertise from the user, and it takes approximately 10 minutes to curate a data set, as opposed to hours of potential work that users would need to spend doing curation by hand. We hope this tool will ease the curation burden for scientists who wish to work with the vast and heterogeneous data in GEO.

P06: Using Cross-Species Analysis to investigate mechanisms of chemotherapy agents

Subject: inference and pattern discovery
Presenting Author: Judith Blake, Jackson Laboratory

ABSTRACT: To better understand how chemotherapy agents impinge upon normal cellular processes, we have taken a cross-species analysis approach to investigate gene functions associated with chemotherapy agents. Chemotherapy agents rely on their abilities to interfere with normal cellular processes of rapidly dividing cells causing them to decrease proliferation. Unfortunately, chemotherapy agents cannot distinguish cancer cells from non-cancer cells and, also, often lead to severe side effects. Furthermore, cancer cells can become refractory to the use of these agents.
In our analysis, we used the Comparative Toxicogenomics Database to retrieve gene sets correlating human, mouse, and rat genes with use of the chemotherapy agent, Cisplatin. We did a cross-species analysis of these gene sets using the GeneWeaver system, followed by GO functional enrichment analysis using the VLAD tool, to identify significant associations of genes with biological processes. Here we present our results showing strong association between Cisplatin-related gene sets and the process of apoptosis. Cross-species analysis of conserved gene sets confirms the association of apoptosis-related genes with the action of Cisplatin. We discuss methods to improve this approach towards understanding the mechanisms of chemotherapy action, resistance, and side-effects in cancer treatment.

This work was supported by NIH NHGRI -R01AA018776 Diversity action Plan for Mouse Genome Database (C. Bult, PI); and NIH NIC grant P30CA034196 Jackson Laboratory Cancer Center (E. Liu, PI).

**P07: Estimating the lower bound of the known unknowns in the scientific literature**

*Subject: Text Mining*

*Presenting Author:* Mayla Boguslav, University of Colorado Anschutz Medical Campus

*Author(s):* Larry Hunter, University of Colorado Anschutz Medical Campus, United States; Sonia Leach, University of Colorado Anschutz Medical Campus, United States; Kevin Cohen, University of Colorado Anschutz Medical Campus, United States

*ABSTRACT:* Generally, people think science is driven by knowledge. In contrast, Stuart Firestein (2012) claims science is driven by ignorance – known unknowns set priorities in scientific research. If this claim is true, then new computational tools for ignorance management, similar to knowledge management, may be useful for scientists. To test Firestein’s claim, we aim to estimate a lower bound of statements regarding ignorance in the scientific literature. One approach to find these known unknowns is using lexical cues to count the number of ignorance statements in different scientific corpora. To identify such cues, a list of ignorance phrases was gathered by annotating some articles in the biomedical literature. These lexical cues were then examined to assess their ambiguity: if a counterexample, a sentence that contains the lexical cue but is not a statement of ignorance, was found, then the cue is designated ambiguous. The remaining cues were designated nearly unambiguous. Counts of these nearly unambiguous lexical cues were made in three corpora: PubMed (biomedical literature), Arxiv (general scientific literature), and Wikipedia (general domain). These counts are a lower bound on the ignorance statements in each corpus. To further characterize statements regarding ignorance in the biomedical literature, we counted biomedical ontology concepts, including proteins and genes, recognized in the statements of ignorance. Our findings
suggest statements of ignorance are much more widespread in the scientific literature than in general factual material, and we have produced an initial characterization of the topics ignorance is expressed about in the biomedical literature specifically.

P08: Multivariate equations increase the accuracy for outcome prediction of children and adolescents with pulmonary hypertension.

Subject: Qualitative modeling and simulation

Presenting Author: Shawna Burgett, University of Colorado Denver Anschutz

Author(s): Uyen Truong, University of Colorado, United States; John Walker, University of Colorado, United States; Dunbar Ivy, University of Colorado, United States; Kendall Hunter, University of Colorado, United States

ABSTRACT: Background • Pulmonary hypertension (PH) is an incurable, progressive disease with poor survival in children. Cardiac catheterization is conventionally used for diagnosis and serial monitoring. Prognostics for pediatric PH are based upon adult standards/guidelines. Given differences in clinical manifestation and natural history between pediatric and adult PH, significant opportunity exists to customize prognostics specific to younger patients allowing for more proactive clinical interventions. Multiple studies have evaluated individual clinical factors. Disparity remains regarding factors most predictive of outcome. This study’s goal was to evaluate multiple factors, in combination, to predict patient outcomes.

Methods • This study includes 78 patients evaluated for PH in the cardiac catheterization laboratory at The Children’s Hospital Colorado (August 2003 – July 2010). This study used measurements of the proximal pulmonary artery along with right heart catheterization and clinical measurements commonly found to be predictive in PH outcomes. We used a multi-covariate proportional odds logistic regression model to explore prediction of patient outcome represented by World Health Organization Functional Classification at time of follow-up, comparing to Random Forest and PCA.

Results • Our findings support the combination of mean pulmonary arterial pressure, systolic pressure, systolic proximal artery diameter and the ratio of mean pulmonary arterial pressure to mean aortic pressure as the model combination that predicts future patient outcome with the highest accuracy.

Conclusions • Single variable predictors when combined provide incremental improvement that greatly surpasses their individual ability to predict future health status. Additionally, mechanical measures when combined with classic right heart catheterization measurements significantly increased prediction accuracy.
P09: Identifying and Annotating Uninvestigated Preeclampsia-Related Genes Using Linked Open Biomedical Data

Subject: inference and pattern discovery

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

Author(s): Adrianne Stefanski, University of Colorado Denver Anschutz Medical Campus, United States; William Baumgartner Jr., University of Colorado Denver Anschutz Medical Campus, United States; Jin-Dong Kim, Database Center for Life Science, Japan; Toyofumi Fujiwara, Database Center for Life Science, Japan; Ann Cirincione, University of Maryland, United States; Maricel Kann, University of Maryland, United States; Lawrence Hunter, University of Colorado Denver Anschutz Medical Campus, United States

ABSTRACT: Preeclampsia is a leading cause of maternal and fetal morbidity and mortality. Currently, there is no cure for preeclampsia except delivery of the placenta, which is central to preeclampsia pathogenesis. Transcriptional profiling of human placenta from pregnancies complicated by preeclampsia and control has been extensively performed to identify differentially expressed genes (DEGs). DEGs are identified using unbiased assays, however the decisions to investigate DEGs experimentally are biased by many factors (e.g., investigator interests, available reagents, knowledge of gene function) causing many DEGs to remain uninvestigated. To address this shortcoming, we utilized existing linked open biomedical resources and publicly available high-throughput transcriptional profiling data to identify and annotate the function of currently uninvestigated preeclampsia-associated DEGs. Using the keyword “preeclampsia”, we identified and reviewed 68 publicly available human gene expression experiments deposited in the Gene Expression Omnibus. Meta-analysis of the 13 experiments meeting our inclusion criteria generated a list of 273 DEGs. We annotated these genes using a knowledge graph constructed with Semantic Web Technologies that contained several Open Biomedical Ontologies and publicly available datasets. The relative complement of the annotation-derived and meta-analysis derived genes were identified as the uninvestigated preeclampsia-associated genes. Experimentally investigated DEGs were then identified from published literature based on semantic and syntactic annotations to PubMed abstracts by PubTator and PubAnnotation. Finally, novel biological relationships between experimentally investigated and uninvestigated preeclampsia-associated genes were identified by learning neuro-symbolic logic embeddings to predict missing edges in the knowledge graph. Detailed documentation and source code can be found on GitHub (https://github.com/callahantiff/ignorenet/wiki).
P10: Host phenotype prediction from differentially abundant microbes

Subject: Metagenomics

Presenting Author: Anna Paola Carrieri, IBM Research UK

Author(s): Niina Haiminen, IBM Research, United States; Laxmi Parida, IBM Research, United States

ABSTRACT: Metagenomic sequencing is increasingly used in human and animal health, food safety, and environmental studies. One of the major research challenges of the current decade is gaining insight into the structure, organization, and function of microbial communities. In these high-dimensional data the phenotype of the host organism may not be obvious to detect, thus the ability to predict it becomes a powerful analytic tool – e.g., predicting the disease status of an individual from their gut microbiome. Since the sequencing methods yield relative gene or species counts, a methodological question of appropriate normalization and scaling of the counts arises. We propose applying RoDEO (Robust Differential Expression Operator) projection as a new pre-processing method for metagenomic counts. We compare the impact of RoDEO w.r.t. various normalization methods on phenotype prediction. By analysing data across human, mouse and environmental samples and applying RoDEO together with the state of the art of machine learning methods, we accurately predict phenotypes or traits. We also address the problem of identifying the most relevant microbial features, i.e., Operational Taxonomy Units (OTUs) that could give insight into the structure of differential bacterial communities observed between phenotypes. In all three real datasets, we obtain similar or better phenotype prediction accuracy using the top few most differentially abundant OTUs as with the complete set of sequenced OTUs. The tool has the potential to help disease diagnostics and improve the future of personalised medicine enabling tailored treatments.

P11: Withdrawn

P12: Characterization of the mechanism of drug-drug interactions from PubMed

Subject: Data management methods and systems

Presenting Author: Feng Cheng, University of South Florida

ABSTRACT: Identifying drug-drug interaction (DDI) is an important topic for the development of safe pharmaceutical drugs and for the optimization of multidrug regimens for complex diseases such as cancer and HIV. There have been about 150,000 publications on DDIs
in PubMed, which is a great resource for DDI studies. In this abstract, we introduced an automatic computational method for the systematic analysis of the mechanism of DDIs using MeSH (Medical Subject Headings) terms from PubMed literature. MeSH term is a controlled vocabulary thesaurus developed by the National Library of Medicine for indexing and annotating articles. Our method can effectively identify DDI-relevant MeSH terms such as drugs, proteins and phenomena with high accuracy. The connections among these MeSH terms were investigated by using co-occurrence heatmaps and social network analysis. Our approach can be used to visualize relationships of DDI terms, which has the potential to help users better understand DDIs. As the volume of PubMed records increases, our method for automatic analysis of DDIs from the PubMed database will become more accurate.

**P13: Pathway Networks Generated from Human Disease Phenome**

*Subject: inference and pattern discovery*

*Presenting Author: Ann Cirincione, University of Maryland, Baltimore County*

*Author(s): Kaylyn Clark, University of Maryland, Baltimore County, United States; Maricel Kann, University of Maryland, Baltimore County, United States*

**ABSTRACT:** Understanding the effect of human genetic variations on disease can provide insight into phenotype-genotype relationships, and has great potential for improving the effectiveness of personalized medicine. While some genetic markers linked to disease susceptibility have been identified, a large number are still unknown. In this paper, we propose a pathway-based approach to extend disease-variant associations and find new molecular connections between genetic mutations and diseases. We used a compilation of over 80,000 human genetic variants with known disease associations from databases including the Online Mendelian Inheritance in Man (OMIM), Clinical Variance database (ClinVar), Universal Protein Resource (UniProt), and Human Gene Mutation Database (HGMD). Furthermore, we used the Unified Medical Language System (UMLS) to normalize variant phenotype terminologies, mapping 87 percent of unique genetic variants to phenotypic disorder concepts. Lastly, variants were grouped by UMLS Medical Subject Heading (MeSH) identifiers to determine pathway enrichment in Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways. Subsequent linking of KEGG pathways through underlying variant associations helped to elucidate connections between the human genetic variant-based disease phenome and metabolic pathways, suggesting novel disease connections not otherwise produced through gene-level analysis. We found many large complex diseases, such as cancer, to be highly linked by their common pathways. This study constitutes an important contribution to extending disease-variant connections and new molecular links between diseases.
P14: Development of a statistical model of the genetics of maternal-fetal dyads in neonatal abstinence syndrome

Subject: inference and pattern discovery

Presenting Author: James Denvir, Marshall University

Author(s): Richard Egleton, Marshall University, United States; Don Primerano, Marshall University, United States; Jun Fan, Marshall University, United States; Vincent Setola, West Virginia University, United States; Laura Lander, West Virginia University, United States

ABSTRACT: As a consequence of the dramatic increase in opioid use in the US in the last decade, a large increase in the number of infants born with Neonatal Abstinence Syndrome (NAS) has been observed. Opioids entering the mother’s bloodstream during pregnancy are metabolized at three primary sites before having the opportunity to influence brain function in the developing fetus: the maternal liver, the placenta, and the fetal liver. We are in the process of performing whole exome sequencing on mother-infant dyads in order to identify genetic variants in either the mother or infant that may be predictors of NAS severity or prognosis, and may inform treatment plans for the neonate. Here, we present the development of a genetic model that regards the mother-infant dyad as a single, quasi-tetraploid, entity that may be used for hypothesis testing in this context.

