Rocky 2013 is an official conference of the International Society for Computational Biology.

Rocky 2013 is supported by the Computational Bioscience Program at the University of Colorado School of Medicine.

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

Dear Rocky 2013 participant

Welcome to the Eleventh Rocky Mountain Regional Bioinformatics Meeting.

The organizers hope that you enjoy the program, and find the meeting a productive opportunity to meet researchers, students and industrial users of bioinformatics technology in our region. We are grateful for your continued interest in the meeting.

We are also grateful for the support of our sponsors. The ongoing sponsorship from IBM has made it possible to continue this conference for the past eleven years and SomaLogic/GoldLab for a fourth year in a row. 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, Suzi Smith from the ISCB team, and Kathy Thomas.

We hope you enjoy the science, the company, the hotel and the spectacular scenery of the Rocky Mountains.

Welcome!

Larry Hunter
WELCOME

Dear Rocky 2013 participant

Welcome to the Eleventh Rocky Mountain Regional Bioinformatics Meeting.

The organizers hope that you enjoy the program, and find the meeting a productive opportunity to meet researchers, students and industrial users of bioinformatics technology in our region. We are grateful for your continued interest in the meeting.

We are also grateful for the support of our sponsors. The ongoing sponsorship from IBM has made it possible to continue this conference for the past eleven years and SomaLogic/GoldLab for a fourth year in a row. 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, Suzi Smith from the ISCB team, and Kathy Thomas.

We hope you enjoy the science, the company, the hotel and the spectacular scenery of the Rocky Mountains.

Welcome!

Larry Hunter

SPONSORS

PLATINUM SPONSOR

IBM DEEP COMPUTING

1 Rogers Street, Cambridge, MA 02142

http://www-03.ibm.com/systems/deepcomputing/index.html

Kirk Jordan - kjordan@us.ibm.com 617-693-4581

IBM’s Technical Computing organization is the high performance computing organization in IBM Systems and Technology Group. This group is responsible for the strategy, marketing and identification of areas that can benefit from IBM’s high end technology. The life sciences is such an area, and IBM is and will continue to bring valued solutions to life sciences.

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

GOLD SPONSORS

GOLD LAB – UNIVERSITY OF COLORADO

Biology-MCD Instruction

University of Colorado at Boulder

http://www.goldlabcolorado.com/about.html

Roger Walz – roger.walz@colorado.edu

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

POSTER PRIZE SPONSOR
# Agenda at-a-glance

## Wednesday – December 11, 2013

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

## Thursday – December 12, 2013

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:00 AM</td>
<td>Registration</td>
</tr>
</tbody>
</table>
| 9:00 AM    | Keynote 1: Pathway Analysis of Genomics Data – from Correlation to Causation  
Gary Bader, PhD, University of Toronto, Canada |
| 9:45 AM    | Oral Presentations 1 - 4                   |
| 10:00 AM   | Break                                      |
| 10:45 AM   | Keynote 2: The Cure: A Game with the Purpose of Gene Selection for Breast Cancer Survival Prediction  
Benjamin Good, PhD, The Scripps Research Institute  
La Jolla, CA, United States |
| 11:15 AM   | Oral Presentations 5 - 8                   |
| 11:55 AM   | Ski Break                                  |
| 4:00 PM    | Keynote 3: A Flux Balance Analysis Model of E. coli K-12 MG1655 Derived From the EcoCyc Database  
Peter Karp, PhD, SRI International, Menlo Park, CA, United States |
| 5:10 PM    | Oral Presentations 9 - 12                  |
| 5:30 PM    | Break                                      |
| 6:15 PM    | Oral Presentations 13 - 15                 |
| 7:00 PM    | Banquet                                    |

*Il Poggio Restaurant, Snowmass Village (Transportation provided from Viceroy Hotel Lobby beginning at 6:45pm)*
### FRIDAY – DECEMBER 13, 2013

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:00 AM</td>
<td>Registration</td>
</tr>
<tr>
<td>9:00 AM</td>
<td>KEYNOTE 5: Tales from the Crypt: Do You Know Where Your Data Has Been?</td>
</tr>
<tr>
<td></td>
<td>Melissa Haendel, PhD, Oregon Health &amp; Science University, Portland, OR, United States</td>
</tr>
<tr>
<td>9:45 AM</td>
<td>ORAL PRESENTATIONS 16 - 19</td>
</tr>
<tr>
<td>10:25 AM</td>
<td>Break</td>
</tr>
<tr>
<td>10:45 AM</td>
<td>ORAL PRESENTATIONS 20 - 24</td>
</tr>
<tr>
<td>11:35 AM</td>
<td>KEYNOTE 6: Modeling Genetic Complexity with Integrated Interaction Analysis across Multiple Phenotypes</td>
</tr>
<tr>
<td></td>
<td>Greg Carter, PhD, The Jackson Laboratory, Bar Harbor, ME, United States</td>
</tr>
<tr>
<td>12:05 PM</td>
<td>Ski Break</td>
</tr>
<tr>
<td>4:00 PM</td>
<td>KEYNOTE 7: Application of Network Analysis to the Human Gut Microbiota and Use of Comparative Genomics to Understand the Driving Factors of Microbial Associations</td>
</tr>
<tr>
<td></td>
<td>Catherine Lozupone, PhD, Anschutz Medical Campus, Aurora, CO, United States</td>
</tr>
<tr>
<td>4:30 PM</td>
<td>ORAL PRESENTATIONS 25 - 30</td>
</tr>
<tr>
<td>5:30 PM</td>
<td>POSTER SESSION</td>
</tr>
</tbody>
</table>

### SATURDAY – DECEMBER 14, 2013

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:00 AM</td>
<td>Registration</td>
</tr>
<tr>
<td>9:00 AM</td>
<td>KEYNOTE 8: Solving Life Sciences Problems Requires Systems that Optimize the Workflow</td>
</tr>
<tr>
<td></td>
<td>Kirk E. Jordan, PhD, IBM T.J. Watson Research, Cambridge, MA, United States</td>
</tr>
<tr>
<td>9:30 AM</td>
<td>ORAL PRESENTATIONS 31 - 36</td>
</tr>
<tr>
<td>10:30 AM</td>
<td>Break</td>
</tr>
<tr>
<td>10:50 AM</td>
<td>ORAL PRESENTATIONS 37 - 39</td>
</tr>
<tr>
<td>11:30 AM</td>
<td>Closing</td>
</tr>
</tbody>
</table>

**Session Locations:** All oral presentations will be held at the Viceroy Hotel Ballroom Salon 1

**Poster Session Location:** Viceroy Hotel Ballroom Salon 3 & 4.
AGENDA

WEDNESDAY – DECEMBER 11, 2013

4:00 PM  6:00 PM  Registration

THURSDAY – DECEMBER 12, 2013
LOCATION: BALLROOM

<table>
<thead>
<tr>
<th>8:00 AM</th>
<th>9:00 AM</th>
<th>Registration</th>
</tr>
</thead>
</table>
| 9:00 AM | 9:45 AM | KEYNOTE 1: Pathway Analysis of Genomics Data – from Correlation to Causation  
GARY BADER, PhD, University of Toronto, Canada |
| 9:45 AM | 9:55 AM | ORAL PRESENTATION 1: AssemblSprint: Examining Genome Assemblies of Ion PGM next generation sequencing data from Mycobacterium species  
Presenting Author: NABEEH HASAN, University of Colorado, Denver, United States |
| 9:55 AM | 10:05 AM | ORAL PRESENTATION 2: Host-pathogen interactions and microRNAs  
Presenting Author: CHRISTIAN FORST, Icahn School of Medicine at Mount Sinai, United States |
| 10:05 AM | 10:15 AM | ORAL PRESENTATION 3: Recalibration of p-values for Multiple Testing Problems in Genomics  
Presenting Author: DEAN PALEJEV, Bulgarian Academy of Sciences, Bulgaria |
| 10:15 AM | 10:25 AM | ORAL PRESENTATION 4: Predicting disease related and pharmacogenes through curated and text-mined annotations  
Presenting Author: CHRISTOPHER FUNK, University of Colorado, United States |
| 10:25 AM | 10:45 AM | Break |
| 10:45 AM | 11:15 AM | KEYNOTE 2: The Cure: A Game with the Purpose of Gene Selection for Breast Cancer Survival Prediction  
BENJAMIN GOOD, PhD, The Scripps Research Institute, La Jolla, CA, United States |
| 11:15 AM | 11:25 AM | ORAL PRESENTATION 5: PAIRpred: A large margin method for partner-specific prediction of protein interfaces  
Presenting Author: FAYYAZ MINHAS, Colorado State University, United States |

Session Locations: All oral presentations will be held at the Viceroy Hotel Ballroom Salon 1  
Poster Session Location: Viceroy Hotel Ballroom Salon 3 & 4.
<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
<th>Title</th>
<th>Presenting Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>11:25 AM</td>
<td>11:35 AM</td>
<td>ORAL PRESENTATION 6: MP2GO: Inferring Gene Function from Phenotype</td>
<td>Presenting Author: JUDITH BLAKE, The Jackson Laboratory, United States</td>
</tr>
<tr>
<td>11:35 AM</td>
<td>11:45 AM</td>
<td>ORAL PRESENTATION 7: Temporal Expression Recognition for Cell Cycle Phase Concepts in Biomedical Literature</td>
<td>Presenting Author: NEGACY HAILU, University of Colorado Computational Bioscience Program, United States</td>
</tr>
<tr>
<td>11:45 AM</td>
<td>11:55 AM</td>
<td>ORAL PRESENTATION 8: Analyzing biological networks using degree-of-interest functions</td>
<td>Presenting Author: CORINNA VEHLOW, Visualization Research Center, University of Stuttgart, Germany</td>
</tr>
<tr>
<td>11:55 AM</td>
<td>4:00 PM</td>
<td>Ski Break</td>
<td></td>
</tr>
<tr>
<td>4:00 PM</td>
<td>4:30 PM</td>
<td>KEYNOTE 3: A Flux Balance Analysis Model of E. coli K-12 MG1655 Derived From the EcoCyc Database</td>
<td>Presenting Author: PETER KARP, PhD, SRI International, Menlo Park, CA, United States</td>
</tr>
<tr>
<td>4:30 PM</td>
<td>4:40 PM</td>
<td>ORAL PRESENTATION 9: The Critical Assessment of Function Annotation experiment: a community-wide effort towards a better functional annotation of genes and genomes</td>
<td>Presenting Author: SEAN MOONEY, Buck Institute for Research on Aging, United States</td>
</tr>
<tr>
<td>4:40 PM</td>
<td>4:50 PM</td>
<td>ORAL PRESENTATION 10: Conceptual comparison through integrative functional genomics in GeneWeaver.org</td>
<td>Presenting Author: ELISSA CHESLER, The Jackson Laboratory, United States</td>
</tr>
<tr>
<td>4:50 PM</td>
<td>5:00 PM</td>
<td>ORAL PRESENTATION 11: Integrative Visualization for Discovery of Phenotype Associations in Clinical and miRNA Data</td>
<td>Presenting Author: MICHAEL HINTERBERG, University of Colorado-Denver, United States</td>
</tr>
<tr>
<td>5:00 PM</td>
<td>5:10 PM</td>
<td>ORAL PRESENTATION 12: Evolution of palmitoyl acyl transferases (PATs) in apicomplexa</td>
<td>Presenting Author: SWAPNA SESHAHDI, Research Institute, Hospital for Sick Children, Canada</td>
</tr>
<tr>
<td>5:10 PM</td>
<td>5:30 PM</td>
<td>Break</td>
<td></td>
</tr>
<tr>
<td>5:30 PM</td>
<td>6:00 PM</td>
<td>KEYNOTE 4: eMBRLitMine and eMBRHelper: Bioinformatics Approaches for Improved Microbial Bioremediation Outcomes</td>
<td>Presenting Author: CHIJIOKE O. ELEKWACHI, PhD, University of Nottingham, United Kingdom</td>
</tr>
<tr>
<td>6:00 PM</td>
<td>6:10 PM</td>
<td>ORAL PRESENTATION 13: From Sample to Answer— Sequencing in the Cloud Era with the Illumina BaseSpace® Bioinformatics Platform</td>
<td>Presenting Author: RAYMOND TECOTZKY, Illumina, United States</td>
</tr>
<tr>
<td>Time</td>
<td>Event</td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>----------------------------------------------------------------------</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| 6:10 PM – 6:20 PM | ORAL PRESENTATION 14: MSProcess – Summarization, Normalization, and Diagnostics for Processing of Mass Spectrometry Based Metabolomic Data  
Presenting Author: KATERINA KECHRIS, University of Colorado Denver, United States |
| 6:20 PM – 6:30 PM | ORAL PRESENTATION 15: Metabolic Reconstruction Identifies Strain-Specific Regulation of Virulence in Toxoplasma gondii  
Presenting Author: NIRVANA NURSIMULU, University of Toronto, Canada |
| 7:00 PM – 10:00 PM | Banquet  
Location: Il Poggio Restaurant, Snowmass Village.  
(Transportation provided from Viceroy Hotel Lobby beginning at 6:45pm) |
| 8:00 AM – 5:30 PM | Registration |
| 9:00 AM – 9:45 AM | KEYNOTE 5: Tales from the Crypt: Do You Know Where Your Data Has Been?  
MELISSA HAENDEL, PhD, Oregon Health & Science University, Portland, OR, United States |
| 9:45 AM – 9:55 AM | ORAL PRESENTATION 16: Introducing Computations in Biology/Bioinformatics – An Undergraduate Perspective  
Presenting Author: MAHEEN KIBRIYA, Chapman University, United States |
| 9:55 AM – 10:05 AM | ORAL PRESENTATION 17: A Corpus-Based Study of Temporal Relations in Clinical Text  
Presenting Author: NATALYA PANTELEYEVA, University of Colorado, United States |
| 10:05 AM – 10:15 AM | ORAL PRESENTATION 18: Building an optimally informative machine-learning model of gene regulatory control  
Presenting Author: MOLLY MEGRAW, Oregon State University, United States |
| 10:15 AM – 10:25 AM | ORAL PRESENTATION 19: Characterizing Unknown Viral Genes Through Metabolomics  
Presenting Author: TIFFANY LIANG, San Diego State University, United States |
<p>| 10:25 AM – 10:45 AM | Break |</p>
<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
</tr>
</thead>
</table>
| 10:45 AM   | ORAL PRESENTATION 20: GeneSeer Aids Drug Discovery by Exploring Evolutionary Relationships Between Genes Across Genomes  
**Presenting Author:** DOUGLAS FENGER, Dart NeuroScience, United States |
| 10:55 AM   | ORAL PRESENTATION 21: The Human Gene Connectome Server: an Online Tool for Prioritizing Genes by Biological Distance  
**Presenting Author:** YUVAL ITAN, The Rockefeller University, United States |
| 11:05 AM   | ORAL PRESENTATION 22: Type I Error Rate Analysis of Methods for Correlated Binary Outcomes  
**Presenting Author:** AARTI MUNJAL, University of Colorado Denver, United States |
| 11:15 AM   | ORAL PRESENTATION 23: An Optimal Metabolic Route Search Tool: RouteSearch  
**Presenting Author:** MARIO LATENDRESSE, SRI International, United States |
| 11:25 AM   | ORAL PRESENTATION 24: Computational and Mathematical Modeling of the segmentation genes of Honeybee (Apis Mellifera)  
**Presenting Author:** MARYAM BAGHER OSKOUEI, University of Otago, New Zealand |
| 11:35 AM   | KEYNOTE 6: Modeling Genetic Complexity with Integrated Interaction Analysis across Multiple Phenotypes  
**GREG CARTER, PhD, The Jackson Laboratory, Bar Harbor, ME, United States** |
| 12:05 PM   | Ski Break                                                               |
| 4:00 PM    | KEYNOTE 7: Application of Network Analysis to the Human Gut Microbiota and Use of Comparative Genomics to Understand the Driving Factors of Microbial Associations  
**CATHERINE LOZUPONE, PhD, Anschutz Medical Campus, Aurora, CO, United States** |
| 4:30 PM    | ORAL PRESENTATION 25: Improved RNAi interference target sequencing (RIT-seq) enables dissection of cellular function in Trypanosome brucei  
**Presenting Author:** JONATHAN WILKES, University of Glasgow, United Kingdom |
| 4:40 PM    | ORAL PRESENTATION 26: What you did not know your transcription factor was doing  
**Presenting Author:** MARY ALLEN, University of Colorado, United States |
| 4:50 PM    | ORAL PRESENTATION 27: Linking genotype and enterotype in inflammatory bowel disease  
**Presenting Author:** DAN KNIGHTS, University of Minnesota, United States |
<table>
<thead>
<tr>
<th>Time</th>
<th>Time</th>
<th>Oral Presentation</th>
</tr>
</thead>
</table>
| 5:00 PM  | 5:10 PM  | Probing chromatin with digestion assays: from static nucleosome positioning to dynamic response  
Presenting Author: MICHAEL TOLSTORUKOV, Massachusetts General Hospital, United States     |
| 5:10 PM  | 5:20 PM  | GIST – an ensemble approach to the taxonomic classification of metatranscriptomic reads  
Presenting Author: JOHN PARKINSON, Hospital for Sick Children, Canada                       |
| 5:20 PM  | 5:30 PM  | Comparative metagenomics by cross-assembly  
Presenting Author: BAS DUTILH, Radboud University Medical Centre, The Netherlands          |
| 5:30 PM  | 7:30 PM  | Poster Session                                                                   |

**Saturday - December 14, 2013**

<table>
<thead>
<tr>
<th>Time</th>
<th>Time</th>
<th>Oral Presentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:00 AM</td>
<td>11:30 AM</td>
<td>Registration</td>
</tr>
</tbody>
</table>
| 9:00 AM  | 9:30 AM  | Keynote 8: Solving Life Sciences Problems Requires Systems that Optimize the Workflow  
KIRK E. JORDAN, PhD, IBM T.J. Watson Research, Cambridge, MA, United States               |
| 9:30 AM  | 9:40 AM  | An Update on KaBOB: towards an integrated knowledge base of biomedicine  
Presenting Author: KEVIN LIVINGSTON, University of Colorado, United States               |
| 9:40 AM  | 9:50 AM  | Hypernetworks: the Future of Network Modeling  
Presenting Author: DEBRA GOLDBERG, University of Colorado Boulder, United States         |
| 9:50 AM  | 10:00 AM | The changing view of protein interaction networks based on data availability  
Presenting Author: BARRETT HOSTETTER-LEWIS, California State University, Chico, United States |
| 10:00 AM | 10:10 AM | Experimental Determination of Useful and Informative Visualizations of Microbial Ecology Data for Public and Scientific Audiences  
Presenting Author: MEGAN PIRRUNG, University of Colorado, United States                  |
| 10:10 AM | 10:20 AM | Modeling Transcriptional Regulation through simulation of the dynamic changes in DNA binding factor configuration  
Presenting Author: DAVID KNOX, University of Colorado Anschutz Medical Campus, United States |
<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
<th>Presentation Title</th>
<th>Presenting Author</th>
<th>Institution</th>
</tr>
</thead>
<tbody>
<tr>
<td>10:20 AM</td>
<td>ORAL PRESENTATION 36</td>
<td>Deletion of COX-2 is associated with reduced expression of rgl1: relevance to chemopreventative effect of COX-2 inhibitors</td>
<td>NICHOLAS KIRKBY</td>
<td>Imperial College, London, United Kingdom</td>
</tr>
<tr>
<td>10:30 AM</td>
<td>Break</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10:50 AM</td>
<td>ORAL PRESENTATION 37</td>
<td>Prioritizing hypotheses for epigenetic mechanisms in Huntington’s Disease using an e-Science approach</td>
<td>ELENI MINA</td>
<td>Leiden University Medical Center, The Netherlands</td>
</tr>
<tr>
<td>11:00 AM</td>
<td>ORAL PRESENTATION 38</td>
<td>Variance component score test in a mixed-effects model framework to map tissue-specific eQTL</td>
<td>CHAITANYA ACHARYA</td>
<td>Duke University, United States</td>
</tr>
<tr>
<td>11:10 AM</td>
<td>ORAL PRESENTATION 39</td>
<td>High Throughput Phenotype Profiling for Bacterial Flux-Balance Model Optimization</td>
<td>DANIEL CUEVAS</td>
<td>San Diego State University, United States</td>
</tr>
<tr>
<td>11:30 AM</td>
<td>Closing</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
KEYNOTE SPEAKERS

GARY BADER, PhD

Associate Professor, University of Toronto, Canada

PATHWAY ANALYSIS OF GENOMICS DATA – FROM CORRELATION TO CAUSATION

ABSTRACT: Genomic data provides a snapshot of one or a few dimensions of information about an organism, such as gene or protein expression, promoter methylation and genome sequence in a set of cells. This information is typically interpreted using correlation based methods. For instance, a genome wide association study correlates a genetic marker with a phenotype and a set of gene expression profiles across multiple experiments can be correlated and clustered to identify groups of genes that act similarly and thus may be part of the same pathway. This approach is extremely valuable, but a major challenge is to better understand causative mechanisms underlying the genomic snapshot. For instance, we would like to know if the activity of a particular transcription factor or microRNA can explain the pattern of gene or pathway activity observed in an mRNA transcript profile. This could then be experimentally tested by perturbing the controlling factor. We are developing computational approaches to collect and use biochemical pathway information to help interpret genomics data and gain a more mechanistic understanding of cellular function.

GREG CARTER, PhD

The Jackson Laboratory, Bar Harbor, ME, United States

MODELING GENETIC COMPLEXITY WITH INTEGRATED INTERACTION ANALYSIS ACROSS MULTIPLE PHENOTYPES

Co-Authors: Anna Tyler, The Jackson Laboratory, United States, Vivek PhilipThe Jackson Laboratory, United States

ABSTRACT: Contemporary studies are revealing the genetic complexity of many traits in humans and model organisms. Two hallmarks of this complexity are epistasis, or gene-gene interaction, and pleiotropy, in which one gene affects multiple phenotypes. Understanding the genetic architecture of complex traits requires addressing these phenomena, but interpreting the biological significance of epistasis and pleiotropy is often difficult. While epistasis reveals dependencies between genetic variants, it is often unclear how the activity of one variant is specifically modifying the other. Epistasis found in one phenotypic context may disappear in another context, rendering the genetic interaction ambiguous. Pleiotropy can suggest either redundant phenotype measures or gene variants that affect multiple biological processes. Here we address these interpretation ambiguities with a method called combined analysis of pleiotropy and epistasis (CAPE). This approach integrates information from multiple related phenotypes to constrain models of epistasis, thereby enhancing the detection of interactions that simultaneously describe all
phenotypes. The networks inferred are readily interpretable in terms of directed influences that indicate suppressive and enhancing effects of individual genetic variants on other variants, which in turn account for the variance in quantitative traits. We demonstrate the utility of this approach by analyzing mouse intercross data, discovering a novel interaction network influencing kidney gene expression and disease. We have implemented this approach in an R package that can be applied to data from both genetic screens and a variety of segregating populations including backcrosses, intercrosses, and natural populations.

