8th Annual RECOMB/ISCB Conference on REGULATORY and SYSTEMS GENOMICS with DREAM CHALLENGES

NOVEMBER 15-18, 2015
SHERATON PHILADELPHIA DOWNTOWN HOTEL, UNITED STATES
4 Credits
6 Welcome Letters
7 Floorplan
8 Agenda
14 Keynote Speakers
24 Presentation Abstracts
   24 DREAM Challenges
   38 Systems Biology and Regulatory Genomics
80 Posters
   80 DREAM Challenges
   92 Systems Biology and Regulatory Genomics
136 Sponsors
138 Authors Index
Credits

CONFERENCE STEERING COMMITTEE
Elise Blaese, IBM T. J. Watson Research Center, Yorktown Heights, United States
Andrea Califano, Columbia University, New York, United States
Manolis Kellis, Massachusetts Institute of Technology, Cambridge, United States
Christina Leslie, Memorial Sloan Kettering Cancer Center, United States
Saurabh Sinha, University of Illinois, Urbana-Champaign, United States
Gustavo Stolovitzky, IBM Research, Yorktown Heights, United States and Icahn School of Medicine at Mount Sinai

PROGRAM COMMITTEE
CO-CHAIRS
Christina Leslie, Memorial Sloan Kettering Cancer Center, United States
Saurabh Sinha, University of Illinois, Urbana-Champaign, United States

International Society for Computational Biology

The Eighth Annual RECOMB/ISCB Conference on Regulatory and Systems Genomics, with DREAM Challenges is an official conference of the International Society for Computational Biology.

PROGRAM COMMITTEE
Stein Aerts, University of Leuven, Belgium
M. Madan Babu, MRC Laboratory of Molecular Biology, United Kingdom
Mark Biggin, Lawrence Berkeley National Laboratory, United States
Serdar Bozdag, Marquette University, United States
Michael R. Brent, Washington University in St. Louis, United States
Harmen Bussemaker, Columbia University, United States
Andrea Califano, Columbia University, United States
Barry Demchak, University of California, San Diego, United States
Finn Drablos, Norwegian University of Science and Technology, Norway
Jason Ernst, University of California, Los Angeles, United States
Dean Felscher, Stanford University, United States
Ana T. Freitas, INESC-ID/IST, Technical University Lisbon, Portugal
Anthony Gitter, University of Wisconsin-Madison, United States
Raluca Gordan, Duke University, United States
Xin He, University of Chicago, United States
Avi Ma’ayan, Mount Sinai School of Medicine, United States
Shaun Mahony, Pennsylvania State University, United States
Satoru Miyano, University of Tokyo, Japan
T. M. Murali, Virginia Tech, United States
Theodore Perkins, Ottawa Health Research Institute, Canada
Alexander Pico, The Gladstone Institutes; University of California, San Francisco, United States
Miguel Angel Pujana, IDIBELL, Spain
Sushmita Roy, University of Wisconsin, Madison, United States
Ron Shamir, Tel Aviv University, Israel
Gustavo Stolovitzky, IBM Computational Biology Center, United States
Pavel Sumazin, Baylor College of Medicine, United States
Wei Wang, University of California, San Diego, United States
Weixiong Zhang, Washington University in St. Louis, United States

CONFERENCE ORGANIZATION
Diane Kovats, Executive Director, International Society for Computational Biology
Steven Leard, International Society for Computational Biology

CONFERENCE WEBSITE
Edna Nerona, Legacy Studios

PROGRAM BOOKLET DESIGN
Carol Dragich Bishop, Dragich Design
Scientific Publication Partners

F1000 RESEARCH
The ISCB Community Journal by F1000Research publishes a range of research formats in one centralized channel. Articles based upon the research presented at select official and affiliated conferences will be published using F1000Research's immediate publication and transparent peer review model.

DREAM CHALLENGES
ORGANIZING COMMITTEE
James Costello, University of Colorado Anschutz Medical Campus
Stephen Friend, Sage Bionetworks
Mike Kellen, Sage Bionetworks
Lara Mangravite, Sage Bionetworks
Pablo Meyer, IBM TJ Watson Research Center, Computational Biology Center
Thea Norman, Sage Bionetworks
Julio Saez-Rodriguez, RWTH-Aachen & EMBL-EBI
Gustavo Stolovitzky, IBM Research, Yorktown Heights, United States and Icahn School of Medicine at Mount Sinai

We gratefully acknowledge the additional challenge designers, data providers, curators, scorers, and sponsors without whom the DREAM challenges would not have happened:

PROSTATE CANCER
Kald Abdallah (Project Data Sphere, LLC), Christopher Bare (Sage Bionetworks), Elise Blaese (IBM Research), Brian Bot (Sage Bionetworks), James Costello (University of Colorado Anschutz Medical Campus), Stephen Friend (Sage Bionetworks), Justin Guinney (Sage Bionetworks), Stephen Jones (Covance), Mike Kellen (Sage Bionetworks), Joseph Morrell (Project Data Sphere, LLC), Thea Norman (Sage Bionetworks), Gustavo Stolovitzky (IBM Research), Tao Wang (UT Southwestern), Yang Xie (UT Southwestern), Liz Zhou (Project Data Sphere, LLC)

OLFACTION PREDICTION
Guillermo Cecchi (IBM Research), Andre Falcao (Universidade de Lisboa), Stephen Friend (Sage Bionetworks), Bruce Hoff (Sage Bionetworks), Andreas Keller (The Rockefeller University), Pablo Meyer (IBM Research), Raquel Norel (IBM Research), Thea Norman (Sage Bionetworks), Gustavo Stolovitzky (IBM Research), Christine Suver (Sage Bionetworks), Leslie Vosshall (The Rockefeller University)

ASTRAZENECA-SANGER DRUG COMBINATION PREDICTION
Christopher Bare (Sage Bionetworks), Rebecca Barrett (AstraZeneca), Graham Bignell (Sanger), Elias Chaibub Neto (Sage Bionetworks), Di Veroli (AstraZeneca), Jonathan Dry (AstraZeneca), Ulrika Edvardsson (AstraZeneca), Manel Esteller (IDIBELL), Stephen Friend (Sage Bionetworks), Simon Forbes (Sanger), Mathew Garnett (Sanger), Zara Ghazoui (AstraZeneca), Justin Guinney (Sage Bionetworks), Thea Norman (Sage Bionetworks), Julio Saez-Rodriguez (RWTH-Aachen & EMBL-EBI), Jo de Schoolmeester (AstraZeneca), Gustavo Stolovitzky (IBM Research), Eric KY Tang (AstraZeneca), Giovanni John P Vincent (AstraZeneca), Dennis Wang (AstraZeneca), Thomas Yu (Sage Bionetworks)

ALS STRATIFICATION PRIZE4LIFE
Hagit Alon (Prize4Life), Nazem Atassi (Massachusetts General Hospital), Merit Cudkowicz (Massachusetts General Hospital), Barbara Di Camillo (University of Padova), Stephen Friend (Sage Bionetworks), Javier Garcia (Pompeu Fabra University), Orla Hardiman (Trinity College Institute of Neuroscience), Bruce Hoff (Sage Bionetworks), Robert Kueffner (Ludwig-Maximilian-University), Guang Li (Origent data Sciences), Lara Mangravite (Sage Bionetworks), Raquel Norel (IBM Research), Thea Norman (Sage Bionetworks), Shay Rishoni (Prize4Life), Gustavo Stolovitzky (IBM Research), Christine Suver (Sage Bionetworks), Amy Truong (Sage Bionetworks), Liuxia Wang (Origent data Sciences), Neta Zach (Prize4Life)
Dear Conference Attendee,


It is an honor to have the privilege of working along side RECOMB and DREAM Challenges to produce this conference. The conference format is specifically designed to bring these focused communities together to share a rapidly growing body of knowledge through presentation of research results, information exchange, and collaboration. By the end of the conference, you should expect to have an increased understanding of gene regulation at the systems level, as well as best practices for predicting the structure of biological networks.

ISCB is dedicated to providing high quality meetings to our members and the scientific community. Each individual on the organizing and program committees have helped advance our mission. We thank them and our sponsors of the conference for their commitment especially the Conference Chairs, Christina Leslie, Memorial Sloan-Kettering Cancer Center and Saurabh Sinha, University of Illinois, and the other members of the Steering Committee: Elise Blaese, IBM T. J. Watson Research Center; Andrea Califano, Columbia University; Manolis Kellis, Massachusetts Institute of Technology; Gustavo Stolovitzky, IBM Computational Biology Center and Icahn School of Medicine at Mount Sinai; Christina Leslie, Memorial Sloan-Kettering Cancer Center; Saurabh Sinha, University of Illinois and ISCB Conferences Manager, Steven Leard. We also want to recognize and thank the reviewers involved in evaluating the scientific research.

If you are not already a member of ISCB, we encourage you to join. We are confident that your membership experience and the benefits will exceed the investment. To our members, thank you for your continued support of the Society. We are a small society and are driven by your passion, your visions, and by your activities. It is great to have you as a member; it is even greater to gain you as an active contributor such as those who made this event happen.

Together with our partners, we encourage you to make the best of every networking, learning, sharing and teaching moment of the conference. An opportunity like this to share high-level science in such a vibrant computational biology and bioinformatics region does not present itself every day. Enjoy it to the fullest!

Sincerely,

Alfonso Valencia  
ISCB President  
Diane E. Kovats  
ISCB Executive Director
Sheraton Philadelphia Downtown Hotel
201 North 17th Street, Philadelphia, PA 19103
Regulatory and Systems Genomics
with DREAM Challenges Agenda*

*SCHEDULE SUBJECT TO CHANGE

**DAY 1: Sunday, November 15, 2015**

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 – 5 pm</td>
<td>REGISTRATION</td>
<td>Independence lobby</td>
</tr>
<tr>
<td>3:15 pm</td>
<td>WELCOME &amp; ANNOUNCEMENTS</td>
<td>Independence A/B</td>
</tr>
<tr>
<td>3:30 pm</td>
<td>INVITED PRESENTATION</td>
<td>Independence A/B</td>
</tr>
<tr>
<td></td>
<td>Project Data Sphere® initiative overview</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Liz Zhou</td>
<td></td>
</tr>
<tr>
<td></td>
<td>DREAM PROSTATE CANCER CHALLENGE</td>
<td></td>
</tr>
<tr>
<td>3:50 pm</td>
<td>Introduction to the DREAM Prostate Cancer Challenge</td>
<td>Independence A/B</td>
</tr>
<tr>
<td></td>
<td>Jim Costello</td>
<td></td>
</tr>
<tr>
<td>4:10 pm</td>
<td>Presentation of Certificates</td>
<td>Independence A/B</td>
</tr>
<tr>
<td>4:20 pm</td>
<td>BEST PERFORMER PRESENTATION</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Predicting patient survival in the DREAM 9.5 mCRPC challenge</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Teemu D. Laajala, Suleiman Khan, Antti Airola, Tuomas Mirtti, Tapio</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pahikkala, Peddinti Gopalacharyulu, Tero Aittokallio</td>
<td></td>
</tr>
<tr>
<td>4:40 pm</td>
<td>BEST PERFORMER PRESENTATION</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Docetaxel adverse event prediction: a boosting method application</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fatemeh Seyednasrollah, Mehrad Mahmoudian, Outi Hirvonen, Sirkku</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Jyrkkiö, and Laura L. Elo</td>
<td></td>
</tr>
<tr>
<td>5:00 pm</td>
<td>BEST PERFORMER PRESENTATION</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Predicting discontinuation due to adverse effect in mCRPC</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Yuanfang Guan</td>
<td></td>
</tr>
<tr>
<td>5:20 pm</td>
<td>Discussion</td>
<td></td>
</tr>
<tr>
<td>5:30 pm</td>
<td>PREPARATION FOR DREAM HACKATHON</td>
<td>Independence C/D</td>
</tr>
<tr>
<td>6:00 pm</td>
<td>DREAM HACKATHON</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Big Data Hackathon</td>
<td></td>
</tr>
<tr>
<td>9:00 pm</td>
<td>ADJOURN</td>
<td></td>
</tr>
</tbody>
</table>