P15: Increasing flexibility of Laboratory Information Management System (LIMS) through extension of relational database model with NoSQL

Subject: Data management methods and systems

Presenting Author: Marcin Domagalski, University of Virginia

Author(s): Przemyslaw Porebski, University of Virginia, United States; David Cooper, University of Virginia, United States; Marcin Cymborowski, University of Virginia, United States; Heping Zheng, University of Virginia, United States; Marek Grabowski, University of Virginia, United States; Wladek Minor, University of Virginia, United States

ABSTRACT: LabDB is a modular, specialized super-LIMS, originally developed to track the macromolecular structure determination pipeline from cloning to structure solution (Zimmerman et al. 2014, Methods Mol. Biol. 1140). The system has modules to automatically or semi-automatically import data from a variety of different types of laboratory equipment, including chromatography and electrophoresis systems, crystallization observation robots, isothermal titration calorimetry, and others. Here we present a novel abstract data model for a new generation of LabDB LIMS, designed to address the extreme complexity of biology workflows. A purely relational database model was reduced to minimum number of database tables capable of storing experiment specific data structures and their predicate schemas (definitions) in JavaScript Object Notation (JSON) form. The model
is suitable for extension with programmatic plugins as well as for manual definition of samples and workflows through user interface. Its "generic" design limits the structural requirements for stored data and provides better usability and higher performance. The data is accessible using a representational state transfer (REST) API, which provides easy interoperability with other software systems. In conclusion, LabDB flexible schema can integrate instruments and manage various experimental samples and workflows from end-to-end, ensuring traceable results, reproducibility of experiments, and last but not least improving overall lab efficiency.

P16: Dissimilarity Matrix Based Clustering for Phylogenetic Analysis

Subject: Machine learning

Presenting Author: Tara Eicher, The Ohio State University

Author(s): Piyali Das, The Ohio State University, United States; Juan Barajas, The Ohio State University, United States; Ewy Mathe, The Ohio State University, United States; Andrew Patt, The Ohio State University, United States; Kevin Coombes, The Ohio State University, United States

ABSTRACT: Clustering is a useful tool for evaluating the relative evolutionary relationships among organisms in a population. It forms the basis of assessing genetic similarity between individuals for phylogenetic analyses. Toward this end, software has been developed that uses Bayesian inference algorithms to obtain a set of clusters that represent the population distribution with respect to genetic similarity. However, these algorithms are computationally intensive, and the software typically requires domain specific knowledge before it can be successfully used. The process also involves running the analysis for several iterations to derive the desired number of clusters. Because this can be cumbersome for researchers, it is advantageous to provide a comprehensive method that allows for quick and simple cluster generation.

In this work, we show that standard clustering methods can be applied to distance matrices of sequences between individuals across multiple loci, obtaining results similar to those of Bayesian algorithms and traditional phylogenetic classification. Our analyses included subspecies of giraffa, species of Ursus, and orders of Aves. We first used the Demerau-Levenshtein string distance metric to compute distances between genetic sequences of individuals at specified loci, then used the Manhattan distance to compute the total distance between individuals across all loci of interest and obtain our final distance matrix. By applying conventional clustering algorithms to these distance matrices, we obtained results that resembled those of previous studies in all three data sets.
P17: A Comparison of Viral Identification Techniques in Longitudinal Metagenomic Datasets

Subject: Metagenomics

Presenting Author: Cody Glickman, University of Colorado Anschutz Medical Campus

ABSTRACT: Viruses that influence bacterial community compositions, also known as bacteriophages, are commonly small genomes that lack a universally conserved genetic marker. As a result, bacteriophages make up a small proportion of reads in traditional DNA shotgun metagenomic experiments. Biological methods to enrich for virus-like particles (VLPs) suffer from the biases of filtering for free floating elements, thus limiting the recovery of endogenous viral elements. The importance of endogenous viral elements in the life cycles and adaptation of pathogenic bacteria is well known (Schuch et al. 2009-Aug-12 ; van et al. 2017-Sep-01). To capture novel endogenous viral elements in DNA shotgun metagenomic data, I propose a novel methodology to extract and pool viral reads across longitudinal patient samples. I propose that this methodology will produce longer viral read assemblies and more accurate taxonomic assignments than assembly of individual samples and full cross-sample assembly.

To test the performance of viral read assembly across time, I performed a simulation study using bacteria with endogenous viral elements. The study tests the lengths of viral read assemblies and the accuracy of the methodologies to recapitulate the elements in the synthetic dataset. In addition, I performed a synthetic spike-in on a real longitudinal metagenomic dataset with rare viral species to measure the sensitivity of the methodologies within noisy data. These simulations are a benchmark for mining viral elements from longitudinal datasets in the publicly available databases. Mining viral reads from metagenomic experiments allows researchers to study endogenous viral elements not typically found in viral enrichment studies.

P18: Locating sites of ribonucleotide incorporation in RNase H2-deficient cells

Subject: other

Presenting Author: Alli Gombolay, Georgia Institute of Technology

Author(s): Francesca Storici, Georgia Institute of Technology, United States; Fredrik Vannberg, Georgia Institute of Technology, United States

ABSTRACT: Ribonucleoside monophosphates (ribonucleotides or rNMPs) that are inadvertently incorporated into DNA can wreak havoc on genome stability. Under normal cellular conditions, the RNase H2 enzyme initiates the removal of rNMPs by efficiently cleaving these toxic nucleotides. However, when left unrepaired, rNMPs can cause breaks in the DNA strand, replication stress, and spontaneous mutagenesis.
To understand the biological consequences of rNMPs and their role in the pathogenesis of disease, we must profile the distribution of rNMPs in the genome. Determining where rNMPs are differentially incorporated will allow us to identify how rNMPs cause genome instability. Recent advances in laboratory techniques and computational methods provide the unique opportunity to capture these non-standard nucleotides and map their locations in the genome. One of these techniques is ribose-seq (Koh et al. Nature Methods 2015). In contrast to other techniques, ribose-seq directly captures rNMPs embedded in DNA and may be applied to any cell type at any stage of the cell cycle. Achieving the full potential of ribose-seq is dependent upon computational methods tailored to analyzing this type of data. The Ribose-Map toolkit is a novel collection of user-friendly, open-source, and well-documented scripts developed to profile the incorporation of rNMPs captured via ribose-seq. Ribose-Map allows the user to determine the genomic coordinates of rNMPs, calculate nucleotide frequencies, locate rNMP genomic hotspots, and create publication-ready figures. By exploring the location and distribution of rNMPs in RNase H2-deficient cells, we may begin to understand the role rNMPs play in genome instability and, ultimately, disease.

P19: Integrated molecular and clinical analysis for understanding human disease relationships

Subject: other

Presenting Author: Winston Haynes, Stanford University

Author(s): Rohit Vashisht, Stanford University, United States; Francesco Vallania, Stanford University, United States; Charles Liu, Stanford University, United States; Gregory Gaskin, Stanford University, United States; Erika Bongen, Stanford University, United States; Shane Lolgren, Stanford University, United States; Timothy Sweeney, Stanford University, United States; Paul J. Utz, Stanford University, United States; Nigam Shah, Stanford University, United States; Purvesh Khatri, Stanford University, United States

ABSTRACT: A detailed understanding of relationships among diseases will enable a deeper understanding of disease causation as well as offer opportunities to reposition drugs. To identify unbiased cluster of molecularly and clinically related diseases, we performed gene expression meta-analysis of 104 diseases using 600 studies with 41,000 samples and electronic health record analysis of over two million patients. Based on molecular data, we observed autoimmune diseases clustering with their specific infectious triggers and brain disorders clustering by disease class. In contrast, clinical data clustered diseases based on clinical practice. Our integrated molecular and clinical analysis spanned vastly different scales to identify robust disease clusters. We identified diseases with under-appreciated, therapeutically actionable relationships in our analysis. We highlighted the relationship between myositis and interstitial cystitis to encourage collaboration by connecting the seemingly disparate research communities.
P20: Modular genomic variant calling workflow in Swift/T

Subject: Data management methods and systems

Presenting Author: Jacob Heldenbrand, University of Illinois at Urbana-Champaign

Author(s): Azza Ahmed, University of Khartoum, Sudan; Yan Asmann, Mayo Clinic, United States; Faisal Fadlelmola, University of Khartoum, Sudan; Daniel Katz, University of Illinois at Urbana-Champaign, United States; Matthew Kendzior, University of Illinois at Urbana-Champaign, United States; Tiffany Li, University of Illinois at Urbana-Champaign, United States; Yingxue Ren, Mayo Clinic, United States; Elliott Rodriguez, University of Illinois at Urbana-Champaign, United States; Matthew Weber, University of Illinois at Urbana-Champaign, United States; Jennie Zermeno, University of Illinois at Urbana-Champaign, United States; Liudmila Mainzer, University of Illinois at Urbana-Champaign, United States

ABSTRACT: Genomic variant discovery is widely performed using the GATK’s Variant Calling Best Practices pipeline, a complex workflow with multiple steps, fans/merges, and conditionals. Managing the workflow can be difficult on a computer cluster, especially when running in parallel on large batches of data. One potential solution is monolithic implementations that replace the multi-stage workflow with a single executable. While such implementations exist, they are not sufficiently flexible to accommodate nuances of analysis particular to different species, types of sequencing, and research objectives.

Here, we present a scalable GATK-based variant calling workflow written in the Swift/T parallel scripting language. Key built-in features include the flexibility to split by chromosome before variant calling, the option to continue the analysis when faulty samples are detected, and the ability to analyze multiple samples in parallel within each node. With its modular design, execution can easily be separated into multiple stages that request the resources optimal for each portion of the pipeline. Swift/T’s ability to operate in multiple cluster scheduling environments (OGE, PBS Torque, SLURM, etc.) enables a workflow to be trivially portable across numerous clusters. Finally, the leaf functions of Swift/T permit developers to swap executables in and out of the workflow, increasing maintainability. With these features, users have a simple, efficient, and portable way to scale up their variant calling analyses to run in many traditional HPC architectures.

P21: The regulatory power of lincRNAs

Subject: Machine learning

Presenting Author: Mikel Hernaez, University of Illinois at Urbana-Champaign

Author(s): Olivier Gevaert, Stanford University, United States

ABSTRACT: A small number of important genes, known as “regulatory” or “driver” genes, play a crucial role at the molecular pathway level and directly influence the expression of several other genes. It seems natural that some of these regulatory genes should be able to explain
the variability of gene expression in genes that appear downstream in these biological pathways.

We propose a new method for module network discovery that works by iteratively linking newly discovered modules of genes with the set of regulatory elements that best express the data. A representative of each module, computed via dimensionality reduction by means of PCA, is used to find the most relevant regulators via linear regression with LASSO regularization. We applied the proposed method to gene expression datasets from 8 cancers obtained from TCGA and linked their expression to that of a well curated set of regulatory genes. In more than half of the discovered modules, the gene expression variance of the genes was explained almost entirely by the combination of very few regulatory elements, and in most of the modules, the correlation between the genes and the regulatory elements was more than 0.8. This pose a significant improvement over the previously proposed methods. In addition, starting from the raw sequencing data obtained from TCGA, we used the proposed method to link modules of protein coding transcript to few lincRNAs, and showed that the lincRNAs have similar regulatory capabilities as the well curated set of driver genes.

P22: IMPACT Web Portal: An Oncology Database Integrating Molecular Profiles with Actionable Therapeutics from Next Generation Sequencing Data

Subject: web services

Presenting Author: Jennifer Hintzsche, University of Colorado Anschutz Medical Campus

Author(s): Minjae Yoo, University of Colorado, United States; Jihye Kim, University of Colorado, United States; Carol Amato, University of Colorado, United States; William Robinson, University of Colorado, United States; Aik Choon Tan, University of Colorado, United States

ABSTRACT: Next generation sequencing (NGS) technology allows researchers to identify important variants and structural changes in DNA and RNA in cancer patient samples. Using this information, we can now correlate specific variants and/or structural changes with known inhibitory actionable therapeutics. We introduce the creation of the IMPACT Web Portal, a new online resource that links molecular profiles from NGS of tumors to approved drugs, investigational therapeutics, and pharmacogenetics associated drugs. IMPACT Web Portal contains a total of 776 drugs connected to 1,326 target genes and 435 target variants, fusion, and copy number alterations. The online IMPACT Web Portal allows users to search for genetic alterations and connects them to three levels of actionable therapeutics. The results are categorized into 3 levels: Level 1 contains approved drugs separated into two groups; Level 1A contains approved drugs with variant specific information while Level 1B contains approved drugs with gene level information. Level 2 contains drugs currently in oncology clinical trials.
Level 3 provides pharmacogenetic associations between approved drugs and genes. The results for each level are ranked by a p-value calculated from a hypergeometric test of all overlapping gene targets. This allows users to understand the specificity of each actionable therapeutic. Each drug also links to another web page containing external information and additional gene targets, allowing further investigation of each actionable therapeutic. This resource is a valuable database for personalized medicine and drug repurposing translational oncology studies. IMPACT Web Portal is freely available for non-commercial use at http://tanlab.ucdenver.edu/IMPACT/.