CHIJIOKE O. ELEKWACHI, PhD
University of Nottingham, United Kingdom

EMBRLITMINE AND EMBRHELPER: BIOINFORMATICS APPROACHES FOR IMPROVED MICROBIAL BIOREMEDIATION OUTCOMES
Co-Author: Charlie Hodgman, University of Nottingham, Multidisciplinary Centre for Integrative Biology (MyCIB)

ABSTRACT: Contamination of ecosystems by xenobiotic substances has negatively impacted affected ecologies and the health and economic livelihoods of human populations in such environments. Bioremediation has proven to be a safe, low-cost and environmentally friendly method for remediation of such areas. However, a lack of complete understanding of the metabolic, enzymatic and cellular processes involved has made it difficult to model or predict outcomes of field processes. Researchers’ ability to make critical decisions capable of influencing the direction and outcome of these processes is hampered, thereby hindering its development. Following a survey that highlighted priorities, practices and needs of the sector the environmental Microbial BioRemediation (eMBR) web-portal was developed. This article describes the structure, algorithms and output of two bioinformatics resources for improved microbial bioremediation outcomes, deployed via the portal.

eMBRLitMine helps in identifying which microorganisms would be suitable for remediating sites contaminated by named compounds. It combines named-entity-recognition algorithms, a mysql database, and graph-rendering technologies to create, from information available in literature, a statistical co-occurrence matrix from which it infers associations among microorganisms and contaminants. This provides insights into possible bacteria/contaminant relationships and highlights microorganisms that may be useful for remediating particular contaminants. Following the construction of a comprehensive metabolic biodegradation network eMBRHelper enables the delineation of possible biodegradation pathways for named contaminants. By integrating chemical, and enzymatic information, it attempts to model the interplay between contaminants, enzymes, microorganisms in degradation pathways, and enables researchers to make informed decisions, capable of improving outcomes of remediation exercises involving bio-augmentation.
Benjamin Good, PhD
Senior Staff Scientist, The Scripps Research Institute, La Jolla, CA, United States

The Cure: A Game with the Purpose of Gene Selection for Breast Cancer Survival Prediction
Co-Authors: Salvatore Loguercio, The Scripps Research Institute, Molecular and Experimental Medicine, Max Nanis, The Scripps Research Institute, Molecular and Experimental Medicine, Andrew Su, The Scripps Research Institute, Molecular and Experimental Medicine

Abstract: Breast cancer has been studied extensively with genomic technologies, with many attempts to devise molecular predictors of clinical outcomes. A key aspect to these studies is the selection of small, informative sets of genes with which to compose predictors. Many groups now apply prior knowledge in forms such as protein-protein interaction networks and pathway databases in the process of gene selection. However, these approaches can not make use of unstructured knowledge.

Since the year 2000, more than 160000 publications related to breast cancer have been added to PubMed. We created a ‘scientific discovery game’, called The Cure, to tap into the knowledge represented in those articles and in the minds of those who can read them. The objective of The Cure is to identify genes that can be used to build improved survival predictors for breast cancer. It is formulated as a card game with genes as cards. Hands are scored based on the value of the genes in creating a decision tree predictor of survival using their expression values. To win a game, players must select more predictive genes than their opponent. In one year, 1,077 players registered and collectively played 9,904 gene-selection games. Aggregating the results of game play, we generated a ranked list of preferentially selected genes. Preliminary analysis of this list indicates that it is non-random, with the higher-ranked genes showing overlaps with prior breast cancer gene lists as well as surfacing some novel predictive genes.

Melissa Haendel, PhD
Assistant Professor, Ontology Development Group, OHSU Library, Department of Medical Informatics and Epidemiology, Oregon Health & Science University, Portland, OR, United States

Tales from the Crypt: Do You Know Where Your Data Has Been?

Abstract: Researchers produce data that bioinformaticists analyze to generate hypotheses and novel discoveries, which feed back into basic or clinical research. It is a beautiful cycle, but as everyone knows, not all data is created equal and this can cause incorrect conclusions at best and broad propagation of errors at worst. We expect science to be reproducible and replicable. What does
this even mean and how can we work towards having greater confidence in our collective scientific knowledge? Here, we explore these fundamental issues that plague bioinformatics: stories about scientific reproducibility and mechanisms for improvement, efforts to understand how we know when a dataset is annotated sufficiently, and the importance of tracking provenance in data transformations through analytical pipelines. The goal is to instill in all of us increased scrutiny on the data we work with and to discuss development of informatics methods that can aid data quality analysis and metrics.

**KIRK E. JORDAN, PhD**

*IBM Distinguished Engineer, Emerging Solutions Executive & Assoc. Program Director, Computational Science Center, IBM T.J. Watson Research, Cambridge, MA, United States*

**SOLVING LIFE SCIENCES PROBLEMS REQUIRES SYSTEMS THAT OPTIMIZE THE WORKFLOW**

*Co-Authors: Mike Mehan, Alex Stewart, (Somalogic, Boulder CO), and Shintaro Kato (NEC corporation of America)*

**ABSTRACT:** The world and the life sciences are awash in data. The problem for computing is no longer the ability to compute but the inability to move the data to the compute. As a consequence, the focus is shifting to the concept can we move the compute to the data. This raises the question of optimizing the entire workflows to solve problems instead of optimizing a compute intensive kernel. In this talk, I will briefly expound on the concept of moving compute to the data. I will describe some of our ongoing investigations of trying to tackle entire workflows presenting some of these efforts and the impact this is having on looking at the entire workflow.

**PETER KARP, PhD**

*Director, Bioinformatics Research Group, Artificial Intelligence Center, SRI International, Menlo Park, CA, USA*

**A FLUX BALANCE ANALYSIS MODEL OF E. COLI K-12 MG1655 DERIVED FROM THE ECOCYC DATABASE**

*Co-Author: Daniel Weaver, SRI International, Bioinformatics Research Group, United States*

**ABSTRACT:** We present EcoCyc-17.5-FBA, a genome-scale model of the Escherichia coli K-12 MG1655 metabolic network. The model is automatically generated from the current state of the EcoCyc database using the MetaFlux component of Pathway
Tools, ensuring reflection of the current state of knowledge of the E. coli metabolic network. EcoCyc-17.5-FBA represents several advances in E. coli genome-scale models, breaking new ground in the number of genes modeled and the accuracy and breadth of its predictions. EcoCyc-17.5-FBA encompasses 1923 genes, 1996 unique metabolic reactions, and 2026 unique metabolites. We demonstrate a three-part validation of EcoCyc-17.5-FBA: (1) A comparison of simulated EcoCyc-17.5-FBA growth in aerobic and anaerobic glucose culture with experimental results from chemostat culture and simulation results from the E. coli modeling literature. (2) Essentiality predictions for all 1923 genes involved in the model, with an accuracy of 94.0%. (3) Nutrient utilization testing for viability on 478 different growth media, with an accuracy of 77.4%. The development of EcoCyc-17.5-FBA as an aspect of EcoCyc has improved the quality and depth of EcoCyc’s representation of E. coli. EcoCyc-17.5-FBA is a literate model in the sense that it is highly accessible to human comprehension: the component reactions, metabolites, and genes of the model are directly inspectable through the EcoCyc website, which includes a broad array of information that enriches the model, such as metabolite chemical structures, pathway diagrams, genome maps, and regulatory interactions.

Catherine Lozupone, PhD
Anschutz Medical Campus, Aurora, CO, United States

APPLICATION OF NETWORK ANALYSIS TO THE HUMAN GUT MICROBIOTA AND USE OF COMPARATIVE GENOMICS TO UNDERSTAND THE DRIVING FACTORS OF MICROBIAL ASSOCIATIONS

ABSTRACT: Application of network analysis to the human gut microbiota and use of comparative genomics to understand the driving factors of microbial associations

Non-random co-occurrence patterns of human gut bacteria across people can indicate important biological relationships between them. A positive association can indicate symbioses, such as a syntrophic interaction in which microbes support each other’s growth by cooperative metabolism. However a positive association can also be driven by indirect associations such as a shared preference for the same type of environment. For instance in a dataset containing samples with diverse oxygen concentration, strict anaerobes may positively occur simply because a shared dislike of oxygen. Similarly, negative associations can be driven by direct competition between microbes for the same resource, or by opposing environmental preferences. I will discuss methods for detecting significant co-occurrence interactions between microbes using data from culture-independent surveys of the human gut, and for predicting potential direct or indirect drivers of co-occurrence patterns by comparing the content of associated genomes.
**OP01: AssemblSprint: Examining Genome Assemblies of Ion PGM Next Generation Sequencing Data from Mycobacterium Species**

**Presenting Author:** NABEEH HASAN, University of Colorado, Denver, CO, United States  
**Co-Author(s):** Rebecca Davidson, National Jewish Health; Paul Reynolds, National Jewish Health; Eveline Farias-Hesson, Life Technologies; Michael Strong, National Jewish Health

**ABSTRACT:** Genome sequencing technologies differ in their speed, accuracy and capacity. These differences present challenges for genome assembly software to optimally perform with data from multiple sequencing platforms. Moreover, sequence assembly and finishing of genomes remains challenging in many instances. Robust methods to evaluate genome assemblies have not been well defined, particularly for data produced by newer sequencers, such as the Ion Torrent Personal Genome Machine (PGM) from Life Technologies, Inc. In our AssemblSprint, we examined the performance of three commonly used de novo genome assemblers: MIRA, Newbler and Velvet in assembling Ion PGM 400bp sequence data from two Mycobacterium species: M. abscessus ssp. massiliense and M. chelonae, along with Ion PGM 200bp sequence data from M. malmoense using varying read depths and read filtering parameters. Each assembly was evaluated based upon assembly length, number of scaffolds, N50 of scaffolds, number of ambiguous base calls, and comparisons to a given reference genome, if available. We find that Newbler performed the best with assembling Ion PGM data, but inter-assembly variation suggests a need for further research into methods to improve de novo genome assembly and evaluation.

**OP02: Host-Pathogen Interactions and MicroRNAs**

**Presenting Author:** CHRISTIAN FORST, Icahn School of Medicine at Mount Sinai, United States

**ABSTRACT:** MicroRNAs (miRNAs) are increasingly recognized to be important in the regulation of biological functions. Although the study of miRNAs in host-pathogen systems is still in its infancy, growing evidence suggests that miRNAs play a key role in the control of infection. We utilize time-series expression data for both miRNAs and mRNAs derived under identical conditions from cell cultures infected with influenza virus. By an integrative approach we combine expression data with context information to construct an miRNA/mRNA network. We have validated the resulting pathways with miRNA inhibitors revealing pro-viral and anti-viral miRNAs. In particular, we can identify miRNAs that are important players in the regulation of innate immune...
response — targets required for viral replication but also host resistance factors as defense against infection.

**OP03: RECALIBRATION OF P-VALUES FOR MULTIPLE TESTING PROBLEMS IN GENOMICS**

*Presenting Author: DEAN PALEJEV, Bulgarian Academy of Sciences, Bulgaria  
Co-Author(s): John Ferguson, University of Limerick*

**ABSTRACT:** Conservative statistical tests are often used in complex multiple testing settings in which computing the type I error may be difficult. In such tests, the reported p-value for a hypothesis can understate the evidence against the null hypothesis and consequently statistical power may be lost. False Discovery Rate adjustments, used in multiple comparison settings, can worsen the unfavorable effect. Despite these effects, the problem seems to be somewhat overlooked within the computational biology and bioinformatics communities, with many practitioners not even aware of the issue.

We present a computationally efficient and test-agnostic calibration technique that can substantially reduce the conservativeness of such tests. As a consequence, a lower sample size might be sufficient to reject the null hypothesis for true alternatives, and experimental costs can be lowered.

As an example, we apply the calibration technique to the results of DESeq, a popular method for detecting differentially expressed genes from high-throughput RNA sequencing data. The increase in power maybe particularly high in small sample size experiments, often used in preliminary experiments and funding applications. In some situations, after correction, statistical power can increase 3 fold without the need of additional experimental costs.

Our results are structured in a way that makes them easily usable by practitioners in the fields of computational biology and bioinformatics. We also provide R code that could be used, or modified by them as needed.

**OP04: PREDICTING DISEASE RELATED AND PHARMACOGENES THROUGH CURATED AND TEXT-MINED ANNOTATIONS**

*Presenting Author: CHRISTOPHER FUNK, University of Colorado, United States  
Co-Author(s): Lawrence Hunter, University of Colorado, Kevin Cohen, University of Colorado*

**ABSTRACT:** Identifying genetic variants that play a role in disease or affect drug response is an important task. Before individual variants can be explored efficiently specific candidate genes must be identified. While many methods rank candidate genes through the
use of sequence features and network topology, only a few exploit the information contained in the biomedical literature. We present a set of enriched pharmacogenic Gene Ontology concepts and train and test a classifier on known pharmacogenes from PharmGKB. Our classifier uses only Gene Ontology concept annotations and simple features mined from the biomedical literature; it is then used to predict pharmacogenes on a genome-wide scale. We achieve performance of $F=0.86$, $AUC=0.860$, on five-fold cross validation. Additionally, the top 10 predicted genes are analyzed.

**OP05: PAIRPRED: A LARGE MARGIN METHOD FOR PARTNER-SPECIFIC PREDICTION OF PROTEIN INTERFACES**

Presenting Author: FAYYAZ MINHAS, Colorado State University, United States
Co-Author(s): Brian Geiss, Colorado State University, Asa Ben-Hur, Colorado State University

**ABSTRACT:** We have developed a novel partner-specific protein-protein interaction site prediction method called PAIRpred that uses the sequences and unbound structures of two proteins in a complex, and is based on support vector machines (SVMs). Unlike most existing machine learning methods for this problem, PAIRpred uses information extracted from both proteins in a complex using pairwise kernels to predict inter-residue contacts. Due to its partner-specific nature, PAIRpred presents a more accurate model of protein binding and is able to generate more detailed predictions. In order to better model the problem, we present an extension of SVMs that can capture the pairwise constraints that two distant residues in a protein cannot simultaneously interact with the other protein in a complex. We demonstrate PAIRpred’s performance on Docking Benchmark 4.0 and recent CAPRI targets. We have compared PAIRpred’s performance to existing methods such as ZDOCK, PPIPP and PredUS. PAIRpred offers state of the art accuracy in predicting binding sites at the protein level as well as inter-protein residue contacts at the complex level. We have studied the contribution of different sequence and structure features along with the effect of binding-associated conformational change on prediction accuracy. As an illustration of potential applications of PAIRpred, we have used it to analyze the nature and specificity of the interface in the interaction of human ISG15 protein with NS1 protein from influenza A virus. More information on PAIRpred is available at: http://combi.cs.colostate.edu/supplements/pairpred/.
**OP06: MP2GO: INFERRING GENE FUNCTION FROM PHENOTYPE**

*Presenting Author: JUDITH BLAKE, The Jackson Laboratory, United States*
*Co-Author(s): Joao Ascensao, Rice University, Mary Dolan, The Jackson Laboratory, David Hill, The Jackson Laboratory*

**ABSTRACT:** Biomedical ontologies, while they have proven to be instrumental in the advancement of biological research through their ability to efficiently consolidate scientific data, are also hampered by the segregation of knowledge domains that results from their independent curation. We have developed a new method to computationally infer gene function, as encoded in the Gene Ontology (GO), from mutant phenotypes, as encoded in the Mammalian Phenotype Ontology (MP), using a set and graph theory-inspired approach. We apply this methodology to laboratory mouse (Mus musculus) data as represented in the Mouse Genome Informatics Resource (MGI). We believe this procedure represents a novel methodology for the inference of gene function, as it examines the emergent structure and relationships between the GO and MP annotations without considering the relationships semantically. This could allow for the discovery of unforeseen associations between gene function and phenotypes that would be overlooked by a semantic-based approach. The technique could be applied to a variety of other organisms and annotation databases, taking full advantage of the abundance of available high quality curated data.

**OP07: TEMPORAL EXPRESSION RECOGNITION FOR CELL CYCLE PHASE CONCEPTS IN BIOMEDICAL LITERATURE**

*Presenting Author: NEGACY HAILU, University of Colorado Computational Bioscience Program, United States*
*Co-Author(s): Kevin Cohen, University of Colorado School of Medicine*

**ABSTRACT:** The number of publications in the biomedical domain is increasing exponentially. Searching for papers specific to a researcher’s interest in this domain is difficult. PubMed allows search using keywords but it doesn’t rank results based on document relevance. We present a recognizer for temporal expressions related to Cell Cycle Phase (CCP) concepts in biomedical literature. This task is one of the fundamental tasks towards building a search engine for queries with temporal components. Our ultimate goal is to build a specialized search engine, which is specific to searches in the CCP using genes and small molecules. We seek to improve search accuracy by allowing searches using semantic indexing instead of keywords. We identified 11
cell cycle related temporal expressions, for which we made extensions to TIMEX3, arranging them in an ontology derived from the Gene Ontology. We annotated 310 abstracts from PubMed. We developed annotation guidelines which are consistent with existing time related annotation guidelines such as TimeML. Two annotators participated in the annotation. We computed inter-annotator Agreement (IAA). We achieved an IAA of 0.79 for exact span match and 0.82 for relaxed constraints. Our approach is a hybrid of machine learning to recognize the temporal expressions and a rule-based approach to classify them. We trained a named entity recognizer using Conditional Random Fields (CRFs) models. We used an off-the-shelf implementation of the linear chain CRF model. We obtained a performance of 0.77 F-score for temporal expression recognition. We achieved 0.79 and 0.78 macro and micro average F-scores for classification.

**OP08: ANALYZING BIOLOGICAL NETWORKS USING DEGREE-OF-INTEREST FUNCTIONS**

*Presenting Author: CORINNA VELOW, Visualization Research Center, University of Stuttgart, Germany*

*Co-Author(s): Carsten Goerg, University of Colorado Anschutz Medical Campus, David Kao, University of Colorado Anschutz Medical Campus*

**ABSTRACT:** Biologists commonly analyze experimental data using biological networks, such as gene-expression correlation networks, to explain disease specific patterns and identify genotype-phenotype relationships. Biomedical knowledge from various databases and the literature can be integrated with these data networks to allow analysts to interpret experimental data in the context of existing knowledge. While these combined networks provide a rich resource and profound basis for data analysis, they are difficult to explore and understand since they are very dense. Using current static visualization approaches, it takes time and expertise to “untangle the hairball” and manually extract sub-networks that can explain a phenomenon or tell a meaningful biological story. To improve this analytical workflow, we developed a visualization approach that applies the concept of degree-of-interest (DOI) functions to highlight or filter particular parts of a network that are relevant for a specific question or task. We also use these DOI functions to automatically extract and lay out sub-networks in a way that DOI-based groups and their intersections become visually apparent, e.g., extracting a sub-network that includes all nodes involved in a set of pathways of interest and visually arranging these nodes based on their pathway information. To facilitate the analysis of extracted sub-networks in the context of the complete network, the network visualizations are linked through a brushing and linking feature. DOI functions can model various analytical facets, including an analyst’s background and interest, properties of the experimental data, and phenotype information. Hence, they provide a generic and powerful approach for analyzing biological networks.
OP09: THE CRITICAL ASSESSMENT OF FUNCTION ANNOTATION EXPERIMENT: A COMMUNITY-WIDE EFFORT TOWARDS A BETTER FUNCTIONAL ANNOTATION OF GENES AND GENOMES

Presenting Author: SEAN MOONEY, Buck Institute for Research on Aging, United States
Co-Author(s): Predrag Radivojac, Indiana University, Iddo Friedberg, Miami University

ABSTRACT: A major challenge of the post-genomic era is understanding the function and disease associations of gene products. We are discovering new proteins far faster than we can characterize them experimentally. Most genome projects and derived databases rely fully on automated functional annotations, making the increase in annotation accuracy and coverage a prime goal for annotation algorithms. Understanding the accuracy of these function prediction algorithms is of primary importance to the process of translating sequence data into biologically meaningful information. Here we present the results of the first Critical Assessment of Function Annotations (CAFA) held during 2010-2011 and the challenge of the second CAFA experiment underway now. Thirty-four research groups worldwide participated in the first experiment, employing over 50 function annotation algorithms. The prediction methods were assessed using ROC curves, precision/recall curves, and variations on semantic similarity as applied to the Gene Ontology. During this presentation, I will discuss the results of the first CAFA experiment, the challenges we faced in assessing the results, and the future of CAFA. I will also describe the new experiment which will include biological process, molecular function, cellular component and human disease prediction tracks. Finally, I will describe ways in which you, the community, can participate.

OP10: CONCEPTUAL COMPARISON THROUGH INTEGRATIVE FUNCTIONAL GENOMICS IN GENEWEAVER. ORG

Presenting Author: ELISSA CHESLER, The Jackson Laboratory, United States
Co-Author(s): Erich J. Baker, Baylor University, Chales Phillips, University of Tennessee, Michael Langston, Unviersity of Tennessee

ABSTRACT: The comparison of biological concepts is fundamental to the improved definition of disease processes, research models, ontology terms, and other descriptors used to define, characterize and categorize biology within and across species. The GeneWeaver system is designed to close the gap between empirically-derived or conceptually defined data sets and the robust data mining tools required to undertake integrative analysis in high dimensional space. To this end, GeneWeaver contains a rich data store of curated functional genomics experimental results, including differential expression studies, term annotations to sets of biological entities, results of genetic mapping, user submitted gene lists, and privately defined gene sets and gene set associations.
The GeneWeaver tool set is designed to efficiently integrate these data sets and enable flexible, comparison-based gene set analytics. Algorithms for rapid identification of set-set relations based on maximal biclique enumeration have enabled discovery of novel sets of genes related to underlying sets of biological processes in common, and new algorithms provide for systematic comparison of biological concepts through the transitive relationships of multiple known gene set associations. The overall aim of our tools is to rapidly identify conceptual similarity and cohesiveness through the enumeration of the biological basis of that similarity. We will present several examples of the flexibility and utility of the GeneWeaver approach to real time analysis of arbitrary collections of publicly available or user-submitted gene sets representing a range of biological processes. Supported by NIH AA018776

**OP11: INTEGRATIVE VISUALIZATION FOR DISCOVERY OF PHENOTYPE ASSOCIATIONS IN CLINICAL AND MIRNA DATA**

*Presenting Author: MICHAEL HINTERBERG, University of Colorado-Denver, United States  
Co-Author(s): Lawrence Hunter, University of Colorado, David Kao, University of Colorado*

**ABSTRACT:** The increasing size and availability of large clinical datasets provides opportunity for discovery of novel, complex phenotypes in patients. Some of these phenotypes, such as drug responsiveness, are important for differential treatment modalities. Complex phenotypes may also be associated with arrays of diagnostic biomarkers; for example, differential expression of mRNA as well as microRNA can segregate different classes of patients.