**WiFiAccess**
Network: Sheraton Meeting Password: ISCB2015
### DAY 2: Monday, November 16, 2015

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>8 am – 6 pm</td>
<td><strong>REGISTRATION</strong></td>
<td>Independence lobby</td>
</tr>
<tr>
<td>8:45 am</td>
<td><strong>MORNING WELCOME &amp; ANNOUNCEMENTS</strong></td>
<td>Independence A/B</td>
</tr>
<tr>
<td>9:00 am</td>
<td><strong>KEYNOTE: LESLIE VOSSHALL</strong></td>
<td>Independence A/B</td>
</tr>
<tr>
<td></td>
<td>The Known and Unknown of Human Smell</td>
<td></td>
</tr>
<tr>
<td>9:45 am</td>
<td><strong>DREAM OLFACITION CHALLENGE</strong></td>
<td>Independence A/B</td>
</tr>
<tr>
<td></td>
<td>Introduction to the DREAM Olfaction Challenge</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Andreas Keller</td>
<td></td>
</tr>
<tr>
<td>10:05 am</td>
<td>Presentation of Certificates</td>
<td></td>
</tr>
<tr>
<td>10:15 am</td>
<td><strong>BEST PERFORMER PRESENTATION</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Predicting olfaction response for each individual</td>
<td>Yuanfang Guan</td>
</tr>
<tr>
<td>10:35 am</td>
<td><strong>BEST PERFORMER PRESENTATION</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>From Shape to Smell: Predicting Olfactory Perceptual</td>
<td>Richard C. Gerkin</td>
</tr>
<tr>
<td></td>
<td>Descriptors using Molecular Structural Information</td>
<td></td>
</tr>
<tr>
<td>10:55 am</td>
<td><strong>DREAM Olfaction Prediction Challenge: Lessons Learned</strong></td>
<td>Amit Dhurandhar</td>
</tr>
<tr>
<td>11:15 am</td>
<td><strong>COFFEE BREAK WITH POSTERS</strong></td>
<td>Liberty A/B</td>
</tr>
<tr>
<td>11:45 am</td>
<td><strong>DREAM ALS STRATIFICATION PRIZE4LIFE CHALLENGE</strong></td>
<td>Independence A/B</td>
</tr>
<tr>
<td></td>
<td><strong>ALS STRATIFICATION PRIZE</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Using the Power of Crowdsourcing to Catalyze Breakthroughs in Amyotrophic Lateral Sclerosis</td>
<td>Neta Zach</td>
</tr>
<tr>
<td>12:05 pm</td>
<td>Presentation of Certificates</td>
<td>Independence A/B</td>
</tr>
<tr>
<td>12:25 pm</td>
<td><strong>BEST PERFORMER PRESENTATION</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>A Boosting Approach to Predicting ALSFRS Slope for the PRO-ACT Database</td>
<td>Wen-Chieh Fang, Chen Yang, Huan-Jui Chang, Hsih-Te Yang, Jung-Hsien Chiang</td>
</tr>
<tr>
<td>12:45 pm</td>
<td><strong>BEST PERFORMER PRESENTATION</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Predicting ALS survival through complete ranking of censored data</td>
<td>Yuanfang Guan</td>
</tr>
<tr>
<td>1:05 pm</td>
<td><strong>BEST PERFORMER PRESENTATION</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Using aggregated weights along paths across random forest to select important features and predict ALS progression</td>
<td>Jinfeng Xiao</td>
</tr>
<tr>
<td>1:25 pm</td>
<td><strong>LUNCH WITH POSTERS</strong></td>
<td>Liberty A/B</td>
</tr>
</tbody>
</table>
## DAY 2: Monday, November 16, 2015

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
<th>Location</th>
</tr>
</thead>
</table>
| 2:15 pm | Keynote: Steve Quake  
Single Cell Genomics                                                          | Independence A/B |
| 3:00 pm | **DREAM SMC CHALLENGE**   
Paul Boutros, Chair  
The ICGC-TCGA DREAM Somatic Mutation Calling Challenge: benchmarking somatic variant detection  
Josh Stuart                                                                   | Independence A/B |
| 3:20 pm | Presentation of Certificates                                             | Independence A/B |
| 3:30 pm | BEST PERFORMER PRESENTATION  
novoBreak: A k-mer targeted assembly algorithm for breakpoint detection in cancer genomes  
Zechen Chong                                                                | Independence A/B |
| 3:50 pm | BEST PERFORMER PRESENTATION  
TBA                                                                            | Independence A/B |
| 4:10 pm | Lessons from the SMC-DNA IS Challenges & Looking Forward  
Paul Boutros                                                                 | Independence A/B |
| 4:30 pm | **COFFEE BREAK WITH POSTERS**                                           | Liberty A/B       |
| 5:00 pm | **DRUG COMBINATION CHALLENGE**   
Julio Saez-Rodriguez, Chair  
Crowdsourcing combinatorial therapies: The AZ-Sanger DREAM synergy prediction challenge  
Michael Menden                                           | Independence A/B |
| 5:20 pm | Reducing data-leakage in leaderboard evaluations: The Ladder and LadderBoot algorithms  
Elias Chaibub Neto                                                        | Independence A/B |
| 5:40 pm | **OTHER**   
Gustavo Stolovitzky, Chair  
The DREAM Challenges channel: Open science publishing for all participating DREAMers  
Michael Markie                                                            | Independence A/B |
| 5:50 pm | **DREAM Brainstorm**   
All DREAM Conference participants                                             | Independence A/B |
| 6:00 pm | **DREAM POSTERS ON DISPLAY**                                           | Liberty A/B       |
| 6:30 pm | **DREAM HACKATHON**   
Big Data Hackathon                                                          | Independence C/D |
| 9:30 pm | **ADJOURN**                                                        |                   |

*Please remove posters at 7:30 pm

---

**WiFiAccess**

Network: Sheraton Meeting  
Password: ISCB2015
## DAY 3: Tuesday, November 17, 2015

<table>
<thead>
<tr>
<th>Time</th>
<th>Session/Activity</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>8 am – 6 pm</td>
<td><strong>REGISTRATION</strong></td>
<td>Independence lobby</td>
</tr>
<tr>
<td>8:45 am</td>
<td><strong>MORNING WELCOME &amp; ANNOUNCEMENTS</strong></td>
<td>Independence A/B</td>
</tr>
<tr>
<td>9:00 am</td>
<td><strong>KEYNOTE: BONNIE BERGER</strong>&lt;br&gt;Scaling with Compressive Algorithms</td>
<td>Independence C/D, Independence A/B</td>
</tr>
<tr>
<td>9:50 am</td>
<td><strong>SB01</strong>: MiSL: a method for mining synthetic lethal partners of recurrent cancer mutations uncovers novel mutation-specific therapeutic targets&lt;br&gt;• Subarna Sinha</td>
<td>Independence C/D</td>
</tr>
<tr>
<td>10:05 am</td>
<td><strong>SB02</strong>: The germline genetic component of drug sensitivity in cancer cell lines&lt;br&gt;• Michael Menden</td>
<td>Independence A/B</td>
</tr>
<tr>
<td>9:50 am</td>
<td><strong>RG01</strong>: Transcription factor binding site prediction in vivo using DNA sequence and shape features&lt;br&gt;• Anthony Mathelier</td>
<td>Independence A/B</td>
</tr>
<tr>
<td>10:20 am</td>
<td><strong>COFFEE BREAK WITH POSTERS</strong></td>
<td>Liberty A/B</td>
</tr>
<tr>
<td>10:45 am</td>
<td><strong>SB03</strong>: Estimating the Number and Diversity of Cancer Mutations In the Overall Population from 5,319 Complete Cancer Genomes&lt;br&gt;• Ramy Arnaout</td>
<td>Independence C/D</td>
</tr>
<tr>
<td>11:00 am</td>
<td><strong>SB04</strong>: Functional, chemical genomic, and super-enhancer screening identify sensitivity to cyclin D1/CDK4 pathway inhibition in Ewing sarcoma&lt;br&gt;• Gabriela Alexe</td>
<td>Independence A/B</td>
</tr>
<tr>
<td>11:15 am</td>
<td><strong>SB05</strong>: From phenotypic to molecular synergy: A transcriptional study of the dynamics of drug combinations based on single drug responses&lt;br&gt;• Mehmet Eren Ahsen</td>
<td>Independence C/D</td>
</tr>
<tr>
<td>10:45 am</td>
<td><strong>RG02</strong>: A PBM-based glossary for motif discovery in regulatory regions&lt;br&gt;• Luca Mariani</td>
<td>Independence A/B</td>
</tr>
<tr>
<td>11:30 am</td>
<td><strong>LUNCH WITH POSTERS (ODD)</strong></td>
<td>Liberty A/B</td>
</tr>
<tr>
<td>12:45 pm</td>
<td><strong>KEYNOTE: TUULI LAPPALAINEN</strong>&lt;br&gt;Functional Variation in the Human Genome: Lessons from the Transcriptome</td>
<td>Independence A/B</td>
</tr>
<tr>
<td>1:35 pm</td>
<td><strong>SB06</strong>: Large-scale models of signal propagation derived from phosphoproteomic data to study kinase inhibitors&lt;br&gt;• Julio Saez-Rodriguez</td>
<td>Independence A/B</td>
</tr>
<tr>
<td>1:50 pm</td>
<td><strong>SB07</strong>: Global chromatin defects and spurious transcription defines a novel subset of human cancers&lt;br&gt;• Kakajan Komurov</td>
<td>Independence A/B</td>
</tr>
<tr>
<td>2:05 pm</td>
<td><strong>SB08</strong>: Network Maximal Correlation to Infer Nonlinear Gene Modules in Cancer&lt;br&gt;• Soheil Feizi</td>
<td>Independence A/B</td>
</tr>
<tr>
<td>2:20 pm</td>
<td><strong>SB09</strong>: Integrating single cell transcriptional signatures and cell-cell interaction network models to study multiscale control of liver regeneration dynamics&lt;br&gt;• Daniel Cook</td>
<td>Independence A/B</td>
</tr>
<tr>
<td>2:20 pm</td>
<td><strong>RG03</strong>: Distinct specificities of the androgen and glucocorticoid receptors revealed using feature-based recognition model analysis of SELEX data&lt;br&gt;• Harmen J. Bussemaker</td>
<td>Independence A/B</td>
</tr>
<tr>
<td>2:20 pm</td>
<td><strong>RG04</strong>: How do closely related transcription factors recognize distinct genomic targets?&lt;br&gt;• Raluca Gordan</td>
<td>Independence A/B</td>
</tr>
<tr>
<td>2:20 pm</td>
<td><strong>RG05</strong>: Quantitative modeling of gene expression from sequence, using DNA shape-based model of binding sites&lt;br&gt;• Pei-Chen Peng</td>
<td>Independence A/B</td>
</tr>
<tr>
<td>2:20 pm</td>
<td><strong>RG06</strong>: Tracking the Evolution of 3D Gene Organization&lt;br&gt;• Alon Diament</td>
<td>Independence A/B</td>
</tr>
<tr>
<td>2:20 pm</td>
<td><strong>RG07</strong>: Three-dimensional analysis of regulatory features reveals functional enhancer-associated loops&lt;br&gt;• Yao Wang</td>
<td>Independence A/B</td>
</tr>
<tr>
<td>2:20 pm</td>
<td><strong>RG08</strong>: An integrated model for detecting significant chromatin interactions from high-resolution Hi-C data&lt;br&gt;• Mark Carty</td>
<td>Independence A/B</td>
</tr>
<tr>
<td>2:20 pm</td>
<td><strong>RG09</strong>: Visualizing three-dimensional organization and long-range interactions of the mammalian genome with the 3D Genome Browser&lt;br&gt;• Yanli Wang</td>
<td>Independence A/B</td>
</tr>
</tbody>
</table>
### DAY 3: Tuesday, November 17, 2015

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>2:35 pm</td>
<td><strong>COFFEE BREAK WITH POSTERS</strong></td>
<td>Liberty A/B</td>
</tr>
<tr>
<td>3:00 pm</td>
<td><strong>KEYNOTE: JAMES (JAY) BRADNER</strong> \ Targeting Cancer Core Regulatory Circuity</td>
<td></td>
</tr>
<tr>
<td>3:50 pm</td>
<td>SB10: Understanding Breast Cancer Heterogeneity through Personalized Drosophila Models • Jennifer EL Diaz</td>
<td>INDEPENDENCE C/D</td>
</tr>
<tr>
<td></td>
<td>RG10: Modeling methyl-sensitive transcription factor motifs with an expanded epigenetic alphabet • Coby Viner</td>
<td>INDEPENDENCE A/B</td>
</tr>
<tr>
<td>4:05 pm</td>
<td>SB11: A linear time-invariant model of phenotype dynamics in breast cancer cell populations • Margaret P. Chapman</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RG11: Genome-wide mapping of histone marks at single-nucleosome resolution • Marcelo Rivas-Astroza</td>
<td></td>
</tr>
<tr>
<td>4:20 pm</td>
<td>SB12: Precision drug rescue and drug repurposing using structural systems pharmacology • Lei Xie</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RG12: Creating a library of genome-wide chromatin state patterns during B lymphopoiesis • Mark Maienschein-Cline</td>
<td></td>
</tr>
<tr>
<td>4:35 pm</td>
<td>SB13: Bringing big genomic data into focus for studying complex diseases in specific biological contexts • Arjun Krishnan</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RG13: AEF: A methodology based on Assortativity of Epigenetic Features in promoter centered chromatin interaction networks identifies Polycomb and RNA Polymerase as main players • Vera Pancaldi</td>
<td></td>
</tr>
<tr>
<td>4:50 pm</td>
<td><strong>KEYNOTE: WILLIAM STAFFORD NOBLE</strong> \ Gene Regulation in 3D</td>
<td>Independence A/B</td>
</tr>
<tr>
<td>5:35 pm</td>
<td><strong>POSTER RECEPTION - SB AND RG</strong></td>
<td>Liberty A/B</td>
</tr>
</tbody>
</table>