P23: A method of reproducing sample clustering by gene-expression with a panel of few genes

Subject: Machine learning

Presenting Author: Christina Horr, University of Notre Dame

Author(s): Steven Buechler, University of Notre Dame, United States

ABSTRACT: Many studies have employed clustering methods with genetic data, however, high feature dimensionality and its effects on downstream genetic analysis still remain a problem to this day. Presumably, using a relatively small number of features would give more insight into the defining characteristics of the clusters. The objective of this research is to introduce a new clustering algorithm that can significantly decrease the number of features used in sample clustering and to predict sample cluster assignments of new samples using a small panel ($\leq$ 10) of functionally-related genes per cluster. The method was applied to estrogen-receptor-positive (ER+) breast cancer microarray and RNA-Seq data. Original clusterings that created two clusters and three clusters, respectively, were created using average linkage hierarchical clustering (HC). Using our method, we identified small panels of genes, that were able to reproduce these two clusterings with accuracies of 77% for both runs. Furthermore, using cross-validation, we have shown that our method can be utilized to predict cluster assignments of new samples.

P24: Antimicrobial Food Additives Influence the Composition

Subject: Metagenomics

Presenting Author: Tomas Hrncir, Czech Academy of Sciences

Co-Author(s): Lucia Hrnicirova, Charles University, Hradec Kralove, Czech Republic; Tomas Hudcovic, Czech Academy of Sciences, Czech Republic; Vladimira Machova, Czech Academy of Sciences, Czech Republic

ABSTRACT: The incidence of allergies and autoimmune diseases is increasing worldwide. Gut microbiota can modulate local but also
systemic immune responses. In this study, we show that environmental factors, specifically antimicrobial food additives modify the composition and diversity of gut microbiota and thus may influence host’s immune responses and susceptibility to autoimmune diseases.

**P25: Simulation study of two-class classification algorithms for focused metabolomics**

*Subject: Simulation and numeric computing*

*Presenting Author: Akira Imaizumi, Ajinomoto, CO., Inc.*

**ABSTRACT: Aims** • In most cases, no markers with clear-cut feature are detected in the metabolome analysis. Multivariate classifiers are often used for that reason. However, metabolome data with limited sample size and a several number of metabolites often makes it difficult to infer robust classifiers and appropriate algorithm(s) for inferring classifiers is still unclear. In this study, effects of sample size and algorithms on performances and robustness of classifiers were evaluated by statistical simulation, using plasma free amino acid (PFAA) profile, a typical focused metabolomics data.

**Methods** • Both training and test data set was generated from the previously published PFAA profiles of cancer patients and controls. Seven algorithms from three categories were implemented. The first category contains three type of generalized linear models (GLMs). The second were typical machine learning algorithms, namely, naïve Bayes (NB), support vector machine (SVM), and random forest (RF). The third was deep learning (DL). Performance and robustness of the estimated classifier were evaluated in terms of the area under curve (AUC) of receiver- operator characteristics (ROC) curve.

**Results** • Using 700 simulated data set with various sample size, the classification performances were evaluated. When sample size was not less than 100 each, performance became stable in each algorithm. Among them, three glms showed both high performance and robustness even in the relatively small size training data than other algorithms.

**Conclusion** • For two-class classification of the focused metabolome data, simple GLM method is to be applicable from a view point of classification performance and its robustness.
P26: m-NGSG: A modified n-gram and skip-gram-based feature generation model from primary protein sequence

Subject: Machine learning

Presenting Author: S M Ashiqul Islam, Baylor University

Author(s): Christopher Kearney, Baylor University, United States; Meron Ghidey, Baylor University, United States; Erich Baker, Baylor University, United States

ABSTRACT: Classification by supervised machine learning greatly facilitates the annotation of protein characteristics from their primary sequence. However, the feature generation step in this process requires detailed knowledge of related chemical properties for protein classification. Lack of this knowledge risks the selection of irrelevant features, resulting in a faulty model. In this study, we introduce a means of automating the work-intensive feature generation step via a Natural Language Processing (NLP)-dependent model which we refer here as m-NGSG. This approach results in a consistent increase in accuracy when compared to contemporary classification and feature generation models. In another separate study, we generated five m-NGSG-based models to predict different functional classifications of disulfide-rich peptides. Performances of those models show better accuracy than PSI-BLAST. Based on the meta-analysis and PSI-BLAST comparison results we expect this model to accelerate the classification of all types of proteins from primary sequence data and increase the accessibility of protein prediction to a broader range of scientists. m-NGSG is freely available at Bitbucket: https://bitbucket.org/sm islam/mngsg/src. We also report an interactive version of m-NGSG through web interface watson.ecs.baylor.edu/ngsg. By using m-NGSG, we hope, researchers will be able to make their own models to classify proteins from two or more groups in a sequence, structure, function or species agnostic fashion.

P27: Efficiently generating unique vector permutations in R

Subject: Simulation and numeric computing

Presenting Author: Samantha Jensen, Brigham Young University

Author(s): Caleb Cranney, Brigham Young University, United States; Stephen Piccolo, Brigham Young University, United States

ABSTRACT: Biologists often depend on statistical methods that use random permutations to quantify significance. These methods produce an empirical null distribution by repeatedly permuting class labels and performing the statistical test; the results obtained using the actual class labels are then compared against the empirical null distribution.
To avoid biases in performing such calculations, the permutations should be sampled without replacement. However, existing software packages sample with replacement due to the computational complexity of ensuring that permuted vectors are unique. Phipson and Smyth have demonstrated that sampling with replacement often results in empirical p-values that are understated, thus increasing the type I error rate. To date, no computationally efficient method for ensuring the uniqueness of permutation vectors is widely used. To address this problem, we used a crowd-sourcing approach, extending the challenge of creating an efficient algorithm for generating unique permutations to 26 undergraduate students in Brigham Young University’s capstone bioinformatics class. Analysis of the time and space efficiency of these methods led us to identify two distinct algorithms that each can generate 1,000,000 unique permutations of a binary vector of 1000 observations in around 30 seconds. Subsequently, we developed uniquerperm, an R package that implements these algorithms. The availability of these functions to bioinformatics researchers will allow more accurate calculation of empirical p-values, especially enabling more accurate multiple-test correction of biological results.

P28: Tracing the Innate Genetic Evolution and Spatial Heterogeneity in Treatment Naïve Lung Cancer Lesions

Subject: other

Presenting Author: Jihye Kim, University of Colorado Anschutz Medical Campus

Author(s): Kenichi Suda, Kindai University Faculty of Medicine, Japan; Isao Murakami, Higashi-Hiroshima Medical Center, Japan; Leslie Rozeboom, University of Colorado Anschutz Medical Campus, United States; Christopher J. Rivard, University of Colorado Anschutz Medical Campus, United States; Tetsuya Mitsudomi, Kindai University Faculty of Medicine, Japan; Fred R. Hirsch, University of Colorado Anschutz Medical Campus, United States; Aik-Choon Tan, University of Colorado Anschutz Medical Campus, United States

ABSTRACT: Extensive intratumor heterogeneity (ITH) have been observed in individual patient tumors by large-scale sequencing analyses. ITH can contribute to drug resistance and cancer metastasis. Distinct microenvironments provide selection advantages to sub-population of cells for cancer metastasis. We hypothesized that ITH can contribute to metastasis by disseminating different sub-populations of cancer cells to distant sites. We collected tumor specimens and non-cancer tissues from treatment naïve autopsied patients to study the innate genetic evolution and spatial heterogeneity. Our cohort consists of four (two adenocarcinoma and two squamous cell carcinoma) NSCLC patients and one SCLC patient. Each patient had 5–9 primary and metastatic lesions. Comprehensive data analyses were performed on the RNA-seq that include gene expression and pathway analyses, fusion detection and somatic variants detection. Global unsupervised clustering of expression data reveals the NSCLC patients clustered together from the SCLC, and the two adenocarcinoma and squamous
cell carcinoma patients formed two clusters. Within each patient, metastatic lesions clustered according to the distant metastatic sites. Pathway analysis and somatic mutation analysis in individual patients revealed that, in general, the primary lesion is distinct from metastatic lesions in NSCLCs. For the SCLC patient, distant metastases and lymph node metastases clustered according to different parts of the primary tumors. We also identified KIF5B-RET fusion as a founder mutation in all tumor specimens obtained from a never-smoking adenocarcinoma patient. This study provides evidence that ITH contributes to distant metastasis sites based on the similarity and the heterogeneity between primary and metastatic lesions in lung cancer patients.

**P29: SALSA: Systematic Alternative Splicing Analysis Pipeline for Detecting Cryptic 3’ Splice Site Usage in SF3B1 Mutant Cancer RNA-seq**

**Subject:** inference and pattern discovery

**Presenting Author:** Hyunmin Kim, University of Colorado Denver School of Medicine

**Author(s):** Kelsey Wuensch, University of Colorado Denver School of Medicine, United States; Jennifer Hintzsche, University of Colorado Denver School of Medicine, United States; Jihye Kim, University of Colorado Denver School of Medicine, United States; AikChoon Tan, University of Colorado Denver School of Medicine, United States

**ABSTRACT:** Alternative splicing (AS) of RNA is an essential function in cells to facilitate proper processing of pre-mRNAs into protein-coding genes, and alterations in AS have been associated with multiple diseases and tumorigenesis. Recent cancer genomics studies identified recurrent spliceosome mutations in multiple cancers, specifically hot-spot mutations of SF3B1 (i.e. R625, K666 and K700) were found at high frequency in myelodysplastic syndrome, chronic lymphocytic leukemia, breast cancer, pancreatic cancer, uveal and mucosal melanomas. SF3B1 mutant induces alternative 3’ splicing through utilization of a cryptic 3’ss and branch point sequence, generates AS that could be: (i) translated into new aberrant proteins; and (ii) RNA degradation through nonsense-mediated mRNA decay (NMD). However, it remains unclear whether the different hot-spot mutations in SF3B1 induce distinct AS patterns in a particular gene set or cancer type. To study the AS induces by SF3B1 mutation, we developed SALSA, a novel AS analysis tool that can detect altered 3’ splice sites in RNA-seq. We applied SALSA for the analysis of SF3B1-mutant RNA-seq data in TCGA cancer patients. We identified a set of genes that are enriched in cryptic 3’ splice sites, and validated these AS in our in-house SF3B1-mutant RNA-seq data. We also predicted SF3B1 splicing rules regulating the AS-NMD in these samples, and examined the functional enrichment of these AS-NMD transcripts. We concluded that SF3B1-mutant cancers are using AS as a mechanism to promote tumor growth and survival. Some of these altered proteins could be exploited as cancer therapy in SF3B1-mutant cancer patients.
P30: Not so noisy after all — large meta-analysis of annotated human microarray data reveals strong evidence for the alternative hypothesis in gene expression studies

Subject: inference and pattern discovery

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

Author(s): Thomas Hampton, Geisel School of Medicine at Dartmouth, United States; Bruce Stanton, Geisel School of Medicine at Dartmouth, United States

ABSTRACT: Differential gene expression (DGE) experiments such as microarrays or RNA-Seq typically include a step to correct for multiple hypothesis testing, consistent with the default hypothesis that no genes are differentially expressed. However, most experiments are performed using designs predicted to cause widespread DGE. We ran ANOVA models for 6,356 well-annotated human genes across 986 curated human microarray studies deposited in the Gene Expression Omnibus to test the frequency of significant differentially expressed genes, predicted to be 5% under the null hypothesis. The median study had 28% of genes DE (p<0.05) and the median gene was significantly DE in 35% of studies. KEGG pathway enrichment analysis revealed that genes DE in >50% of studies were associated with the Adherens junction and p53 signaling pathways. Genes DE in <15% of studies, which might be candidate reference genes for normalization, were associated with Olfactory transduction, Neuroactive ligand-receptor interaction, and Taste transduction. Unfortunately, these genes have high within-group CV, making them unsuitable as reference genes. Genes with low within-group CV, conversely, were often DE between experimental groups, and therefore not ideal as references either. The median log2 FC for classic reference genes was 0.31, which would skew results by 24% when using these genes for normalization. Hence, we predict that treatment effects <25% cannot be reliably detected by methods like qPCR or NanoString, which use normalization to traditional reference genes. Moreover, our results suggest that gene expression studies may be systematically under-reporting effects based on excessively stringent criteria for selecting DE genes.

P31: Application of Machine Learning based classification approaches to Identify HAT Inhibitors as Potent Anticancer Agents

Subject: Machine learning

Presenting Author: Shagun Krishna, CSIR-Central Drug Research Institute

ABSTRACT: The histone acetyltransferases are involved in acetylation of histones. In this study, we have generated predictive chemoinformatics models for virtual screening based on machine learning algorithms. The
hits were then docked into the active site of enzyme. Finally, a set of 10 molecules were procured and subjected to biological evaluation.