In datasets with thousands of clinical features, testing hypotheses for associations between clinical phenotype and genetic expression can be a tedious process. Furthermore, slight modifications in patient stratification may have dramatic effects on biomarker association, but these differences may not be readily apparent. In ongoing work, we present a novel web-based visualization tool that allows the user to view and modify tree-based representations of clinical phenotypes and examine associations with microRNA and mRNA expression, with visible transitions that show the effect of modifying phenotype definition. A specific motivating application to drug-responsiveness in non-ischemic dilated cardiomyopathy is presented as well.
OP12: EVOLUTION OF PALMITOYL ACYL TRANSFERASES (PATs) IN APICOMPLEXA

Presenting Author: SWAPNA SESHADRI, Research Institute, Hospital for Sick Children, Canada
Co-Author(s): John Parkinson, Research Institute, Hospital for Sick Children, Tim Gilberger, McMaster University

ABSTRACT: Protein palmitoylation is the only reversible post-translational mechanism utilising a hydrophobic anchor known to dynamically regulate a protein’s function by influencing its subcellular localization, stability, and interaction. Although this process is ubiquitous in eukaryotes, a recent study uncovered hundreds of palmitoylated proteins in P. falciparum. Therefore, characterizing the suite of enzymes catalyzing this process (Palmitoyl Acyl Transferases (PATs)) in apicomplexan parasites is essential for understanding various aspects of parasite biology. We conducted a comprehensive survey to identify and classify PATs from complete genomes of 16 parasitic apicomplexans and 2 closely related free-living protists (ciliates). Using HMMER, 159 and 138 PATs were identified in apicomplexans and ciliates, respectively. Classification is confounded due to lower resolution stemming from short (~50aa) conserved catalytic domain combined with presence of ankyrin repeats in many sequences. Analysis revealed a ~170aa region with sufficient information to distinguish them into 7 major clades and 14 sub-clades, using Bayesian and maximum likelihood phylogenetic methods. The sub-clades demonstrate distinct patterns of sequence conservation and indels, providing molecular signatures for possible sub-functionalisation. A structural model of the catalytic domain was generated, providing a molecular perspective of these signatures. Overall, 5 sub-clades are apicomplexa-specific, containing members localized to rhoptries and inner membrane complex, organelles unique to apicomplexa that are involved in host cell invasion. Further, 2 clades and 2 sub-clades contain yeast and human orthologs indicating a role in secretory pathway. In summary, apicomplexans have evolved PATs to serve as an integral part of the biological machinery required to facilitate their parasitic life-style.

OP13: FROM SAMPLE TO ANSWER — SEQUENCING IN THE CLOUD ERA WITH THE ILLUMINA BASESPACE® BIOINFORMATICS PLATFORM

Presenting Author: RAYMOND TECOTZKY, Illumina, United States

ABSTRACT: Genomics research promises to revolutionize public and human health. Yet, extracting meaningful information from an enormous collection of sequence data is at risk without the development of new and scalable bioinformatics approaches. Illumina’s genomics cloud platform delivers a suite of industry-leading analysis tools, ensures secure data storage, and simplifies collaboration.
with integrated, push-button sharing and analysis options. Learn how BaseSpace has become a valuable platform for bioinformatics developers, and how rapid data access enables new genomics research.

**OP14: MSProcess — Summarization, Normalization, and Diagnostics for Processing of Mass Spectrometry Based Metabolomic Data**

_Presenting Author: KATERINA KECHRIS, University of Colorado Denver, United States_  
_Co-Author(s): Grant Hughes, University of Colorado Denver_

**ABSTRACT:** The initial processing of Liquid Chromatography coupled with Mass Spectrometry (LC/MS) metabolomic data is covered by a variety of software packages provided by instrument manufacturers and a number of open source packages such as xMSAnalyzer, XCMS and MzMine. While these manage the initial data pre-processing steps of peak detection, chromatogram building, alignment, and quantification, they often lack functions for further processing. We designed the MSProcess package in R to complement existing software by providing additional processing tools and statistical and graphical tools for evaluation of different methods. As there are no universally accepted procedures, the package provides implementation of a variety of novel and previously published methods. The primary functions of the MSProcess package are: summarization of replicates, filtering, imputation of missing data, normalization and/or batch effect adjustment, and dataset diagnostics. The output is in a format ready for input to leading software such as MetaboAnalyst to perform clustering and other downstream analyses. In summary, we developed the MSProcess package to complement other packages by providing additional pre-processing steps, implementing a selection of popular normalization algorithms and generating diagnostics to help guide investigators in their analyses of LC/MS based metabolomic data.

**OP15: Metabolic Reconstruction Identifies Strain-Specific Regulation of Virulence in Toxoplasma gondii**

_Presenting Author: NIRVANA NURSIMULU, University of Toronto, Canada_  
_Co-Author(s): Carl Song, University of Toronto, Stacy Hung, University of Toronto, Melissa Chiasson, NIAID, National Institutes of Health, James Wasmuth, University of Calgary, Michael Grigg, NIAID, National Institutes of Health, John Parkinson, University of Toronto_

**ABSTRACT:** Estimated to infect at least a third of the world’s population, the Apicomplexan parasite, Toxoplasma gondii, represents a major threat to immunocompromised individuals and pregnant women, especially due to the limited efficacy of current therapeutic interventions. Since metabolism plays an essential role in providing energy and the basic building blocks required for growth, drug-development programs are now focussing more on targeting
metabolic enzymes. We hypothesize that metabolic potential plays a key role in determining the virulence of different strains. Given often nonintuitive relationships between enzymes and pathways, constraints based models such as flux balance analysis (FBA), have emerged as indispensable tools to study the organization and operation of metabolic networks. Here we present a novel application of FBA that leverages microarray data to explore the impact of differential enzyme expression observed between virulent and avirulent strains of T. gondii. Our model correctly predicts the increased growth rate of the more virulent type I strain, relative to type II; further analysis predicts the increase in growth rate to result from increased energy production via upregulation of the glycolytic, pentose phosphate and TCA-cycle pathways. These findings highlight a regulatory route which, in addition to conferring growth rate plasticity, may impact the parasite’s outstanding ability to infect a broad range of hosts. Moreover, drug assays confirm strain-specific sensitivities of several reactions, as predicted by in silico single knock-out experiments. This study demonstrates how expression data can be integrated into a model to give robust strain-specific predictions.

**OP16: INTRODUCING COMPUTATIONS IN BIOLOGY/BIOINFORMATICS – AN UNDERGRADUATE PERSPECTIVE**

Presenting Author: MAHEEN KIBRIYA, Chapman University, United States  
Co-Author(s): Shehzein Khan, Chapman University, Louis Ehwerhemuepha, Chapman University

**ABSTRACT:** The importance of computing in biological and life sciences cannot be overemphasized. It is imperative for students in biological sciences to be introduced to computing at the undergraduate level, and our work presents the view of undergraduates toward this shift to an interdisciplinary field. We developed a simple nucleotide sequence analysis program written in Python and discuss our experience in learning and using Python to solve simple biological problems. The aforementioned sequence analysis program was tested using sequence data from the tuberculosis database (www.tbdb.org), while some high level functions freely available in BioPython are briefly discussed.

**OP17: A CORPUS-BASED STUDY OF TEMPORAL RELATIONS IN CLINICAL TEXT**

Presenting Author: NATALYA PANTELEYEVA, University of Colorado, United States  
Co-Author(s): Lawrence Hunter, University of Colorado, Kevin Cohen, University of Colorado

**ABSTRACT:** A corpus of clinical data was used to investigate the hypothesis that there are correlations between pairs of event types and the temporal links between them. A corpus of about 98,000 words that had been annotated with events, event types, TIMEX3 expressions, and
temporal links was examined for such associations. It was found that in fact most pairs of event types show a strong preference for or against a particular type of temporal link. It was also noted that all possible pairs of event types occur even in this relatively small corpus. The preference of specific pairs of event types for particular types of temporal links has implications for natural language processing systems, including establishing baselines for their performance and providing a priori knowledge that can be used to inform the construction of both rule-based and machine-learning-based systems for labeling temporal links in clinical documents. More basic questions about the linguistic expression of temporal relations in clinical text are examined, such as the extent to which they are sequential or not and the extent to which they are intersentential versus intrasentential. Whether surface linguistic cues from morphology, syntax, and lexicon enhance accuracy in establishing temporal link types is addressed.

**OP18: BUILDING AN OPTIMALLY INFORMATIVE MACHINE-LEARNING MODEL OF GENE REGULATORY CONTROL**

*Presenting Author: MOLLY MEGRAW, Oregon State University, United States*

**ABSTRACT:** Gene promoter prediction has long been a difficult challenge, particularly in organisms for which little high-throughput data is available for building and testing accurate computational models. Our lab has recently produced a large-scale transcription start site (TSS) dataset using a sequencing-based method for analysis of 5’ ends of mRNA transcripts in plants. Using this dataset, we first categorize the different shapes taken on by the TSS location distributions into TSS “tag clusters”. For example, some gene upstream regions have very narrow, high clusters and others have a more broad shape. We then design a high-resolution machine-learning model that predicts the presence of TSS tag cluster with an AUROC near 0.98 for each cluster shape. We use this model to analyze the transcription factor binding site content of different promoter shapes. We find that while canonical notions of sharp narrow peak TATA-containing promoters vs more broad “TATA-less” promoters have some merit, the model shows that a large compendium of elements is actually necessary and sufficient for accurate promoter prediction in the case of all tag cluster shapes. In this talk I will demonstrate how a machine learning model can suggest sets of gene interactions which have the potential to “turn on” a particular gene, and briefly discuss one possible approach for dissecting which of those sets are optimal predictors of gene up-regulation.
OP19: CHARACTERIZING UNKNOWN VIRAL GENES THROUGH METABOLICOMICS

Presenting Author: TIFFANY LIANG, San Diego State University, United States
Co-Author(s): Savannah Sanchez, San Diego State University (SDSU), Jason Rostron, SDSU, Jeremy Frank, SDSU, Daniel Cuevas, SDSU, Anca Segall, SDSU, Forest Rohwer, SDSU, Robert Edwards, SDSU, Daniel Garza, Evandro Chagas Institute

ABSTRACT: Viruses are the most diverse biological entities on earth. However, they also have the least characterized genetic, taxonomic, and functional diversity. In metagenomic analyses of viral communities from various environments, most sequences are unrelated to any known sequences; for example, about 90% of the viral sequences found in marine environments are unknown. The goal of this study is to characterize the function of unknown viral genes and identify those that alter host metabolism.

Viral metagenomes were collected from filtered seawater from Pacific coral reefs, sequenced by Roche 454 technology, and open reading frames were predicted from those sequences. Genes were synthesized and cloned into E. coli. These clones have been characterized in several different ways. To investigate these clones that affected metabolic processes, the metabolites were identified by gas chromatography-coupled time-of-flight (GC/TOF) mass spectrometry. In total 423 metabolites were found, however only 15% of those matched known compounds. We are identifying the specific metabolites produced or affected by the over expression of phage proteins to predict physiological roles for these proteins that can then be tested experimentally. We have also analyzed metabolic changes associated with expression of proteins with known functions that are involved in central metabolism; and clustering of these changes allows us to predict functions for other proteins. We are building a systematic analysis pipeline that can process metabonomics data for downstream analysis of metabolomics and related data sets.

OP20: GENESEEER AIDS DRUG DISCOVERY BY EXPLORING EVOLUTIONARY RELATIONSHIPS BETWEEN GENES ACROSS GENOMES

Presenting Author: DOUGLAS FENGER, Dart NeuroScience, United States
Co-Author(s): Matthew Shaw, Dart NeuroScience, Philip Cheung, Dart NeuroScience, Tim Tully, Dart NeuroScience

ABSTRACT: Homologous relationships facilitate drug discovery by mapping gene/protein function between and within species, allowing functional predictions of novel or unknown genes. Additional benefits of cross-species mapping include the following: use of paralogs for selectivity/specificity screens to eliminate drug side effects, translation of pathway information from model organisms to humans, and allowing comparison and combination of data from different species.
GeneSeer (http://geneseer.com) is a publicly available tool that leverages public sequence data, gene metadata information, and other publicly available data to calculate and display orthologous and paralogous gene relationships for all genes from several species, including yeast, insects, worms, vertebrates, mammals, and primates including humans. GeneSeer calculates homology relationships and its interface is designed to help scientists quickly predict important attributes such as additional closely related family members and paralogous relationships. It is a useful tool for cross-species translational mapping and enables scientists to easily translate hypotheses about gene identity and function from one species to another. We have validated GeneSeer versus Homologene, the homolog prediction tool from NCBI. The results show that GeneSeer is as good as, if not better than, Homologene. Finally, a comparison of features shows GeneSeer to be the most feature rich when compared to alternative homology tools.

**OP21: THE HUMAN GENE CONNECTOME SERVER: AN ONLINE TOOL FOR PRIORITIZING GENES BY BIOLOGICAL DISTANCE**

Presenting Author: YUVAL ITAN, The Rockefeller University, United States  
Co-Author(s): Mark Mazel, The Rockefeller University, Benjamin Mazel, The Rockefeller University, Avinash Abhyankar, New York Genome Center, Bertrand Boisson, The Rockefeller University, Patrick Nitschke, INSERM, Lluis Quintana-Murci, Pasteur Institute, Laurent Abel, INSERM, Shen-Ying Zhang, The Rockefeller University, Jean-Laurent Casanova, The Rockefeller University

**ABSTRACT:** To determine the disease-causing allele(s) underlying human disease, high-throughput genomic methods are applied and provide thousands of gene variants per patient. We recently developed a novel approach, the “human gene connectome” (HGC) – a concept and method that describe the set of all in silico-predicted biologically plausible routes and distances between all pairs of human genes. With the HGC, we generated a “gene-specific connectome” for each human gene – the set of all human genes ranked by their predicted biological proximity to the core gene of interest, available at: http://lab.rockefeller.edu/casanova/HGC/. We demonstrated that the HGC is the most powerful approach for prioritizing high-throughput genetic variants in Mendelian disease studies. However, there is currently no effective gene-centric online interface for ranking genes by biological distance. We describe here the human gene connectome server (HGCS): a powerful, easy-to-use interactive online tool that enables researchers to prioritize any list of genes according to their biological proximity to core genes (i.e. genes that are known to be associated with the phenotype), and to predict novel gene pathways. We demonstrated the effectiveness of the HGCS for detecting herpes simplex encephalitis-predisposing genes in patients’ whole exome sequencing data. The HGCS is freely available to use for non-commercial users at: http://hgc.rockefeller.edu/.
OP22: TYPE I ERROR RATE ANALYSIS OF METHODS FOR CORRELATED BINARY OUTCOMES

Presenting Author: AARTI MUNJAL, University of Colorado Denver, United States
Co-Author(s): Jacqueline Johnson, University of North Carolina at Chapel Hill, Sarah Kreidler, University of Colorado Denver, Deborah Glueck, University of Colorado Denver, Keith Muller, University of Florida

ABSTRACT: We describe a simulation study to estimate type I error rate, calculate power, and investigate optimization algorithm convergence properties for multilevel and longitudinal designs with correlated binary outcomes.

Multilevel and longitudinal studies with binary outcomes are common throughout the biomedical literature. These studies allow scientists to investigate the causes of disease, determine the efficacy of drugs, and conduct basic biomedical research to improve human health. Our simulations studies show that using standard multivariate models with binary outcomes data controls type I error rate for many important designs. The approach produces reasonable results even with upto 10% missing data and very small sample sizes. We discuss the implications for scientists working on oral cancer detection.

OP23: AN OPTIMAL METABOLIC ROUTE SEARCH TOOL: ROUTESEARCH

Presenting Author: MARIO LATENDRESSE, SRI International, United States

ABSTRACT: RouteSearch is a new Web accessible metabolic engineering tool available as part of BioCyc since March 2013. It enables searching for optimal metabolic linear routes between a start compound and a goal compound. The optimality criteria are the weighted sum of the costs of the reactions used, and the weighted sum of the costs of atoms that are lost in the transformation from the start compound to the goal compound. These costs and the number of minimum cost routes to find and display are user selectable. The routes are displayed as a series of connected enzymatic reactions including chemical structures of the substrates, where the conserved moieties within each metabolite are shown using colors. By using a graphical interface, the user can also easily identify each atom conserved or lost along each route. RouteSearch uses two algorithms to search for optimal routes: the Bellman-Ford algorithm that finds the least cost route, and a more general branch and bound search algorithm that can find several minimum cost routes. RouteSearch also uses a preferred organism to search -- a chassis in metabolic engineering terms, such as E. coli -- and a library of additional reactions, which is the MetaCyc database. The cost of using a reaction from MetaCyc is usually set higher than using a reaction from the chassis. In this way, new and more productive metabolic routes can be found for the chassis by adding reactions from MetaCyc. We will also briefly describe the computation of atom
mappings for MetaCyc. Atom mappings are used by RouteSearch to track the atoms conserved and lost in a route.

**OP24: COMPUTATIONAL AND MATHEMATICAL MODELING OF THE SEGMENTATION GENES OF HONEYBEE (APIS MELLIFERA)**

Presenting Author: MARYAM BAGHER OSKOEI, University of Otago, United States
Co-Author(s): Brendan McCane, University of Otago, Peter Dearden, University of Otago

**ABSTRACT:** Drosophila and Honeybee embryos are two examples that develop a segmented body plan during their early development. The basic body plan consists of distinct segments along their anterior-posterior axis established via a segmentation process. The process subdivides the embryos into segments, which is controlled by interactions between segmentation genes. Many experimental and computational works have been tested to reveal which interactions cause this process in Drosophila embryos, but few have been done for Honeybee embryos. The Honeybee genome has some aspects that make it worth studying. Honeybees are excellent comparative model systems that help to understand evolutionary pathways behind the segmentation process considering that the insects diverged ~350 million years ago. Here, we present a method using ordinary differential equations (ODEs) to model segmentation genes in Honeybee embryos. The initial and target models for ODEs were configured with data collected in Peter Dearden’s lab. The computational modeling was carried out in order to explore how likely each gene is regulated by other genes positively or negatively. The simulations were performed in two phases, first as a Pre-stripes Networks and then the striped pattern forming Networks. The main findings predict gene networks that are more likely to pattern different parts of embryos along their anterior-posterior axis during early developmental stages. These results are comparable with Drosophila embryos. Importantly, the predicted networks provide hypotheses that can be tested experimentally.

**OP25: IMPROVED RNAI INTERFERENCE TARGET SEQUENCING (RIT-SEQ) ENABLES DISSECTION OF CELLULAR FUNCTION IN TRYPANOSOME BRUCEI.**

Presenting Author: JONATHAN WILKES, University of Glasgow, United Kingdom
Co-Author(s): Graham Hamilton, University of Glasgow, Richard McCulloch, University of Glasgow

**ABSTRACT:** The protozoan parasite Trypanosoma brucei utilises a RNA interference (RNAi) pathway, widely conserved with other eukaryotes. This can be adapted to regulate expression of the poly-cistronically transcribed genes of T. brucei, utilising gene-specific sequences within a tetracyclin inducible cassette, allowing RNAi ‘knockdown’; now an important research tool. RIT-seq has been developed, which enabled
The parallel analysis of >8000 genes in T.brucei in life-cycle and differentation stages (1). The original RITseq methodology possesses a number of shortcomings which compromise its potential: semi-specific PCR produces small enrichments of the inserted sequences, produces inconsistent amplified sequences, and contains significant genomic sequence unrelated to the inducible fragments.

We have designed an adaptation of the methodology involving a specific PCR to amplify sequences between common primer sites flanking inserted genomic fragments in the RNAi cassette. Preparing the sequencing library from his amplified material requires 10fold less material (500ng of DNA), produces a higher proportion (3-10fold) of reads unequivocally derived from the cassettes, utilises standard protocols for library preparation and permits sample multiplexing. To validate this RITseq approach, we have screened for T.brucei genes that act in DNA damage repair by measuring read abundance after RNAi in the presence or absence of the SN2 alkylator methyl methanesulphonate. A number of previously characterised T. brucei DNA repair genes are revealed, and several novel pathways that have not been examined to date. The system was adapted to produce a comprehensive panel of protein kinase (kinome) probes.


OP26: WHAT YOU DID NOT KNOW YOUR TRANSCRIPTION FACTOR WAS DOING

Presenting Author: MARY ALLEN, University of Colorado, United States
Co-Authort(s): Justin Freeman, University of Colorado, Hestia Mellert, University of Colorado, Joaquin Espinosa, University of Colorado, Robin Dowell, University of Colorado

ABSTRACT: A transcription factor (TF) protein binds to DNA and regulates transcription of a target gene. The guardian of the genome, p53, is a transcription factor important in cancer and aging, and activates transcription of many genes involved in apoptosis and cell cycle arrest. I have used a novel technique to discover over 200 annotated genes that are direct transcriptional targets of p53. This technique, GRo-seq, captures nascent transcription. Moreover, my work shows that short bidirectional transcripts are produced from p53 binding sites when the sites are within, nearby, or distant from protein coding genes. Additionally, I demonstrate that when p53 is activated, transcription at its binding sites increases. Finally, I show the p53 binding sites have high levels of transcription when they are located near p53 targets genes (protein coding genes). The novel discovery that binding sites are transcribed and that transcription levels of binding sites correlate with TF activity leads to new questions about how transcription of binding sites affects TF binding and activation of target genes.
OP27: LINKING GENOTYPE AND ENTEROTYPE IN INFLAMMATORY BOWEL DISEASE

Presenting Author: DAN KNIGHTS, University of Minnesota, United States
Co-Author(s): Mark Silverberg, University of Toronto, Rinse Weersma, University Medical Center Groningen, Dirk Gevers, Broad Institute of Harvard and MIT, Gerard Dijkstra, University Medical Center Groningen, Hailiang Huang, Massachusetts General Hospital, Andrea Tyler, Mount Sinai Hospital, Suzanne von Sommern, University Medical Center Groningen, Floris Imhann, University Medical Center Groningen, Joanne Stempak, Mount Sinai Hospital, Caitlin Russell, Massachusetts General Hospital, Jenny Sauk, Massachusetts General Hospital, Jo Knight, University of Toronto, Mark Daly, Massachusetts General Hospital, Curtis Huttenhower, Harvard School of Public Health, Ramnik Xavier, Massachusetts General Hospital

ABSTRACT: Human genetics and host-associated microbiomes have each been associated with inflammatory bowel disease (IBD), however IBD risk cannot be fully explained by either factor alone. Recent findings implicate genotype-enterotype crosstalk as a contributor to IBD pathogenesis. However, there has been no large study of complex genome-microbiome interactions in humans. We have performed such a study using bacterial 16S ribosomal RNA enterotyping and Immunochip genotyping from intestinal mucosal biopsies in three independent cohorts totalling more than 500 individuals. We present methodology, validated internally between cohorts, to test for host genetic locus interaction with taxonomic and functional components of the microbiome. In a targeted analysis integrating fine mapping of causal variants, we find nucleotide oligomerization domain 2 (NOD2)-specific risk associated with known IBD-related imbalances in bacterial taxa, including increased Gammaproteobacteria and Escherichia. NOD2 has known roles in management of commensal bacteria, and a strong genetic signal for increased IBD risk. Using imputed bacterial metagenomes we also find NOD2 risk linked to increased sulfur reduction and lipopolysaccharide biosynthesis. These findings point to pathobiont expansion and bacterial production of genotoxic agent hydrogen sulfide, both involved in inflammation and IBD pathogenesis. In a novel omnibus tests we demonstrate links between host innate and adaptive immune pathways and broad enterotype composition. Our analysis demonstrates the ability to uncover novel interactions from paired genotype-enterotype data and that host genetics is linked to microbial dysbiosis in IBD.