### DAY 4: Wednesday, November 18, 2015

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>8 – 11 am</td>
<td><strong>REGISTRATION</strong></td>
<td>Independence lobby</td>
</tr>
<tr>
<td>8:45 am</td>
<td><strong>MORNING WELCOME &amp; ANNOUNCEMENTS</strong></td>
<td>Independence A/B</td>
</tr>
<tr>
<td>9:00 am</td>
<td><strong>KEYNOTE: MICHAEL SHEN</strong> \ Systems Analyses of Prostate Development and Cancer</td>
<td></td>
</tr>
<tr>
<td>9:50 am</td>
<td>SB14: Integrating transcriptomic and proteomic data with predictive regulatory network models of host response to pathogens • Deborah Chasman</td>
<td>INDEPENDENCE C/D</td>
</tr>
<tr>
<td></td>
<td>RG14: Nucleotide Sequence Composition Adjacent to Intronic Splice Sites Improves Splicing Efficiency and Reduces Translation Costs in Fungi • Zohar Zafrir</td>
<td>INDEPENDENCE A/B</td>
</tr>
<tr>
<td>10:05 am</td>
<td>SB15: Microbes are STICKY – Large-scale Inference and Topological Analysis of Microbial Interaction Networks • Christian L. Müller</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RG15: Lobular Scale Spatio-temporal Modeling of Calcium Signal Propagation in the Liver • Aalap Verma</td>
<td></td>
</tr>
<tr>
<td>10:20 am</td>
<td>SB16: Multi-omics learning and optimal experimental design for microbial organisms • Minseung Kim</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RG16: Dysregulated transcription factor networks and clusters in breast cancer subtypes, identified by coexpression and cistromic data integration • Qian Zhu</td>
<td></td>
</tr>
</tbody>
</table>
## DAY 4: Wednesday, November 18, 2015

<table>
<thead>
<tr>
<th>Time</th>
<th>Event Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>10:35 am</td>
<td>COFFEE BREAK WITH POSTERS</td>
</tr>
</tbody>
</table>
| 11:00 am | KEYNOTE: JULIA ZEITLINGER  
Combinatorial Regulation of Enhancers during Drosophila Development |
| 11:50 am | RG17: Transcription and differential DNA repair underlies promoter mutation hotspots in cancer genomes • Dilmi Perera  
RG18: The regulation of distal enhancers and silencers from the Cebpα locus during hematopoiesis, inferred using a transcriptional model • Manu |
| 12:05 pm | RG19: Comparison of Methods to Predict Impact of Regulatory Variants • Michael Beer  
RG20: Super-enhancers delineate disease-associated regulatory nodes in T cells • Golnaz Vahedi |
| 12:20 pm | RG21: Evaluating Genetic Variation Impact on Transcription Factor Binding Sites • Wengiang Shi  
RG22: Identifying condition specific transcription factor binding with ATAC-seq • Roger Pique-Regi |
| 12:35 pm | LUNCH WITH POSTERS (EVEN)                                                        |
| 1:30 pm  | KEYNOTE: ANDRE LEVCHENKO  
How Noisy is Noise? Analysis and Interpretation of Variability in Signaling Networks |
| 2:20 pm  | SB17: tRNA-derived fragments in Drosophila and their potential targets  
• Andrey Grigoriev  
RG23: ReMiCs: Regularized Regression Model for Inference of Condition-specific Gene Regulatory Networks from Multiple Data Sets • Nooshin Omranian |
| 2:35 pm  | SB18: High-throughput allele-specific expression across 250 environmental conditions • Gregory Moyerbraileann  
RG24: Transcriptional regulatory network inference for rare immune cell populations from gene expression and chromatin accessibility measurements • Emily Miraldi |
| 2:50 pm  | SB19: DREISS: dynamics of gene expression driven by external and internal regulatory networks based on state space model • Daifeng Wang  
RG25: An Experimentally Supported Model of the Bacillus subtilis Global Transcriptional Regulatory Network • Mario L. Arrieta |
| 3:05 pm  | SB20: SplashRNA: accurate prediction of potent shRNAs with a sequential classification strategy • Lauren Fairchild  
RG26: Network model of normal gene expression predicts gene perturbation fold changes • Sudhir Varma |
| 3:20 pm  | KEYNOTE: OLGA TROYANSKAYA  
From Piles of Data to Understanding Human Disease |
| 4:05 pm  | CONFERENCE CLOSING COMMENTS                                                       |

WiFiAccess
Network: Sheraton Meeting  Password: ISCB2015
Keynote Speakers

Monday, November 16, 2015

9:00 AM – 9:45 AM

Leslie Vosshall
Robin Chemers Neustein Professor, Laboratory of Neurogenetics and Behavior, The Rockefeller University, New York, United States

The Known and Unknown of Human Smell

ABSTRACT: Of all of the human senses, smell is the least understood. We do not know how many possible smells exist, and there is no reliable means to predict what a new molecule will smell like. Humans have 400 genes encoding odorant receptors, which are tasked with detecting all odorants and transmitting this information to the brain. How the brain interprets information from these receptors to yield the perception of a smell is one of the most challenging problems in neuroscience.

BioGraphy:
Leslie B Vosshall is the Robin Chemers Neustein Professor and Head of the Laboratory of Neurogenetics and Behavior at the Rockefeller University, and an investigator of the Howard Hughes Medical Institute. Vosshall received her A.B. in Biochemistry from Columbia University and a Ph.D. from Rockefeller University. Her molecular neurobiology research program aims to understand the genetic basis of behavior, with particular emphasis on how organisms perceive and respond to external sensory stimuli and how these responses are modulated by internal physiological state. She is an elected fellow of the American Association for the Advancement of Science, and was elected to the National Academy of Sciences in 2015.

2:15 PM – 3:00 PM

Steve Quake
Professor of Bioengineering, Stanford University, Investigator, Howard Hughes Medical Institute, Stanford, United States

Single Cell Genomics

ABSTRACT: An exciting emerging area revolves around the use of microfluidic tools for single-cell genomic analysis. We have been using microfluidic devices for both gene expression analysis and for genome sequencing from single cells. In the case of gene expression analysis, it has become routine to analyze hundreds of genes per cell on hundreds to thousands of single cells per experiment. This has led to many new insights
into the heterogeneity of cell populations in human tissues, especially in the areas of cancer and stem cell biology. These devices make it possible to perform “reverse tissue engineering” by dissecting complex tissues into their component cell populations, and they are also used to analyze rare cells such as circulating tumor cells or minor populations within a tissue.

We have also used single-cell genome sequencing to analyze the genetic properties of microbes that cannot be grown in culture—the largest component of biological diversity on the planet—as well as to study the recombination potential of humans by characterizing the diversity of novel genomes found in the sperm of an individual. We expect that single cell genome sequencing will become a valuable tool in understanding genetic diversity in many different contexts.

Stephen Quake studied physics (BS 1991) and mathematics (MS 1991) at Stanford University, after which he earned a doctorate in theoretical physics from Oxford University (1994) as a Marshall Scholar. He then returned to Stanford University, where he spent two years as a postdoc in Steven Chu’s group.

Quake joined the faculty of the California Institute of Technology in 1996, where he rose through the ranks and was ultimately appointed the Thomas and Doris Everhart Professor of Applied Physics and Physics. At Caltech, Quake received “Career” and “First” awards from the National Science Foundation and National Institutes of Health and was named a Packard Fellow. These awards supported a research program that began with single molecule biophysics and soon expanded to include the inventions of single molecule sequencing and microfluidic large scale integration, and their applications to biology and human health. He moved back to Stanford University in 2005 to help launch a new department in Bioengineering, where he is now the Lee Otterson Professor and an investigator of the Howard Hughes Medical Institute.

Quake’s contributions to the development of new biotechnology at the interface between physics and biology have been widely recognized. Honors include the Human Frontiers of Science Nakasone Prize, the MIT-Lemelson Prize, the Raymond and Beverly Sackler International Prize in Biophysics, the American Society for Microbiology Promega Biotechnology Research Award, the Royal Society of Chemistry Publishing Pioneer of Miniaturization Award, and the NIH Director’s Pioneer Award. He is an elected fellow of the American Academy of Arts and Sciences, the National Academy of Inventors, the National Academy of Sciences, the National Academy of Engineering, the Institute of Medicine, the American Institute for Medical and Biological Engineering and of the American Physical Society.
Tuesday, November 17, 2015

9:00 AM – 9:45 AM

Bonnie Berger

Professor of Applied Math and Computer Science,
Head of the Computation and Biology Group, Computer Science and AI Lab,
Massachusetts Institute of Technology, Cambridge, United States

Scaling with Compressive Algorithms

ABSTRACT: The last two decades have seen an exponential increase in genomic and biomedical data, which will soon outstrip advances in computing power. Extracting new science from these massive datasets will require not only faster computers; it will require algorithms that scale sublinearly in the size of the datasets. We introduce a novel class of algorithms that are able to scale with the entropy and low fractal dimension of the dataset by taking advantage of the unique structure of massive biological data to operate directly on compressed data. These algorithms can be used to address large-scale challenges in genomics, metagenomics and chemogenomics.

Bonnie Berger is a Professor of Applied Mathematics and Computer Science at MIT, and head of the Computation and Biology group at MIT’s Computer Science and AI Lab. After beginning her career working in algorithms at MIT, she was one of the pioneer researchers in the area of computational molecular biology and, together with the many students she has mentored, has been instrumental in defining the field. Professor Berger has won numerous awards including a National Science Foundation Career Award and the Biophysical Society’s Dayhoff Award for research. In 1999 Professor Berger was named one of Technology Review Magazine’s inaugural TR100 as a top young innovator of the twenty-first century, in 2003, was elected as a Fellow of the Association for Computing Machinery, and in 2010, received the RECOMB Test of Time Award. She was recently elected to the American Academy of Arts and Sciences, received the Margaret Pittman Director’s Award at the NIH, and was elected as a Fellow of the International Society for Computational Biology (ISCB). She currently serves as Vice President of the ISCB, Head of the steering committee for RECOMB, and on the NIGMS Advisory Council. In addition, Professor Berger is an Associate Member of the Broad Institute, Faculty member of HST, and Affiliated Faculty of Harvard Medical School.
Tuuli Lappalainen
Junior Investigator | Assistant Professor, Department of Systems Biology
New York Genome Center, Columbia University, New York, United States

Functional Variation in the Human Genome: Lessons from the Transcriptome

ABSTRACT: Detailed characterization of cellular effects of genetic variants is essential for understanding biological processes that underlie genetic associations to disease, as well as basic genome function. In this talk, I will discuss recent advances in integrated analysis of genome and transcriptome variation as a powerful approach to understanding functional genetic variants and genome function.

Tuuli Lappalainen is a Junior Investigator and Core Member at the New York Genome Center, and holds a joint position as Assistant Professor at the Department of Systems Biology at Columbia University. Prior to joining the New York Genome Center in January 2014, Tuuli did postdoctoral research at Stanford University in California and University of Geneva, Switzerland. She got her PhD from University of Helsinki, Finland in 2009. Her research focuses on functional genetic variation in human populations and its contribution to human traits and diseases. She has pioneered in integrating large-scale genome and transcriptome sequencing data to understand how genetic variation affects gene expression, which gives insight to biological mechanisms underlying genetic risk to disease. She has an important role in the most important international research consortia in human genetics, the 1000 Genomes Project, and Genotype Tissue Expression (GTEx) Project, and she led the RNA-sequencing work of the European Geuvadis Consortium.

James (Jay) Bradner
Investigator and Associate Professor, Harvard University, Dana Farber Cancer Institute, Boston, United States

Targeting Cancer Core Regulatory Circuitry

ABSTRACT: Post-translational modification of lysine by acetylation occurs at regions of the epigenome involved in active gene expression, both promoters (transcription start sites) and enhancers (cis-regulatory elements). Genome-wide measurement of asymmetry in the localization of acetyl-lysine allows insights into cell specification, cell state, cell function and, in cancer, unrecognized dependencies. Our laboratory has innovated computational and experimental approaches to understand the pathobiology of cancer and the molecular pharmacology of chromatin-active small molecules. Here, I will discuss our studies of advanced adult and pediatric cancers through the lens of integrated epigenomic analysis.
Jay Bradner is a Physician-Scientist in the Department of Medical Oncology at the Dana-Farber Cancer Institute and an Associate Professor in the Department of Medicine at Harvard Medical School. The research focus of the Bradner laboratory concerns the chemical modulation of chromatin structure and function. The clinical objective of the Bradner group is to deliver novel cancer therapeutics for human clinical investigation.