**P32: DeSigN: Exploiting Gene Expression Signatures to Identify Potential Therapeutic Agents**

*Subject: inference and pattern discovery*

*Presenting Author: Bernard Lee, Cancer Research Malaysia*

*Author(s): Chai Phei Gan, Cancer Research Malaysia; Zainal Ariff Abdul Rahman, Faculty of Dentistry, University of Malaya, Malaysia; Tsung Fei Khang, Faculty of Science, University of Malaya, Malaysia; Aik Choon Tan, School of Medicine, University of Colorado Anschutz Medical Campus, United States; Sok Ching Cheong, Cancer Research Malaysia*

*ABSTRACT:* The availability of public pharmacogenomics databases such as the Genomics of Drug Sensitivity in Cancer (GDSC) and the Cancer Cell Line Encyclopedia (CCLE) databases has opened up possibilities for computational method development for in silico drug repurposing analysis. Previously, using DeSigN v.1 (PMID: 28198666), we predicted that oral squamous cell carcinoma would be susceptible to bosutinib, and experimental validation was confirmatory. In the light of a two-fold increase in drug candidates (265 vs. 140) and 1/3 increase in cell lines (800+ vs. 600), we have updated the ranking algorithm in DeSigN v.2 to improve performance. For each drug candidate, the method now defines sensitive (bottom 5%) and resistant (top 95%) lines based on the area under the curve (AUC) values of drug responses. Differentially expressed genes between the two conditions were obtained using limma. The eXtreme Sum (XSum) scoring algorithm was implemented as the pattern-matching algorithm for similarity search, and results are now associated with p-values. We tested and validated a series of previously published gene expression signatures of drug sensitivity against DeSigN v.2. In one particular example, we demonstrated that DeSigN v.2 correctly predicted the sensitivity of gefitinib on a panel of non-small cell lung cancer cell lines (GSE4342). In conclusion, we demonstrated DeSigN v.2 is a simple yet practical tool of connecting similar gene expression signature to drug sensitivity metric to prioritize drug candidates for drug repurposing studies. DeSigN v.2 is freely available at http://design-v2.cancerresearch.my/query.

**P33: Linking Genes to Cell Function**

*Subject: Machine learning*

*Presenting Author: Marc Maldaver, Michigan State University*

*Author(s): Arjun Krishnan, Michigan State University, United States*

*ABSTRACT:* Tissue type and cell function is not always black and white. Although each cell in our body has all of our genetic information, only a small portion of that code is utilized in any given cell. Cell
differentiation is not always clear and it can be difficult to definitively link genes to specific tissues. As innovations regarding stem cells and CRISPR emerge, understanding the genes that are used in a given tissue is becoming increasingly relevant. Using machine learning techniques, we will attempt to explicitly associate genes to tissue types. This will be done by analyzing gene expression data to identify which genes are upregulated and downregulated. We will be utilizing ArrayExpress, one of the largest repositories of gene expression data in the world. By automatically curating ArrayExpress samples using a modified annotating pipeline (metaSRA), we will be able to employ a large training set.

### P34: Classification of glioblastoma subtypes by integrating genomic and histopathological image features with probabilistic graphical models

**Subject:** System integration  
**Presenting Author:** Dimitris Manatakis, University of Pittsburgh  
**Author(s):** Panayiotis Benos, University of Pittsburgh, United States; Akif Tosun, University of Pittsburgh, United States; Chakra Chennubhotla, University of Pittsburgh, United States

**ABSTRACT:** Integrating multi-modal biomedical data types under the same analytic framework is an important step towards harvesting the existing, fragmented knowledge that is collected in a clinical setting. H&E staining has been used extensively in pathology for diagnosing various diseases or disease subtypes, but it has limitations. Computational pathology, i.e. the processing and analysis of H&E stained images with computational methods augmented by high-throughput data collection from the same patients aims in improving disease diagnosis by combining multiple, complementary sources of information. Currently, the methods for integrating clinical image and –omics features have used mainly second order methods. These methods, however, cannot properly capture the complexity of the direct and indirect relations between these features. In this paper, we present an alternative method for data integration: the use of directed graphs, a form of probabilistic graphical models. Using new methods we have developed to (1) parse tissue heterogeneity information from H&E images and (2) learn directed graphs over mixed data types, we test the hypothesis that there is a relationship between gene expression and spatial information in the tissue organization in glioblastoma.
P35: Optimal equation selection for graphlet counting

Subject: Graph Theory

Presenting Author: Ine Melckenbeeck, Ghent University

Author(s): Pieter Audenaert, Ghent University, Belgium; Mario Pickavet, Ghent University, Belgium; Didier Colle, Ghent University, Belgium; Piet Demeester, Ghent University, Belgium

Abstract: Since their introduction by Pržulj, graphlets have been used to compare the structure of different networks, including biological interaction networks, quantitatively. These graphlets are small induced subgraphs of a larger graph. Counting how many times each graphlet appears within a graph allows quantifying that graph’s structure. Counting the graphlets in which a specific node takes part, called that node’s graphlet degrees, gives information about the function of that node in the graph.

Hočevar and Demšar simplified computing a node’s graphlet degrees, by using equations in their ORCA tool. The set of equations they introduced relate a node’s graphlet degrees to the number of common neighbours of nodes that are in smaller graphlets, together with the chosen node. More specifically, they allow calculating the graphlet degrees of order 5 while only finding graphlets of order 4. Graphlets of order greater than 5, however, are beyond ORCA’s capabilities.

In previous work, we introduced an algorithm that can automatically generate a set of equations to calculate graphlet degrees of any order, and another algorithm that allows using them to count the graphlet degrees. However, we generate a linearly dependent system with more equations than we need. Any maximal linearly independent subset of these equations can be used by our counting algorithm, although a running time reduction of a factor 2 is possible by selecting specific equations. Therefore, we will attempt to identify which equation properties speed up the calculation, as well as how to generate only the optimal set of equations for any graphlet order.

P36: Codon Aversion: An alignment-free method to recover phylogenies

Subject: inference and pattern discovery

Presenting Author: Justin Miller, Brigham Young University

Author(s): Perry Ridge, Brigham Young University, United States; Michael Whiting, Brigham Young University, United States; Lauren McKinnon, Brigham Young University, United States

Abstract: Codon bias refers to the non-random usage of synonymous codons, and differs between organisms, between genes, and even within a gene. We previously identified a strong phylogenetic signal,
based on codon usage preferences, in 72 tetrapod species, focusing on stop codon usage preferences. Here we report the expansion of our previous work into >20,000 species across all kingdoms of life, and the development of tools to streamline phylogenetic inference based on codon usage preferences, and here specifically codon non-usage (or codon aversion). For each organism, we constructed a set of tuples, where each tuple contains a list of unused codons for a given gene. We define the pairwise distance between two species, A and B, as the ratio of total possible overlaps to direct overlaps. Total possible overlaps is the number of tuples in the set, for A or B, containing the fewest tuples, and direct overlaps is the intersection of tuples in the two sets. This approach allows us to calculate pairwise distances, even though there are substantial differences in the number of genes for each species. Finally, we use neighbor-joining to recover phylogenies. Using the Open Tree of Life and NCBI Taxonomy Database as expected phylogenies, our approach compares well, recovering phylogenies that largely match expected trees.

**P37: DySE – Dynamic System Explanation Framework, Application in Cancer**

*Subject: Qualitative modeling and simulation*

*Presenting Author: Natasa Miskov-Zivanov, University of Pittsburgh*

**ABSTRACT:** Biomedical research results are being published at a high rate, and with existing search engines, the vast amount of published work is usually easily accessible. To accurately reuse this voluminous knowledge that is fragmented and sometimes inconsistent, one can extract and assemble published information into models. However, the creation of models often relies on intense human effort: model developers have to read hundreds of published papers, and conduct discussions with domain experts. This laborious process results in slow development of models, as it includes steps such as model validation with experimental data and model extension with new available information. To automate the process of explaining biological observations and answering biological questions, we have built a Dynamic System Explanation (DySE) framework. DySE automatically assembles models, extends them with the interactions extracted by automated reading engines, analyzes the models for various conditions and scenarios, and iteratively improves assembled models. The framework includes techniques such as stochastic simulation, statistical model checking, static and dynamic sensitivity analysis, and hypothesis generation. With the automated process of reading, assembly, and reasoning, our framework allows for rapidly designing and conducting
thousands of in silico experiments, and thus, can speed up discoveries from the order of decades to the order of hours or days. We have applied DySE to studying cancer microenvironment, as well as on explaining mechanisms of several melanoma drugs. The techniques and the modeling approach incorporated within the framework are not disease specific, and therefore, DySE can be used for explaining systems in many other domains.

P38: An Interactome Analyses to Reveal Chaperone Proteins as Effective Drug Targets against Salmonellosis

Subject: other

Presenting Author: Shama Mujawar, Sunway University

Author(s): Chandrajit Lahiri, Sunway University, Malaysia; Derek Gatherer, Lancaster University, United Kingdom

ABSTRACT: Strains of Salmonella that are resistant to antimicrobial agents have become a worldwide health problem. The severity of the disease depends upon host factors and the serotypes of Salmonella. The infection mechanism comprises a series of secretory proteins whose folding and quality are being regulated by the molecular chaperones. This takes us to the juncture of utilising chaperones as potential antibiotic therapeutic strategies targeting the host-pathogen interactions. We have found one such chaperones SicA as indispensable from an interactome analysis of the proteins from Salmonella pathogenicity islands (SPI) and two component signal transduction systems. This was done by implementing the network parameters like centrality and k-core measures followed by validation through microarray analyses. In a broader perspective, a whole genome interactome analyses of Salmonella proteins revealed DnaK to be one of the crucial proteins playing an important role in pathogen survival under stress conditions as encountered during antibiotic therapies. As DnaK is a heat-shock protein from Hsp70 class, we have selected another Hsp70 chaperone SigE which is required for stabilization and secretion of the secretory proteins. Moreover, its presence in the highly indispensable sub-network to which SicA belongs brings us to the three representative proteins to work upon. This was done by studying their binding interaction with XR770, a phenalenone derivative for a therapeutic intervention against emerging significant drug resistance conferred by the pathogen Salmonella Typhimurium strain LT2. Our result showed potential susceptible predictors suitable for a series of trials, testing and analysis in the wet-lab environment.
P39: A Bioinformatic Framework to Assess the Transcriptomic Response of Species to Environmental Toxins

Subject: inference and pattern discovery

Presenting Author: Siavash Nazari, University of Guelph

Author(s): Mehrdad Hajibabaei, University of Guelph, Canada

ABSTRACT: The current ecotoxicological approach to identify sites of biological concern uses measures like median lethal dose and lethal concentration analysis which do not link the effects of toxins to any other level of biological organization. These traditional methods, would not inform of the presence of hazardous chemicals not before a significant environmental damage has already been dealt. Being motivated by the idea that essential pathways and genes are conserved across the taxa, we present a framework that incorporates transcriptomic data of inhabitant species to monitor the environment. This framework mines chemical-gene interaction data from publicly available databases like the Comparative Toxicology Database, retrieving an initial set of affected genes in the input organisms. Next, it retrieves translated protein sequences related to the obtained genes and performs homology-based searches in the genome of target organisms to predict a more thorough list of affected sequences and clusters these protein sequences to obtain homolog groups. Lastly, retrieving supplementary data, like pathways and Gene Ontology terms, it learns a Bayesian Network to demonstrate meaningful correlations between shared pathways and input chemicals.

We ran the pipeline on a group of phylogenetically distant taxa, namely Mus musculus, Danio rerio, Drosophila melanogaster and Caenorhabditis elegans, with input chemical groups of heavy-metals and dioxins. A total number of 8700 and 34102 affected genes were retrieved for heavy-metals and dioxins respectively. These genes were formed into 1866 and 7740 homologous gene groups in 101 and 125 pathways. The highly affected pathways suggested by our pipeline correspond with previous studies on toxic effects of these chemicals.

P40: Reducing the footprint of mass spectrometry data

Subject: Data management methods and systems

Presenting Author: Idoia Ochoa, University of Illinois at Urbana-Champaign

ABSTRACT: High-resolution mass spectrometry (MS) is a powerful technique used to identify and quantify molecules in simple and complex mixtures by separating molecular ions on the basis of their mass and charge. MS has become invaluable across a broad range of fields and applications, including biology, chemistry and physics, but
also clinical medicine and even space exploration. As a result, there has
been a rapid growth in the volume of MS data being generated. It is
therefore crucial to design better compression algorithms to lessen the
pressure of data storage and transmission.

The MS data is mainly composed of the mass-to-charge ratio (m/z)-
intensity pairs, which can take several GB of space per experiment.
We present a lossless compressor for these data that utilizes an
adaptive scheme for compression. Specifically, the proposed algorithm
compresses the floating-point pairs efficiently by calculating the
hexadecimal difference of consecutive m/z values, and by searching for
parts of the intensity values that match previous ones. The proposed
algorithm delivers an average compression ratio improvement of 40%
when compared to the lossless compressors GZIP, the current de-facto
compressor for MS data, and FPC, a state-of-the-art compressor for
floating point data. In particular, the proposed algorithm reduces the
MS file sizes by 56% on average.