OP28: PROBING CHROMATIN WITH DIGESTION ASSAYS: FROM STATIC NUCLEOSOME POSITIONING TO DYNAMIC RESPONSE

Presenting Author: MICHAEL TOLSTORUKOV, Massachusetts General Hospital, United States

ABSTRACT: The read-out of the genetic information occurs in cell in the context of chromatin and in the recent years chromatin structure has been in the focus of intensive research. The basic repeating unit
of chromatin, termed nucleosome, comprises eight histone proteins and about 150 bp of genomic DNA. Nucleosomes are subject to repositioning, histone modification and variant exchange, which constitute potent tools of epigenetic regulation of gene expression. Furthermore, multiple diseases, including developmental disorders and cancers, are associated with dysregulation of these pathways.

Digestion assays often in combination with chromatin immunoprecipitation and followed by next generation sequencing are widely used for profiling primary chromatin structure. However, processing of the data produced in such experiments constitutes a substantial challenge due to both complexity of the underlying causes influencing the chromatin response to a digestion agent and internal biases in the technique. In my presentation I will describe a comprehensive set of the computational procedures developed to address these challenges and to extract multiple layers of information on chromatin structure from a single experimental series. In addition to mapping stable positions of bulk and epigenetically modified nucleosomes, our approach allows identification of the nucleosomes of non-canonical sizes and the sites where DNA accessibility is regulated by the changes in nucleosome physical properties. While still in development, this approach has already produced novel results on distribution of the variants of histone H2A and, importantly, provided mechanistic insights into the role of these variants in gene expression regulation.

**OP29: GIST – AN ENSEMBLE APPROACH TO THE TAXONOMIC CLASSIFICATION OF METATRANSCRIPTOMIC READS**

*Presenting Author: JOHN PARKINSON, Hospital for Sick Children, Canada*

*Co-Author(s): Geoffrey Halliday, University of Toronto*

**ABSTRACT:** Whole-microbiome gene expression profiling (‘metatranscriptomics’ or ‘RNA-seq’) has emerged as a powerful means of gaining a mechanistic understanding of the complex inter-relationships that exist in microbial communities. However, due to the inherent complexity of microbial communities and the lack of a comprehensive set of reference genomes, currently available computational tools for metatranscriptomic analysis are limited in their ability to functionally classify and organize these sequence datasets. To meet this challenge we have been developing methods that combine accurate transcript annotation with systems-level functional interrogation of metatranscriptomic datasets. As part of these methods, we present GIST (Generative Inference of Sequence Taxonomy), which combines several statistical and machine learning methods for compositional analysis of both nucleotide and amino acid content with the output from the Burroughs-Wheeler Aligner to produce robust taxonomic assignments of metatranscriptomic RNA reads. In
addition to identifying taxon-specific pathways within the context of a pan-microbial functional network, linking taxa with specific functions in a microbiome will produce deeper understanding of how their loss or gain alters microbiome functionality. Applied to real as well as synthetic datasets, generated using an inhouse simulation tool termed GENEPUDDLE, we demonstrate an improved performance in taxonomic assignments over existing methods.

OP30: COMPARATIVE METAGENOMICS BY CROSS-ASSEMBLY

Presenting Author: BAS DUTILH, Radboud University Medical Centre, The Netherlands
Co-Author(s): Rob Edwards, San Diego State University

ABSTRACT: Determining the interrelationships between metagenomes from different biomes or different time points is important to understand the microbial world around us. Mapping metagenomic sequences to a reference database of known genes is a feasible approach to transfer taxonomical and functional annotations to sequence reads. However, it can limit the amount of data that can be analyzed because the majority of the sequencing reads in difficult-to-annotate datasets, such as viral metagenomes from biomes other than the human microbiome, lack known homologs. A promising alternative is reference-independent comparative metagenomics by cross-assembly.

Cross-assembly of different metagenomes is a fast and insightful way to obtain information about sequences that are shared between the samples, represented by cross-contigs. Importantly, cross-assembly is independent of an annotated reference database, providing a way to also handle unknown sequences. The cross-assembly tool crAss allows a rapid analysis of these cross-contigs. First, it provides cross-contig-based similarity scores between all metagenome pairs. Second, crAss creates insightful images displaying the inter-relationships between samples. Third, it generates occurrence profiles of the cross-contig sequences across metagenomes that can be used to discover related sequences, aiding further assembly and interpretation.

OP31: AN UPDATE ON KABOB: TOWARDS AN INTEGRATED KNOWLEDGE BASE OF BIOMEDICINE

Presenting Author: KEVIN LIVINGSTON, University of Colorado, United States
Co-Author(s): Michael Bada, University of Colorado, William Baumgartner, University of Colorado, Lawrence Hunter, University of Colorado

ABSTRACT: A vast wealth of information currently exists in the form of curated biomedical databases with information about genes, proteins, SNPs, drugs, diseases, pathways, interactions, etc. Unfortunately there are many hurdles in the way of researchers being able to view their data in a context that spans multiple of these sources. These problems
include idiosyncratic file formats, multiple unique identifiers for the same entities, and varying semantics of the curated data.

The Knowledge Base of Biomedicine (KaBOB) aims to overcome these problems through a series of technical solutions. We have produced a RDF model for representing the contents of databases as extension of the Information Artifact Ontology (IAO), thus providing a uniform foundation for the incoming data. Curated mappings are used to link identifiers across the data sources. The transitive closure of these linkages is computed creating a set of all the identifiers in the source databases for each biomedical entity, thus resolving terminology problems. Biomedical entities are created to represent the real concept (genes, proteins, drugs, etc.) being referred to by each set of database identifiers. These biomedical entities form the foundation for representing knowledge extracted from the source databases. Declarative rules are used to translate the source data into OWL representations that are integrated under the Open Biomedical Ontologies (OBOs). These rules also function to preserve the provenance of the newly created representations. These representations and rules are currently under active development and evaluation.

**OP32: HYPERNETWORKS: THE FUTURE OF NETWORK MODELING**

*Presenting Author: DEBRA GOLDBERG, University of Colorado Boulder, United States*

**ABSTRACT:** Hypernetworks (hypergraphs) are an extension to network models that allows any number of nodes in an edge (called a hyperedge). Many classes of problems currently modeled with networks involve data that are not intrinsically binary. These would be more naturally captured by a hypernetwork model. There are not (yet) many analysis tools available for biological hypernetworks. We are developing such tools in the context of protein interactions. However, many measures and algorithms must be tailored specifically to the data being analyzed and the questions being asked. Here I describe some of the domains currently modeled by networks that could benefit from analysis with hypernetworks.

**OP33: THE CHANGING VIEW OF PROTEIN INTERACTION NETWORKS BASED ON DATA AVAILABILITY**

*Presenting Author: BARRETT HOSTETTER-LEWIS, California State University, Chico, United States*

*Co-Author(s): Todd A Gibson, California State University, Chico*

**ABSTRACT:** The end of the 20th century marked the beginning of the era of large-scale studies identifying protein interactions. These large data sets catalyzed a renaissance in protein interaction network
research. The mainstay of this research has been in elucidating the biological and evolutionary factors that affect the network’s topological features. As the quantity of data increases and the quality of data improves we have been regularly refining our understanding of these biological and evolutionary influences. Now, as interaction data becomes available for recently sequenced organisms, there is a great opportunity for research on today’s nascent interactomes to benefit from the analytical steps and missteps taken on fledgling interactome analysis 10-15 years ago.

Here we describe how the researcher’s view of the Saccharomyces cerevisiae protein interaction network has changed since the first publication of large-scale yeast data in late 1999. By creating network snapshots using increasing amounts of interaction data constrained by date-of-publication and various quality criteria, we identify trends in the researcher’s view of network topology, and compare this to the interactomes of organisms which still have early, incomplete interaction data sets.

**OP34: EXPERIMENTAL DETERMINATION OF USEFUL AND INFORMATIVE VISUALIZATIONS OF MICROBIAL ECOLOGY DATA FOR PUBLIC AND SCIENTIFIC AUDIENCES**

**Presenting Author:** MEGAN PIRRUNG, University of Colorado, United States

**ABSTRACT:** Sequencing technologies are getting cheaper and producing vast amounts of data, especially in the field of microbial ecology. Proper visualization of biological data is key to informative analysis and insight. Data analysis by users through informative, useful, and responsive visualizations is key to harnessing the true potential of big data. We propose an experiment that will help to determine which types of visualization techniques are most informative for microbial ecology data in both public and expert scientific audiences. To perform this experiment on a large number of subjects in a systematic way, we have created a modular system with easily substituted visualization methods. We have created dynamic visualizations that parallel the visualizations available in QIIME using the d3 (Data Driven Documents) javascript library, and a visualization-testing framework that will be used to display the results of the American Gut Project. A visualization technique for one particular analysis as performed in QIIME will be randomly selected from the set of visualizations appropriate for the data selected and shown to the user. The user will also be presented with a questionnaire that will let us determine which visualizations allow users to answer the most questions correctly. We expect that the scores will indicate that certain visualization techniques are more appropriate for certain types of data, and that certain visualizations may be found informative for one audience over another, in public and scientific audiences.
OP35: MODELING TRANSCRIPTIONAL REGULATION THROUGH SIMULATION OF THE DYNAMIC CHANGES IN DNA BINDING FACTOR CONFIGURATION

Presenting Author: DAVID KNOX, University of Colorado Anschutz Medical Campus, United States
Co-Author(s): Robin Dowell, University of Colorado Boulder

ABSTRACT: Transcriptional regulation is the complex system behavior arising from the interaction of numerous regulators with DNA. Experimental efforts have unraveled the function of many individual components of the process, but the systems level behavior remains unpredictable. Growing evidence indicates that the transcriptional response of the system emerges not solely from the individual components, but rather by their collective behavior – including competition and cooperation. The environment surrounding DNA undergoes millions of molecular interactions every second, resulting in continuous changes to the configuration of physically bound molecular components. It is from these stochastic, temporal, and spatial interactions of regulatory components that transcriptional regulation arises within each cell. Encapsulating our understanding of these interactions into computational models is essential for a full understanding of transcriptional regulation.

Our goal was to create biologically realistic computational models of Transcriptional Regulation that not only capture the behavior of several individual components, but also describe the dynamic and stochastic behavior of competing components. To this end we have developed an automated rule builder to not only create stochastic simulation rule sets, but also basic visualizations of the resultant simulations. Our modeling framework captures the competition between regulatory proteins, and more importantly, the dynamics of regulatory events occurring within individual cells.

OP36: DELETION OF COX-2 IS ASSOCIATED WITH REDUCED EXPRESSION OF RGL1: RELEVANCE TO CHEMOPREVENTATIVE EFFECT OF COX-2 INHIBITORS

Presenting Author: NICHOLAS KIRKBY, Imperial College London, United Kingdom
Co-Author(s): Jane Mitchell, Imperial College London

ABSTRACT: Non-steroidal anti-inflammatory drugs (NSAIDs) such as aspirin, ibuprofen, rofecoxib and meloxicam inhibit cyclo-oxygenase (COX)-2 and are used for the treatment of arthritis. NSAIDs and COX-2 gene deletion have also been shown to prevent cancer in animal models, but research into the chemopreventative benefits of NSAIDs in man has been restricted following association of these drugs with cardiovascular toxicity. As such, the mechanisms by which COX-2 regulates cancer progression and cardiovascular health are incompletely understood. Here we have performed transcriptomic analysis of COX-2
knock out mouse aorta in order to identify gene pathways that may help to explain the role of COX-2 in these systems.

Microarray analysis of COX-2/-/- vs COX-2+/+ aorta illustrated 29 differentially expressed genes. The most greatly altered of these was Ral guanine nucleotide dissociation stimulator-like 1 (Rgl1) which was down-regulated in COX-2/-/- mice (-1.73-fold, p=1.03x10^-8). To validate these observations, we mined data from ArrayExpress, which revealed studies showing that chronic dosing of rats with several NSAIDs including meloxicam has been associated with a reduction in Rgl1 expression. This is in line with our findings and supports our observations in gene-deleted mice.

These data suggest Rgl1 gene expression may be regulated by COX-2 activity. Rgl1 is a guanine nucleotide exchange factor, activated by Ras, the most common oncogene in human cancer, and catalyses the activation of, Ral. Ras and Ral are known to be crucial for cancer transformation and progression. Although preliminary, our observations may implicate Rgl1 down-regulation as a novel mechanism by which NSAIDs exert chemoprentative effects.

**OP37: PRIORITIZING HYPOTHESES FOR EPIGENETIC MECHANISMS IN HUNTINGTON’S DISEASE USING AN E-SCIENCE APPROACH**

*Presenting Author: ELENI MINA, Leiden University Medical Center, The Netherlands*
*Co-Author(s): Willeke van Roon-Mom, Leiden University Medical Center (LUMC), Peter A.C. ‘t Hoen, LUMC, Mark Thompson, LUMC, Reinout van Schouwen, LUMC, Rajaram Kaliyaperumal, LUMC, Kristina M. Hettne, LUMC, Erik Schultes, LUMC, Barend Mons, LUMC, Marco Roos, LUMC*

**ABSTRACT:** The amount of data from high throughput experiments requires novel approaches for effective data analysis, dissemination of methods and collaborative research. We followed an e-Science approach to elucidate molecular mechanisms involved in chromatin-mediated regulation of gene expression in Huntington’s Disease (HD), particularly for prioritizing hypotheses that result from the analysis and integration of publicly available datasets.

Our e-Science approach involves implementation of computational experiments by scientific workflows, and communication with different experts at all stages. By mining literature information, we prioritized mechanisms and proteins that are likely to be involved in HD, such as SLIT2 (involved in axonal transport during neural development), or the neuronal adaptor APBA1. We discovered enrichment of differentially expressed genes in the poised chromatin state of the caudate nucleus, suggesting a functional link with neuronal development. Moreover, we observed overrepresentation of CpG islands among promoters of differentially expressed genes, implicating DNA methylation in gene deregulation in HD.
A computational approach can produce more predictions than we can test ourselves. To preserve predictions and facilitate data integration, we expose them in a machine readable format (nanpublications), together with provenance information on authorship and the origin of the data. We demonstrate interoperability by linking the nanopublication provenance to the Research Object (RO) model to preserve our methods and resources.

In conclusion, by applying our methodology we can extract meaningful biological associations for generating novel hypotheses that can be tested and validated by wet lab experimentation. Exposure as nanopublications and ROs enhances reproducibility and reuse of our methodology.

**OP38: VARIANCE COMPONENT SCORE TEST IN A MIXED-EFFECTS MODEL FRAMEWORK TO MAP TISSUE-SPECIFIC EQTL**

_Presenting Author: CHAITANYA ACHARYA, Duke University, United States_
_Co-Author(s): Andrew Allen, Duke University_

**ABSTRACT:** Expression quantitative trait loci (eQTL) analysis associates putative regulatory variants (SNPs) with gene expression levels, which are treated as quantitative traits. Until recently, eQTL analysis is performed in a tissue-by-tissue basis followed by an examination of overlap of eQTLs across all tissues. However, most of those methods fall short in their ability to jointly analyze data across multiple tissues. Such type of joint analyses of tissue-types have been shown to improve power to identify eQTLs that have similar effects across tissues. We propose a variance component score test approach in a mixed-effects framework in order to jointly analyze multiple tissue types and assess the power of such tests. Using Monte Carlo simulations, we show that the new score test performs much better than the traditional likelihood ratio method in terms of statistical power. Using real data sets, we show that the new score test not only preserves power but also is computationally very efficient. We think that this method will particularly be very useful in prioritizing variants when analyzing heterogeneous disease model systems especially for any downstream genomic analysis including but not restricted to next-generation sequencing analysis.

**OP39: HIGH THROUGHPUT PHENOTYPE PROFILING FOR BACTERIAL FLUX-BALANCE MODEL OPTIMIZATION**

_Presenting Author: DANIEL CUEVAS, San Diego State University, United States_
_Co-Author(s): Rob Edwards, San Diego State University, Daniel Garza, Evandro Chagas Institute, Savannah Sanchez, San Diego State University_

**ABSTRACT:** Advances in large-scale genomic sequencing allow researchers to create accurate computational models of organisms
through the use of gene annotation software, such as RAST (Rapid Annotation using Subsystem Technology). These bioinformatics software deduce gene function through homology-based distinctions that are dependent on previously verified information; thus new discoveries cannot be easily extrapolated from current analysis tools without experimental examination. Recent developments using phenotype microarrays (PMs) provide a high throughput, large-scale technique in profiling bacterial characteristics and their phenotypes. PMs have the potential to experimentally test various growth conditions and then provide bacterial yield in real-time. By coupling PM experiments with the advances of genomic sequencing and annotation, more robust and accurate computational models can be developed and confirmed.

Here we present a combined biological and computational approach that (1) uses optical density data from a PM system as input to evaluate various growth curves, and (2) optimize the flux-balance analysis (FBA) models by using the PM results as a base for in silico growth simulations. The bacterium Citrobacter sedlakii was sequenced and studied in the PM-FBA pipeline to assess the capabilities of our approach. RAST annotations produced a base computational model consisting of 1,367 enzymatic reactions. After PM-FBA optimization a total of 44 reactions were added to, or modified within, the model. The model correctly predicted the outcome on 89% of growth experiments.
**POSTER LIST**

P01: VARIANCE COMPONENT SCORE TEST IN A MIXED-EFFECTS MODEL FRAMEWORK TO MAP TISSUE-SPECIFIC EQTL  
*Presenting Author: CHAITANYA ACHARYA, Duke University, United States*

P02: WHAT YOU DID NOT KNOW YOUR TRANSCRIPTION FACTOR WAS DOING  
*Presenting Author: MARY ALLEN, University of Colorado, United States*

P03: COMPUTATIONAL AND MATHEMATICAL MODELING OF THE SEGMENTATION GENES OF HONEYBEE (APIS MELLIFERA)  
*Presenting Author: MARYAM BAGHER OSKOUEI, University of Otago, United States*

P04: TOWARDS REALIZING CLOUD BASED BIOLOGICAL ECHO SYSTEM FOR BIOLOGICAL WORKFLOWS  
*Presenting Author: JANAKA BALASORIIYA, Arizona State University, United States*

P05: DETECTING THE CRISPR/CAS SYSTEM IN FRESHWATER ENVIRONMENTS  
*Presenting Author: DAVID BALTRUSAITIS, Loyola University Chicago, United States*

P06: PREDICTING GENOMIC 3D CONTACTS  
*Presenting Author: SVEN BILKE, National Cancer Institute, United States*

P07: MP2GO: INFERRING GENE FUNCTION FROM PHENOTYPE  
*Presenting Author: JUDITH BLAKE, The Jackson Laboratory, United States*

P08: HIGH THROUGHPUT PHENOTYPE PROFILING FOR BACTERIAL FLUX-BALANCE MODEL OPTIMIZATION  
*Presenting Author: DANIEL CUEVAS, San Diego State University, United States*

P09: GENOME SEQUENCE OF AN OUTBREAK STRAIN OF MYCOBACTERIUM ABSCESSUS FROM RIO DE JANEIRO, BRAZIL AND PHYLOGENOMIC RELATIONSHIPS TO GLOBALLY DIVERSE STRAINS  
*Presenting Author: REBECCA DAVIDSON, National Jewish Health, United States*

P10: ANNOTATION OF HYPOTHETICAL PROTEINS IN HUMAN ADENOVIRUSES  
*Presenting Author: SHANE DORDEN, University of Tampa, United States*

P11: SYSTEMATIC CLASSIFICATION OF COMMON DISEASE-ASSOCIATED SNPS BY THEIR EPIGENOMIC RELATIONSHIP  
*Presenting Author: MIKHAIL DOZMOROV, Oklahoma Medical Research Foundation, United States*

P12: COMPARATIVE METAGENOMICS BY CROSS-ASSEMBLY  
*Presenting Author: BAS DUTILH, Radboud University Medical Centre, The Netherlands*

P13: GENEOER AID DISCOVER BY EXPLORING EVOLUTIONARY RELATIONSHIPS BETWEEN GENES ACROSS GENOMES  
*Presenting Author: DOUGLAS FENGER, Dart NeuroScience, United States*

P14: CLUSTERING COEFFICIENTS IN PROTEIN INTERACTION HYPERSONETWORKS  
*Presenting Author: SUZANNE GALLAGHER, University of Colorado Boulder, United States*

P15: IDENTIFICATION OF TISSUE-SPECIFIC GENE EXPRESSION SIGNATURES OF AGING VIA MICROARRAY META-ANALYSIS  
*Presenting Author: CORY GILES, Oklahoma Medical Research Foundation, United States*
P16: MARK2CURE: A CROWDSOURCING PLATFORM FOR BIOMEDICAL LITERATURE ANNOTATION
Presenting Author: BENJAMIN GOOD, The Scripps Research Institute, United States

P17: WITHDRAWN

P18: TEMPORAL EXPRESSION RECOGNITION FOR CELL CYCLE PHASE CONCEPTS IN BIOMEDICAL LITERATURE
Presenting Author: NEGACY HAILU, University of Colorado, United States

P19: ONTOLOGY TRANSLATION: A CASE STUDY ON TRANSLATING THE GENE ONTOLOGY FROM ENGLISH TO GERMAN
Presenting Author: NEGACY HAILU, University of Colorado, United States

P20: INTEGRATIVE VISUALIZATION FOR DISCOVERY OF PHENOTYPE ASSOCIATIONS IN CLINICAL AND MIRNA DATA
Presenting Author: MICHAEL HINTERBERG, University of Colorado-Denver, United States

P21: THE CHANGING VIEW OF PROTEIN INTERACTION NETWORKS BASED ON DATA AVAILABILITY
Presenting Author: BARRETT HOSTETTER-LEWIS, California State University, Chico, United States

P22: INTRODUCING COMPUTATIONS IN BIOLOGY/BIOINFORMATICS — AN UNDERGRADUATE PERSPECTIVE
Presenting Author: MAHEEN KIBRIYA, Chapman University, United States

P23: DELETION OF COX-2 IS ASSOCIATED WITH REDUCED EXPRESSION OF RGL1: RELEVANCE TO CHEMOPREVENTATIVE EFFECT OF COX-2 INHIBITORS
Presenting Author: NICHOLAS KIRKBY, Imperial College London, United Kingdom

P24: LINKING GENOTYPE AND ENTEROTYPE IN INFLAMMATORY BOWEL DISEASE
Presenting Author: DAN KNIGHTS, University of Minnesota, United States

P25: MODELING TRANSCRIPTIONAL REGULATION THROUGH SIMULATION OF THE DYNAMIC CHANGES IN DNA BINDING FACTOR CONFIGURATION
Presenting Author: DAVID KNOX, University of Colorado Anschutz Medical Campus, United States