Dr. Bradner’s awards and honors include the Damon Runyon-Rachleff Innovation Award, the Smith Family Award for Excellence in Biomedical Research, the Dunkin Donuts Rising Star Award and the HMS Distinguished Excellence in Teaching Award. He was elected into the American Society of Clinical Investigation in 2011 and the Alpha Omega Alpha medical society in 2013. His recent research has been published in Nature, Cell, Nature Chemical Biology and the Journal of the American Chemical Society. He has authored more than thirty United States Patent applications and is a scientific founder of Acetylon Pharmaceuticals, SHAPE Pharmaceuticals, Tensha Therapeutics and Syros Pharmaceuticals. Three first-in-class molecules arising from his research are presently studied in open Phase I and Phase II clinical trials. Dr. Bradner also serves on the Board of Directors for the Leukemia & Lymphoma Society and the American Society of Hematology.

Dr. Bradner received his AB from Harvard University, his MD from the University of Chicago, and an MMS from Harvard Medical School. He completed his postgraduate training in Internal Medicine at Brigham & Women’s Hospital, followed by fellowships in Medical Oncology and Hematology at Dana-Farber Cancer Institute. Following additional post-doctoral training in Chemistry with Prof. Stuart Schreiber at Harvard University and the Broad Institute, where he now serves as the Associate Director for the Science of Therapeutics. Dr. Bradner joined the research faculty of Dana-Farber in 2008.

4:45 PM – 5:30 PM

William Stafford Noble
Professor, Department of Genome Sciences, Department of Computer Science and Engineering, University of Washington, Seattle, United States

Gene Regulation in 3D

ABSTRACT: The three-dimensional conformation of DNA in the nucleus impacts functional processes such as gene regulation and DNA replication. High-throughput sequencing assays like Hi-C provide detailed snapshots of DNA 3D configuration. In this talk, I will discuss several computational methods we have developed to make sense of the resulting DNA-DNA contact maps, both by projecting the maps into 3D and by relating the maps to data derived from traditional, 1D genomics assays such as RNA-seq or histone modification ChIP-seq.
William Stafford Noble (formerly William Noble Grundy) was raised in Naperville, IL, and graduated from Stanford University in 1991 with a degree in Symbolic Systems. Between undergraduate and graduate school, he worked in the speech group at SRI International in Menlo Park, CA, and at Entropic Research Laboratory in Palo Alto, CA. He also spent two years teaching high school math, physics and English literature with the US Peace Corps in Lesotho, Africa. In 1994, he entered graduate school at the University of California, San Diego, where he studied with Charles Elkan. He received the Ph.D. in computer science and cognitive science in 1998. He then spent one year as a Sloan/DOE Postdoctoral Fellow with David Haussler at the University of California, Santa Cruz. From 1999 until 2002, Noble was an Assistant Professor in the Department of Computer Science at Columbia University, with a joint appointment at the Columbia Genome Center. In 2002, he joined the faculty of the Department of Genome Sciences at the University of Washington, where he has adjunct appointments in the Department of Computer Science and Engineering and in the Department of Medicine. His research group develops and applies statistical and machine learning techniques for modeling and understanding biological processes at the molecular level. Noble is the recipient of an NSF CAREER award and is a Sloan Research Fellow.
Michael Shen

Professor, Departments of Medicine, Genetics & Development, Urology, and Systems Biology, Columbia University, New York, United States

Systems Analyses of Prostate Development and Cancer

ABSTRACT: Experimental analyses of genetically-engineered mouse models have been widely pursued for the study of human cancer, yet face significant challenges due to intrinsic differences between rodent and human biology. I will describe how computational systems approaches can address these challenges through the generation and analysis of regulatory networks for mouse and human prostate cancer to identify and validate conserved drivers of prostate cancer malignancy. In addition, recent studies have extended this experimental paradigm to identify drugs and drug combinations with potential therapeutic efficacy, as well as drivers of cellular reprogramming to recapitulate prostate tissue specification.

BIOGRAPHY: Michael M. Shen, Ph.D. is a Professor in the Departments of Medicine, Genetics & Development, Urology, and Systems Biology, and a member of the Herbert Irving Comprehensive Cancer Center at Columbia University Medical Center. He received his B.A. from Harvard University in 1984, and received his Ph.D. from Cambridge University in 1988. Following his post-doctoral training with Dr. Philip Leder at Harvard Medical School, he started his independent laboratory at UMDNJ – Robert Wood Johnson Medical School in 1994, and moved to Columbia University Medical Center in 2007. During the past twenty years, Dr. Shen has investigated the molecular mechanisms of mammalian development and cancer using in vivo analyses of genetically-engineered mouse models. Ongoing work in his laboratory includes systems analyses of embryonic stem cell pluripotency and prostate cancer progression, investigation of mechanisms of prostate epithelial lineage specification and cell-type differentiation, and generation of patient-derived organoid models of bladder cancer.

Julia Zeitlinger

Associate Investigator, Stowers Institute of Medical Research, Kansas City, United States

Combinatorial Regulation of Enhancers during Drosophila Development

ABSTRACT: My goal is to understand how genome-wide transcription programs are established during embryonic development and how this information is encoded in the DNA sequence. Taking advantage of
the well-studied regulatory networks in Drosophila, we use genomics approaches to gain a deeper understanding of the mechanisms and sequence requirements by which enhancers are regulated in a stage-specific and tissue-specific fashion. In my talk, I will focus on the mechanisms by which early Drosophila enhancers are primed by the pioneer transcription factor Zelda and are activated or repressed by signaling pathways.

**Julia Zeitlinger** is an Associate Investigator at the Stowers Institute for Medical Research and Assistant Professor at the University of Kansas Medical Center. She obtained her Ph.D. at EMBL in Germany, where she was trained in Drosophila developmental genetics and transcription. During her postdoctoral work with Rick Young at Whitehead/MIT, she pioneered the use of genomics approaches to understand transcription during development, initially in yeast and then in Drosophila. After starting her own lab in 2007, she has been named Pew scholar and received the NIH New Innovator and Bill Neaves Award.

**Andre Levchenko**

*John C. Malone Professor of Biomedical Engineering, Director of Yale Systems Biology Institute, Yale University, New Haven, United States*

**How Noisy is Noise? Analysis and Interpretation of Variability in Signaling Networks**

**ABSTRACT:** Cell-cell variability in isogenic populations is a finite feature of all known single cell and multicellular organisms. Much of this variability is contained in the signaling networks mediating cell response to external stimuli. Can this variability be ‘mapped’ across signaling networks to evaluate the degree to which the information conveyed by them can be distorted? Can the effects of this variability on functional responses be assessed and interpreted? I will present a general methodology for the analysis and interpretation of noise or variability in the signaling networks, and discuss how this new analysis affects our understanding of the systems properties of biological regulatory networks.

**Andre Levchenko**, John C. Malone Professor of Biomedical Engineering, combines experimental results with computational models to learn about the interactions of proteins and cells in healthy and disease states.

Levchenko, who joined the Yale faculty on July 1, 2013 is engaged in multi-disciplinary research to define molecular and cellular interactions from single-cell to multi-cell levels. In addition to his endowed post, Levchenko was also appointed as the inaugural director of the Yale Systems Biology Institute on the West Campus.

After growing up in Siberia, Russia, Levchenko received a Master of Science degree in biophysics from the Moscow Institute of Physics and
Technology. After arriving in the United States as a refugee, he went on to earn a Master of Science and a Doctor of Science degree in bioengineering from Columbia University. He was then a postdoctoral scholar in the California Institute of Technology’s biology division.

Levchenko joined the faculty of the Department of Biomedical Engineering at Johns Hopkins University as an assistant professor in 2001, rising through the ranks to associate professor (with tenure) in 2007 and full professor in 2011. He held affiliations with the Whitaker Institute for Biomedical Engineering, the Institute for Cell Engineering, the Johns Hopkins Medical School Epigenetics Center, the John Hopkins Institute for NanoBio Technology, and the Center for Cell Dynamics.

The Yale systems biologist is the author of more than 90 published research articles and several book chapters. He serves on the editorial boards of PLOS Biology and Science Signaling, among numerous others. He is a member of the New York Academy of Science, the Biophysical Society, the American Society for Biochemistry and Molecular Biology, the American Chemical Society, and the Biomedical Engineering Society.

He is a recipient of the American Asthma Foundation Early Excellence Award and has been elected a fellow of the American Institute for Medical and Biological Engineering.

3:20 PM – 4:05 PM

**Olga Troyanskaya**

*Professor, Department of Computer Science, Lewis-Sigler Institute of Integrative Genomics, Princeton University, United States*

*Deputy Director for Genomics, Simons Center for Data Analysis, Simons Foundation, New York, United States*

**From Piles of Data to Understanding Human Disease**

**ABSTRACT:** An immense molecular complexity forms the foundation of human disease. This complexity must be interpreted and distilled through accurate modeling of molecular networks and pathways whose malfunction promotes the emergence of complex human diseases. Although cell-lineage-specific gene expression and function underlie the development, function, and maintenance of diverse cell types within an organism and are critical to understanding molecular basis of disease, high-throughput data are rarely resolved with respect to specific cell lineages. In this talk, I will focus on our recent work developing integrative approaches that leverage functional genomics data collections to study how cellular pathways function in diverse cell types, enabling molecular-level understanding of human disease. I will describe how integrated analysis of functional genomics data can be leveraged to study tissue-lineage-specific protein function and interactions and to identify genes involved in disease in a novel approach for re-prioritizing quantitative genetics studies results.
Olga Troyanskaya is a Professor in the Lewis-Sigler Institute for Integrative Genomics and the Department of Computer Science at Princeton University and Deputy Director for Genomics at the Simons Center for Data Analysis at the Simons Foundation. Dr. Troyanskaya is a member of the Board of Directors of the International Society for Computational Biology and Associate Editor for Bioinformatics and PLOS Computational Biology. She received her Ph.D. from Stanford University and is a recipient of the Sloan Research Fellowship, the NSF CAREER award, the Howard Wentz faculty award, and the Blavatnik Finalist Award. She has also been honored as one of the top young technology innovators by the MIT Technology Review and is the 2011 recipient of the Overton Prize from the International Society for Computational Biology and the 2014 Ira Herskowitz Award from the Genetic Society of America.
DREAM Challenges Oral Presentations

Sunday, November 15, 2015

DREAM Prostate Cancer Challenge Session

**Project Data Sphere® initiative overview**

**Liz Zhou**  
*Director, US Medical Affairs, Sanofi*

The Project Data Sphere® initiative (PDS) is an independent not-for-profit initiative of the Life Sciences Consortium of the CEO Roundtable on Cancer, with the vision to broadly share, integrate, and analyze historical comparator-arm cancer trial data sets from any source (academic, government, industry, etc.) with protocols, CRFs, and data descriptors, freely online at www.projectdatasphere.org. The goal is to accelerate innovation to improve cancer care.

Launched in April 2014, the PDS website provides registered users access to de-identified patient level raw data from the control arms of Phase 3 oncology clinical trials. At the time of the launch, there were 4,000 patient lives across 9 datasets from 7 industry and academic data providers. As of August 2015, the number of datasets had increased from 9 to 52, with nearly 30,000 patient lives across multiple tumor types.

Data from the PDS platform was used for the Prostate Cancer DREAM Challenge in the summer of 2015 – this challenge had the greatest level of registration of any DREAM Challenge to date. Laborious effort was put behind preparing data for the Challenge, which resulted in relatively smooth Challenge process for solvers working with clinical trial data.

PDS provides a model to share clinical trial data from oncology; the rapid growth of the number of users, downloads, and publications demonstrates a viable approach to accelerating research, including exploring innovative ways to analyze the data (e.g. crowdsourcing) without exposing patients to new clinical trials.

**Introduction to the Prostate Cancer DREAM Challenge**

**James C Costello**  
*Department of Pharmacology, University of Colorado Anschutz Medical Campus, Colorado, USA*

Prostate cancer is the most common cancer among men in developed countries and ranks third in terms of mortality after lung cancer and colorectal cancer. Nearly 15% of prostate cancer patients have metastatic disease (Stage IV) at the time of diagnosis. The mainstay of treatment for metastatic disease is androgen deprivation therapy (ADT), though
inevitably many patients develop resistance resulting in metastatic castrate-resistant prostate cancer (mCRPC). To gain a better understanding of mCRPC, the Prostate Cancer DREAM Challenge was developed to address two sub-challenges: 1) Predict overall survival for mCRPC patients based on clinical variables, and 2) Predict treatment discontinuation for mCRPC patients treated with docetaxel due to adverse events at early time points. The underlying data was collected from 4 separate clinical trials and annotated by Project Data Sphere, LLC. The prognostic calculators for sub-challenge 1 were scored using the integrated Area Under the Curve (iAUC) and sub-challenge 2 was scored using the Area Under the Precision Recall Curve (AUPRC). The Challenge had over 50 teams with 180 individual researchers actively participating and more than half of the teams outperformed the standard model in the field for sub-challenge 1. Sub-challenge 2 represents a novel set of results as the amount of data needed to address discontinuation due adverse events has not been compiled until the Prostate Cancer DREAM Challenge. The results of the Prostate Cancer DREAM Challenge, specifically the prognostic calculators developed by participating teams, will be made available to clinicians to aide in patient treatment decisions. The methods developed to identify patients likely to discontinue can be used for patient selection in future clinical trial design.