P41: Identifying the molecular phenotypes of urothelial carcinoma of the bladder

Subject: other

Presenting Author: Oluwole Olaku, Morgan State University

ABSTRACT: Bladder cancer is the ninth most common cancer
worldwide, with an average of 150,000 cases per year and an
estimated 5-year survival rate makes it one of the major cause of
disease morbidity and mortality. Advances in the identification of
genomic alterations that lead to urothelial carcinoma of the bladder
(BCa), and recent approval of immune check point inhibitors have
provided durable systemic treatment options for patients with advanced
or metastatic bladder cancer. In view of positive clinical responses to
cabozantinib an immunotherapy drug in a recent phase II clinical trial
at the National Cancer Institute, a broad panel of putative cabozantinib
targeted receptor tyrosine kinases (RTKs) was examined to identify new
therapeutic targets. Alterations in The Cancer Genome Atlas (TCGA)
BCa datasets revealed that patients with muscle invasive disease had at
least one RTK gene amplified and/or mRNA upregulated. Furthermore,
many TCGA patient samples displayed both RTK and cognate ligand
overexpression, specifically GAS6/AXL, MST1/MST1R and CSF1/CSF1R
pathways, creating the potential for autocrine RTK signaling to drive
BCa oncogenesis in these patients. Recent studies on other bladder
cancer provisional datasets have identified groups of receptor tyrosine
kinases, DNA repair genes and transcriptional activators in the epithelial
mesenchymal transition (EMT) pathway. The goal of this research is to
identify the cohorts that have these genetic alterations in these groups
and observe the impact on overall and disease-free survival on those
with altered and unaltered genes.
P42: The use of signature genes and motifs to improve annotations of viruses

Subject: Metagenomics

Presenting Author: Dnyanada Pande, San Diego State University

Author(s): M. Ben Turner, San Diego State University, United States; Giselle Cavalcanti, San Diego State University, United States; Bhavya Papudeshi, National Center for Genome Analysis Support, United States; Elizabeth Dinsdale, San Diego State University, United States

ABSTRACT: Viruses are important in the biogeochemical cycling of oceans and drive bacterial evolution; however, the majority of viruses cannot be cultivated. Metagenomics, a culture independent method helps in studying these undescribed viral communities. Taxonomic and functional characterization of viruses and phages still remains a challenge because a significant proportion of viral metagenomes have low similarity with known viruses. In this study I investigate viral ‘dark matter’ within environmental metagenomes. I aim to improve current viral annotation methods through identification of virus-specific genetic motifs. Studies have identified certain conserved sequences called signature genes in some taxa which could be used to identify the presence of those taxa in unknown sequences. I first utilized these viral taxonomy-inferring signature genes and their corresponding motifs to identify virus families using homology searches by BLAST and MAST. For the reads that did not have homology with the known viral sequences in databases, I identified de novo motifs using the MEME tool. By implementing these two distinct approaches, I increased the number of annotated viral reads. I was able to decrease the percentage of unknown sequences from 90 % to 70 %. The next aim in this study is to reconstruct near-complete viral genomes from the metagenomic data, through extracting, assembling and binning all the sequences. The motifs that I identified for the specific viral family or genus will be searched in the binned sequences to help with the reconstructions of novel viruses. Overall, my study improves upon current bioinformatics methods in identification of viral sequences.

P43: National Center for Genome Analysis Support (NCGAS) as a platform for metagenomic analysis

Subject: Metagenomics

Presenting Author: Bhavya Papudeshi, Indiana University

Author(s): Sheri Sanders, Indiana University, United States; Carrie Ganote, Indiana University, United States; Philip Blood, Carnegie Mellon University, United States; Thomas Doak, Indiana University, United States

ABSTRACT: A metagenome contains all the microbial genetic material from an environmental sample, and could be from a human fecal sample to study human gut microbiome, a hospital environmental sample to study the microbes present within neonatal intensive
care units, a freshwater sample to monitor invasive species, or water from Great Barrier Reef to study coral disease. The vast applications of these techniques in biology has led to a drastic increase in the number of metagenomic-related projects, generating large datasets that exceed the ability of a personal computer to analyze. High Performance Computing (HPC) is being applied to these challenges, but most biologists do not have the necessary background to run analyses on HPC clusters. NCGAS is a collaborative project between Indiana University and Pittsburgh Supercomputing Center to enable the genomics community to analyze sequence data by providing HPC computational resources, data storage, and training and project consultation. Metagenomic analysis continues to evolve rapidly, with a wide array of bioinformatics software to select from at each step in an analysis. Selection of tools therefore remains a concern, and is being addressed by the metagenomics community through critical assessment of metagenomic interpretation (CAMI), to derive best practices and pipelines. NCGAS will continue providing training in HPC, data management and consultation to the metagenomics community, as well as work with users to accumulate user feedback on metagenomic tools, and feedback to the metagenomics community through CAMI and conferences, working towards establishing pipelines and reproducibility. I will use marine metagenomes to illustrate this process.

P44: Using functional interaction networks as prior knowledge in expression epistasis network centrality feature selection improves classification and enrichment of biologically relevant pathway for major depressive disorder

Subject: Machine learning

Presenting Author: Saeid Parvandeh, University of Tulsa

Author(s): Brett McKinney, University of Tulsa, United States

ABSTRACT: Recently, we developed an expression-epistasis network centrality method (SNPrank), analogous to Google’s PageRank. SNPrank captures the importance of gene-gene interactions and main effect of genes in a matrix we call the Genetic Association Interaction Network (GAIN). SNPrank includes a damping factor, which is a constant that balances the influence of main effects and interaction effects in the gene x gene statistical GAIN matrix. In this study we generalize the damping factor to be a vector that allows us to include gene-specific prior knowledge. We incorporate functional gene interactions from the Integrative Multi-species Prediction (IMP) database as prior knowledge via the damping vector. We used multiple microarray studies of major depressive disorder (MDD), to compare accuracies and pathway enrichments with and without prior knowledge. We used three different centralities: PageRank, Katz centrality, and SNPrank, and three different
classification methods: glmnet, random forest, and support vector machines. We find that prior knowledge improves enrichment of genes in relevant neural pathways and improves classification of MDD without overfitting.

**P45: Knowtator: A graphical text annotation plugin for Protegé**

*Subject: Graphics and user interfaces*

*Presenting Author: Harrison Pielke-Lombardo, University of Colorado*

**ABSTRACT:** We introduce an update to Knowtator, a text annotation plugin for Protegé, which features some new tools for more complex literature annotation and also brings it up to date with the current version of Protegé. Biological literature often describes complex processes and mechanisms which are difficult to capture using the available text annotation platforms. We aim to provide a tool that scales to arbitrarily complex annotations. Knowtator interacts with the Web Ontology Language (OWL) model built into Protegé to create annotation schemas. It also has tools for multi-annotator support and determining inter-annotator agreement. A major new tool is a graphical annotation pane which allows annotations to be manipulated outside of the original text as a graph structure. Annotations are linked with edges from the OWL model making it possible to annotate complex assertions. More features under development that we hope will assist annotators include link restrictions, annotation prediction, built in Part of Speech annotators, and annotation prediction.

**P46: Quality Assurance of Bioinformatics Software**

*Subject: Text Mining*

*Presenting Author: Morteza Pourreza Shahri, Montana State University*

*Author(s): Madhusudan Srinivasan, Montana State University, United States; Indika Kahanda, Montana State University, United States*

**ABSTRACT:** Bioinformatics software play a very important role in making critical decisions within many areas including medicine and health care. However, most of the research is directed toward developing tools, and little time and effort is spent on testing the software to assure their quality. The main challenge associated with conducting systematic testing on bioinformatics software is the oracle problem. An oracle is used to determine whether a test has passed or failed during testing, and unfortunately, for many of bioinformatics software, the exact expected outcomes are not well defined.

Metamorphic testing (MT) is a technique used to test programs that face the oracle problem. MT uses metamorphic relations (MRs) to determine whether a test has passed or failed and specifies how the
output should change according to a specific change made to the input. In this work, we use MT to test LingPipe, a tool for processing text using computational linguistics, often used in bioinformatics for bio-entity recognition from biomedical literature.

First, we identify a set of MRs for testing any bio-entity recognition program. Then we develop a set of test cases that can be used to test LingPipe's bio-entity recognition functionality using these MRs. To evaluate the effectiveness of this testing process, we automatically generate a set of faulty versions of LingPipe. According to our preliminary analysis of the results, we observe that our MRs can detect the majority of these faulty versions, which shows the utility of this testing technique for quality assurance of bioinformatics software.

P47: Evaluating the Effectiveness of Using Literature Features for Automated Protein Phenotype Prediction using PHENOstruct

Subject: Machine learning

Presenting Author: Morteza Pourreza Shahri, Montana State University

ABSTRACT: Human Phenotype Ontology (HPO) is a recently introduced standard vocabulary for describing disease-related phenotypic abnormalities in human. Experimental determination of HPO categories for human proteins is a highly-resource-consuming task. Therefore, developing automated tools that can accurately predict HPO categories for a given protein has gained interest over the last couple of years. In our previous work, we developed PHENOstruct, an automated phenotype prediction tool that uses features generated from protein-protein interactions, disease variants, experimentally validated functional annotations, and biomedical literature. In terms of literature features, we used a simple bag-of-word model which uses all the words of sentences that contain protein names as features. In this work, we introduce novel co-mention (CoM) features for PHENOstruct in which features are the HPO categories themselves and the feature values are the frequencies of each HPO term occurring within sentences that contain protein names. We use LingPipe and NCBO Annotator for extracting protein and phenotype names from a very large corpora of biomedical literature composed of 27 million Medline abstracts and 1.6 million PubMed full text articles, respectively. We use macro-AUROC (Area Under the Receiver Operating Characteristic curve) values obtained through five fold cross validation for assessing the effectiveness of the newly introduced CoM features on human data. Our preliminary results indicate that the addition of CoM features to the original set of features improves performance in the organ abnormality and mode of inheritance subontologies. These findings have implications for practitioners interested in developing automated biocuration pipelines for phenotypes.
P48: A rank-based hypergeometric test to extend GSEA for discerning specific effects of small molecule treatment

Subject: inference and pattern discovery

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

Author(s): Harrison Pielke-Lombardo, University of Colorado Anschutz Medical Campus, United States; Aik Choon Tan, University of Colorado Anschutz Medical Campus, United States; James Costello, University of Colorado Anschutz Medical Campus, United States

ABSTRACT: Small molecule drugs are low molecular weight compounds that rapidly diffuse across cell membranes to reach their molecular target, for example, inhibiting the activity of a specific kinase. When investigating the effect of a treatment on in vitro cultures, researchers commonly compare the genome-wide mRNA levels of drug-treated cells and vehicle-treated control cells. The resulting differentially expressed gene list can be analyzed with gene set enrichment analysis (GSEA), an algorithm which performs hypergeometric tests with curated gene sets to find biological processes that are more or less active in the drug-treated cells. However, even a highly-specific small molecule drug could result in non-specific effects in the cell, such as an up-regulation of generic stress response pathways. We hypothesized that this could overshadow crucial effects specific to the drug under investigation. To test this, we curated a collection of over 500 gene expression experiments using small molecule drugs. We used the GSEA Preranked algorithm to identify enriched gene sets in each experiment and found that gene sets related to DNA replication and hypoxia pathways were consistently overrepresented, regardless of the drug or cell line used. Therefore, to identify alterations more specific to a given experiment, we developed an extension of GSEA that uses our curated collection of small molecule experiments to inform the hypergeometric testing procedure. Evaluating our algorithm on experiments using drugs with known targets, we showed that it successfully prioritizes gene expression changes that are more likely due to direct drug effects, thus providing valuable experimental and biological insight.

P49: A systems biology approach using cell-to-cell heterogeneity to map cell fate decisions in developmental processes

Subject: Machine learning

Presenting Author: Jens Preussner, Max Planck Institute for Heart and Lung Research

Author(s): Guangshuai Jia, Max Planck Institute for Heart and Lung Research, Germany; Stefan Guenther, Max Planck Institute for Heart and Lung Research, Germany; Xuejun Yuan, Max Planck Institute for Heart and Lung Research, Germany; Michael Yekelchyk, Max Planck Institute for Heart and Lung Research, Germany; Carsten Kuenne, Max Planck Institute for Heart and Lung Research, Germany; Mario Looso, Max Planck Institute for Heart and Lung Research, Germany; Yonggang Zhou, Max Planck Institute for Heart and Lung Research, Germany; Thomas Braun, Max Planck Institute for Heart and Lung Research, Germany
ABSTRACT: Challenges for understanding developmental processes are the characterization of cell fate transitions, specification and hierarchical lineage descendants of progenitor cells. Differentiating cell populations typically manifest in profound cell-to-cell heterogeneity, justified by gene expression changes and gradual remodeling of underlying gene regulatory networks (GRNs). RNA sequencing at the single cell level allows the comprehensive characterization and analysis of cell population heterogeneity, especially if large ensembles of identical systems are profiled. Here we present computational approaches to harness statistical properties of heterogeneity with the objective of in-silico reconstruction of lineage trajectories and cell fate decision mapping in ~1500 differentiating cardiac progenitor cells. We define suitable measures of cell-to-cell heterogeneity and employ self-organizing maps and a density-based, hierarchical clustering to identify heterogeneous sub-populations. Reconstruction of the developmental trajectory revealed a bifurcation event when cells segregated into their terminal fates. Next, we show how a correlation-based analysis of cells in a transiently unstable state can be used to identify key regulatory elements in the underlying gene regulatory network that drive differentiation. We additionally show that those determinants progressively overcome epigenetic barriers to achieve open chromatin states associated with elevated expression in differentiated cell populations. Our approach not only comprehensively exploits heterogeneity and emerging correlations among large numbers of cells and genes to study early cardiogenesis at single-cell resolution, but also sets a general systems biology framework of transcriptional and epigenetic regulations in cell fate decisions.