P26: HIPPO: A TOOL FOR CONSTRUCTING QUALITY-CONSCIOUS META-ASSEMBLIES
Presenting Author: IRENA LANC, University of Notre Dame, United States

P27: FLOW-BASED NETWORK ALIGNMENT
Presenting Author: RYAN LANGENDORF, University of Colorado, Boulder, United States

P28: AN OPTIMAL METABOLIC ROUTE SEARCH TOOL: ROUTESEARCH
Presenting Author: MARIO LATENDRESSE, SRI International, United States

P29: CHARACTERIZING UNKNOWN VIRAL GENES THROUGH METABOLOMICS
Presenting Author: TIFFANY LIANG, San Diego State University, United States

P30: LINKAGE ANALYSIS: GENOMIC REGIONS CONTRIBUTING TO THE EXPRESSION OF TYPE 1 DIABETES MICROVASCULAR COMPLICATIONS
Presenting Author: ETTIE LIPNER, University of Colorado-Denver/National Jewish Health, United States

P31: A BIOINFORMATICS APPROACH TO IDENTIFY AND CLASSIFY EXTENSINS FROM SELECTED PLANT GENOMES
Presenting Author: XIAO LIU, Ohio University, United States
P32: PREDICTING GENE MUTATIONS DURING CANCER EVOLUTION: A NEW TOOL IN SEARCHING NOVEL TARGETS FOR CANCER TREATMENT  
Presenting Author: JUAN MARTINEZ, Florida International University, United States

P33: PAIRPRED: A LARGE MARGIN METHOD FOR PARTNER-SPECIFIC PREDICTION OF PROTEIN INTERFACES  
Presenting Author: FAYYAZ MINHAS, Colorado State University, United States

P34: SOMATIC HYPERMUTATIONS IN IMMUNOGLOBULIN VARIABLE REGION IN CHRONIC LYMPHOCYTIC LEUKEMIA B-CELLS  
Presenting Author: ALEXANDRA MIRINA, Albert Einstein College of Medicine of Yeshiva University, United States

P35: THE CRITICAL ASSESSMENT OF FUNCTION ANNOTATION EXPERIMENT: A COMMUNITY-WIDE EFFORT TOWARDS A BETTER FUNCTIONAL ANNOTATION OF GENES AND GENOMES  
Presenting Author: SEAN MOONEY, Buck Institute for Research on Aging, United States

P36: MACHINE LEARNING ANALYSIS OF HIGH-THROUGHPUT TRANSCRIPTION START SITE DATASETS  
Presenting Author: TAJ MORTON, Oregon State University, United States

P37: COMPARING DE NOVO ASSEMBLERS FOR METAGENOMIC DATA  
Presenting Author: AMIR MUHAMMADZADEH, University of Saskatchewan, Canada

P38: METABOLIC RECONSTRUCTION IDENTIFIES STRAIN-SPECIFIC REGULATION OF VIRULENCE IN TOXOPLASMA GONDII  
Presenting Author: NUR SIMULU, University of Toronto, Canada

P39: IDENTIFICATION OF SPLICE-MODULATING GENETIC VARIANTS FROM RNA-SEQUENCING DATA  
Presenting Author: ANELIA HORVATH, McCormick Genomics and Proteomics Center, The George Washington University, United States

P40: A CORPUS-BASED STUDY OF TEMPORAL RELATIONS IN CLINICAL TEXT  
Presenting Author: NATALYA PANTELEYEVA, University of Colorado, United States

P41: GIST – AN ENSEMBLE APPROACH TO THE TAXONOMIC CLASSIFICATION OF METATRANSCRIPTOMIC READS  
Presenting Author: JOHN PARKINSON, Hospital for Sick Children, United States

P42: NOVEL STRATEGY FOR ASSEMBLING GENOMES FROM PROKARYOTES EXHIBITING STRONG SYMBIOSIS WITH OTHER BACTERIAL POPULATIONS IN NATURE  
Presenting Author: ROBIN PAUL, Arizona State University, United States

P43: EXPERIMENTAL DETERMINATION OF USEFUL AND INFORMATIVE VISUALIZATIONS OF MICROBIAL ECOLOGY DATA FOR PUBLIC AND SCIENTIFIC AUDIENCES  
Presenting Author: MEGAN PIRRUNG, University of Colorado, United States

P44: HUMAN METABOLIC ATLAS: AN ONLINE RESOURCE FOR HUMAN TISSUE-SPECIFIC GENOME SCALE METABOLIC MODELS  
Presenting Author: NATAPOL PORNPUTTAPONG, Chalmers University of Technology, Sweden

P45: A POLYGLOT APPROACH TO BIOINFORMATICS DATA INTEGRATION: PHYLOGENETIC ANALYSIS OF HIV-1  
Presenting Author: STEVEN REISMAN, Loyola University of Chicago, United States
P46: WEBGOOSE: INTEGRATION OF DIVERSE BIOLOGICAL SOFTWARE AND DATA ON THE WEB
Presenting Author: DIEGO SALVANHA, University of Sao Paulo and Institute for Systems Biology, United States

P47: EVOLUTION OF PALMITOYL ACYL TRANSFERASES (PATS) IN APICOMPLEXA
Presenting Author: SWAPNA SESHADRI, Research Institute, Hospital for Sick Children, United States

P48: DETECTING ALTERED METHYLATION STATES USING HIGH THROUGHPUT DNA SEQUENCING
Presenting Author: MEENAKSHI SHARMA, University of Houston, United States

P49: HISTORECEPTOMIC SIGNATURE OF THE ATYPIA OF CLOZAPINE
Presenting Author: EVGENY SHMELKOV, New York University School of Medicine, United States

P50: FOCUS: AN ALIGNMENT-FREE MODEL TO IDENTIFY ORGANISMS IN METAGENOMES USING NON-NEGATIVE LEAST SQUARES
Presenting Author: GENIVALDO SILVA, San Diego State University, United States

P51: USING DIFFERENTIAL CORRELATION TO IDENTIFY SIGNIFICANT MOLECULAR INTERACTIONS
Presenting Author: CHARLOTTE SISKA, University of Colorado Denver, United States

P52: INFERRING HIGH-RESOLUTION INTERACTION NETWORK FROM SINGLE-CPG CO-METHYLATION ANALYSIS
Presenting Author: QIANG SONG, University of Southern California, United States

P53: ANALYZING BIOLOGICAL NETWORKS USING DEGREE-OF-INTEREST FUNCTIONS
Presenting Author: CORINNA VEHLOW, Visualization Research Center, University of Stuttgart, Germany

P54: YSTRESS: YEAST STRESS MICROARRAY DATABASE
Presenting Author: KWANJEERA WANCHITANARAK, Chalmers University of Technology, Sweden

P55: IMPROVED RNAI INTERFERENCE TARGET SEQUENCING (RIT-SEQ) ENABLES DISSECTION OF CELLULAR FUNCTION IN TRYPANOSOME BRUCEI
Presenting Author: JONATHAN WILKES, University of Glasgow, United States

P56: CORRELATES OF GENETIC DIVERSITY IN ACTINOPTERYGII S7 RIBOSOMAL PROTEIN GENE
Presenting Author: TANGJIE ZHANG, Yangzhou University, China
P01: VARIANCE COMPONENT SCORE TEST IN A MIXED-EFFECTS MODEL FRAMEWORK TO MAP TISSUE-SPECIFIC eQTL

SUBJECT: QUALITATIVE MODELING AND SIMULATION

Presenting Author: CHAITANYA ACHARYA, Duke University, United States
Author(s): Andrew Allen, Duke University, United States

ABSTRACT: Expression quantitative trait loci (eQTL) analysis associates putative regulatory variants (SNPs) with gene expression levels, which are treated as quantitative traits. Until recently, eQTL analysis is performed in a tissue-by-tissue basis followed by an examination of overlap of eQTLs across all tissues. However, most of those methods fall short in their ability to jointly analyze data across multiple tissues. Such type of joint analyses of tissue-types have been shown to improve power to identify eQTLs that have similar effects across tissues. We propose a variance component score test approach in a mixed-effects framework in order to jointly analyze multiple tissue types and assess the power of such tests. Using Monte Carlo simulations, we show that the new score test performs much better than the traditional likelihood ratio method in terms of statistical power. Using real data sets, we show that the new score test not only preserves power but also is computationally very efficient. We think that this method will particularly be very useful in prioritizing variants when analyzing heterogeneous disease model systems especially for any downstream genomic analysis including but not restricted to next-generation sequencing analysis.

P02: WHAT YOU DID NOT KNOW YOUR TRANSCRIPTION FACTOR WAS DOING

SUBJECT: OTHER

Presenting Author: MARY ALLEN, University of Colorado, United States
Author(s): Justin Freeman, University of Colorado, United States, Hestia Mellert, University of Colorado, United States, Joaquin Espinosa, University of Colorado, United States, Robin Dowell, University of Colorado, United States

ABSTRACT: A transcription factor (TF) protein binds to DNA and regulates transcription of a target gene. The guardian of the genome, p53, is a transcription factor important in cancer and aging, and activates transcription of many genes involved in apoptosis and cell cycle arrest. I have used a novel technique to discover over 200 annotated genes that are direct transcriptional targets of p53. This technique, GRO-seq, captures nascent transcription. Moreover, my work shows that short bidirectional transcripts are produced from p53 binding sites when the sites are within, nearby, or distant from protein
coding genes. Additionally, I demonstrate that when p53 is activated, transcription at its binding sites increases. Finally, I show the p53 binding sites have high levels of transcription when they are located near p53 targets genes (protein coding genes). The novel discovery that binding sites are transcribed and that transcription levels of binding sites correlate with TF activity leads to new questions about how transcription of binding sites affects TF binding and activation of target genes.

**P03: COMPUTATIONAL AND MATHEMATICAL MODELING OF THE SEGMENTATION GENES OF HONEYBEE (APIS MELLIFERA)**

**SUBJECT: OPTIMIZATION AND SEARCH**

_Presenting Author: MARYAM BAGHER OSKOUEI, University of Otago, United States
Author(s): Brendan McCane, University of Otago, New Zealand, Peter Dearden, University of Otago, New Zealand_

**ABSTRACT:** Drosophila and Honeybee embryos are two examples that develop a segmented body plan during their early development. The basic body plan consists of distinct segments along their anterior-posterior axis established via a segmentation process. The process subdivides the embryos into segments, which is controlled by interactions between segmentation genes. Many experimental and computational works have been tested to reveal which interactions cause this process in Drosophila embryos, but few have been done for Honeybee embryos. The Honeybee genome has some aspects that make it worth studying. Honeybees are excellent comparative model systems that help to understand evolutionary pathways behind the segmentation process considering that the insects diverged ~350 million years ago. Here, we present a method using ordinary differential equations (ODEs) to model segmentation genes in Honeybee embryos. The initial and target models for ODEs were configured with data collected in Peter Dearden’s lab. The computational modeling was carried out in order to explore how likely each gene is regulated by other genes positively or negatively. The simulations were performed in two phases, first as a Pre-stripes Networks and then the striped pattern forming Networks. The main findings predict gene networks that are more likely to pattern different parts of embryos along their anterior-posterior axis during early developmental stages. These results are comparable with Drosophila embryos. Importantly, the predicted networks provide hypotheses that can be tested experimentally.
**P04: TOWARDS REALIZING CLOUD BASED BIOLOGICAL ECHO SYSTEM FOR BIOLOGICAL WORKFLOWS**

**SUBJECT:** NETWORKING, WEB SERVICES, REMOTE APPLICATIONS  
*Presenting Author: JANAKA BALASOORIYA, Arizona State University, United States*

**ABSTRACT:** Biological computations involving genes and proteins more often than not require tremendous amount of computing resources and the discovery of the required services is very time consuming. One of our earlier papers titled “Cloud Computing Infrastructure for Biological Echo-Systems” proposed a framework on a cloud environment. This poster provides a proof of concept application above towards realizing our framework. Web services for gene analysis are combined in such a way that the output of one web service serves as the input of another. The web services when deployed on a cloud infrastructure would facilitate scalability, improved response time and virtualization of required resources (which could be in the form of storage elements like databases or computing elements like servers).

The proof of concept application provides functionalities corresponding to Single Nucleotide Polymorphisms (SNPs), proteins and genes. SNP functions include finding consequence type, regulatory feature etc. Retrieval of protein features and protein interactions are functions of the application that come under proteins functions.

**P05: DETECTING THE CRISPR/CAS SYSTEM IN FRESHWATER ENVIRONMENTS**

**SUBJECT:**  
*Presenting Author: DAVID BALTRUSAITIS, Loyola University Chicago, United States  
Author(s): Mike Shaffer, UC-Denver, United States, Catherine Putonti, Loyola University Chicago, United States*

**ABSTRACT:** Advances in next-generation sequencing technologies have led to increased numbers of metagenomic studies for a wide variety of environmental niches. Although first characterized in culturable species, the clustered regulated interspaced short palindromic repeats (CRISPR) system is just starting to be studied in these complex data sets. The CRISPR system, shown to provide bacteria with adaptive immunity to foreign genomic material, consists of loci of associated genes that code for pertinent proteins in conjunction with arrays of spacer and direct repeat sequences. These spacer sequences match a subsequence within the invading virus/plasmid and thus confer immunity. A handful of tools have been developed to detect these arrays, typically within long contig sequences or assembled genomes, with varied success. Furthermore, as previous research has shown, these tools are ill-equipped to examine shorter sequencing reads. We recently extended our existing tool, SpacerSeeker, to evaluate its performance in array detection, optimized for unassembled short read metagenomic data.
Instead of only detecting spacer sequences, the program will now detect repeat and spacer units. This was performed in conjunction with a phylogenetic analysis of the direct repeats thus far observed in nature. Our own molecular work has indicated a high prevalence of the CRISPR system within freshwater environments. Although previous studies have detected CRISPR arrays within salt-water environments, little research has been devoted to exploring the prevalence of the CRISPR system in these communities. As such, we tested this new functionality on reads from bacteria isolated from freshwater samples.

**P06: PREDICTING GENOMIC 3D CONTACTS**

**SUBJECT:** QUALITATIVE MODELING AND SIMULATION

*Presenting Author: SVEN BILKE, National Cancer Institute, United States*

**ABSTRACT:** Evidence for a non-random spatial 3d organization of the cells DNA content and its relevance for gene regulation has been accumulating in recent years. In a recent study [1], Dekker and co-workers introduced a novel method, HiC, allowing for an unbiased genome wide study of 3d conformations producing a “probability map” of DNA-DNA contacts in an ensemble of cells. Here we aim to identify genomic parameters correlating with the 3d-structure described in [1].

We developed a model based on DNA sequence and chromatin structure related observables and a set of mixing parameters. Using Monte Carlo optimization techniques, we identify features correlating with the contact matrix. The resulting model reproduces the empirical consensus contact probability map described in [1] with Pearson’s correlation $r > 0.71$.


**P07: MP2GO: INFERRING GENE FUNCTION FROM PHENOTYPE**

**SUBJECT:** OTHER

*Presenting Author: JUDITH BLAKE, The Jackson Laboratory, United States
Author(s): Joao Ascensao, Rice University, United States, Mary Dolan, The Jackson Laboratory, United States, David Hill, The Jackson Laboratory, United States*

**ABSTRACT:** Biomedical ontologies, while they have proven to be instrumental in the advancement of biological research through their ability to efficiently consolidate scientific data, are also hampered by the segregation of knowledge domains that results from their independent curation. We have developed a new method to computationally infer gene function, as encoded in the Gene Ontology (GO), from mutant phenotypes, as encoded in the Mammalian Phenotype Ontology (MP), using a set and graph theory-inspired approach. We
apply this methodology to laboratory mouse (Mus musculus) data as represented in the Mouse Genome Informatics Resource (MGI). We believe this procedure represents a novel methodology for the inference of gene function, as it examines the emergent structure and relationships between the GO and MP annotations without considering the relationships semantically. This could allow for the discovery of unforeseen associations between gene function and phenotypes that would be overlooked by a semantic-based approach. The technique could be applied to a variety of other organisms and annotation databases, taking full advantage of the abundance of available high quality curated data.

**P08: HIGH THROUGHPUT PHENOTYPE PROFILING FOR BACTERIAL FLUX-BALANCE MODEL OPTIMIZATION**

**SUBJECT: QUALITATIVE MODELING AND SIMULATION**

*Presenting Author: DANIEL CUEVAS, San Diego State University, United States*  
*Author(s): Rob Edwards, San Diego State University, United States, Daniel Garza, Evandro Chagas Institute, Brazil, Savannah Sanchez, San Diego State University, United States*

**ABSTRACT:** Advances in large-scale genomic sequencing allow researchers to create accurate computational models of organisms through the use of gene annotation software, such as RAST (Rapid Annotation using Subsystem Technology). These bioinformatics software deduce gene function through homology-based distinctions that are dependent on previously verified information; thus new discoveries cannot be easily extrapolated from current analysis tools without experimental examination. Recent developments using phenotype microarrays (PMs) provide a high throughput, large-scale technique in profiling bacterial characteristics and their phenotypes. PMs have the potential to experimentally test various growth conditions and then provide bacterial yield in real-time. By coupling PM experiments with the advances of genomic sequencing and annotation, more robust and accurate computational models can be developed and confirmed.

Here we present a combined biological and computational approach that (1) uses optical density data from a PM system as input to evaluate various growth curves, and (2) optimize the flux-balance analysis (FBA) models by using the PM results as a base for in silico growth simulations. The bacterium Citrobacter sedlakii was sequenced and studied in the PM-FBA pipeline to assess the capabilities of our approach. RAST annotations produced a base computational model consisting of 1,367 enzymatic reactions. After PM-FBA optimization a total of 44 reactions were added to, or modified within, the model. The model correctly predicted the outcome on 89% of growth experiments.
P09: GENOME SEQUENCE OF AN OUTBREAK STRAIN OF MYCOBACTERIUM ABSCESSUS FROM RIO DE JANEIRO, BRAZIL AND PHYLOGENOMIC RELATIONSHIPS TO GLOBALLY DIVERSE STRAINS

SUBJECT: OTHER
Presenting Author: REBECCA DAVIDSON, National Jewish Health, United States
Author(s): Nabeeh Hasan, National Jewish Health, United States, Benjamin Garcia, University of Colorado Denver, United States, Paul Reynolds, National Jewish Health, United States, Eveline Farias-Hesson, National Jewish Health, United States, Rafael Silva Duarte, Universidade Federal do Rio de Janeiro, Brazil, Mary Jackson, Colorado State University, United States, Michael Strong, National Jewish Health, United States

ABSTRACT: Multiple isolates of Mycobacterium abscessus subsp. bolletii, collectively called BRA100, were associated with outbreaks of post-surgical skin infections across various regions of Brazil from 2003 to 2009. To investigate the genome content of these clinically important isolates, we sequenced, assembled and annotated the genome of one BRA100 strain called CRM0020 that was isolated from a patient in Rio de Janeiro, Brazil in 2006. The 4.8Mb draft genome contains 4794 predicted genomic features including 3224 (67.2%) genes with functional annotations, 1524 (31.8%) genes classified as hypothetical proteins and 46 tRNA. CRM0020 also contains a 56.4kb plasmid sequence that encodes for 63 predicted proteins and shows 99% sequence identity to the previously described plasmids, pMAB01 and BRA100, which were derived from distinct Brazilian M. abscessus subsp. bolletii strains. Given the recent report of an outbreak of M. abscessus subsp. bolletii strains infecting cystic fibrosis patients in the United Kingdom (UK), we used a phylogenomics approach to compare the CRM0020 genome to multiple UK outbreak isolates as well as other globally diverse strains with publically available genomes. Analyses of genome-wide single nucleotide polymorphisms (SNPs) revealed that the Brazilian-derived CRM0020 strain is more closely related to UK outbreak isolates than to strains derived from patients in the United States, Europe or Malaysia. Our study merges new genome sequence data with existing genomic information to explore the global diversity of infectious M. abscessus isolates and to compare outbreak strains from different continents.

P10: ANNOTATION OF HYPOTHETICAL PROTEINS IN HUMAN ADENOVIRUSES

SUBJECT: OTHER
Presenting Author: SHANE DORDEN, University of Tampa, United States

ABSTRACT: Adenoviruses are double stranded DNA viruses that have a genome size of approximately 35kb. These viruses infect all vertebrates and human adenoviruses are associated with various illnesses such as acute respiratory disease, conjunctivitis, and gastroenteritis. Human adenoviruses are divided into seven species, A through G. Each of these
species is further divided into types that are numbered numerically. Numerous adenovirus genomes have been sequenced and are available in GenBank. While most of the proteins in these adenovirus genomes have been annotated, there are several hypothetical proteins whose functions are unknown. Assignment of function to these proteins will yield greater insight into adenovirus pathogenicity and epidemiology. We extracted these hypothetical proteins from these genomes and used sequence analysis and structure prediction to infer the functions of these proteins. We found that some of these hypothetical proteins can be annotated with a greater degree of confidence than other proteins where broader functional predictions can be assigned.

P11: SYSTEMATIC CLASSIFICATION OF COMMON DISEASE-ASSOCIATED SNPS BY THEIR EPIGENOMIC RELATIONSHIP

SUBJECT: SYSTEM INTEGRATION
Presenting Author: MIKHAIL DOZMOROV, Oklahoma Medical Research Foundation, United States
Author(s): Cory Giles, Oklahoma Medical Research Foundation, United States, Jonathan Wren, Oklahoma Medical Research Foundation, United States

ABSTRACT: The success of genome-wide association studies (GWASs) in finding causative SNPs for Mendelian phenotypes is contrasted with their inability to accurately elucidate complex patterns and biological roles of mutations underlying non-Mendelian inheritance. Our motivation was to find common epigenomic elements enriched with sets of disease-specific SNPs, and to systematically classify the diseases by their epigenomic background.

Human disease-specific SNPs were extracted from the UCSC GWAS catalog. We used our method, GenomeRunner (http://www.genomerunner.org) to test them for statistically significant associations with epigenomic data from the UCSC genome database. Disease-specific epigenomic associations were compared with random associations, obtained by testing random sets of SNPs. P-values of enriched associations were calculated using Fisher’s exact test, and corrected for multiple testing using Benjamini-Hochberg procedure. 212 disease and 363 trait/phenotype associated sets of SNPs were tested for associations with >4,000 genome annotation data. We identified that diseases/traits of similar origin (immunological, neurological, metabolic) tend to be located within similar epigenomic features. Our results suggest that alterations of specific epigenomic regulators may underlie disease susceptibility, guiding future epigenomic drug design and therapeutic targets.

The vast and growing amount of genome annotation data contains enormous potential to interpret sets of disease-associated mutations within a common, unifying theme of epigenomic regulators.
Considering these themes will empower us to interpret the results of GWASs in terms of unifying mechanisms, complementing SNP-gene-pathway approaches. Conversely, similarities and differences in epigenomic context of disease- and trait-associated SNPs provide a new means to classify phenotypes and understand their common epigenomic denominators.