**Predicting patient survival in the DREAM 9.5 mCRPC challenge**

Teemu D. Laajala1,2, Suleiman Khan2, Antti Airola3, Tuomas Mirtti2,4, Tapio Pahikkala3, Peddinti Gopalacharyulu2, Tero Aittokallio1,2

1 Department of Mathematics and Statistics, University of Turku, Finland
2 Institute for Molecular Medicine Finland, University of Helsinki, Finland
3 Department of Information Technology, University of Turku, Finland
4 Department of Pathology, HUSLAB, Helsinki University Hospital, Finland

We present the top performing ensemble of models for predicting patient survival in the context of metastatic castration-resistant prostate cancer (mCRPC) patients, originating from several clinical trials (subchallenge 1a of the DREAM 9.5 Prostate Cancer Challenge). By coupling unsupervised learning with survival-analysis-based supervised learning, we constructed an ensemble of batch-wise optimized penalized regression *coxnet*-models. The final ensemble models were simultaneously optimized for the penalized regression through $L_1/L_2$-norm parameter $\alpha$ along with the penalization coefficient $\lambda$. Model-based imputation of missing values as well as incorporating clinical á priori knowledge of variables is discussed, along with practical lessons learned from processing such challenging clinical data that required wide multidisciplinary expertise. Lastly, we offer our personal view of the clinical novelty of the model coefficients and its ensemble structure including interactions among the clinical variables.
The aim of this study was to predict the adverse event of docetaxel at early stage in patients with metastatic castration-resistant prostate cancer (mCRPC). More specifically, our objective was to provide more precise insights for clinicians on deciding whether to continue or discontinue docetaxel within three months of starting the treatment using baseline clinical features. To address this question, boosting method, a class of supervised machine learning techniques was utilized.

The analysis was commenced with feature selection and data preprocessing. Preliminary predictors were selected after filtering out features with insufficient clinical relevance to the study question (suggested by our clinical members) and features with high rates of missing values and/or collinearity. For selected feature categories, including lesions, prior medications and diseases, we used an arithmetic sum of presence of features in the corresponding categories. For the selected laboratory values, they were transformed, scaled or truncated based on their reference ranges and distributions. Finally, these preprocessed features were used to develop the final predictive model.

In the model building step, we focused on the R package: “gbm” (Generalized Boosted Regression Modeling). The package utilizes gradient boosting approach which iteratively tests the importance of features and aims to combine several weak models into a powerful high performance ensemble prediction. Also, the gbm package contains capacity of handling missing values and requires lower run time. In addition to this DREAM challenge, they have been shown to be successful in various practical applications.

In conclusion, the model can assist clinicians to capture suitable candidates to continue/discontinue docetaxel treatment.
Predicting discontinuation due to adverse effect in mCRPC

Yuanfang Guan
Department of Computational Medicine and Bioinformatics; Department of Internal Medicine, Ann Arbor, United States

In the second sub-challenge of the DREAM 2015 prostate cancer challenge, participants were asked to predict which patients cannot tolerate docetaxel therapy, i.e. early adverse effect (AE) defined as termination within 91.5 days. Multi-task comparison showed that early AE is tightly connected to early death, while the patients that experienced early AE but not early death showed heterogeneous characteristics, preventing them from used as gold standard. Thus predicting adverse effect was transformed into the problem of predicting early death. I used 3 months as the cutoff, where deaths prior to 3 months were used as the gold standard positives versus the rest as negatives. This method turned out to the best performing method of this sub-challenge.
Monday, November 16, 2015

DREAM Olfaction Challenge Session

Introduction to the DREAM olfaction challenge

Andreas Keller
Laboratory of Neurogenetics and Behavior, The Rockefeller University, New York, United States

The complexity of olfactory stimuli as well as of the olfactory perceptual space makes the question of what determines a molecule’s smell an ideal topic for the collaborative approach provided by DREAM challenges. I will briefly discuss previous attempts to predict how a molecule smells based on its physical properties and then present the psychophysical data set that was collected at Rockefeller University as the basis for the DREAM olfaction challenge. I will point out how this dataset differs from data that has been traditionally used for this type of project. I will also discuss how the influence of genetic variability between individuals and of previous experiences on odor perception complicates the stimulus-percept-correlation in olfaction.

Predicting olfaction response for each individual

Yuanfang Guan
Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, United States

This abstract describes the method I wrote for the 2015 DREAM Olfaction Challenge- sub-challenge 1: building models to predict olfactory response for each individual. I used decision tree as a base-learner of this chemical structural data. There are two reasons that I chose decision tree: 1) the dimension of the structure data is high, which contained over 4000 parameters; decision tree helped to reduce the dimension. 2) The data matrix is sparse; a decision boundary can be put between zeros and the rest of the values. Olfactory responses reported by individuals were noisy. Thus I used the global response across all to balance the individual response, in order to 1) capture the personalized features, and b) stabilize the predictions. For example, a chemical reported to be ‘sweet’ by individual A is trusted more when other peoples also report this chemical to be ‘sweet’. Finally, I used 0.2*individual score + 0.8* global average for each chemical as predictions. But the above parameter can be rather flexible to achieve decent performance. Similar technique also resulted in one of the best performing algorithms in 2014 DREAM Broad Institute Gene Essentiality Challenge. No external data was used. This technique turns out to be the best-performing method for this sub-challenge.
From Shape to Smell: Predicting Olfactory Perceptual Descriptors using Molecular Structural Information

Richard C. Gerkin
School of Life Sciences, Arizona State University, United States

The DREAM Olfaction Prediction Challenge asked participants to predict 21 different olfactory perceptual features (i.e. smell descriptors) of single molecules using a library of several thousands structural features of those molecules. One sub-challenge asked participants to make predictions for individual human subjects, and one for the mean and variance of responses across subjects. Here I describe the winning submission to the latter sub-challenge.

After using missing value imputation to fill out the structural feature dataset, I trained Random Forest Regression models (using Python’s scikit-learn package) to predict mean (across subjects) responses for each of the perceptual features. I made extensive use of cross-validation to optimize these models. While I also constructed similar models to predict the variance (across subjects) of the responses, I found that prediction was improved by exploiting the relationship between the variance and the mean that was guaranteed by basic psychometric considerations. Consequently, I obtained decisive improvements in my prediction of the variance by pooling results from models trained only on the variance with a theoretically motivated non-linear transformation of results from models trained only on the mean. This technique proved decisive in constructing the winning submission for the sub-challenge.

Collaboration with other challenge participants further improved olfactory prediction by utilizing additional molecular features; this provided scientific insight into the categories and origins of features that provide useful olfactory information about molecules.

DREAM Olfaction Challenge: Lessons Learned

Amit Dhurandhar, Pablo Meyer, Guillermo Cecchi
IBM TJ Watson Research, New York, United States

We present the insights gained from the DREAM Olfaction challenge run earlier this year. Our sense of smell critically affects our emotions and hence decisions, making it an important part in human cognition. We report analysis of results from the challenge where only molecular structure was used to design learning algorithms that were trained to predict individual and average judgment of smell for 49 commoners (not experts) based on 21 descriptors. The results were promising in the sense that the best models achieved greater than 0.75 correlation, with linear models being competitive with the best. The challenge showed that in terms of predictability not only generalization across odors for specific individuals is possible, but that generalization across individuals for the same odors is also possible, which is highly encouraging from a science and application point of view.
DREAM ALS Stratification Prize4Life Challenge Session

Using the Power of Big Data and Crowdsourcing for Catalyzing Breakthroughs in Amyotrophic Lateral Sclerosis (ALS)

Neta Zach¹, Robert Kueffner², Nazem Atassi³, Venkat Balagurusamy⁴, Barbara di Camillo⁵, Merit Cudkowicz³, Donna Dillenberger⁴, Javier Garcia-Garcia⁶, Orla Hardiman⁷, Bruce Hoff⁸, Joshua Knight⁴, Melanie Leitner⁹, Guang Li¹⁰, Lara Mangravite⁸, Raquel Norel¹, Thea Norman⁸, Liuxia Wang¹⁰, Gustavo Stolovitzky⁴

² Prize4Life, Israel; 2 Ludwig-Maximilian-University, Germany; 3 Massachusetts General Hospital, United States; 4 IBM Research, New York, United States; 5 University of Padova, Italy; 6 Pompeu Fabra University, Spain; 7 Trinity College Institute of Neuroscience, Ireland; 8 Sage Bionetworks; 9 Biogen Idec, Massachusetts, United States; 10 Origent, Virginia, United States

Amyotrophic lateral sclerosis (ALS) is a fatal neurodegenerative disease with significant heterogeneity in its progression. In order to address this heterogeneity and spur ALS research, clinical care and drug development we need sufficient clinical data and suitable analysis approaches. To address this heterogeneity for the first time we launched the DREAM ALS Stratification Prize4Life Challenge in summer 2015, using clinical Data from the PRO-ACT database of ALS clinical trials, as well as data from National ALS registries from Italy and Ireland.

In the challenge, we asked participants to derive meaningful subgroups of ALS patients relative to disease progression and survival. The challenge drew in 75+ submissions from 31 teams. We will discuss the different approaches used by participants, as well as the baseline algorithms, and how patient classification affected performance in predicting disease outcomes. We will also discuss the different predictive features and patient subgroups that the challenge helped unveil.

DREAM ALS STRATIFICATION CHALLENGE: BEST PERFORMER PRESENTATION

A Boosting Approach to Predicting ALSFRS Slope for the PRO-ACT Database

Wen-Chieh Fang¹, Chen Yang¹, Huan-Jui Chang², Hsih-Te Yang¹,³, Jung-Hsien Chiang¹,³

¹ Department of Computer Science and Information Engineering, National Cheng Kung University, Taiwan; 2 Department of Economics, National Cheng Kung University, Taiwan; 3 Institute of medical informatics, National Cheng Kung University, Taiwan

Amyotrophic lateral sclerosis (ALS) is a progressive neurological disease that leads muscle weakness and gradually impacts on the functioning of the body, leading to eventual death. It greatly reduces an individual’s life expectancy. Currently, experts do not know precisely
what causes ALS. There is no known cure for ALS. The DREAM ALS Stratification Prize4Life Challenge is held for the purpose of enabling better understanding of patient profiles and application of personalized ALS treatments.

In our approach to DREAM ALS Challenge, we first neglect those features with high percent of missing values. For the remaining, we replace any missing value with the mean of that feature for all other cases. Meanwhile, we apply equal frequency binning that divides the response variable into three groups such that each group contains approximately same number of values. There are two kinds of features in the data set: static features and ‘time-resolved’ features (those with different values when time varies). For the latter, we try two designated measurements, the minimum and the maximum as additional features. Then for both two kinds of features, we apply feature selection based on information gain to select the top-six features. In order to select optimal features, we run cross validation on the feature candidates. In prediction, we apply Gradient Boosted Regression Trees (GBRT) to predict the ALSFRS slopes. GBRT computes a sequence of simple decision trees, where each successive tree is built for the prediction residuals of the preceding tree.

In this challenge, we think that the feature selection is one of the most important steps and we believe that the most appropriate features dominate the performance of the model. In the final submission round, our team attained the best performance, outperforming the methods of all other teams.

**DREAM ALS STRATIFICATION CHALLENGE: BEST PERFORMER PRESENTATION**

**Predicting ALS survival through complete ranking of censored data**

Yuanfang Guan  
*Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, United States*

Prize4life ALS sub-challenge 2 and 4 asked participants to predict survival for two separate cohorts, both as typical censored data problems. In this talk, I will review existing methods (cox-related, survival-random-forest, etc.) and their potential limitation as an incomplete ranking of training patients, and the resulting limitation in the choices of base-learners. Then, I will describe the method I wrote for this challenge, which provides a probabilistic comparison between two censored data points derived from the K-M curves (in addition to non-censored data point pairs, and censored-non-censored pairs). A complete ranking of all patients is thus given, which allows incorporation of any base-learners. This generalizes a censored data prediction problem to a standard regression problem. Finally, I will talk about one (GPR) of the many base-learners that are capable of achieving similar performance under the above modeling framework. This method turns out to be the best-performing one for both sub-challenges.
Amyotrophic lateral sclerosis (ALS) is typically a rapidly progressing neurodegenerative disease. In many cases it leads to death within 3-5 years from onset of symptoms, but the rate of progression across the patient population can vary by an order of magnitude. Unwinding such underlying heterogeneity can hopefully shed light on disease mechanisms and drug development, and reliable prediction on progression rate can assist clinical decisions.

We developed a novel random forest based method to select 6 important clinical variables from 68 and predict ALS progression based on those 6. After training a random forest $F$ with all $n$ available features (whose missing rate < 50%), the patients used for training were dropped down the forest and their paths across each tree were tracked. Along each path, nodes were assigned different weights based on their positions along the path. Then node weights from all paths were aggregated so that each patient was represented by a point in an $n$-dimensional space $\mathbb{R}^n$, where the coordinates are the aggregated weights of the $n$ clinical features. All patients for training were then clustered in $\mathbb{R}^n$, and within the $i^{th}$ cluster ($i$ went from 1 to the number of clusters) the 6 clinical variables with the highest aggregated weights were used to train a new random forest $F_i$. When a new patient came in, based on the aggregated node weights along his/her paths across $F$, an $F_i$ and the 6 corresponding clinical variables were selected for predicting his/her ALS progression rate.