P50: Correcting for non-random fragmentation allows for more accurate estimation of relative abundance in NGS metagenomic data

Subject: Metagenomics

Presenting Author: Elmar Pruesse, University of Colorado Denver

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

ABSTRACT: Accurately estimating the relative abundances of genes, contigs or entire genomes is essential to metagenomic studies. Variation in contig abundance between samples is used for de-novo in-silico organism isolation (binning); per genome or gene abundances allow comparative microbial community studies to establish associations of genes or microbes with disease states or environmental factors. However, current methods are surprisingly naive. Using simple averaging, most tools assume that reads are randomly sampled from the source DNA material.

This assumption, while acceptable in the past, is broken by modern NGS sequencing methods. Tagmerase based library preparation
protocols such as Nextera exhibit distinct biases in fragmentation point distribution. Sonication based fragmentation protocols exhibit less, but still noticeable, bias. These biases result in read depth variation at gene-scale windows far greater than to be expected from a Poisson distribution. This inhibits the binning of small contigs which often comprise the majority of an assembly. Moreover, the association of bias with DNA patterns potentially distorts relative abundances within gene families, which can lead to false positives or false negatives in comparative analyses.

We further present an algorithm for abundance estimation in the presence of non-random fragmentation. By learning fragmentation point preferences of the employed library preparation protocols from all mapped reads, the read depths can be adjusted for the positional likelihood of observing a read. Positions at which no fragmentation was observed are accounted for by incorporating the ratio of expected and observed zero observation positions. Overall, our Cython implementation demonstrates far lower variance in predicted abundance in small (1kb) window sizes than naive approaches, compensating for the effects of non-random fragmentation nearly completely.

**P51: Integrating Protein Localization Information in Signaling Pathway Reconstructions**

**Subject:** Graph Theory

**Presenting Author:** Anna Ritz, Reed College

**Author(s):** Ibrahim Youssef, Reed College, United States

**ABSTRACT:** Understanding the components of intracellular signaling pathways is an important task in systems biology. Computational methods have been developed to automatically reconstruct signaling pathways from large networks of protein-protein interactions (interactomes). These pathway reconstruction methods can accelerate the time-consuming manual curation of pathway databases and generate hypothesis that aid in the discovery of novel signaling components. PathLinker is a pathway reconstruction method that computes many short paths within an interactome that connect known receptors to known transcriptional regulators specific to a particular pathway (e.g., Wnt or EGFR). PathLinker has previously been shown to outperform other state-of-the-art methods (e.g., Steiner trees, network flow, and random walk approaches), and RNAi experiments have confirmed a PathLinker prediction of CFTR’s involvement in canonical Wnt signaling. Despite the improved performance over other methods, PathLinker reconstructions still contain erroneous signaling interactions. Here, we utilize information about protein localization within the cell to improve pathway reconstructions. By adding information about cellular compartments, we preserve the spatial hierarchy of signaling flow by explicitly finding paths that begin at the membrane, terminate at the
nucleus, and pass through intermediate compartments such as the cytosol. The new pathway reconstructions contain fewer false positives than the original PathLinker reconstructions, based on a benchmark dataset of pathways associated with cancer and immune signaling from the NetPath database. This work illustrates the utility of using additional biological information about protein interactions to improve predictions about signaling pathway interactions.

**P52: Design and evaluation of peptides that recognize the Tn antigen.**

*Subject: other*

*Presenting Author: Daniel Armando Romero Frenchy, National University of Colombia*

*Author(s): Edgar Antonio Reyes Montaño, National University of Colombia; Nohora Angelica Vega Castro, National University of Colombia*

**ABSTRACT:** The Tn antigen (GalNac α-Ser / Thr) was first described in patients with a rare hemolytic anemia (Tn syndrome) and since its presence was discovered in 90% of human carcinomas in the mid-1980s and has been the subject of an extensive study. It has been found that a high detection of this antigen is associated with more aggressive cancers and it appears in the early stages of different carcinomas being an alternative for the early diagnosis of these diseases. Currently, lectins such as B4 isoelectin from Vicia villosa, A4 isoelectin from Griffonia simplicifolia, as well as lectins from Salvia sclarea and Moluccella laevis among others, have been used for the identification of the Tn antigen. In the present study, the three-dimensional structures of Vicia villosa (PDB 1N47), Vatairea macrocarpa (PDB 4XTP), Psophocarpus tetragonolobus (PDB 2D3S), Glycine max (PDB 4D69) and Bauhinia forticata (PDB 5T5J) were used to identify the amino acids presented in the interaction of these lectins with the Tn antigen, then based on the results obtained we design and modeling peptides that were evaluated using molecular docking tests proving the Autodock VINA program, and then identify the more promising peptides to be used in vivo essays.

**P53: Measuring chromosome conformation by fluorescence microscopy**

*Subject: other*

*Presenting Author: Brian Ross, University of Colorado*

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

**ABSTRACT:** How to directly measure in-vivo chromosome conformation is an outstanding problem in structural biology. Whereas global conformational information can be inferred from DNA-DNA contact frequencies obtained using 3C-derived methods, direct measurements
of chromosomal positioning using fluorescence microscopy are limited to a very few loci that can be distinguished by color. One possible route to obtaining large-scale conformations directly by microscopy is to label many more loci than can be distinguished by color, and then computationally infer the identity of each imaged locus using the known color ordering and spacing of the labels along the chromosome. Here we report on improvements to one such reconstruction algorithm, and present experimental validation of the method from a 3-color in-situ hybridization (FISH) labeling of 10 loci on a 4 MB stretch of human chromosome 4. Our results show that we can both generate likely conformations as well as give unbiased statistical measures of the reconstruction quality.

P54: Discovering the Contribution of the Gut Microbiome to the Plasma Metabolome

Subject: Metagenomics

Presenting Author: Michael Shaffer, University of Colorado Denver Anschutz Medical Campus

Author(s): Catherine Lozupone, University of Colorado Denver Anschutz Medical Campus, United States; Nichole Reisdorph, University of Colorado Denver Anschutz Medical Campus, United States

ABSTRACT: The human body contains roughly the same number of bacterial cells as human cells and these bacteria encode for 150 times more genes. These microbes primarily live in the gut, produce metabolites that are transported all over the human body and have to potential to influence disease. While untargeted metabolomics can be used to investigate influence of microbial metabolites, determining whether disease-associated metabolites come from microbes, the host or the environment can be challenging. We have developed methods that use the KEGG and PICRUSt databases to predict the origin of metabolites and applied them to a set of 54 paired plasma metabolome and stool 16S microbiome samples. Using human KEGG genes we predicted 1376 unique compounds could be produced by the host and using PICRUSt and KEGG we predicted 1321 compounds produced by the detected gut microbial community. Of 1018 KEGG annotated compounds in the plasma metabolome, 155 were predicted human metabolites and 135 were bacterial. This cohort contains individuals with HIV and with HIV and lipodystrophy (a metabolic comorbidity); of compounds that differed with lipodystrophy 2 were predicted to only be produced by bacteria, 11 only by humans and 11 by both. However, the majority of metabolites in the plasma metabolome could not be assigned to KEGG IDs, limiting these techniques to only a small subset of metabolites. Our results highlight both the promise and challenges of using metabolic networks to predict bacterial origin of metabolites.
P55: Unsupervised discovery of phenotype specific heterogeneous gene networks

Subject: Machine learning

Presenting Author: Jenny Shi, University of Colorado Anschutz Medical Campus

Author(s): Pamela Russell, University of Colorado Anschutz Medical Campus, United States; Pratyaydipta Rudra, University of Colorado Anschutz Medical Campus, United States; Brian Vestal, National Jewish Health, United States; Brian Hobbs, Brigham and Women’s Hospital, United States; Craig Hersh, Brigham and Women’s Hospital, United States; Laura Saba, University of Colorado Anschutz Medical Campus, United States; Katerina Kechris, University of Colorado Anschutz Medical Campus, United States

ABSTRACT: Complex diseases often have a wide spectrum of symptoms. Better understanding the biological mechanism behind each symptom (a.k.a. phenotype) promotes targeted and effective treatment plans. We propose to utilize a machine learning technique, sparse canonical correlation analysis (SCCA), to integrate messenger RNA (mRNA) and microRNA (miRNA) expression data, taking a phenotype of interest into account. With canonical weights, mRNA-miRNA (i.e. heterogeneous) subnetworks that are specific to the phenotype can be constructed. Unlike traditional pairwise target prediction, the SCCA approach allows identification of associations that can be missed based on marginal correlations. We applied the method to a recombinant inbred mouse panel with endophenotypes that are associated with alcohol use disorders in human, and constructed heterogeneous subnetworks that are specific to drinking behavior. The leading subnetwork identified included three mRNA and four miRNAs, including two miRNAs from the same family. Most of the SCCA detected network connections were not predicted using pairwise methods, yielding novel associations. The strong associations discovered will be validated through biological experiments, such as knock-out. We also applied the method to a chronic obstructive pulmonary disease pilot study. The preliminary results revealed three subnetworks, which contain candidate features for more focused studies. The proposed SCCA method is not limited to expression data. It can be easily generalized to other data types, such as copy number variation, etc. It can also be applied to integrating more than three data types. The versatility of the approach will be useful in many other applications.

P56: Visual analysis of disease-associated multi-omics relationships in the human gut

Subject: Graphics and user interfaces

Presenting Author: Janet Siebert, University of Colorado Denver Anschutz Medical Campus

Author(s): Charles Neff, University of Colorado School of Medicine, United States; Brent Palmer, University of Colorado School of Medicine, United States; Catherine Lozupone, University of Colorado School of Medicine, United States; Carsten Görg, University of Colorado School of Medicine, United States
ABSTRACT: HIV-1 infection is associated with alterations in the gut microbiome and immune cell repertoire. However, it is unknown if these alterations drive or impact each other. Preliminary research suggests mechanistic relationships between immune cell subsets, gut microbes, and disease. Exploring these relationships demands expertise from both immunologists and microbiologists, and tools that allow these experts to navigate this team science context. Using paired 16S rRNA microbiome and CyTOF immune cell repertoire from mucosal biopsy of HIV-infected individuals and controls, we performed linear regressions coupled with permutation testing to identify pairs of microbe genera and immune cell subsets differentially related by disease state. This yielded a “Top N” list of statistically significant associations. Vetting these for scientific relevance is integral to honing hypotheses.

To support this vetting in a multi-omics domain, we conducted a visualization design study with our collaborating microbiologist and immunologist. We identified key tasks and derived associated tool requirements. High-level tasks included discovering, identifying, and comparing. For example, a result that connects an immune cell population with an emerging role (e.g. PD-1+ regulatory T cells) to a particular microbe may lead to a meaningful discovery.

We designed an interactive visualization that incorporated glyphs representing the linear models, a network illustrating communities within the results, and person-level parallel coordinates for demographics and selected analytes. We present results from the analysis of 19 human gut biopsies as well as feedback on an early prototype, solicited from our collaborators.

P57: Knowledge Network-guided analysis of genomics data on the Cloud

Subject: Machine learning

Presenting Author: Nahil Sobh, University of Illinois

ABSTRACT: We have developed ‘KnowEnG’ (Knowledge Engine for Genomics, pronounced ‘knowing’), an analytics engine and cyberinfrastructure for genomics data analysis. This has been the result of research and development conducted at the BD2K Center of Excellence at the University of Illinois, in partnership with the Mayo Clinic, Rochester, MN. The KnowEnG system is deployed on a public Cloud infrastructure (currently, AWS) to provide easy access to state-of-the-art and computationally intensive genomics analysis in a scalable and decentralized manner. KnowEnG supports new as well as established machine learning algorithms and statistical methods for ‘knowledge-guided’ analysis of genomics data sets in spreadsheet form. These analyses are powered by a massive ‘knowledge network’ encoding public domain knowledge of gene
properties and relationships. We have also made significant progress toward interoperability with other Cloud-based data repositories such as Cancer Genomics Cloud to provide KnowEnG’s analytical capabilities to users of those external repositories. The main driver (application) project for the technology development has been ‘cancer pharmacogenomics’ – the study of molecular mechanisms underlying cytotoxic drug response, using genomics data. We have also developed and applied new computational tools to address major problems in cancer biology and cancer treatment, as well as answer questions related to the evolution of social behavior using neurogenomics. This talk will present a brief overview of the KnowEnG system from the user’s perspective, while also touching upon some of the key features of the underlying technologies.