P12: COMPARATIVE METAGENOMIC BY CROSS-ASSEMBLY

SUBJECT:

Presenting Author: BAS DUTILH, Radboud University Medical Centre, The Netherlands
Author(s): Rob Edwards, San Diego State University, United States

ABSTRACT: Determining the interrelationships between metagenomes from different biomes or different time points is important to understand the microbial world around us. Mapping metagenomic sequences to a reference database of known genes is a feasible approach to transfer taxonomical and functional annotations to sequence reads. However, it can limit the amount of data that can be analyzed because the majority of the sequencing reads in difficult-to-annotate datasets, such as viral metagenomes from biomes other than the human microbiome, lack known homologs. A promising alternative is reference-independent comparative metagenomics by cross-assembly.

Cross-assembly of different metagenomes is a fast and insightful way to obtain information about sequences that are shared between the samples, represented by cross-contigs. Importantly, cross-assembly is independent of an annotated reference database, providing a way to also handle unknown sequences. The cross-assembly tool crAss allows a rapid analysis of these cross-contigs. First, it provides cross-contig-based similarity scores between all metagenome pairs. Second, crAss creates insightful images displaying the inter-relationships between samples. Third, it generates occurrence profiles of the cross-contig sequences across metagenomes that can be used to discover related sequences, aiding further assembly and interpretation.

P13: GENESEEER AIDS DRUG DISCOVERY BY EXPLORING EVOLUTIONARY RELATIONSHIPS BETWEEN GENES ACROSS GENOMES

SUBJECT: NETWORKING, WEB SERVICES, REMOTE APPLICATIONS

Presenting Author: DOUGLAS FENGER, Dart Neuroscience, United States
Author(s): Matthew Shaw, Dart Neuroscience, United States, Philip Cheung, Dart Neuroscience, United States, Tim Tully, Dart Neuroscience, United States

ABSTRACT: Homologous relationships facilitate drug discovery by mapping gene/protein function between and within species, allowing functional predictions of novel or unknown genes. Additional benefits
of cross-species mapping include the following: use of paralogs for selectivity/specificity screens to eliminate drug side effects, translation of pathway information from model organisms to humans, and allowing comparison and combination of data from different species.

GeneSeer (http://geneseer.com) is a publicly available tool that leverages public sequence data, gene metadata information, and other publicly available data to calculate and display orthologous and paralogous gene relationships for all genes from several species, including yeast, insects, worms, vertebrates, mammals, and primates including humans. GeneSeer calculates homology relationships and its interface is designed to help scientists quickly predict important attributes such as additional closely related family members and paralogous relationships. It is a useful tool for cross-species translational mapping and enables scientists to easily translate hypotheses about gene identity and function from one species to another. We have validated GeneSeer versus Homologene, the homolog prediction tool from NCBI. The results show that GeneSeer is as good as, if not better than, Homologene. Finally, a comparison of features shows GeneSeer to be the most feature rich when compared to alternative homology tools.

**P14: CLUSTERING COEFFICIENTS IN PROTEIN INTERACTION HYPERNETWORKS**

**SUBJECT:** GRAPH THEORY

*Presenting Author: SUZANNE GALLAGHER, University of Colorado Boulder, United States*

**ABSTRACT:** Biological networks are usually modeled using graphs where nodes represent molecules (e.g., genes, proteins, metabolites), and pairs are connected by an edge to indicate an association (e.g., co-expression, regulation, binding). However, this model is insufficient for some types of data, such as affinity purification protein interaction data, which captures the interaction amongst many proteins rather than just two. To model this non-binary data, an extension of graphs known as hypergraphs has been proposed. However, due to the relative newness of the study of large complex hypergraphs, many of the statistics we use to study biological networks have not been well defined.

We examine a commonly-used network statistic, the clustering coefficient, in the context of protein interaction networks. We examine previous suggestions on how to extend this statistic to hypergraphs and look at the physical meaning of these in terms of protein interactions. We also present several novel definitions for the clustering coefficient that may better capture the intent of the clustering coefficient in binary graphs. We determined how well the various statistics can predict proteins in complexes or co-complexed pairs of proteins. Our results show that many of the hypergraph clustering coefficients perform...
better at these tasks than the clustering coefficient on the usual binary graph representation. We conclude that hypergraphs do represent an improvement over graphs and provide recommendations on which clustering coefficient definitions perform best.

**P15: IDENTIFICATION OF TISSUE-SPECIFIC GENE EXPRESSION SIGNATURES OF AGING VIA MICROARRAY META-ANALYSIS**

**SUBJECT:**

Presenting Author: CORY GILES, Oklahoma Medical Research Foundation, United States
Author(s): Jonathan Wren, Oklahoma Medical Research Foundation, United States, Mikhail Dozmorov, Oklahoma Medical Research Foundation, United States, Xiavan Roopnarinesingh, Oklahoma Medical Research Foundation, United States

**ABSTRACT:** Previous meta-analyses of gene expression changes that occur with aging have provided insights into individual genes and functional modules associated with aging, but have typically been either limited to analyses of one tissue, or have pooled different tissues together in an attempt to determine a “common” signature of aging. Because different tissues have different baseline signatures of gene expression and show distinct age-associated pathologies, we hypothesized that a comparative meta-analysis of various tissues would reveal tissue-specific signatures of aging.

Using text-mining, we identified approximately 18,000 human microarray samples from GEO, from various tissues, which had been annotated with the patient’s age. We then normalized these samples, and imputed the most probable tissue for each sample based on gene expression data. Using a linear regression model, we identified signatures of genes significantly associated with age within major tissues. We also searched for putative differences in epigenomic regulation of these differentially expressed genes by searching for enrichment or depletion among epigenomic tracks from UCSC in the promoters of differentially expressed genes.

**P16: MARK2CURE: A CROWDSOURCING PLATFORM FOR BIOMEDICAL LITERATURE ANNOTATION**

**SUBJECT:** TEXT MINING

Presenting Author: BENJAMIN GOOD, The Scripps Research Institute, United States
Author(s): Max Nanis, The Scripps Research Institute, United States, Andrew Su, The Scripps Research Institute, United States

**ABSTRACT:** Identifying concepts and relationships in biomedical text enables knowledge to be applied in computational analyses, such as gene set enrichment evaluations, that would otherwise be impossible. As such, there is a long and fruitful history of BioNLP projects that apply natural language processing to address this challenge. However, the state of the art in BioNLP still leaves much room for improvement in
terms of precision, recall and the complexity of knowledge structures that can be extracted automatically. Expert curators are still vital to the process of knowledge extraction but are in short supply.

Recent studies have shown that workers on microtasking platforms such as Amazon’s Mechanical Turk (AMT) can, in aggregate, generate high-quality annotations of biomedical text. In addition, several recent volunteer-based citizen science projects have demonstrated the public’s strong desire and ability to participate in the scientific process even without any financial incentives. Based on these observations, the mark2cure initiative is developing a Web interface for engaging large groups of people in the process of manual literature annotation. The system will support both microtask workers and volunteers. These workers will be directed by scientific leaders from the community to help accomplish ‘quests’ associated with specific knowledge extraction problems. In particular, we are working with patient advocacy groups such as the Chordoma foundation to identify motivated volunteers and to develop focused knowledge extraction challenges. We are currently evaluating the first prototype of the annotation interface using the AMT platform.

P17: WITHDRAWN

P18: TEMPORAL EXPRESSION RECOGNITION FOR CELL CYCLE PHASE CONCEPTS IN BIOMEDICAL LITERATURE

SUBJECT: TEXT MINING

Presenting Author: NEGACY HAILU, University of Colorado Computational Bioscience Program, United States
Author(s): Kevin Cohen, University of Colorado School of Medicine, United States

ABSTRACT: The number of publications in the biomedical domain is increasing exponentially. Searching for papers specific to a researcher’s interest in this domain is difficult. PubMed allows search using keywords but it doesn’t rank results based on document relevance. We present a recognizer for temporal expressions related to Cell Cycle Phase (CCP) concepts in biomedical literature. This task is one of the fundamental tasks towards building a search engine for queries with temporal components. Our ultimate goal is to build a specialized search engine, which is specific to searches in the CCP using genes and small molecules. We seek to improve search accuracy by allowing searches using semantic indexing instead of keywords. We identified 11 cell cycle related temporal expressions, for which we made extensions to TIMEX3, arranging them in an ontology derived from the Gene Ontology. We annotated 310 abstracts from PubMed. We developed annotation guidelines which are consistent with existing time related annotation guidelines such as TimeML. Two annotators participated in the annotation. We computed inter-annotator Agreement (IAA).
We achieved an IAA of 0.79 for exact span match and 0.82 for relaxed constraints. Our approach is a hybrid of machine learning to recognize the temporal expressions and a rule-based approach to classify them. We trained a named entity recognizer using Conditional Random Fields (CRFs) models. We used an off-the-shelf implementation of the linear chain CRF model. We obtained a performance of 0.77 F-score for temporal expression recognition. We achieved 0.79 and 0.78 macro and micro average F-scores for classification.

P19: Ontology Translation: A Case Study on Translating the Gene Ontology from English to German

Subject: Text Mining

Presenting Author: NEGACY HAILU, University of Colorado Computational Bioscience Program, United States

Author(s): Kevin Cohen, University of Colorado School of Medicine, United States, Lawrence Hunter, University of Colorado School of Medicine, United States

Abstract: Ontologies should be accessible to people in multiple languages, since multilingualism is inevitable in any scientific work. Furthermore, in theory the concepts of an ontology should not change their relationships regardless of language, but this hypothesis has not previously been tested. Due to resource scarcity, most ontologies of the biomedical domain are available only in English at present. We present techniques to translate Gene Ontology terms from English to German. We used three tools—a dictionary, Wikipedia, and the Google Translate API to translate the terms. Ontology terms can be a single word or as long as a complex phrase. Translating the short terms—terms which have word length less than or equal to 3—is difficult, since they are ambiguous. We constructed sentences using the ontologies so that the terms would have some contextual cues as to word sense. The sentence construction was done using the terms as key words in a PubMed/MEDLINE search. Evaluation was done by human experts. The evaluators were given a Likert scale ranging from 1 through 5, where 1 is poorly translated and 5 is fluent. They were asked to assess adequacy and fluency. On average, the adequacy Likert scale scores were 4.86, 4.09, and 4.90 for the three methods i.e. translation using Wikipedia, Google translator for short terms and terms within sentence. Average fluency scores for the three methods were 4.02, 4.39, and 4.43, respectively. Our results showed that performance for translation of short terms was improved when the terms are in a sentential context.
**P20: INTEGRATIVE VISUALIZATION FOR DISCOVERY OF PHENOTYPE ASSOCIATIONS IN CLINICAL AND miRNA DATA**

**SUBJECT: MACHINE LEARNING, INFERENCE AND PATTERN DISCOVERY**

*Presenting Author: MICHAEL HINTERBERG, University of Colorado-Denver, United States*

*Author(s): Lawrence Hunter, University of Colorado, United States, David Kao, University of Colorado, United States*

**ABSTRACT:** The increasing size and availability of large clinical datasets provides opportunity for discovery of novel, complex phenotypes in patients. Some of these phenotypes, such as drug responsiveness, are important for differential treatment modalities. Complex phenotypes may also be associated with arrays of diagnostic biomarkers; for example, differential expression of mRNA as well as microRNA can segregate different classes of patients.

In datasets with thousands of clinical features, testing hypotheses for associations between clinical phenotype and genetic expression can be a tedious process. Furthermore, slight modifications in patient stratification may have dramatic effects on biomarker association, but these differences may not be readily apparent. In ongoing work, we present a novel web-based visualization tool that allows the user to view and modify tree-based representations of clinical phenotypes and examine associations with microRNA and mRNA expression, with visible transitions that show the effect of modifying phenotype definition. A specific motivating application to drug-responsiveness in non-ischemic dilated cardiomyopathy is presented as well.

**P21: THE CHANGING VIEW OF PROTEIN INTERACTION NETWORKS BASED ON DATA AVAILABILITY**

**SUBJECT: GRAPH THEORY**

*Presenting Author: BARRETT HOSTETTER-LEWIS, California State University, Chico, United States*

*Author(s): Todd A Gibson, California State University, Chico, United States*

**ABSTRACT:** The end of the 20th century marked the beginning of the era of large-scale studies identifying protein interactions. These large data sets catalyzed a renaissance in protein interaction network research. The mainstay of this research has been in elucidating the biological and evolutionary factors that affect the network’s topological features. As the quantity of data increases and the quality of data improves we have been regularly refining our understanding of these biological and evolutionary influences. Now, as interaction data becomes available for recently sequenced organisms, there is a great opportunity for research on today’s nascent interactomes to benefit from the analytical steps and missteps taken on fledgling interactome analysis 10-15 years ago.
Here we describe how the researcher’s view of the Saccharomyces cerevisiae protein interaction network has changed since the first publication of large-scale yeast data in late 1999. By creating network snapshots using increasing amounts of interaction data constrained by date-of-publication and various quality criteria, we identify trends in the researcher’s view of network topology, and compare this to the interactomes of organisms which still have early, incomplete interaction data sets.

**P22: INTRODUCING COMPUTATIONS IN BIOLOGY/ BIOINFORMATICS – AN UNDERGRADUATE PERSPECTIVE**

**SUBJECT: QUALITATIVE MODELING AND SIMULATION**

*Presenting Author: MAHEEN KIBRIYA, Chapman University, United States*

*Author(s): Shehzein Khan, Chapman University, United States, Louis Ehwerhemuepha, Chapman University, United States*

**ABSTRACT:** The importance of computing in biological and life sciences cannot be overemphasized. It is imperative for students in biological sciences to be introduced to computing at the undergraduate level, and our work presents the view of undergraduates toward this shift to an interdisciplinary field. We developed a simple nucleotide sequence analysis program written in Python and discuss our experience in learning and using Python to solve simple biological problems. The aforementioned sequence analysis program was tested using sequence data from the tuberculosis database (www.tbdb.org), while some high level functions freely available in BioPython are briefly discussed.

**P23: DELETION OF COX-2 IS ASSOCIATED WITH REDUCED EXPRESSION OF RGL1: RELEVANCE TO CHEMOPREVENTATIVE EFFECT OF COX-2 INHIBITORS**

**SUBJECT: OTHER**

*Presenting Author: NICHOLAS KIRKBY, Imperial College London, United Kingdom*

*Author(s): Jane Mitchell, Imperial College London, United Kingdom*

**ABSTRACT:** Non-steroidal anti-inflammatory drugs (NSAIDs) such as aspirin, ibuprofen, rofecoxib and meloxicam inhibit cyclo-oxygenase (COX)-2 and are used for the treatment of arthritis. NSAIDs and COX-2 gene deletion have also been shown to prevent cancer in animal models, but research into the chemopreventative benefits of NSAIDs in man has been restricted following association of these drugs with cardiovascular toxicity. As such, the mechanisms by which COX-2 regulates cancer progression and cardiovascular health are incompletely understood. Here we have performed transcriptomic analysis of COX-2 knock out mouse aorta in order to identify gene pathways that may help to explain the role of COX-2 in these systems.

Microarray analysis of COX-2-/- vs COX-2+/+ aorta illustrated 29 differentially expressed genes. The most greatly altered of these was
Ral guanine nucleotide dissociation stimulator-like 1 (Rgl1) which was down-regulated in COX-2-/- mice (-1.73-fold, p=1.03x10^-8). To validate these observations, we mined data from ArrayExpress, which revealed studies showing that chronic dosing of rats with several NSAIDs including meloxicam has been associated with a reduction in Rgl1 expression. This is in line with our findings and supports our observations in gene-deleted mice.

These data suggest Rgl1 gene expression may be regulated by COX-2 activity. Rgl1 is a guanine nucleotide exchange factor, activated by Ras, the most common oncogene in human cancer, and catalyses the activation of, Ral. Ras and Ral are known to be crucial for cancer transformation and progression. Although preliminary, our observations may implicate Rgl1 down-regulation as a novel mechanism by which NSAIDs exert chemoprentative effects.

**P24: LINKING GENOTYPE AND ENTEROTYPE IN INFLAMMATORY BOWEL DISEASE**

**SUBJECT: METGENOMICS**

*Presenting Author: DAN KNIGHTS, University of Minnesota, United States*

*Author(s): Mark Silverberg, University of Toronto, Canada, Rinse Weersma, University Medical Center Groningen, Netherlands, Dirk Gevers, Broad Institute of Harvard and MIT, United States, Gerard Dijkstra, University Medical Center Groningen, Netherlands, Hailiang Huang, Massachusetts General Hospital, United States, Andrea Tyler, Mount Sinai Hospital, Canada, Suzanne von Sommeren, University Medical Center Groningen, Netherlands, Floris Imhann, University Medical Center Groningen, Netherlands, Joanne Stempak, Mount Sinai Hospital, Canada, Caitlin Russell, Massachusetts General Hospital, United States, Jenny Sauk, Massachusetts General Hospital, United States, Jo Knight, University of Toronto, Canada, Mark Daly, Massachusetts General Hospital, United States, Curtis Huttenhower, Harvard School of Public Health, United States, Ramnik Xavier, Massachusetts General Hospital, United States*

**ABSTRACT:** Human genetics and host-associated microbiomes have each been associated with inflammatory bowel disease (IBD), however IBD risk cannot be fully explained by either factor alone. Recent findings implicate genotype-enterotype crosstalk as a contributor to IBD pathogenesis. However, there has been no large study of complex genome-microbiome interactions in humans. We have performed such a study using bacterial 16S ribosomal RNA enterotyping and Immunochip genotyping from intestinal mucosal biopsies in three independent cohorts totalling more than 500 individuals. We present methodology, validated internally between cohorts, to test for host genetic locus interaction with taxonomic and functional components of the microbiome. In a targeted analysis integrating fine mapping of causal variants, we find nucleotide oligomerization domain 2 (NOD2)-specific risk associated with known IBD-related imbalances in bacterial taxa, including increased Gammaproteobacteria and Escherichia. NOD2 has known roles in management of commensal bacteria, and a strong genetic signal for increased IBD risk. Using imputed bacterial metagenomes we also find NOD2 risk linked to increased sulfur
reduction and lipopolysaccharide biosynthesis. These findings point to pathobiont expansion and bacterial production of genotoxic agent hydrogen sulfide, both involved in inflammation and IBD pathogenesis. In a novel omnibus tests we demonstrate links between host innate and adaptive immune pathways and broad enterotype composition. Our analysis demonstrates the ability to uncover novel interactions from paired genotype-enterotype data and that host genetics is linked to microbial dysbiosis in IBD.

**P25: Modeling Transcriptional Regulation Through Simulation of the Dynamic Changes in DNA Binding Factor Configuration**

**SUBJECT: Simulation and Numeric Computing**

*Presenting Author: DAVID KNOX, University of Colorado Anschutz Medical Campus, United States*

*Author(s): Robin Dowell, University of Colorado Boulder, United States*

**ABSTRACT:** Transcriptional regulation is the complex system behavior arising from the interaction of numerous regulators with DNA. Experimental efforts have unraveled the function of many individual components of the process, but the systems level behavior remains unpredictable. Growing evidence indicates that the transcriptional response of the system emerges not solely from the individual components, but rather by their collective behavior -- including competition and cooperation. The environment surrounding DNA undergoes millions of molecular interactions every second, resulting in continuous changes to the configuration of physically bound molecular components. It is from these stochastic, temporal, and spatial interactions of regulatory components that transcriptional regulation arises within each cell. Encapsulating our understanding of these interactions into computational models is essential for a full understanding of transcriptional regulation.

Our goal was to create biologically realistic computational models of Transcriptional Regulation that not only capture the behavior of several individual components, but also describe the dynamic and stochastic behavior of competing components. To this end we have developed an automated rule builder to not only create stochastic simulation rule sets, but also basic visualizations of the resultant simulations. Our modeling framework captures the competition between regulatory proteins, and more importantly, the dynamics of regulatory events occurring within individual cells.
P26: HIPPO: A TOOL FOR CONSTRUCTING QUALITY-CONSCIOUS META-ASSEMBLIES

**SUBJECT: OPTIMIZATION AND SEARCH**

Presenting Author: IRENA LANC, University of Notre Dame, United States
Author(s): Aaron Steele, University of Notre Dame, United States, Scott Emrich, University of Notre Dame, United States

**ABSTRACT:** The ability to extract meaningful results from genomic data depends on access to a well-constructed genome assembly. However due to limitations of time and money, manual finishing and validation are sometimes neglected. This issue is compounded by the fallibility of assemblers, which struggle with repeats, chimeric reads and contaminants and can vary widely in the caliber of their assemblies. In addition, some assemblers fare better than their competitors on specific datasets. Consequently, multiple assemblies are often generated to compensate for these issues. These assemblies can be the result of using different software, varying parameters, or incorporating new libraries. Deciding which of them is best can be daunting, as heuristics like N50 or number of contigs are too broad to adequately capture the overall quality of the resulting assemblies. HIPPO is designed to alleviate these problems by automatically merging multiple assemblies based on their sequence quality. The measure of quality derives from our previous work in assembly validation, where we gauged the correctness of an assembled sequence by generating a vector of quantifiable values for consecutive windows across the assembly. The two-step approach begins with a full genome alignment which is fed into a bipartite matching to identify candidate sections for improvement. These sections are woven together by a simplified de Bruijn path process to produce a meta-assembly consisting of only the highest-quality sections. The tool can extend, fill gaps, and replace dubious sections of sequence. HIPPO allows users to combine multiple assemblies and incorporate high-quality supplemental regions such as fosmids.

P27: FLOW-BASED NETWORK ALIGNMENT

**SUBJECT: GRAPH THEORY**

Presenting Author: RYAN LANGENDORF, University of Colorado, Boulder, United States

**ABSTRACT:** Network alignment has proven insightful in settings ranging from protein interaction networks to ontology matching. In ecology, the ability of networks to holistically describe systems and their dynamics has proven useful to community ecology as well as its applications to conservation biology and political ecology. However, analyses of networks have historically entailed correlating theoretical properties with ecological ones, such as the relationship between complexity and stability. While conceptually informative, such approaches have a limited ability to mechanistically account for
these relationships and only a superficial means by which to compare networks. In light of this, a flow algorithm was developed to align networks conceptually spreading water along interactions resulting in accumulation within species. Through the distributions of accumulated flow at nodes, networks of varying size and constitution can be directly and quantitatively compared addressing the relationship between a system’s structure and its emergent ecological properties. Compared to singular, often static metrics this technique allows for indirect interactions and network dynamics to be captured. Moreover, systems can be compared to themselves through time allowing for questions of complexity, scale-dependence, and thresholds to be addressed more robustly by quantifying the structural differences underlying temporal changes in a system. This technique allows the functional importance of a system’s structure, be it ecological or otherwise, to be approached more mechanistically.