Our method was the top performer in sub-challenge 3, which was to predict ALS progression of patients from two national registries. Several other methods were locally tested, and our method turned out the best. For example, we tried representing each patient with a point in $\mathbb{R}^n$ whose coordinates were the values of the $n$ clinical features instead of their aggregated weights along paths across random forest, calculated the $z$-score (aggregated from concordance index, Pearson correlation and root-mean-square deviation) and found that the $z$-score of our submitted method was 41% higher. We also tried ranking the importance of features using the permutation setting of the `importance` function of R package `randomForest`, and the $z$-score of our submitted method was 30% higher. Inspired by those preliminary results, we are currently further developing and testing our aggregated weights method.
The ICGC-TCGA DREAM Somatic Mutation Calling Challenge: Combining accurate tumour genome simulation with crowd-sourcing to benchmark somatic variant detection

Joshua M. Stuart\textsuperscript{2,11}, Anna Y. Lee\textsuperscript{1,10}, Kathleen E. Houlahan\textsuperscript{1,10}, Adam D. Ewing\textsuperscript{2,3,10}, Kyle Ellrott\textsuperscript{2,10}, Yin Hu\textsuperscript{4}, J. Christopher Bare\textsuperscript{4}, Shadielle Espiritu\textsuperscript{1}, Vincent Huang\textsuperscript{1}, Kristen Dang\textsuperscript{4}, Cristian Caloi\textsuperscript{1}, Takafumi N. Yamaguchi\textsuperscript{1}, ICGC-TCGA DREAM Somatic Mutation Calling Challenge Participants: Michael R. Kellen\textsuperscript{4}, Thea C. Norman\textsuperscript{4}, Stephen H. Friend\textsuperscript{4}, Justin Guinney\textsuperscript{4}, Gustavo Stolovitzky\textsuperscript{5}, David Haussler\textsuperscript{2}, Adam A. Margolin\textsuperscript{4,7,11}, Paul C. Boutros\textsuperscript{1,8,9,11}

1 Informatics and Biocomputing Program; Ontario Institute for Cancer Research; Toronto, Canada; 2 Department of Biomolecular Engineering; University of California, Santa Cruz, United States; 3 Mater Research Institute; University of Queensland; Woolloongabba, Australia; 4 Sage Bionetworks; Seattle, United States; 5 IBM Computational Biology Center; T.J. Watson Research Center; Yorktown Heights, New York, United States; 6 Computational Biology Program; Oregon Health & Science University; Portland, United States; 7 Department of Biomedical Engineering; Oregon Health & Science University; Portland, United States; 8 Department of Medical Biophysics; University of Toronto, Canada; 9 Department of Pharmacology & Toxicology; University of Toronto, Canada; 10 These authors contributed equally; 11 Corresponding authors

The identification of somatic mutations in cancer genomes via next-generation sequencing will transform our understanding and treatment of cancer. Unfortunately, accurate identification of somatic mutations of all types – point-mutations and structural variants – remains challenging, with many anecdotal, small-scale reports of discordance across methods. One underlying reason for this problem is the lack of robust, impartial benchmarking studies and widely-accepted gold-standards. The Cancer Genome Atlas (TCGA) and the International Cancer Genomics Consortium (ICGC) launched the ICGC-TCGA DREAM Somatic Mutation Calling Challenge: a crowd-sourcing effort to identify the best pipelines for detecting mutations in the high-throughput sequencing reads of cancer genomes (https://www.synapse.org/#!Synapse:syn312572).

To benchmark variant calling approaches, a novel simulator called BAMSurgeon was developed to synthesize cancer genomes \textit{in silico}. The results of 248 single nucleotide variant and 204 structural variant analyses run on five synthetic tumors will be presented. Different algorithms exhibit characteristic error profiles, and, intriguingly, false positives show a trinucleotide mutation signature often reported in human tumors. Although the three simulated tumors differ in sequence contamination (deviation from normal cell sequence) and in sub-clonality, an ensemble of pipelines outperforms the best individual pipeline in all cases. We will discuss several findings from the analysis of these methods including the ability of methods to improve performance without overfitting, that SNV but not SV callers benefit from a “wisdom of the crowds” ensemble, and the first clear picture of the origins of methodological errors in SNV and SV calling. The leaderboard for this Challenge remains open and is continually attracting new entries, serving as a living-benchmark for comparing new algorithms.
DREAM SMC CHALLENGE: BEST PERFORMER PRESENTATION (TBA)

DREAM SMC CHALLENGE: BEST PERFORMER PRESENTATION

novoBreak: a k-mer targeted assembly algorithm for breakpoint detection in cancer genomes

Zechen Chong and Ken Chen
Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, United States

Somatic structural variations (SVs) are major driving forces for tumor development and progression. Sporadic and recurrent chromosomal aberrations have been observed in most cancer types, including breast, lung, brain, leukemia, pancreatic and prostate cancers. The advent of high-throughput next generation sequencing (NGS) technologies has made it possible to perform genome-wide detection of SVs at base pair resolution. However, current sequencing-based computational methods are limited in sensitivity and comprehensiveness due to the challenges of acquiring sufficient information to characterize different types of SVs. Here, we present novoBreak, a novel k-mer targeted local assembly algorithm that discovers somatic and germline structural variation breakpoints in whole genome sequencing data. NovoBreak can directly identify breakpoints from clusters of reads that share a set of k-mers uniquely present in a subject genome (e.g., a tumor genome) but not in the human reference genome or any control data (e.g., a matched normal genome). In synthetic data from the ICGC-TCGA DREAM 8.5 Somatic Mutation Calling Challenge and real data from a cancer cell-line, novoBreak consistently outperformed existing algorithms due mainly to more effective utilization of reads spanning breakpoints. NovoBreak also demonstrated great sensitivity in identifying short INDELs and gene fusions. The wider application of novoBreak is expected to reveal comprehensive structural landscape that can be linked to novel mechanistic signatures in cancer genomes.

Lessons from the SMC-DNA IS Challenges & Looking Forward

Paul C. Boutros1,2, Adam A. Margolin3, Kyle Ellrott3, Quaid D. Morris2, Paul Spellman3, David Wedge4, Peter Van Loo5, Gustavo Stolovitzky6, Joshua D. Stuart7
1 Ontario Institute for Cancer Research, Toronto, Canada; 2 University of Toronto, Canada; 3 Oregon Health & Sciences University, Portland, United States; 4 Wellcome Trust Sanger Institute, Hinxton, United Kingdom; 5 Crick Research Institute, London, United Kingdom; 6 IBM Research, New York, United States; 7 University of California, Santa Cruz, United States

We present a summary of the key lessons learnt through the in silico challenges of the ICGC-TCGA DREAM Somatic Mutation Calling Challenge (SMC-DNA), with a particular focus on the challenges in simulating, scoring and integrating structural variant (SV) predictions.
We discuss the remarkable failure of ensemble models to improve upon SV prediction, and note the significant differences to other types of genomic data. Finally we outline the future of the SMC series of DREAM Challenges, giving the official launch of a tumour subclonality reconstruction challenge (SMC-Het) and talking about the progress towards an RNA-Seq Challenge (SMC-RNA).

**DREAM Drug Combination Challenge Session**

**Crowdsourcing combinatorial therapies: The AZ-Sanger DREAM synergy prediction challenge**

Menden MP\(^1,1\), Wang D\(^2,2\), Chaibub Neto E\(^3\), Ghazoui Z\(^2\), Jang IS\(^3\), Giovanni Di Veroli\(^4\), Gustavo Stolovitzky\(^5\), Dry JR\(^2,\#\), Guinney J\(^3,\#\), Saez-Rodriguez J\(^1,6,\#\)

1 European Molecular Biology Laboratory – European Bioinformatics Institute; 2 Oncology Innovative Medicines, AstraZeneca; 3 Sage Bionetworks; 4 Early Clinical Development – Innovative Medicine, AstraZeneca; 5 IBM Research; 6 RWTH Aachen University Hospital, Joint Research Center for Computational Biomedicine (JRC-Combine); * First authors; # Corresponding authors (alphabetically ordered)

In the last 20 years targeted therapies and personalized treatments have been the most promising assets to treat cancer. However, the success is often lessened by secondary resistance. As a solution to increase the therapeutic search space and to overcome resistance, combinations of drugs are currently extensively investigated. A major current limitation is the lack of effective strategies to search the virtually intractable combinatorial space. To tackle this issue and advance the development of combinatorial therapies DREAM, Sage Bionetworks and AstraZeneca are jointly hosting a drug combination challenge, which is open to the scientific community to participate.

The challenge aims to address two specific questions: (i) to predict synergies in cancer cell lines from molecular data, and (ii) to identify biomarkers that discriminate between synergistic and non-synergistic behaviors. Towards this goal, AstraZeneca provides a combinational cell line screening, which comprehends a total of \(\sim 11.5k\) experimental tested drug combinations, as well as all mono therapies as baselines. The previous DREAM drug combination challenge focused on \(\sim 100\) experimental tested combinations and explored synergy signatures based on before and after treatment gene expression. Besides the smaller scale, those signatures are not practical/ethical to retrieve from patients. In addition our proposed challenge focus on revealing mechanistic insights from any biomarker selection and model build. By providing a framework where any group in the world can participate, and whereby methods can be evaluated in an unbiased and double-blinded way, we aim to move forward in this complex problem of discovering novel and effective drug combinations.

Here we present results from the first challenge round and will show glimpses of knowledge gained in this crowd-sourcing drug combination challenge.
Reducing data-leakage in leaderboard evaluations: the Ladder and LadderBoot algorithms

Elias Chaibub Neto, Sage Bionetworks

Over-fitting is a common issue in machine learning challenges. Because participants rely on the public leaderboard to evaluate and refine their models, there is always the danger they might start to over-fit their models to the holdout data supporting the leaderboard. Standard remedies to this problem include limiting the number of allowed submissions per participant and rounding the released public scores. Recently, Hardt and Blum (2015) proposed the Ladder algorithm, which reduces over-fitting by preventing the participant from exploiting minor fluctuations in public leaderboard scores during their model refinement activities. Mechanistically, the Ladder only releases the actual (rounded) score of a new submission if the score presents a statistically significant improvement over the previously best submission of the participant. If not, the Ladder releases the score of the best submission so far.

In this talk, we present evaluations of the Ladder algorithm under two adversarial attacks. Both are inspired by Freedman’s paradox, where the selection of the features entering a multiple regression model is guided by the public leaderboard. In the first attack, we simply select the top features according to the public leaderboard scores of the univariate regression models. Our experiments show the effectiveness of the Ladder algorithm in this context. Our second attack, on the other hand, is based on a more aggressive step-forward variation of this first attack, and can lead to severe over-fitting. This attack explores the fact that the Ladder leaks too much information about the holdout data when it releases the public leaderboard score of the best model so far. To circumvent this problem, we propose a variation of the Ladder mechanism, called LadderBoot algorithm, which releases a bootstrapped estimate of the public leaderboard score, instead of the actual rounded score. In our experiments, the LadderBoot mechanism tended to compare favorably to the Ladder.


Other

The DREAM Challenges channel: Open science publishing for all participating DREAMers

Michael Markie, F1000Research, London, United Kingdom

F1000Research is an Open Science publishing platform that offers the immediate publication of posters, slides and articles with no editorial bias. All published articles benefit from a collaborative, transparent peer review process and the inclusion of all source data and code. F1000Research partners with the International Society for Computational Biology through the publication of the ISCB Community Journal, and have recently built the DREAM Challenges channel, a central venue to publish peer-reviewed
method articles from participants of DREAM challenges. Developments in science most often build upon previous findings, insights, and data, so it is important to make this information accessible to enable easy reuse. Both DREAM challenges and F1000Research have core values that are underpinned by open, reproducible science and this collaboration aspires to advance the questions posed in DREAM Challenges through the open sharing of data-analyses and computational methods. Participants of the recently concluded Prostate Cancer challenge will be the first DREAMers to make use of the channel. Participants will take advantage of a dynamic, post-publication peer review model and have the opportunity to work with expert reviewers from the biomedical research community to help improve and refine their methods after the challenge has completed. In addition to encouraging participants to publish their findings, the DREAM Challenges channel is also open to the wider DREAM community to publish research around challenge topics and the theory of how the challenges work.

The aim of the collaboration is to make further progress beyond the challenges, help answer important biomedical and biological questions and pave the way to improving clinical practice, such as the prognostic calculators developed for the recent Prostate Cancer DREAM Challenge.