P58: Harmonization and Preparation of Clinical Data using Shiny

Subject: Data management methods and systems

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

Author(s): Carsten Görg, University of Colorado, Anschutz Medical School, United States; David Kao, University of Colorado, Anschutz Medical School, United States

ABSTRACT: Clinical observation data are produced at unprecedented rates. While the capacity to collect and store data rapidly grows, the ability to analyze these data volumes increases only incrementally. Depending on the size, quality, heterogeneity, and completeness of collected data, preparing the data for clinical research and statistical analyses can be a time-consuming process. We hypothesize that visual and interactive contexts can streamline data preparation through facilitating the identification of poor quality and heterogeneous data as well as providing intuitive mechanisms for modifying the data for conducting analyses. We created an interactive, web-based visual application for data cleaning and preparation. Focused on clinical trial, electronic medical record, and survey data, it enables researchers to assess the data quality and distribution, merge, remove or replace missing data, create subsets, and make transformations, all through the click of a button. We integrated tabular and linked interactive plotting views to aid in these tasks. To account for high-dimensional and big data, we utilized Apache Spark to minimize compute time and improve interactivity. Using four clinical cohort study datasets and two clinical survey datasets, we conducted three case studies to evaluate the tool’s utility to effectively prepare data in the context of different research questions. In addition, we gathered feedback from tool demonstrations to a variety of clinical researchers. Initial findings indicate that visual analysis and interactive contexts can simplify and enrich clinical data preprocessing, expediting the transition to data analytics and research.
ABSTRACT: Drug adverse events (AEs) are a major health threat to patients seeking medical treatment and a significant barrier in drug discovery and development. We performed systematic analysis of kinase inhibitors and their associated adverse events by querying clinical trial results reported in Clinicaltrials.gov. We obtained 368 kinase inhibitors from the Drug Repurposing Hub, and queried against our Kinase Inhibition Experiments Omnibus (KIEO) database to obtain kinase inhibition profiles. In total, we collected kinase inhibition and clinical trials data for 83 kinase inhibitors (35 approved and 48 investigational therapeutics) for this study. We extracted >1500 serious adverse events from 758 clinical trials covering 224,200 patients tested with kinase inhibitors. We also collected > 300 kinase inhibition data for these 83 kinase inhibitors. To determine the kinase inhibition – adverse events, we performed various computational and statistical analyses. In particular, we developed a computational method that integrates proportional reporting frequency of adverse events and kinase inhibition data for identifying kinase inhibition-adverse event (KI-AE) relationships. To validate the co-occurrence of kinase inhibition-adverse event detected in the clinical drug trials data, we queried the co-occurrence of KI-AE in PubMed. We computed the co-occurrence of the KI-AE correlation pairs using Fisher’s exact test, corrected by multiple comparisons. We also compared the KI-AE associations against SIDER web resources. Finally, we developed an interaction network for predicting new associations between kinase inhibition and adverse events. We envision that the KI-AE network will assist future drug discovery and development in reducing drug adverse events.
take into account the species info of the sequences in the multiple sequence alignment.

Here we introduce a phylogeny based approach, which is general to be incorporated with previous conservation measures. We first construct a reference taxonomic tree with leaves being taxonomy nodes of rank “class”, like Mammalia and Sauropsida. Then we group sequences in multiple sequence alignment into these class nodes, and compute conservation score of choice for sequences of each class. The scores are then added up together weighting for the distance between classes in the taxonomic tree. To validate, we use the previously published experimental mutations effects. We define positions with deleterious mutations as functionally important residues and positions with only neutral mutations as unimportant residues.

We find our phylogeny based approach to improve predictions of functionally important residues while using different conservation measures, including Shannon entropy, Jensen Shannon divergence and the Evolution Trace method, of which the Evolution Trace method has the best performance. The schematic improvement could come from reduced weights for sequences that are taxonomically distant from the query sequence. Our analysis also reveals that window heuristic which averages sequentially neighboring sites does not improve prediction of residues with mutation effects, the improved overlap with catalytic and ligand binding sites could due to the fact that neighboring sites are often labelled together.

**P61:** “Parent” of duplication gene pairs tends to be more conserved than “daughter” copy, and is more likely to be further duplicated

*Subject: inference and pattern discovery*

*Presenting Author: Haiming Tang, University of Southern California*

*Author(s): Paul Thomas, University of Southern California, United States*

*ABSTRACT:* Gene duplication is a major mechanism through which new genetic material is generated. Paralogs of a genome may come from multiple duplication events at different evolutionary periods, making deciphering the “Parent-Daughter” relationship of each duplication event a forbidden task.

The hypothesized 2 rounds of whole genome duplications (2R WGDs) at early vertebrates provides a unique perspective for decoding the “Parent-Daughter” relationships. The whole genome duplications result in sequential homologous genes with conserved gene orders. Using 17 extant vertebrates, we extensively examine the within and between genomes synteny evidences by extracting genes that are descended
from duplication at early vertebrates. Through additional phylogenetic analysis, we reconstruct duplication events at periods younger than early vertebrates, like Mammalia, or Primates. By examining descendants of these younger duplication events, we could identify the “Parent” gene from the later “daughter” gene as the “Parent” gene is from the 2R WGDs and should be located within synteny blocks.

Our study reveals that the “Parent” copy has significantly smaller branch length compared with the “Daughter” copy, indicating that the “parent” copy is more resistant to mutations. When there are several rounds of duplications after early vertebrates, in 892 of 1198 (74.5%) cases, the “Parent” copy continues to be the “Parent” of the younger duplication events which lead to “grand-daughters”. In 110 cases, the “Parent” copy at WGD remains the role of “Parent” for each of the later duplication events. This study is a first attempt to reveal the actual history of gene duplications during the genome evolution.

**P62: Meeting today’s challenging computational biology demands by integrating a focused IT team into the research process**

*Subject: Data management methods and systems*

*Presenting Author: Dan Timmons, University of Colorado Boulder/BioFrontiers Institute*

*Author(s): Matt Hynes-Grace, University of Colorado Boulder, United States; Jon Demasi, University of Colorado Boulder, United States; Robin Dowell, University of Colorado Boulder, United States; Mary Allen, University of Colorado Boulder, United States; Cassidy Thompson, University of Colorado Boulder, United States*

**ABSTRACT:** BioFrontiers IT (BIT) team is taking steps towards meeting today’s rapidly changing HPC needs to support life science and health informatics research. This initiative includes team members immersing themselves in researchers’ work through one-on-one support and consultation, workshops to introduce tools to aid in reproducibility and data integrity, and the facilitation of hackathons to bring together researchers and programmers working through innovative, interdisciplinary projects. Through these personal interactions, BioFrontiers IT members are able to ascertain practical knowledge from researchers that helps to engineer large-scale data and project management-oriented tools to support the community at large. To further this goal, and recognizing that computational research in academia benefits from collaboration with industry partners, BIT strives to create environments to foster mutually beneficial connections. These efforts have led to an increased integration between BIT and supported researchers which allows a more complete involvement in experimental design and execution, thus allowing BIT to step beyond a traditional troubleshooting role to provide more value with direct partnership. *
P63: Knowledge-based chemical relation extraction

Subject: Text Mining

Presenting Author: Ignacio Tripodi, University of Colorado, Boulder

Author(s): Lawrence Hunter, University of Colorado, Denver, United States; Mayla Boguslav, University of Colorado, Denver, United States; Negacy Hailu, University of Colorado, Denver, United States

ABSTRACT: Prior knowledge about how a chemical interacts with genes or proteins is valuable in predictive computational toxicology. Many relationships between chemicals and proteins (or genes) have been catalogued in various databases. However, these databases are incomplete; some information can be found only in the literature. Here, we describe a text mining approach that leverages information in databases to improve the quality of text mining with the goal of identifying relationships missing from those databases. Automated relation extraction from text is difficulty due to the many ambiguities in natural language. The current state of the art consists of selecting features such as words, word stems and syntactic information, and using them as inputs to a machine learning classifier. Here, we demonstrate that automatic identification of relationships between chemicals and proteins found in publications can be enriched by adding prior knowledge about the chemicals and proteins found in existing databases to the features used in machine learning. We integrate knowledge from many different databases using the KaBOB [1] knowledge-base, to automatically identify a set of five possible relations (“upregulation”, “downregulation”, “antagonist”, “agonist”, and “substrate of”) between a chemical and a protein mentioned in PubMed abstracts. The knowledge-base incorporates information about the chemicals and proteins (i.e. “participates in kinase activity”, “has N aromatic rings”, “it’s lipoxygenase activating”, etc). We tested our approach on an extensive manually annotated set of relations from the ChemProt [2] database (including therapeutics), using this prior knowledge in conjunction with text-derived features.

Feature selection algorithms and post-hoc analysis of the machine learning results identifies the aspects of the prior knowledge that were most helpful. Furthermore, these results can now be used to estimate the probability of each of these relationships between any chemical-protein pair, based on their attributes in the knowledge-base.


P64: Statistical Integration and Feature Selection for Candidate Biomarker Discovery

Subject: Machine learning

Presenting Author: Bobbie-Jo Webb-Robertson, Pacific Northwest National Laboratory

Author(s): Lisa Bramer, Pacific Northwest National Laboratory, United States; Sarah Reehl, Pacific Northwest National Laboratory, United States; Brigette Frohnert, University of Colorado, United States; Marian Rewers, University of Colorado, United States

ABSTRACT: High-throughput technologies currently have the capability to capture information at both global and targeted scales for the transcriptome, proteome, metabolome, and lipidome as well as determining functional aspects of these biomolecules. The promise of data integration is that by utilizing these disparate data streams, in combination with low-throughput clinical information, a more complete or accurate estimate of system behavior can be obtained. In the case of biomarker discovery to better diagnose and predict outcomes of disease one goal is to identify the best subset of molecules that can separate specific phenotypes of interest. However, in a space of tens of thousands of variables (e.g., genes, proteins), feature selection approaches often yield over-trained models with poor predictive power. Moreover, feature selection algorithms are typically focused on single sources of information and do not evaluate the effect on downstream statistical integration models. We present an ensemble-based feature selection approach that optimizes the outcome of interest in the context of the integrated posterior probability. We demonstrate that this approach improves sensitivity and specificity over simple selection routines based on individual datasets and present the application of the approach on a juvenile diabetes cohort.

P65: Pattern Recognition in Full-Text Biomedical Articles

Subject: Text Mining

Presenting Author: Elizabeth White, University of Colorado-Denver

ABSTRACT: Full-text biomedical publications have long sentences with complex structures, and statistical parsers are error-prone on these texts. Our lab has built the Colorado Richly Annotated Full-Text (CRAFT) annotation project, which contains over 90 full-text articles on mouse biology, marked up not only for biologically relevant concepts, but also for syntax, and hand-checked by linguists. Mistaken parses bedevil much work in biological NLP, but CRAFT is a gold standard that represents a rich resource for natural language processing of full-text articles.

Our group focuses on modeling knowledge from the scientific literature. Part of this effort leads us to trace out contradictions and uncertainty in scientific writing; many of these controversial areas indicate a fertile
zone for future work. We already have a rich library of words and phrases indicating hedging, disagreement, and certainty, but with a syntactic context added, we expect to see far fewer mistaken matches. Adding syntactic information to these patterns refines them at little search cost. We show a sample of patterns that rely on curated parses of CRAFT sentences and compare them to patterns that use the statistical language parses from annotators.

**P66: MeTeOR: a Literature Network for Hypothesis Generation and Precision Medicine**

*Subject: Graph Theory*

*Presenting Author: Stephen Wilson, Baylor College of Medicine*

*Author(s):* Angela Dawn, Baylor College of Medicine, United States; Matthew Holt, Baylor College of Medicine, United States; Daniel Konecki, Baylor College of Medicine, United States; Chih-Hsu Lin, Baylor College of Medicine, United States; Amanda Koire, Baylor College of Medicine, United States; Yue Chen, Baylor College of Medicine, United States; Yi Wang, Baylor College of Medicine, United States; Brigitta Wastuwidyaningtyas, Baylor College of Medicine, United States; Jun Qin, Baylor College of Medicine, United States; Olivier Lichtarge, Baylor College of Medicine, United States

**ABSTRACT:** The paradigm of hypothesis-driven discovery depends on an understanding of the literature to guide experiments. However, as the number of publications grows exponentially, the ability to read the literature increasingly falls short. To relieve this bottleneck, we introduce MeTeOR, a network for automated knowledge summarization and hypothesis generation, which aggregates PubMed articles by connecting MeSH indexing terms curated by the National Library of Medicine. MeTeOR tallies more novel associations among genes, drugs and diseases than other databases and is more reliable than existing natural language processing algorithms. When combined with automated hypothesis generation, MeTeOR analyses of past literature predicted associations discovered afterwards. In a prospective example, immunoprecipitation mass spectrometry supported both known and novel epidermal growth factor receptor associations predicted by MeTeOR. We conclude that MeTeOR generates valuable integrative hypotheses through a uniquely broad and accurate summarization of PubMed knowledge.