**P28: AN OPTIMAL METABOLIC ROUTE SEARCH TOOL: ROUTESEARCH**

**SUBJECT: OPTIMIZATION AND SEARCH**

*Presenting Author: MARIO LATENDRESSE, SRI International, United States*

**ABSTRACT:** RouteSearch is a new Web accessible metabolic engineering tool available as part of BioCyc since March 2013. It enables searching for optimal metabolic linear routes between a start compound and a goal compound. The optimality criteria are the weighted sum of the costs of the reactions used, and the weighted sum of the costs of atoms that are lost in the transformation from the start compound to the goal compound. These costs and the number of minimum cost routes to find and display are user selectable. The routes are displayed as a series of connected enzymatic reactions including chemical structures of the substrates, where the conserved moieties within each metabolite are shown using colors. By using a graphical interface, the user can also easily identify each atom conserved or lost along each route. RouteSearch uses two algorithms to search for optimal routes: the Bellman-Ford algorithm that finds the least cost route, and a more general branch and bound search algorithm that can find several minimum cost routes. RouteSearch also uses a preferred organism to search -- a chassis in metabolic engineering terms, such as E. coli -- and a library of additional reactions, which is the MetaCyc database. The cost of using a reaction from MetaCyc is usually set higher than using a reaction from the chassis. In this way, new and more productive metabolic routes can be found for the chassis by adding reactions from MetaCyc. We will also briefly describe the computation of atom mappings for MetaCyc. Atom mappings are used by RouteSearch to track the atoms conserved and lost in a route.
P29: CHARACTERIZING UNKNOWN VIRAL GENES THROUGH METABOLOMICS

SUBJECT: MACHINE LEARNING, INFERENTIAL AND PATTERN DISCOVERY

Presenting Author: TIFFANY LIANG, San Diego State University, United States
Author(s): Savannah Sanchez, San Diego State University (SDSU), United States, Jason Rostron, SDSU, United States, Jeremy Frank, SDSU, United States, Daniel Cuevas, SDSU, United States, Anca Segall, SDSU, United States, Robert Edwards, SDSU, United States, Daniel Garza, Evandro Chagas Institute, Brazil

ABSTRACT: Viruses are the most diverse biological entities on earth. However, they also have the least characterized genetic, taxonomic, and functional diversity. In metagenomic analyses of viral communities from various environments, most sequences are unrelated to any known sequences; for example, about 90% of the viral sequences found in marine environments are unknown. The goal of this study is to characterize the function of unknown viral genes and identify those that alter host metabolism.

Viral metagenomes were collected from filtered seawater from Pacific coral reefs, sequenced by Roche 454 technology, and open reading frames were predicted from those sequences. Genes were synthesized and cloned into E. coli. These clones have been characterized in several different ways. To investigate these clones that affected metabolic processes, the metabolites were identified by gas chromatography-coupled time-of-flight (GC/TOF) mass spectrometry. In total 423 metabolites were found, however only 15% of those matched known compounds. We are identifying the specific metabolites produced or affected by the over expression of phage proteins to predict physiological roles for these proteins that can then be tested experimentally. We have also analyzed metabolic changes associated with expression of proteins with known functions that are involved in central metabolism; and clustering of these changes allows us to predict functions for other proteins. We are building a systematic analysis pipeline that can process metabolomics data for downstream analysis of metabolomics and related data sets.

P30: LINKAGE ANALYSIS: GENOMIC REGIONS CONTRIBUTING TO THE EXPRESSION OF TYPE 1 DIABETES MICROVASCULAR COMPLICATIONS

SUBJECT: OTHER

Presenting Author: ETTIE LIPNER, University of Colorado-Denver/National Jewish Health, United States
Author(s): David Greenberg, Nationwide Children’s Hospital, United States, Yaron Tomer, Mount Sinai, United States, Janelle Noble, Children’s Hospital Oakland Research Institute, United States, Cristina Monti, University of Pavia, Italy, Barbara Corso, University of Pavia, Italy, John Lonsdale, National Disease Research Interchange, United States

ABSTRACT: We conducted a linkage analysis to identify susceptibility loci for microvascular complications of type 1 diabetes (T1D). Using 402
SNP markers, our analysis used the phenotypes: 1) any microvascular complication, 2) retinopathy, 3) nephropathy, 4) neuropathy. When using “any complication” as the phenotype, we identified two linkage peaks: one located at HLA (HLOD=2.90) and another, novel locus telomeric to HLA (HLOD=3.13). These peaks were also evident when retinopathy was the phenotype (HLODs of HLA=2.69, telomeric locus=3.30). We did not find evidence for linkage for nephropathy or neuropathy. Previously published evidence suggest that DRB1 locus alleles affect complications’ expression, we stratified on families whose probands were positive for DRB1*03:01 and DRB1*04:01. Using the phenotype “any complication” and including only DRB1*03:01-positive families, the HLA peak decreased (HLOD=1.82) from the unstratified analysis (HLOD=2.90), and a peak centromeric to HLA appeared (HLOD=1.27). When stratifying on DRB1*04:01-positive families, the linkage evidence for HLA (HLOD=3.83) and the telomeric locus (HLOD=3.69) went up, despite the drop in sample size with stratification. When using the phenotype retinopathy, we observed the same increase in linkage peaks (HLOD at HLA=3.62, telomeric locus=3.76). These observations suggest that DRB1*04:01 interacts with the telomeric locus to produce complications’ susceptibility. Simultaneously, the drop in linkage evidence for DRB1*03:01 confirms a protective effect seen in our previously reported analysis (Lipner et al, 2013). Based on large differences in the HLOD scores, we argue that the DRB1*03:01-positive and DRB1*04:01-positive groups are genetically distinct, a finding in accordance with the observation that DRB1*03:01 is protective for retinopathy.

P31: A BIOINFORMATICS APPROACH TO IDENTIFY AND CLASSIFY EXTENSINS FROM SELECTED PLANT GENOMES

SUBJECT: TEXT MINING
Presenting Author: XIAO LIU, Ohio University, United States
Author(s): Richard Wolfe, Ohio University, United States, Lonnie Welch, Ohio University, United States

ABSTRACT: The cell wall plays crucial roles in plants and serves as a significant source for agricultural and industrial products. A deep understanding of the components of the cell wall will facilitate its utilization and better meet human needs. One important group of glycoproteins in the cell wall is the extensins (EXTs). EXTs have been involved in plant growth, development, and defense mechanisms, yet the research of EXTs had only been focused on a few model plants. A bioinformatics study was conducted to identify and classify EXTs using a bioinformatics software, BIO OHIO/Prot-class, developed at Ohio University. A total of 795 EXTs were identified from 14 selected
plant genomes. Based on their sequence information, these identified EXTs were divided into several subgroups, including classical EXTs, proline-rich extensin-like receptor kinases (PERks), leucine-rich repeat extensins (LRXs), formin-homologs (FHs) EXTs, short EXTs, long EXTs, and other chimeric EXTs. The identified EXTs revealed a more complete picture of the abundance and distribution of EXT family in plants, and provide insight for the evolution of EXTs in the plant kingdom.

**P32: PREDICTING GENE MUTATIONS DURING CANCER EVOLUTION: A NEW TOOL IN SEARCHING NOVEL TARGETS FOR CANCER TREATMENT**

**SUBJECT: MACHINE LEARNING, INFERENCE AND PATTERN DISCOVERY**

*Presenting Author: JUAN MARTINEZ, Florida International University, United States*

*Author(s): Sitharama Iyengar, Florida International University, United States, Nelson LopezJimenez, University of Miami, United States*

**ABSTRACT:** Classification of cancer based on gene expression has provided insights into its complex landscape of multiple interactions between gene networks. Next generation sequencing (NGS) strategies provide a genome-wide coverage at a single nucleotide resolution which has enabled scientists to interrogate cancer-specific genomic variants and compare them with the normal variants in the same patient. Determining the molecular signatures of genes mutated in cancer may help to predict the clinical outcome and to carry out therapeutic modifications in treating the patients. The search for reliable molecular signatures has provided a fertile field for computational approaches.

Our goal is to test a computational framework utilizing published data from a longitudinal study of patients with acute myeloid leukemia (AML), whose genome was subjected to NGS procedure at various points in time. In our framework we designed a combination of statistical analysis and machine learning approaches to process the available sequencing data. Chromosomes are analyzed separately and a prediction model is generated individually for each one of them. A total of 24 prediction models are created per patient. We test the framework by comparing the predictions of mutations based on the sequencing data generated at the time of cancer discovery with the actual regions that suffered mutational modifications at the time of relapse. The accuracy of the predicted number of mutations ranged from 75 to 84%; the accuracy of the locations of mutations ranged from 69% to 88% in the patients analyzed. Our predictions are in good agreement with the reported data.
P33: PAIRPRED: A LARGE MARGIN METHOD FOR PARTNER-SPECIFIC PREDICTION OF PROTEIN INTERFACES

SUBJECT: MACHINE LEARNING, INFERENCE AND PATTERN DISCOVERY

Presenting Author: FAYYAZ MINHAS, Colorado State University, United States
Author(s): Brian Geiss, Colorado State University, United States, Asa Ben-Hur, Colorado State University, United States

ABSTRACT: We have developed a novel partner-specific protein-protein interaction site prediction method called PAIRpred that uses the sequences and unbound structures of two proteins in a complex, and is based on support vector machines (SVMs). Unlike most existing machine learning methods for this problem, PAIRpred uses information extracted from both proteins in a complex using pairwise kernels to predict inter-residue contacts. Due to its partner-specific nature, PAIRpred presents a more accurate model of protein binding and is able to generate more detailed predictions. In order to better model the problem, we present an extension of SVMs that can capture the pairwise constraints that two distant residues in a protein cannot simultaneously interact with the other protein in a complex. We demonstrate PAIRpred’s performance on Docking Benchmark 4.0 and recent CAPRI targets. We have compared PAIRpred’s performance to existing methods such as ZDOCK, PPiPP and PredUS. PAIRpred offers state of the art accuracy in predicting binding sites at the protein level as well as inter-protein residue contacts at the complex level. We have studied the contribution of different sequence and structure features along with the effect of binding-associated conformational change on prediction accuracy. As an illustration of potential applications of PAIRpred, we have used it to analyze the nature and specificity of the interface in the interaction of human ISG15 protein with NS1 protein from influenza A virus. More information on PAIRpred is available at: http://combi.cs.colostate.edu/supplements/pairpred/.

P34: SOMATIC HYPERMUTATIONS IN IMMUNOGLOBULIN VARIABLE REGION IN CHRONIC LYMPHOCYTIC LEUKEMIA B-CELLS

SUBJECT: MACHINE LEARNING, INFERENCE AND PATTERN DISCOVERY

Presenting Author: ALEXANDRA MIRINA, Albert Einstein College of Medicine of Yeshiva University, United States
Author(s): Thomas MacCarthy, Stony Brook University, United States, Lirong Wei, Albert Einstein College of Medicine of Yeshiva University, United States, Matthew Scharff, Albert Einstein College of Medicine of Yeshiva University, United States, Aviv Bergman, Albert Einstein College of Medicine of Yeshiva University, United States

ABSTRACT: B-cell Chronic Lymphocytic Leukemia (CLL) is the most common adult leukemia in the western hemisphere. It is characterized by an excessive proliferation of one B-cell clone over others. Previous studies suggest that antibody selection by some unknown antigen(s) plays significant role in CLL. Therefore, the analysis of mutations in
immunoglobulin (Ig) variable region sequences derived from B-cells of CLL patients may give us valuable insights into the disease mechanism. However, it is not a trivial task as there is a need for methods to detect selection pressure in antibodies. For this purpose we developed a novel approach for detecting selection in Ig. Our methodology is based on comparing in vivo data in question obtained from B-cells of CLL patient to a reference dataset. We construct such a dataset based on in vitro data obtained from a biochemical assay, which allows us to exclude the possibility of selection presence in reference sequences. We applied our method to datasets of two different human immunoglobulin heavy chain (IGHV) variable regions: IGHV4-34 and IGHV3-23. Comparison of CLL B-cell sequences of these regions to sequences of healthy donors’ B-cells, which were under antigen selection, indicated a correlation and, thus, suggesting presence of antibody selection in CLL. Interestingly, different IGHV regions of CLL B-cells correlate with different subtypes of healthy donors’ B-cells, which suggests that mechanism of the disease may vary between patients depending on IGHV regions in prevailing B-cell clones.

**P35: THE CRITICAL ASSESSMENT OF FUNCTION ANNOTATION EXPERIMENT: A COMMUNITY-WIDE EFFORT TOWARDS A BETTER FUNCTIONAL ANNOTATION OF GENES AND GENOMES**

**SUBJECT: MACHINE LEARNING, INERENCE AND PATTERN DISCOVERY**

*Presenting Author: SEAN MOONEY, Buck Institute for Research on Aging, United States*  
*Author(s): Predrag Radivojac, Indiana University, United States, Iddo Friedberg, Miami University, United States*

**ABSTRACT:** A major challenge of the post-genomic era is understanding the function and disease associations of gene products. We are discovering new proteins far faster than we can characterize them experimentally. Most genome projects and derived databases rely fully on automated functional annotations, making the increase in annotation accuracy and coverage a prime goal for annotation algorithms. Understanding the accuracy of these function prediction algorithms is of primary importance to the process of translating sequence data into biologically meaningful information. Here we present the results of the first Critical Assessment of Function Annotations (CAFA) held during 2010-2011 and the challenge of the second CAFA experiment underway now. Thirty-four research groups worldwide participated in the first experiment, employing over 50 function annotation algorithms. The prediction methods were assessed using ROC curves, precision/recall curves, and variations on semantic similarity as applied to the Gene Ontology. During this presentation, I will discuss the results of the first CAFA experiment, the challenges we faced in assessing the results, and the future of CAFA. I will also describe the new experiment which will include biological process, molecular function, cellular component and human disease prediction.
tracks. Finally, I will describe ways in which you, the community, can participate.

**P36: MACHINE LEARNING ANALYSIS OF HIGH-THROUGHPUT TRANSCRIPTION START SITE DATASETS**

**SUBJECT: MACHINE LEARNING, INFERENCE AND PATTERN DISCOVERY**

*Presenting Author: TAJ MORTON, Oregon State University, United States*

*Author(s): Molly Megraw, Oregon State University, United States*

**ABSTRACT:** High-throughput sequencing protocols are now able to provide vast and detailed quantitative data on RNA polymerase II transcription start sites (TSSs). Combining these datasets with machine learning techniques can provide valuable new insights into the regulation and production of these mRNA transcripts. Additionally, these models can be used to build high-accuracy predictive models which rely solely on sequence content. The use of sequence content-based models is attractive because they can predict transcriptional events even in species with sparsely annotated genomes or those which lack genome-wide experimental data. The resulting models can be used to suggest or infer regulatory mechanisms which may control transcript production. Here we provide an overview of a sequence content-centric approach to the use of machine learning tools in TSS data analysis. We present a machine learning model for highly accurate TSS prediction, present preliminary results on an mRNA recapped product classifier in Arabidopsis thaliana, and show how the Elastic Nets technique can be used to improve the interpretability of TSS prediction models.

**P37: COMPARING DE NOVO ASSEMBLERS FOR METAGENOMIC DATA**

**SUBJECT: METAGENOMICS**

*Presenting Author: AMIR MUHAMMADZADEH, University of Saskatchewan, Canada*

*Author(s): Brett Trost, University of Saskatchewan, Canada, Vanessa Pittet, University of Saskatchewan, Canada, Stephen Johnson, University of Saskatchewan, Canada, Anthony Kusalik, University of Saskatchewan, Canada*

**ABSTRACT:** The main goal of metagenomics is to characterize the structure and dynamics of communities of non-clonal microorganisms. One step in metagenomic analysis is reconstruction of genomes by assembling sequence reads. Unlike a traditional sequencing project, which aims to determine the complete genome sequence of a single organism, metagenomic analyses require thousands of (partial) genomes from a microbial community to be sequenced and assembled simultaneously. Over the past few years, different methods have been developed or revised specifically for the de novo assembly of next generation sequencing data; however, there are only a few tools that specifically focus on metagenomic data. Given the considerable
difficulties involved in assembling such data, including inadequate and partial sampling of some genomes, different organism compositions and relationships, and the presence of repetitive fragments, reconstructing the full metagenome is a very demanding task. Here we provide an evaluation of current de novo short read assembly tools on metagenomic data. We test a number of state-of-the-art assemblers that were designed specifically for metagenomic data, as well as some that were not. The accuracy, performance, and computational requirements of these assemblers were evaluated using three datasets of simulated sequence reads, each having a different community complexity (low, medium, or high), as well as real reads obtained from the sequencing of environmental samples using Ion Torrent technology. Our evaluation of assemblers suggested that although no single assembler performed best on all of our criteria, MIRA slightly outperformed the other programs.

P38: METABOLIC RECONSTRUCTION IDENTIFIES STRAIN-SPECIFIC REGULATION OF VIRULENCE IN TOXOPLASMA GONDII

SUBJECT: QUALITATIVE MODELING AND SIMULATION

Presenting Author: NIRVANA NURSIMULU, University of Toronto, Canada

Author(s): Carl Song, University of Toronto, Canada, Stacy Hung, University of Toronto, Canada, Melissa Chiasson, NIAID, National Institutes of Health, United States, James Wasmuth, University of Calgary, Canada, Michael Grigg, NIAID, National Institutes of Health, United States, John Parkinson, University of Toronto, Canada

ABSTRACT: Estimated to infect at least a third of the world’s population, the Apicomplexan parasite, Toxoplasma gondii, represents a major threat to immunocompromised individuals and pregnant women, especially due to the limited efficacy of current therapeutic interventions. Since metabolism plays an essential role in providing energy and the basic building blocks required for growth, drug-development programs are now focussing more on targeting metabolic enzymes. We hypothesize that metabolic potential plays a key role in determining the virulence of different strains. Given often nonintuitive relationships between enzymes and pathways, constraints based models such as flux balance analysis (FBA), have emerged as indispensable tools to study the organization and operation of metabolic networks. Here we present a novel application of FBA that leverages microarray data to explore the impact of differential enzyme expression observed between virulent and avirulent strains of T. gondii. Our model correctly predicts the increased growth rate of the more virulent type I strain, relative to type II; further analysis predicts the increase in growth rate to result from increased energy production via upregulation of the glycolytic, pentose phosphate and TCA-cycle pathways. These findings highlight a regulatory route which, in addition to conferring growth rate plasticity, may impact the parasite’s outstanding ability to infect a broad range of hosts. Moreover,
drug assays confirm strain-specific sensitivities of several reactions, as predicted by in silico single knock-out experiments. This study demonstrates how expression data can be integrated into a model to give robust strain-specific predictions.

P39: IDENTIFICATION OF SPlice-MODULATING GENETIC VARIANTS FROM RNA-SEQUENCING DATA

SUBJECT: METOGENOMICS

Presenting Author: ANELIA HORVATH, McCormick Genomics and Proteomics Center, United States

Author(s): Prakriti Mudvari, McCormick Genomics and Proteomics Center (MGPC), Mercedeh Movassagh, MGPC, Kamran Kowsari, MGPC, Ali Seyfi, MGPC, Maria Kokkinaki, Georgetown University, Nady Golestaneh, Georgetown University

ABSTRACT: Among the major mechanisms that affect the splicing process are nucleotide changes, which disrupt or create binding sites for the spliceosome components. Most of the currently available tools for splice variants annotation employ preexisting knowledge, and may miss variants acting through unknown mechanisms. We have developed a novel, experimentally based, directly observational approach to identify potential cis-acting splice modulating variants from second generation sequencing RNA datasets. Our approach is based on screening for co-existence of SNP-call and junction abrogation within a single uninterrupted sequencing read, which represents a copy of the original RNA molecule. Our strategy employs the assumption that SNPs frequently found on reads spanning into the intron (vs continuing in the next exon) can indicate splice-modulating potency of the nucleotide change. Applying our pipeline on five in-house transcriptomes highlighted known and novel splice-modulating SNPs, located both within and outside canonical splice sites. Selected alternatively spliced alleles have been confirmed through wet lab studies.

This is the first experimentally based, high-throughput pipeline to identify cis-acting splice-modulating SNPs; it highlights novel splice implicated genetic variants and provides an innovative strategy to re-visit the splice-modulating potential of SNPs located in consensus sequences, and traditionally considered critical for the splicing regulation.

P40: A CORPUS-BASED STUDY OF TEMPORAL RELATIONS IN CLINICAL TEXT

SUBJECT: TEXT MINING

Presenting Author: NATALYA PANTELEYEVA, University of Colorado, United States

Author(s): Lawrence Hunter, University of Colorado, United States, Kevin Cohen, University of Colorado, United States

ABSTRACT: A corpus of clinical data was used to investigate the hypothesis that there are correlations between pairs of event types and the temporal links between them. A corpus of about 98,000 words that
had been annotated with events, event types, TIMEX3 expressions, and temporal links was examined for such associations. It was found that in fact most pairs of event types show a strong preference for or against a particular type of temporal link. It was also noted that all possible pairs of event types occur even in this relatively small corpus. The preference of specific pairs of event types for particular types of temporal links has implications for natural language processing systems, including establishing baselines for their performance and providing a priori knowledge that can be used to inform the construction of both rule-based and machine-learning-based systems for labeling temporal links in clinical documents. More basic questions about the linguistic expression of temporal relations in clinical text are examined, such as the extent to which they are sequential or not and the extent to which they are intersentential versus intrasentential. Whether surface linguistic cues from morphology, syntax, and lexicon enhance accuracy in establishing temporal link types is addressed.