**DREAM Hackathon**

**The DREAM Challenges Hackathon: mining big data from a Parkinson’s Disease mobile research study**

DREAM Challenges Hackathon Team  
* Sage Bionetworks, University of Rochester, DREAM Challenges*

The DREAM Challenges will be sponsoring a Hackathon on Sunday and Monday evenings, based on data from the mPower study. mPower, is an App-based clinical study, focused on Parkinson Disease (PD), developed by Sage Bionetworks and the University of Rochester. As of August 2015, mPower enrolled 18,000 participants — the largest PD study ever. Typically symptoms in PD patients are evaluated and recorded twice a year, when the patient goes to the doctor. Between these doctor’s visits, a patient’s disease status is left unmonitored, and important decisions about interventions are often not made in a timely manner. mPower is built to allow patients to continuously track their PD signs and symptoms, generating data on the patient’s voice (pitch and tremor), balance, walking and finger speed and dexterity. The data being collected from this study will be made available for this year’s DREAM Conference Hackathon’s participants.

The goal of this hackathon is to use the mPower data to generate insights about PD. The data is collected using Sage Bionetworks Bridge Server, and Sage scientists have already proved that valuable information can be extracted from the mPower study. However there is a lot more work to be done. In this hackathon participants will bring their expertise and creative energies to develop ideas, models or proposals on how best to glean insights from this powerful study.
Synthetic lethality, in which a single gene defect leads to dependency on a second gene that is otherwise not essential, is an attractive paradigm to identify targeted therapies for cancer-specific mutations. Current methods to detect synthetic lethal (SL) partners for somatic mutations rely on large-scale shRNA screens in cell-lines or use human orthologs of yeast SL interactions, both of which are not necessarily representative of primary tumors and have incomplete coverage.

We have developed MiSL, a novel Boolean implication-based algorithm that utilizes large pan-cancer patient datasets (mutation, copy number and gene expression) to identify SL partners for cancer mutations. The underlying assumption of our approach is that, across multiple cancers, SL partners of a mutation will be amplified more frequently or deleted less frequently, with concordant changes in expression, in primary tumor samples harboring the mutation. Pan-cancer analysis discovers robust biological relationships that are likely to be independent of cancer subtype and increases statistical power.

First, we sought to validate MiSL using existing knowledge and large-scale shRNA data. Consistent with prior knowledge, MiSL candidates for BRCA1 mutation (mut) in breast cancer were enriched for DNA repair genes (p=.0.006). We also found: (1) significant overlap (p=0.002) between leukemia IDH1mut MiSL candidates and essential genes in IDH1mut cells determined by a DECIPHER shRNA screen we performed in doxycycline-inducible IDH1(R132) THP-1 cells, and (2) for multiple mutations in colorectal cancer, MiSL candidates were enriched (p<0.05) with genes that were selectively essential in mutated colorectal cell-lines in Achilles data.

Secondly, we experimentally confirmed novel SL partners that are druggable in acute myeloid leukemia (AML) and breast cancer. MiSL predicted a novel SL interaction in AML between IDH1mut and ACACA,
the rate-limiting enzyme that controls lipid biosynthesis. Consistent with our prediction, inhibition of ACACA with shRNA or a small molecule inhibitor TOFA prevented cell proliferation in IDH1mut (but not wildtype) AML cell-lines and primary blasts. MiSL also predicted that AKT1 is a SL partner of PIK3CAmut in breast cancer which we experimentally confirmed using 8 breast cancer lines. All four PIK3CAmut (but not wildtype) breast cancers were sensitive to AKT1 inhibition in viability and colony assays.

In conclusion, MiSL is a scalable computational solution that finds novel SL interactions. Using primary patient data allows it to capture in vivo tumor evolution, revealing SL interactions missed by existing methods. It can be widely applicable and can greatly accelerate novel target discovery for precision medicine in cancer.

9:50 AM - 10:05 AM

**RG01: Transcription factor binding site prediction in vivo using DNA sequence and shape features**

**Anthony Mathelier¹, Lin Yang², Tsu-Pei Chiu², Remo Rohs², Wyeth Wasserman¹**

¹ University of British Columbia, Canada; ² University of Southern California, United States

The interaction between transcription factors (TFs) and DNA results from a complex interplay between the DNA structure and nucleotide-amino acid contacts. We previously developed a sequence-based flexible hidden Markov model approach for TF binding site (TFBS) prediction, the Transcription Factor Flexible Model (TFFM). We describe a novel approach which builds upon the TFFMs by integrating four DNA shape features derived from the DNAshape prediction method: minor groove width, roll, propeller twist, and helix twist. The model is an ensemble machine learning framework combining the sequence and shape information of the DNA at TFBSs. Our results from 400 human ENCODE ChIP-seq data-sets show that adding DNA shape features to the TFFM scores in the machine learning framework consistently improves the prediction of TFBSs. Namely, combining DNA sequence and shape information perform better than current methods to discriminate ChIP-seq from background sequences. We observe significant improvements in TFBS predictions, which are not specific to TFFMs as they are reproduced when using the classical position-specific scoring matrices in the same framework. From the TFs under study, the results highlight that incorporating DNA shape information is most beneficial when applied to the E2F and MADS-domain TF families. A further validation on plant ChIP-seq data underline the importance of the propeller twist at a specific position of the TFBSs for seven plant MADS-domain TFs. Our approach corresponds to the in vivo counterpart of previous in vitro studies underscoring the importance of DNA shape for TFBS recognition.
SB02: The germline genetic component of drug sensitivity in cancer cell lines

Michael P. Menden1, Francesco P. Casale1, Johannes Stephan1, Graham R. Bignell2, Francesco Iorio1, Ultan McDermott2, Mathew J. Garnett2, Julio Saez-Rodriguez3, Oliver Stegle1

1 European Bioinformatics Institute (EMBL-EBI), United Kingdom, 2 Wellcome Trust Sanger Institute, United Kingdom, 3 RWTH Aachen University Hospital, Germany

Understanding differential therapeutic efficiency of cancer drugs is critical to deliver personalised treatments. While the utility of somatic mutations as drug sensitivity biomarker has been extensively studied, much less is known about the relevance of the germline genetic background. Herein, we develop a joint analysis approach that leverages both types of mutations in the context of drug susceptibility, before applying it to the systematic genetic analysis of 993 cell lines and 265 drugs from the Genomics of Drug Sensitivity in Cancer resource (GDSC). We find that the germline contribution to drug efficacy can be as large or larger than effects due to somatic mutations and demonstrate clear improvements in the prediction of drug response. In addition, we identify epistatic interactions between germline variants and somatic mutation signatures, revealing widespread interconnections between both variant types. Finally, we unpick the molecular mechanism of a mutation in NQO1, which affects drug susceptibility of the HSP90 inhibitor 17-AAG, thereby demonstrating the value of germline genetic variations as an additional dimension to inform treatment decisions.

RG02: A PBM-based glossary for motif discovery in regulatory regions

Luca Mariani and Martha Bulyk
Harvard Medical School, United States

Gene regulation depends on the binding of Transcription Factors (TFs) to specific DNA sequences. Recently, several genome-wide methods (e.g., ChIP-seq and DNase-seq) have identified many putative gene regulatory regions, but the exact underlying TF:DNA interactions remain elusive. Protein binding microarrays (PBMs) can profile these interactions in vitro and furnish position weight matrices (PWMs), which represent the standard model for TF binding. PWMs intuitively describe TF:DNA specificity, but cannot capture some of its features (e.g., alternative binding modes or interdependency between binding residues), unlike more complex, kmer-based models. Previous performance comparisons in motif discovery favored PWM models, but were assessed on few ChIP-seq datasets (6 in DREAM5).
Here, we develop a motif discovery approach that bypasses the limitations of PWMs through the joint use of 8mers and a non-redundant description of the TF specificity landscape, and test it extensively on ENCODE datasets. We first analyze 640 non-redundant metazoan TFs from the two main PBM repositories (UniPROBE and CisBP) and from unpublished experiments (all together 25 TF families). By developing an iterative dynamical clustering method over the entire dataset, we were able to associate the main TF (sub)families to small (<200) subsets of unique 8mers, which collectively form a “glossary.”

In contrast to previous kmer and PWM models for single TFs, our glossary exploits larger, currently available PBM repositories to globally evaluate the TF specificity repertoire. Moreover, it recapitulates and expands our knowledge of the heterogeneous specificity between, and within, TF families. The glossary outperforms PWMs in identifying enriched motifs for cofactors mediating indirect TF:DNA binding in a large collection of putative regulatory regions (240 ChIP-seq datasets for site-specific TFs from ENCODE). Gene expression profiles (10 RNA-seq datasets from ENCODE) support this enrichment, showing significant up-regulation of the cofactors identified by the glossary. In accessible chromatin regions (15 DNase-seq datasets from ENCODE) the glossary reveals cell-type specific patterns of motif enrichment, corroborated by the up-regulation of the associated TFs. Besides recapitulating well-known master TFs (e.g., SOX2/OCT4 in stem cells, and GATA1 in erythrocytes), the glossary highlights new candidates (e.g., TEAD2/FOXH1 in stem cells or FOS/ELF3 in colon cells).

In conclusion, for motif discovery our glossary improves both interpretability for the field’s “neophytes”, since it explicitly describes a manageable number (114) of modules of specifically recognized 8mers and their associated TFs, and performance, since it minimizes PBM experimental variability and identifies distinguishing 8mers that are unique to the associated TFs.

10:45 AM - 11:00 AM

SB03: Estimating the Number and Diversity of Cancer Mutations In the Overall Population from 5,319 Complete Cancer Genomes

Prathik Naidu1, Joseph Kaplinsky2, Ramy Arnaout2,3
1 Thomas Jefferson High School for Science and Technology, United States, 2 Beth Israel Deaconess Medical Center, United States, 3 Harvard Medical School, United States

Cancer is a genetic disease. To understand the link between cancer and genetics, large-scale efforts such as The Cancer Genome Atlas (TCGA) have begun to catalog cancer-related mutations, using hundreds of samples across many cancer types and subtypes. However, because even this sample size is small compared to the tens of millions of people with cancer
worldwide, there is a risk for substantial sampling bias: the number and diversity of cancer mutations in the sample may not reflect their number and diversity in the overall population. To address this issue, we used the Recon (Reconstruction of Estimated Communities from Observed Numbers) algorithm to estimate the overall number and diversity of coding and non-coding cancer mutations for 14 common cancer types, including breast cancer, prostate cancer, and glioblastoma, and four clinically important subtypes of breast cancer (e.g., luminal A, luminal B, Her2, and Basal). Our results suggest that while most common mutations have been discovered, the majority (78,186 – 1,887,539) remain unknown. Interestingly, the number of undiscovered mutations is not obvious from observations for a given cancer. For example, although glioblastoma and prostate-cancer samples exhibit relatively few mutations, our results suggest that overall, glioblastoma—but not prostate cancer—is likely to have as many mutations as all breast cancers. Thus, our algorithm may reveal aspects of cancer that are not obvious from direct observations (e.g., the potential presence of genetically different subtypes).

10:45 AM - 11:00 AM

RG03: Distinct specificities of the androgen and glucocorticoid receptors revealed using feature-based recognition model analysis of SELEX data

Liyang Zhang², Gabriella Martini¹, H. Tomas Rube¹, Vincent D. Fitzpatrick¹, Harmen J. Bussemaker¹, Miles A. Pufall²
¹ Columbia University, United States; 2 University of Iowa, United States

The androgen (AR) and glucocorticoid (GR) nuclear hormone receptors are closely related transcription factors. They are believed to bind to DNA as homodimers with indistinguishable specificity through identical DNA binding surfaces, and yet each occupies distinct genomic loci to drive distinct gene expression programs. How this functional difference ensues is not well understood. Here, by combining SELEX-seq assays on the DNA binding domain of AR and GR with statistical modeling, we show that the intrinsic DNA binding preference of the two factors differ substantially. We present an iterative algorithm that can accurately quantify the free energy parameters of a biophysically motivated recognition model over DNA footprints of unprecedented length (~30bp) by fitting a feature-based generalized linear model. Use of this algorithm allows us to analyze contributions to the binding specificity well outside the 15bp core region. In these outer flanks AR, but not GR, shows a preference for poly-A sequences. Isothermal titration calorimetry measurements confirm the difference in intrinsic specificity, and point to an AR-specific enthalpy-driven binding mechanism that derives additional binding energy from a narrowed minor groove. Our analysis shows that this mode of recognition restricts AR from binding GR sites, although the converse is not true. This contrast provides a basis for the differential genomic occupancy exhibited by AR and GR in LnCaP cells, helping to explain the finding that GR
can functionally substitute for AR in androgen independent prostate cancers. Taken together, our results demonstrate that differences in the intrinsic DNA binding specificity between closely related steroid hormone receptors exist and are functionally relevant. Our computational approach is general and widely applicable.