**P67: A comparison of error correction algorithms in T-cell receptor sequencing experiments**

*Subject: inference and pattern discovery*

*Presenting Author: Nicolle Witte, University of Colorado Anschutz Medical Campus*

*Author(s):* Debashis Ghosh, University of Colorado Anschutz Medical Campus, United States

**ABSTRACT:** Classifying T-cell receptor (TCR) repertoires is a fundamental step in interpreting the diversity of immune responses between
different patient cohorts and disease states. Somatic recombination is the driving mechanism that leads to immense TCR diversity. With the emergence of high throughput sequencing, genomic sequences of a sample’s TCR repertoire are available. However, the intrinsic nature of the recombination product results in closely homologous sequences with sparse mismatches due to differing alleles, making it difficult to distinguish whether a mismatch in a sequence alignment procedure is a novel gene/allele or an error originating from the pcr or sequencing process. Many efforts have been made to classify the cause of these mismatches, ultimately aiming to correctly classify groups of TCRs correlated to a disease state. Some algorithms use a predefined threshold of mismatch frequency, which can lead to a large fraction of sequences being omitted when they realistically could contain novel genes or alleles. Other algorithms model the likelihood of a mismatch being a novel gene, allele, or error using biological principles of the recombination event, resulting in more efficient use of the data. In this study, we compare the performance of several types of TCR data processing tools and discuss the fundamental differences of each of the algorithms. Deep sequencing reads of human glioma tumor tissue samples (SRP071932) are used in each of the tools, highlighting the strengths and weaknesses of each. Upon these results, we deliberate on future avenues to strengthen the utility and precision of TCR processing algorithms.

P68: Withdrawn

P69: A Pilot Systematic Genomic Comparison of Recurrence Risks of Hepatitis B Virus-associated Hepatocellular Carcinoma with low and high degree of liver fibrosis

Subject: inference and pattern discovery

Presenting Author: Seungyeul Yoo, Icahn School of Medicine at Mount Sinai

Author(s): Wenhui Wang, Icahn School of Medicine at Mount Sinai, United States; Qin Wang, Icahn School of Medicine at Mount Sinai, United States; Maria Fiel, Icahn School of Medicine at Mount Sinai, United States; Eunjee Lee, Icahn School of Medicine at Mount Sinai, United States; Spiros Hiotis, Icahn School of Medicine at Mount Sinai, United States; Jun Zhu, Icahn School of Medicine at Mount Sinai, United States

ABSTRACT: Chronic Hepatitis B virus (HBV) infection leads to liver fibrosis which is a major risk factor in Hepatocellular carcinoma (HCC). HBV genome can be inserted into human genome, and chronic inflammation may trigger somatic mutations. However, how HBV integration and other genomic changes contribute to the risk of tumor recurrence with regard to different degree of liver fibrosis is not clearly understood. We performed comprehensive genomic analyses of our
RNaseq data from 21 HBV-HCC patients treated at Mount Sinai Medical Center, and public available HBV-HCC sequencing data. Using a robust pipeline we developed, consistently more HBV integrations were identified in non-neoplastic liver than in tumor tissues. HBV host genes identified in non-neoplastic liver tissues significantly overlapped with known tumor suppressor genes. More significant enrichment of tumor suppressor genes was observed among HBV host genes identified from patients with tumor recurrence, indicating potential risk of tumor recurrence driven by HBV integration in non-neoplastic liver tissues. Pathogenic SNP loads in non-neoplastic liver were consistently higher than ones in normal liver tissues. And HBV host genes identified in non-neoplastic liver tissues significantly overlapped with pathogenic somatic mutations, suggesting HBV integration and somatic mutations targeting the same set of genes important to tumorigenesis. HBV integrations and pathogenic mutations showed distinct patterns between low and high liver fibrosis patients with regard to tumor recurrence. The results suggest that HBV integrations and pathogenic SNPs in non-neoplastic tissues are important for tumorigenesis and different recurrence risk models are needed for patients of low and high liver fibrosis.

**P70: Inference of protein networks and protein function in nontuberculous mycobacterial pathogens**

*Subject: inference and pattern discovery*

*Presenting Author: Diana Zajac, University of Colorado, Denver*

**ABSTRACT:** Pulmonary nontuberculous mycobacterial (NTM) infections pose an increasing threat to human health. Although NTM bacteria are phylogenetically related to the pathogenic bacteria that cause tuberculosis disease, much less is known about NTM and their mechanisms of pathogenesis. The primary species of NTM that cause disease are Mycobacterium abscessus and members of the Mycobacterium avium complex. In order to gain a better understanding of the coding capacity of the pathogen genome, how the encoded proteins function together, and to infer function for previously non-annotated genes, we performed a progressive analysis strategy to construct genome-wide protein networks using literature derived and computationally inferred functional linkages. We utilized linkages inferred in the STRING database, constructed and visualized networks using cytoscape, and identified novel network clusters containing previously non-annotated proteins that may be involved in virulence or enable the bacteria to invade human cells, including non-annotated proteins linked to mammalian cell entry proteins. As our laboratory continues to sequence hundreds of clinical NTM strains, these and other computational methods will become increasingly important to increasing our understanding of the functional coding capacity of NTM genomes.
P71: Enabling Deep Learning on Structural Biological Data

Subject: Machine learning
Presenting Author: Rafael Zamora, Lawrence Berkeley National Laboratory
Author(s): Tom Corcoran, Lawrence Berkeley National Laboratory, United States

Abstract: Convolutional Neural Network (CNN)-based machine learning has made noticeable breakthroughs in feature extraction tasks, but its applications in protein research are constrained by limited training data availability, as well as by the disparity between the 2D-oriented functionality of mainstream CNN technology and the inherently 3D nature of protein structures. We present a mapping algorithm that converts 3D structures to 2D data grids by first traversing the 3D space with a space-filling curve, encoding the 3D structural information into a 1D vector, and then projecting that vector into 2D via a reverse process with a complementary curve. For comparison against state-of-the-art CNN-based classification methods, we explore the performance of 2D CNNs trained on data encoded with our method versus comparable volumetric CNNs operating upon raw 3D data. Our results indicate that our mapping process preserves sufficient locality information across the transformation to be useful for training 2D CNNs on classification tasks between proteins and other structural models. We show that 2D CNNs trained on data generated using our method perform equivalently to state-of-the-art volumetric methods in terms of accuracy and generalizability while offering the potential for decreased training time cost compared to their 3D counterparts. We discuss several experiments that show the effectiveness of our approach, including classifying between everyday 3D object models from the ModelNet10 benchmarking dataset and between protein models sourced from the Protein Data Bank such as KRas and HRas. An implementation of our encoding process and neural network architectures is available for download at GitHub.

P72: Genotype-phenotype Association Discovery through Accurately Dissecting Phenotype-associated Genetic Causal Effects

Subject: Optimization and search
Presenting Author: Patrick Xuechun Zhao, Noble Research Institute
Author(s): Wenchao Zhang, Noble Research Institute, United States; Bongsong Kim, Noble Research Institute, United States; Xinbin Dai, Noble Research Institute, United States; Shizhong Xu, University of California, Riverside, United States

Abstract: Phenotypes, often called traits, are controlled by many genes (G), the interactions among genes (GxG) and genes with environment (GxE). The GxG and GxE effects are thought to contribute
significantly to the phenotypic variations of the complex traits. Generally, genetic variances (G; GxG) can be partitioned into 1) main effect that represents the cumulative effect of both individual genes/loci (additive effects) and dominant allelic interactions (dominance effects) to a given phenotype and 2) epistasis effect that refers to a trait that can be attributed to the interactions between multiple genetic loci. We developed a trio of genotype-phenotype association analysis tools, namely 1) GWASPRO (bioinfo.noble.org/GWASPRO/), which is designed to analyze main effects in large-scale genome-wide association studies (GWASs); 2) PEPIS (bioinfo.noble.org/PolyGenic_QTL/), which adopts a full polygenic linear mixed model (LMM) to analyze the additive, dominance and epistatic effects in GWASs and quantitative trait loci mapping; and 3) PATOWAS (bioinfo.noble.org/PATOWAS/), which further extends the LMMs for broader associative ‘omics’ studies, i.e. can be applied to not only GWASs, but also transcriptomics-wide association studies (TWASs), and metabolomics-wide association studies (MWASs). In our case analysis of a publically available Immortalized F2 (IMF2) associative omics study data, we found only about 66% of the total phenotypic variance could be explained in GWAS results, while about 99% of the phenotypic variances were accounted for in the TWAS results, suggesting that the GxG and GxE involved in transcriptomic level. Our case studies demonstrated the high performance of our tools for genetic variances analysis, enabling genotype-phenotype association discovery.

P73: Utilizing DNA methylation for Genome Annotation through Deep Learning

Subject: Machine learning
Presenting Author: Gregory Zynda, Texas Advanced Computing Center

ABSTRACT: Historically, genome assemblies and annotations were the product of large collaborations and years of work. The first Eukaryotic genomes were first assembled using Sanger sequencing methods and long insert library preparation. Modern, single-molecule sequencing technologies generate reads on the kilobase scale. These long reads simplify the assembly, but Eukaryotic genome annotation is not a push-button process. Annotations usually begin from predictions, which are then used in pipelines and eventually the application of manual heuristics to classify regions of protein-coding RNA, non-coding RNA, and repetitive elements.

We believe the prediction process can be improved by incorporating DNA methylation data, which adds another dimension to the genomic landscape. We have previously shown that Hidden Markov Models can be trained on known annotations and yield a 96% sensitivity and 79% precision rates at gene classification on A. thaliana. We are now exploring the use of deep neural networks to classify genomic regions
based on their sequence, strand, and methylation frequency. We plan on training our model on the Araport 11 annotation and version 3 of the Maize annotation, and then validating the results using version 4 of the Maize annotation. In the future, we hope our methods can utilize the methylation detected by long-read technologies to generate a reliable annotation directly from the data used for the genome assembly.

P74: Transcriptional regulatory network inference from gene expression and chromatin accessibility measurements

Subject: inference and pattern discovery

Presenting Author: Ronald Taylor, Cincinnati Children’s Hospital Medical Center

Co-Author(s): Emily Miraldi, Cincinnati Children’s Hospital Medical Center, Immunobiology, United States; Jason Hall, NYU School of Medicine, Skirball Institute, United States; Dayanne Martins de Castro, New York University, Center for Genomics and Systems Biology, United States; Ren Yi, Flatiron Institute, Center for Computational Biology, United States; Nick De Veaux, Flatiron Institute, Center for Computational Biology, United States; Aaron Watters, Flatiron Institute, Center for Computational Biology, United States; Dan Litman, NYU School of Medicine, Skirball Institute, United States; Richard Bonneau, Flatiron Institute, Center for Computational Biology, United States; Nicholas Carriero, Flatiron Institute, Center for Computational Biology, United States

ABSTRACT: Gene expression profiling provides a global readout on what cells are doing, and transcriptional regulatory network (TRN) models provide mechanistic insight into how these behaviors are regulated. The Assay for Transposase Accessible Chromatin (ATAC)-seq provides a unique opportunity for inference of the TRNs, especially for cell types and contexts where sample material is limiting and a priori knowledge of transcriptional regulation is scarce. Integration of accessible chromatin regions with TF motif database provides a large initial network (~10^6 edges), where putative interactions are based on TF motif occurrences in accessible regions cis to gene loci. Integration with parallel gene expression measurements further refines TRNs, yielding ~10^4 - ~10^5 edges, depending on inference procedure employed (e.g., model selection). We first validate our methods in T Helper 17 (Th17) cells, integrating new ATAC-seq data with published RNA-seq data, and use of TF knockout and ChIP-seq to evaluate model performance. Then we infer TRNs for intestinal innate lymphoid cells (ILCs), where very few transcriptional regulatory interactions are known. We validate the ILC TRN models both computationally (using gene expression prediction in new tissues) and experimentally through TF perturbation response measurements. We rigorously demonstrate the strength of our method to learn predictive transcriptional regulatory network models from ATAC-seq and RNA-seq experimental designs. The inferred regulatory networks in these cell types could provide guidance in regard to control points for human gene therapy.
SPONSORS

GOLD

IBM’s OpenPOWER Group is the organization that includes high performance computing (HPC) within the IBM Systems 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 is a partner with IBM’s HPC and OpenPOWER on developing the next generation of high performance data centric computers. In addition, the Research Division has many groups investigating numerous application areas in collaboration with IBM’s customers and partners in the life sciences. This includes IBM Research’s Data Centric Solutions and Computational Biology Center.

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.aspx#sthash.Fj3ONBCS.dpuf
Biodesix® is a fully-integrated molecular diagnostics company with multi-omic (DNA, RNA, Protein) test development capabilities leading to robust, clinically validated, blood-based tests for oncology. Our design and development teams use a specific machine learning approach to develop tests for researchers, bio-pharma partners, and clinicians. Recent clinical data show that the Biodesix test development approach, which utilizes the circulating proteome, provides unique and important information on the efficacy of immunotherapies and targeted agents. On market GeneStrat® and VeriStrat® tests are performed in a CAP/CLIA/CLEP certified laboratory with results in 72 hours for patients with non-small cell lung cancer. For more information, please visit www.biodesix.com.
JOIN US AT THE PREMIER COMPUTATIONAL BIOLOGY MEETING OF THE YEAR!
DISCOVER, INNOVATE AND CONNECT
MARK YOUR CALENDAR
Rocky 2017 is supported by the Computational Bioscience Program at the University of Colorado School of Medicine