P41: GIST – AN ENSEMBLE APPROACH TO THE TAXONOMIC CLASSIFICATION OF METATRANSCRIPTOMIC READS

SUBJECT: METGENOMICS

Presenting Author: JOHN PARKINSON, Hospital for Sick Children, Canada
Author(s): Geoffrey Halliday, University of Toronto, Canada

ABSTRACT: Whole-microbiome gene expression profiling ('metatranscriptomics' or 'RNA-seq') has emerged as a powerful means of gaining a mechanistic understanding of the complex inter-relationships that exist in microbial communities. However, due to the inherent complexity of microbial communities and the lack of a comprehensive set of reference genomes, currently available computational tools for metatranscriptomic analysis are limited in their ability to functionally classify and organize these sequence datasets. To meet this challenge we have been developing methods that combine accurate transcript annotation with systems-level functional interrogation of metatranscriptomic datasets. As part of these methods, we present GIST (Generative Inference of Sequence Taxonomy), which combines several statistical and machine learning methods for compositional analysis of both nucleotide and amino acid content with the output from the Burroughs-Wheeler Aligner to produce robust taxonomic assignments of metatranscriptomic RNA reads. In addition to identifying taxon-specific pathways within the context of a pan-microbial functional network, linking taxa with specific functions in a microbiome will produce deeper understanding of how their loss or gain alters microbiome functionality. Applied to real as well as synthetic datasets, generated using an inhouse simulation tool termed GENEPUDDLE, we demonstrate an improved performance in taxonomic assignments over existing methods.
P42: NOVEL STRATEGY FOR ASSEMBLING GENOMES FROM PROKARYOTES EXHIBITING STRONG SYMBIOSIS WITH OTHER BACTERIAL POPULATIONS IN NATURE

SUBJECT: METOGENOMICS

Presenting Author: ROBIN PAUL, Arizona State University, United States
Author(s): Robert Jinkerson, Colorado School of Mines, United States, Jason Steel, Arizona State University, United States, Kristina Buss, Arizona State University, United States, Petra Fromme, Arizona State University, United States

ABSTRACT: Leptolyngbya Heron Island (L.HI) is a newly isolated strain of cyanobacteria isolated from the Great Barrier Reefs in Australia. This strain exhibits the phenomenon of chromatic acclimation in which the cyanobacteria selectively express light harvesting proteins according to the wavelength of light its exposed to. To explore this phenomenon, sequencing the genome of the cyanobacteria is essential. However, this strain of cyanobacteria has strong symbiosis with other bacteria in its natural habitat making it impossible to obtain axenic cultures. The L.HI genome was sequenced using Illumina and reads were assembled using the Abyss software package. Initial %GC analysis of assembled scaffolds showed multiple peaks confirming the presence of heterotroph scaffolds. Initial selection of the cyanobacterial scaffolds involved devising a BLAST algorithm such that it only selected scaffolds which contained a gene which matched with some gene in Leptolyngbya sp. PCC 7375, a closely related cyanobacteria. Tetranucleotide frequencies were calculated for the BLAST positive scaffolds followed by principal component analysis. The cyanobacterial scaffolds which were clustered together were selected. %GC analysis showed one major peak followed by few smaller peaks which were removed. Genome annotation was carried out using NCBI Prokaryotic Genomes Annotation Pipeline followed by validation against the NCBI non-redundant database. This study shows that a complete genome sequence for a prokaryote can be obtained from nature which may have a very strong symbiotic relationship with other contaminating bacterial species by a combination of analyzing BLAST, tetranucleotide frequencies and %GC results. NCBI scaffold accession no. AWNH01000001-AWNH01000119.

P43: EXPERIMENTAL DETERMINATION OF USEFUL AND INFORMATIVE VISUALIZATIONS OF MICROBIAL ECOLOGY DATA FOR PUBLIC AND SCIENTIFIC AUDIENCES

SUBJECT: GRAPHICS AND USER INTERFACES

Presenting Author: MEGAN PIRRUNG, University of Colorado, United States

ABSTRACT: Sequencing technologies are getting cheaper and producing vast amounts of data, especially in the field of microbial ecology. Proper visualization of biological data is key to informative analysis and insight. Data analysis by users through informative, useful, and responsive visualizations is key to harnessing the true potential of big
data. We propose an experiment that will help to determine which types of visualization techniques are most informative for microbial ecology data in both public and expert scientific audiences. To perform this experiment on a large number of subjects in a systematic way, we have created a modular system with easily substituted visualization methods. We have created dynamic visualizations that parallel the visualizations available in QIIME using the d3 (Data Driven Documents) javascript library, and a visualization-testing framework that will be used to display the results of the American Gut Project. A visualization technique for one particular analysis as performed in QIIME will be randomly selected from the set of visualizations appropriate for the data selected and shown to the user. The user will also be presented with a questionnaire that will let us determine which visualizations allow users to answer the most questions correctly. We expect that the scores will indicate that certain visualization techniques are more appropriate for certain types of data, and that certain visualizations may be found informative for one audience over another, in public and scientific audiences.

P44: HUMAN METABOLIC ATLAS: AN ONLINE RESOURCE FOR HUMAN TISSUE-SPECIFIC GENOME SCALE METABOLIC MODELS

SUBJECT: NETWORKING, WEB SERVICES, REMOTE APPLICATIONS

Presenting Author: NATAPOL PORNPUTTAPONG, Chalmers University of Technology, Sweden
Author(s): Intawat Nookaew, Chalmers University of Technology, Sweden, Jens Nielsen, Chalmers University of Technology, Sweden

ABSTRACT: In recent years human tissue-specific genome-scale metabolic (h-tGEM) modeling has been provided much new information about human metabolism, in particular in connection with disease development. In order to efficiently manage and utilize this kind of data, we built the Human Metabolic Atlas (HMA) website as an online resource to provide comprehensive human metabolic information as models and a database for further specific analysis as well as to communicate with the wider research community. The metabolic models are illustrated using a web based metabolic map visualization system and provided in SBML file formats, which can be opened in most pathway and analysis software. With the visualization system, a summary of the provided h-tGEMs is overlaid on KEGG metabolic pathway maps with a zoom/pan user interface. Besides the models, users can easily access human reaction data, gathered from all h-tGEMs through a data query interface. The reaction data are standardized and organized by an internal developed Object-oriented database management system. Connecting to the database, users can use the provided web interface to easily retrieve reaction data with specific keywords or by using gene, protein, compound and cross reference in JSON and CSV format. This online resource is a useful tool
for studying human metabolism at the specific cell level, organ level and the for the overall human body.

**P45: A POLYGLOT APPROACH TO BIOINFORMATICS DATA INTEGRATION: PHYLOGENETIC ANALYSIS OF HIV-1**

**SUBJECT: NETWORKING, WEB SERVICES, REMOTE APPLICATIONS**

*Presenting Author: STEVEN REISMAN, Loyola University of Chicago, United States*

*Author(s): Catherine Putonti, Loyola University of Chicago, United States, George Thiruvathukal, Loyola University of Chicago, United States, Konstantin Läufer, Loyola University of Chicago, United States*

**ABSTRACT:** The rapid mutation rates of retroviruses such as HIV prove challenging when developing molecular therapies. RNA-interference (RNAi) has been recently developed as a means to destroy a known targeted sequence, and shows potential as an HIV-1 therapy. The designed small interfering RNA molecule used for the interference mechanism must be highly accurate, as it will only bind to a target with near perfect complementarity which current research suggests must be from 18-25 nucleotides in length. Therefore, in order to avoid viral escapes, the siRNAs must target the most highly conserved non-variant regions. Identifying such regions necessitates a multifaceted approach, considering functional and structural constraints. We have developed a data repository to facilitate such analyses. Using the wealth of sequence data publicly available, sequences are parsed, allowing to expose this data in the form of a RESTful web service, allowing for users to query the data based on several parameters, including country of isolation, collection date, and gene. We are now able to observe how sequence conservation varies with respect to distribution throughout the world. Scripts have been developed to align user-selected sequences such that the most non-variant regions can be identified by implementation of a Longest Common Subsequence (LCS) algorithm. In doing so, we have provided a method for future research to identify potential RNAi targets for specific subpopulations rather than attempting to find a non-specific global solution. While focusing here on HIV, the tools developed can be applied to any viral species of interest.

**P46: WEBGOOSE: INTEGRATION OF DIVERSE BIOLOGICAL SOFTWARE AND DATA ON THE WEB**

**SUBJECT: DATA MANAGEMENT METHODS AND SYSTEMS**

*Presenting Author: DIEGO SALVANHA, University of Sao Paulo and Institute for Systems Biology, United States*

*Author(s): Aaron Brooks, University of Washington / Institute for Systems Biology, United States, Ricardo Vencingo, University of Sao Paulo, Brazil, Nitin Baliga, Institute for Systems Biology, United States*

**ABSTRACT:** Many analysis and visualization tools have emerged as a result of the “Big Data” revolution in the biological sciences. Unfortunately, the enthusiasm to mine these data sets led to the
development of highly-specialized software with few resources dedicated to linking diverse tools and datasets. In systems biology, investigators typically want to observe biological systems across scales (e.g. molecular types or time scales). Since each analysis in this workflow typically involves its own custom-designed software, a centralized data integration tool would be invaluable. Here we describe the WebGoose, a data manager designed to integrate existing software and experimental data on the web. WebGoose is a browser-independent, HTML5-compliant data manager integrated into the Java-based Gaggle framework (http://gaggle.systemsbiology.net/). Like Gaggle, WebGoose is a light-weight data service that provides interoperability between web applications. WebGoose implements two distinct but related modules: A front end interface allowing data source/target selection and a back-end module which is responsible for transferring data to Gaggle. Once an independent web-resource is integrated to Gaggle using WebGoose, it becomes a full-fledged Gaggle-goose -- automatically receiving the capability to share data between all other developed geese. This allows third-party web-applications to access Gaggle-enabled databases (such as KEGG or STRING) as well as the suite of Gaggle-enabled software (such as R and MeV) with relatively little configuration. The WebGoose makes it easy to integrate diverse software applications on the web. The application is open-source and can be download at http://labpib.fmrp.usp.br/~dmartinez/webgoose

P47: EVOLUTION OF PALMITOYL ACYL TRANSFERASES (PATs) IN APICOMPLEXA

SUBJECT:
Presenting Author: SWAPNA SESHADRI, Research Institute, Hospital for Sick Children, Canada
Author(s): John Parkinson, Research Institute, Hospital for Sick Children, Canada, Tim Gilberger, McMaster University, Canada

ABSTRACT: Protein palmitoylation is the only reversible post-translational mechanism utilising a hydrophobic anchor known to dynamically regulate a protein’s function by influencing its subcellular localization, stability, and interaction. Although this process is ubiquitous in eukaryotes, a recent study uncovered hundreds of palmitoylated proteins in P. falciparum. Therefore, characterizing the suite of enzymes catalyzing this process (Palmitoyl Acyl Transferases (PATs)) in apicomplexan parasites is essential for understanding various aspects of parasite biology. We conducted a comprehensive survey to identify and classify PATs from complete genomes of 16 parasitic apicomplexans and 2 closely related free-living protists (ciliates). Using HMMER, 159 and 138 PATs were identified in apicomplexans and ciliates, respectively. Classification is confounded due to lower resolution stemming from short (≈50aa) conserved catalytic domain combined with presence of ankyrin repeats in many sequences. Analysis revealed a ~170aa region with sufficient information to
distinguish them into 7 major clades and 14 sub-clades, using Bayesian and maximum likelihood phylogenetic methods. The sub-clades demonstrate distinct patterns of sequence conservation and indels, providing molecular signatures for possible sub-functionalisation. A structural model of the catalytic domain was generated, providing a molecular perspective of these signatures. Overall, 5 sub-clades are apicomplexa-specific, containing members localized to rhoptries and inner membrane complex, organelles unique to apicomplexa that are involved in host cell invasion. Further, 2 clades and 2 sub-clades contain yeast and human orthologs indicating a role in secretory pathway. In summary, apicomplexans have evolved PATs to serve as an integral part of the biological machinery required to facilitate their parasitic life-style.

**P48: DETECTING ALTERED METHYLATION STATES USING HIGH THROUGHPUT DNA SEQUENCING**

**SUBJECT: OTHER**

*Presenting Author: MEENAKSHI SHARMA, University of Houston, United States*

*Author(s): William Widger, University of Houston, United States, Yuriy Fofanov, University of Texas Medical Branch, United States*

**ABSTRACT:** Disruption of methylation patterns has been associated to genomic instability and is a hallmark of cancer. To identify “methylation signatures” in a genome, affinity-based methods like immuno-precipitation (IP) along with high-throughput sequencing (HTS) are used in comparative and genome-wide studies. Mapping reads (subsequences) generated by HTS to human genome results in large numbers of alignments in the repeat-rich regions. The repetitive elements introduce bias and interfere with the accurate identification of differentially methylated regions (DMRs). We present novel computational approaches to detect DMRs in both unique and repetitive segments of the entire genome.

To eliminate/minimize bias introduced by repetitive DNA regions which cause coverage spikes (pile ups) affecting coverage analysis, we created “sequence length specific maps” of all the repeatable and unique locations in the human genome. This was accomplished by the “disassembly” of the human genome into all possible n-mers equal to the read length and alignment (mapping) of all the subsequences present in more than one copy in the reference genome. All the identified unique locations can now be used to estimate the differences in the “reads coverage” among the samples.

Log-transformations, MA-plots, and Z-scores were used to identify several genomic regions containing at least 44 genes that were hyper-methylated in dexamethasone (dex)-resistive and hypo-methylated in dex-sensitive Acute Lymphoblastic Leukemia cell lines. Our novel approach enables researchers to detect methylation alterations on a global scale and select candidate genes for locus-specific studies.
ABSTRACT: Historically, the mechanism of drug action is conceptualized via its interaction with a single cognate receptor, agnostic to the genetic expression of the latter. However, the entire pharmacologic activity of a drug, including both its beneficial and adverse effects, derives also from its “off-target” actions (polypharmacology). Additionally, the expression pattern of all the drug’s receptors is an essential factor that localizes the effect of a drug to a particular tissue. Thus, the true molecular signature of a drug consists of a complete vector of all its physiologically relevant receptor interactions across the spectrum of all receptors expressed in various tissues (histoReceptomic signature). Here, we defined a novel histoReceptomic signature for the atypical pharmacologic action (“atypia”) of the antipsychotic drug clozapine, i.e. its beneficial effects that the typical antipsychotic drug chlorpromazine does not exhibit.

Specifically, we derived the atypia signature by subtracting signatures of chlorpromazine and clozapine, obtained by integrating drug:receptors affinities with receptors gene-expression data. The generalized extreme Studentized deviate test was used to identify only physiologically significant tissue-specific drug:receptor interactions. Our results suggest that the common antipsychotic effects of clozapine and chlorpromazine are mediated through the 5-HT2a and 5-HT2c receptors in prefrontal cortex and caudate nucleus respectively, histamine H1 receptors in superior cervical ganglion, and muscarinic acetylcholine M3 receptors in prefrontal cortex. In contrast, targets exclusive to clozapine are dopamine D4 receptors in pineal gland, and muscarinic acetylcholine M1 receptors in prefrontal cortex. These results provide novel perspectives on mechanisms of action of antipsychotics and drug discovery in schizophrenia.

ABSTRACT: Microbes are more abundant than any other organism, and it is important to understand what those organisms are doing and who they are. In many environments a large majority of the members of
the microbial community cannot be cultured. Metagenomics uses high throughput sequencing, a fast and cheap sequencing method provided by the next generation of sequencing technologies. One of the major goals in metagenomics is to identify the presence of organisms in the microbial community from a huge set of unknown DNA sequences. This profiling has valuable applications in multiple important areas of medical research such as disease diagnostics. Nevertheless, it is not a simple task, and many approaches that have been developed are slow and depend on the read length of the DNA sequences. We designed FOCUS, an innovative and agile composition based model using non-negative least squares to profile and report the organisms present in metagenomic samples and their relative abundance without sequence length dependencies. The program was tested with simulated and real metagenomes, and the results show that our approach accurately predicts the organisms present in random communities faster than the available tools. The code and web-sever of FOCUS is freely available at http://edwards.sdsu.edu/FOCUS.

P51: USING DIFFERENTIAL CORRELATION TO IDENTIFY SIGNIFICANT MOLECULAR INTERACTIONS

SUBJECT: OTHER

Presenting Author: CHARLOTTE SISKA, University of Colorado Denver, United States
Author(s): Katerina Kechrís, University of Colorado Denver, United States

ABSTRACT: Different types of molecular features such as transcripts, proteins and metabolites can be measured using various –omics platforms and techniques. As it is becoming more common to generate –omics data on the same samples, methods are being developed that integrate the different types of data. We propose the use of differential correlation to identify pairs of molecular features (e.g. protein and transcript) with correlation that differs between disease groups. Molecular features that have differential correlation between groups are assumed to be involved in biological processes that are associated with disease status. We apply differential correlation to –omics data from NCI-60 cell lines to investigate cancer types and from human blood samples to study Chronic Obstructive Pulmonary Disorder (COPD). Results are validated using pathway-finding algorithms, where it is assumed that pairs of molecular features with significant differential correlation will be close to each other in a biological network. We also evaluate differential correlation using experimentally validated miRNA-mRNA interactions. We find that pairs of molecules that show differential correlation are close in biological networks compared to unrelated, randomly chosen pairs. We also discovered that differentially correlated pairs are enriched for experimentally validated interactions. In summary, we demonstrate how differential correlation can be used to predict novel molecular interactions associated with disease status, in addition to confirming the role of previously known molecular interactions.
P52: INFERRING HIGH-RESOLUTION INTERACTION NETWORK FROM SINGLE-CPG CO-METHYLATION ANALYSIS

SUBJECT: GRAPH THEORY

Presenting Author: QIANG SONG, University of Southern California, United States
Author(s): Andrew Smith, University of Southern California, United States

ABSTRACT: With the progress in genome annotation, a great proportion of the human genome has been assigned biological functions. How different genomic regions interact with each other, especially over long distances, becomes increasingly interesting. The whole-genome bisulfite sequencing are used to generate single-base resolution methylomes for multiple cell types, revealing a detailed picture of methylation dynamics during cell differentiation. In the human genome, functional regulatory regions, such as promoters, enhancers and transcription factor binding sites, are characterized by hypo-methylation. Functionally related regions tend to show similar methylation variation, and we call them co-methylated regions. We performed a genome-wide co-methylation analysis at single-base resolution using a fast algorithm based on locality-sensitive hashing. This analysis results in a high-resolution interaction network, revealing detailed information for several types of interactions, including gene module co-regulation, exon-promoter interaction, and physical proximity.

P53: ANALYZING BIOLOGICAL NETWORKS USING DEGREE-OF-INTEREST FUNCTIONS

SUBJECT: GRAPHICS AND USER INTERFACES

Presenting Author: CORINNA VEHLOW, Visualization Research Center, University of Stuttgart, Germany
Author(s): Carsten Goerg, University of Colorado Anschutz Medical Campus, United States, David Kao, University of Colorado Anschutz Medical Campus, United States

ABSTRACT: Biologists commonly analyze experimental data using biological networks, such as gene-expression correlation networks, to explain disease specific patterns and identify genotype-phenotype relationships. Biomedical knowledge from various databases and the literature can be integrated with these data networks to allow analysts to interpret experimental data in the context of existing knowledge. While these combined networks provide a rich resource and profound basis for data analysis, they are difficult to explore and understand since they are very dense. Using current static visualization approaches, it takes time and expertise to “untangle the hairball” and manually extract sub-networks that can explain a phenomenon or tell a meaningful biological story. To improve this analytical workflow, we developed a visualization approach that applies the concept of degree-of-interest (DOI) functions to highlight or filter particular parts of a network that are relevant for a specific question or task. We also use these DOI functions to automatically extract and lay out sub-networks...
in a way that DOI-based groups and their intersections become visually apparent, e.g., extracting a sub-network that includes all nodes involved in a set of pathways of interest and visually arranging these nodes based on their pathway information. To facilitate the analysis of extracted sub-networks in the context of the complete network, the network visualizations are linked through a brushing and linking feature. DOI functions can model various analytical facets, including an analyst's background and interest, properties of the experimental data, and phenotype information. Hence, they provide a generic and powerful approach for analyzing biological networks.

**P54: YSTRESS: YEAST STRESS MICROARRAY DATABASE**

**SUBJECT: OTHER**

*Presenting Author: Kwanjeera Wanichthanarak, Chalmers University of Technology, Sweden*

*Author(s): Dina Petranovic, Chalmers University of Technology, Sweden*

**ABSTRACT:** Unicellular organisms, as other cells, such as the model organism yeast Saccharomyces cerevisiae have to develop stress-response strategies in order to deal with various stresses they may encounter in a dynamically changing environment. In the last decade many have studied stress responses in yeast, at the level of genome wide DNA transcriptional response using DNA microarray technology, mostly focused on environmental changes such as aeration, temperature, pH, nutrients and osmolarity. All these data are very interesting and useful however, the data is scattered and difficult to compare so there remains the challenge of having a unifying bioinformatics resource where integrating and effectively querying data from numerous sources are available. Here we present yStress, a Yeast stress microarray database aimed to facilitate exploration of cross-platform and cross-laboratory stress microarray data. In addition, our platform allows meta-analyses, combining microarray data from related studies to identify differentially expressed genes, which can enhance statistical power, reliability and generalization of the results. The database collects the results from differential expression analysis and gene set analysis for both single microarray analysis and meta-analysis. A user-friendly web interface and interactive visualization are provided to display the queried data and results.
**P55: IMPROVED RNAI INTERFERENCE TARGET SEQUENCING (RIT-SEQ) ENABLES DISSECTION OF CELLULAR FUNCTION IN TRYPANOSOME BRUCEI**

**SUBJECT: OTHER**

*Presenting Author: JONATHAN WILKES, University of Glasgow, United Kingdom*

*Authors*: Graham Hamilton, University of Glasgow, United Kingdom, Richard McCulloch, University of Glasgow, United Kingdom

**ABSTRACT**: The protozoan parasite Trypanosoma brucei utilises a RNA interference (RNAi) pathway, widely conserved with other eukaryotes. This can be adapted to regulate expression of the poly-cistronically transcribed genes of T. brucei, utilising gene-specific sequences within a tetracyclin inducible cassette, allowing RNAi ‘knockdown’; now an important research tool. RIT-seq has been developed, which enabled the parallel analysis of >8000 genes in T. brucei in life-cycle and differentiation stages (1). The original RITseq methodology possesses a number of shortcomings which compromise its potential: semi-specific PCR produces small enrichments of the inserted sequences, produces inconsistent amplified sequences, and contains significant genomic sequence unrelated to the inducible fragments.

We have designed an adaptation of the methodology involving a specific PCR to amplify sequences between common primer sites flanking inserted genomic fragments in the RNAi cassette. Preparing the sequencing library from his amplified material requires 10-fold less material (500ng of DNA), produces a higher proportion (3-10-fold) of reads unequivocally derived from the cassettes, utilises standard protocols for library preparation and permits sample multiplexing. To validate this RITseq approach, we have screened for T. brucei genes that act in DNA damage repair by measuring read abundance after RNAi in the presence or absence of the SN2 alkylator methyl methanesulphonate. A number of previously characterised T. brucei DNA repair genes are revealed, and several novel pathways that have not been examined to date. The system was adapted to produce a comprehensive panel of protein kinase (kinome) probes.


---

**P56: CORRELATES OF GENETIC DIVERSITY IN ACTINOPTERYGII S7 RIBOSOMAL PROTEIN GENE**

**SUBJECT: METAGENOMICS**

*Presenting Author: TANGJIE ZHANG, Yangzhou University, China*

**ABSTRACT**: To research relationship between genetic variation and life-history variables of Actinopterygii, as indicated by common length, maximum length, maximum weight and longevity, and environmental variation, as indicated by three different fishes’ living environments, we applied analysis of independent regression and phylogenetically-
independent contrasts methods to evaluate life-history variables correlations with rps7 neutral genetic diversity. Polymorphism datasets of rps7 gene, belonging to 48 genera, 25 families and 9 orders, of Actinopterygii, was obtained from Polymorphix and Popset of GenBank. Life-history variables were obtained from the AnAge database and fishbase. The results showed that neutral genetic diversity of fishes is significantly negatively associated with common length. No strong level of correlation was found between fish’s neutral genetic diversity and maximum size, maximum weigh or maximum longevity. No correlation was found among neutral genetic diversity and fishes’ habits (marine, freshwater and marine-freshwater) either.
Rocky 2013 is an official conference of the International Society for Computational Biology.

Rocky 2013 is supported by the Computational Bioscience Program at the University of Colorado School of Medicine.

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