11:00 AM - 11:15 AM

SB04: Functional, chemical genomic, and super-enhancer screening identify sensitivity to cyclin D1/CDK4 pathway inhibition in Ewing sarcoma

Alyssa Kennedy¹, Mounica Vallrupalli¹, Liying Chen¹, Brian Crompton¹, Glenn Cowley², Paquita Vazquez², Barbara Weir², Aviad Tsherniak², S Parasuraman³, Sun Kim³, Gabriela Alexe¹, Kimberly Stegmaier¹

¹ Harvard Medical School, United States
² Broad Institute of MIT and Harvard, United States
³ Novartis Institute for Biomedical Research, United States

Ewing sarcoma is an aggressive bone and soft tissue tumor in children and adolescents, with treatment remaining a clinical challenge. This disease is mediated by somatic chromosomal translocations of the EWS gene and a gene encoding an ETS transcription factor, most commonly, FLI1. While direct targeting of aberrant transcription factors remains a pharmacological challenge, identification of dependencies incurred by EWS/FLI1 expression would offer a new therapeutic avenue. We used a combination of super-enhancer profiling, near-whole genome shRNA-based and small-molecule screening to identify cyclin D1 and CDK4 as Ewing sarcoma-selective dependencies. We revealed that super-enhancers mark Ewing sarcoma specific expression signatures and EWS/FLI1 target genes in human Ewing sarcoma cell lines. Particularly, a super-enhancer regulates cyclin D1 and promotes its expression in Ewing sarcoma. We demonstrated that Ewing sarcoma cells require CDK4 and cyclin D1 for survival and anchorage-independent growth. Additionally, pharmacologic inhibition of CDK4 with selective CDK4/6 inhibitors led to cytostasis and cell death of Ewing sarcoma cell lines in vitro and growth delay in an in vivo Ewing sarcoma xenograft model. These results demonstrated a dependency in Ewing sarcoma on CDK4 and cyclin D1 and support exploration of CDK4/6 inhibitors as a therapeutic approach for patients with this disease.

11:00 AM - 11:15 AM

RG04: How do closely related transcription factors recognize distinct genomic targets?

Ning Shen, Joshua Schipper, Tristan Bepler, John Horton, Raluca Gordan

Duke University, United States

Most eukaryotic transcription factors (TFs) are part of large protein families, with several TF family members (i.e. paralogous TFs) being expressed at the same time in the cell but targeting different sets of genes
and performing different regulatory functions. Closely related TFs, with amino acid similarity of 70% or more in the DNA binding domain (DBD), are generally believed to have identical DNA binding specificities. However, their in vivo genomic binding patterns are markedly different. Currently, we do not have a good understanding of the general molecular mechanisms by which TFs with highly similar DBDs select distinct in vivo targets.

Here, we show that closely related TFs interact differently with their putative genomic targets even in vitro, in the absence of any additional factors. Our study is focused on nine paralogous factors from three protein families: bHLH, E2F, and ETS. For each pair of related TFs, we used genomic-context protein-binding microarray (gcPBM) assays to compare the binding affinities of the two factors for ~25,000 putative genomic binding sites. We found that for most pairs of paralogous TFs, the two factors interact differently with their genomic sites in vitro, despite having identical PWMs. The only exception were TFs E2F1 and E2F3, which showed identical specificities. Interestingly, these factors play similar regulatory roles and can partially substitute for each other in the cell.

The way in which paralogous TFs differ is specific to each protein family: E2F1 and E2F4 prefer the same core GCGC/GCGG and differ in their flanking preferences for high affinity sites, ETS factors ETS1 and ELK1 different in specificity for medium and low affinity sites, while bHLH factors c-Myc and Mad1 prefer different flanks for their highest affinity site CACGTG, and differ significantly in their affinity for alternative cores CACATG/CACGCG. Overall, differences in genomic binding specificity between paralogous TFs are due both to direct recognition of DNA bases in the core binding sites (i.e. base readout), and to indirect recognition of different structural features in the flanking regions (i.e. shape readout).

Importantly, the differences in intrinsic binding specificity between paralogous TFs, as identified in vitro by gcPBM, can partly explain differential in vivo binding, measured by ChIP-seq. While we cannot expect the in vitro specificities of paralogous TFs to completely explain their in vivo binding patterns, our work shows that intrinsic TF preferences for genomic sites represent an important mechanism by which closely related factors achieve their regulatory specificity.

SB05: From phenotypic to molecular synergy: A transcriptional study of the dynamics of drug combinations based on single drug responses

Mehmet Eren Ahsen1,2, Jennifer E.L. Diaz1,2, Xintong Chen1, Bojan Losic1, Gustavo Stolovitzky2

1 Icahn School of Medicine at Mount Sinai, United States; 2 IBM Research, United States

Drug combination therapies have proven to be good strategies in cancer treatment in that they may elicit less adverse effects than single drugs
while overcoming the resistance to individual drugs that cancer cells tend to develop. Since screening all possible drug pairs is impractical, accurate methods for predicting synergistic drug combinations are needed. Attempts to predict the effect of drug combinations based on the transcriptional response of cells to single drugs have succeeded only partially because not enough data exists on how those transcriptional responses combine in the cellular environment. In this work we study the mechanisms whereby transcriptional responses combine to give rise to synergistic or additive responses to combined therapies. Specifically, we used RNAseq to study the transcriptional response over time (0, 3, 6, 9, 12, and 24 h) and for three drugs (A, B and C) and their combinations (AB, AC and BC) in MCF-7 breast cancer cells. Cell viability measurements show that one of the combinations (AB) is strongly synergistic, whereas the other two (AC and BC) are additive. The number of differentially expressed genes for the synergistic combination AB was at least one order of magnitude larger than the number of the differentially expressed genes resulting from each of the individual drugs A or B, and increased over time. For the additive combinations the number of differential expressed genes was about the same as for the single drugs, and was dominated by one of the drugs (C). To explain the massive transcriptional response of the synergistic combination, we extended the concept of additivity from the phenotypic to the transcriptional level. We found that most of the genes differentially expressed in AB but not in A nor B are non-additive. Using this information in the MCF-7 specific gene regulatory network we looked for transcriptional cascades that could explain the transcriptional program in AB based on that in A and B. We found that the majority of transcription factors that get activated at a given time point remain active at later time points. We studied how the activation of transcriptional regulators in A and B activate synergistic genes explaining much of the transcriptional response to AB. These analyses can pave the way for the design of algorithms to predict the response of cells to drug combinations based on RNAseq data from single drugs.

11:15 AM - 11:30 AM

RG05: Quantitative modeling of gene expression from sequence, using DNA shape-based model of binding sites

Pei-Chen Peng and Saurabh Sinha
University of Illinois at Urbana Champaign, United States

Motivation: Prediction of gene expression levels driven by regulatory sequences is one of the major challenges of genomic biology. A major current focus in transcriptional regulation is sequence-to-expression modeling, which interprets the enhancer sequence in light of transcription factor concentrations and DNA binding specificities and predicts precise gene expression levels in varying cellular contexts. Such models have so far exclusively relied on the position weight matrix (PWM) model for
transcription factor (TF)-DNA binding. Several reports have pointed out deficiencies in the PWM model and presented alternative models, including DNA shape-based models, that are claimed to be in greater agreement with TF-DNA binding data. However, it is not known if alternative models of DNA binding, such as DNA shape models, can also improve prediction of gene expression.

**Results:** Here, we adapted a statistical thermodynamics model to develop a quantitative model of gene expression interprets enhancer sequences using DNA shape features of binding sites, as opposed to a PWM-based scoring of sites. We used rigorous methods to evaluate the fits of expression readouts of more than 35 enhancers regulating spatial gene expression patterns in the blastoderm-stage Drosophila embryo, and we show that DNA shape-based models perform at least as well as, and arguably better than PWM-based models. We objectively characterized the relationship between DNA shape-based models and PWM models of binding site affinity, and observed that DNA shape features carry information that is complementary to the PWM and useful for sequence-to-expression modeling. In addition, we combined DNA shape and PWM into a single model and tested if it would achieve better predictions than using either binding model independently. The integrative model did not perform consistently better than either DNA shape or PWM based model alone.

**Conclusion:** Our work shows that quantification of TF binding site affinity using DNA shape is not only justified by binding affinity data, it is also effective in interpreting enhancer sequence to accurately predict gene expression. With the growing availability of data sets describing TF-DNA binding affinities comprehensively, we expect that it will be possible to train such models more accurately and to utilize them to better predict gene expression and the functional effects of single nucleotide polymorphisms in the non-coding genome.

**SB06: Large-scale models of signal propagation derived from phosphoproteomic data to study kinase inhibitors**

**Julio Saez-Rodriguez**¹,², Camille Terfve¹, Edmund Wilkes³, Pedro Casado³, Pedro Cutillas³  
¹ EMBL-EBI, United Kingdom; ² RWTH Aachen University Hospital, United Kingdom; ³ Queen Mary University of London, United Kingdom

Mass spectrometry is widely used to probe the proteome and its modifications in an untargeted manner, with unrivalled coverage. Applied to phosphoproteomics, it has tremendous potential to interrogate phospho-signalling and its therapeutic implications. However, this task is complicated by issues of undersampling of the phosphoproteome and challenges stemming from its high-content but low-sample-throughput nature. Hence, methods using such data to reconstruct signalling networks have been limited to restricted data sets and insights.
We propose PHONEmeS, a new method to handle high-content discovery phosphoproteomics data on perturbation by putting it in the context of kinase/phosphatase-substrate knowledge, from which we derive and train logic models [1]. We show, on a data set obtained through perturbations of cancer cells with small-molecule inhibitors [2], that this method can study the targets and effects of kinase inhibitors, and reconcile insights obtained from multiple data sets, a common issue with these data [1].

REFERENCES

1:35 PM - 1:50 PM

RG06: Tracking the Evolution of 3D Gene Organization
Alon Diament and Tamir Tuller
Tel Aviv University, Israel

One of the most fundamental open biological questions is what determines the eukaryotic genomic organization. It has been shown that the distribution of genes in eukaryotic genomes is not random; however, formerly reported large scale relations between gene function and genomic organization were relatively weak.

Previous studies have demonstrated that codon usage bias is related to all stages of gene expression and to protein function. Here we apply a novel tool for assessing functional relatedness, codon usage frequency similarity (CUFS), which measures similarity between genes in terms of codon and amino acid usage. By analyzing Hi-C data, describing the three dimensional conformation of the DNA, we show that the functional similarity between genes captured by our metric is directly and very strongly correlated with their three dimensional (3D) distance in five eukaryotes (r > 0.74; p<1e-323 in all cases; Diament et al. Nature Commun. 2014).

We utilize this result to propose a novel approach for improving the accuracy of 3D genome reconstructions by introducing additional predicted physical interactions to the model, based on orthologous interactions in an evolutionary-related organism and based on predicted functional interactions between genes (e.g. based on CUFS). We demonstrate in the eukaryote S. cerevisiae that this approach indeed leads to the reconstruction of improved models (Diament et al. PloS Comput. Biol. 2015).

We have previously shown that some level of conservation of genomic organization exists between organisms. However, almost all studies of 3D genomic organization analyzed each organism independently from others. Here we propose a novel approach for inter-organismal analysis of the organization of genes. By utilizing Hi-C data from two fungi – S. cerevisiae and S. pombe – we detect orthologous gene families that
underwent changes in their 3D co-localization during evolution. We show that this approach enables identifying various biologically meaningful modules of co-evolving genes with shared function (Diament et al. under-review 2015).

Our results emphasize the importance of three-dimensional genomic organization in eukaryotes and suggest that the evolutionary mechanisms that shape the 3D organization of genes are affected by their functionality and expression pattern. In addition, we provide novel algorithms for 3D genome reconstructions and for deciphering gene function and organization.

SB07: Global chromatin defects and spurious transcription defines a novel subset of human cancers

Kakajan Komurov, Navneet Singh, Kwangmin Choi, Nathan Salomonis, Matthew Weirauch, Kakajan Komurov
Cincinnati Children’s Hospital Medical Center, United States

Through a comprehensive computational pan-cancer analysis of The Cancer Genome Atlas mRNA-seq datasets, we found that a subset (ranging in 2% to 30%) of almost every human cancer is characterized by severely defective gene body chromatin remodeling and RNA Polymerase II (Pol II)-mediated transcription, resulting in extensive genome-wide aberrant shortening of mRNA transcripts (GTS: Global Transcript Shortening). GTS was characterized by extensive cryptic spurious transcription and intron splicing defects, such that, more than half of all expressed genes were producing truncated transcripts. Strikingly, GTS mostly affected large genes with specific chromatin structures, while small genes involved in housekeeping processes were mainly overexpressed in these tumors. The mRNA-level observations, including pathway enrichments and gene length bias in expression, were also reproduced at the protein level in these tumors. Importantly, pro-inflammatory cell death pathways, which are primarily regulated by large genes, were significantly repressed in all GTS+ tumors due to excessive truncation of their mRNAs. Accordingly, GTS+ melanoma and clear cell renal cell carcinoma (ccRCC) patients had significantly poor response to immunotherapy with interferon (IFN), interleukin (IL) and the newer drugs such as ipilimumab in the clinic, but not to other forms of therapy. We found that some cancer cell lines in vitro exhibit the GTS phenotype. GTS+ cell lines had severe defects in genic histone methylation and acetylation, as well as in Pol II transcription initiation and elongation functions, and had a transcriptional and pathway activation profile highly consistent with GTS+ clinical samples. Consistent with defective immune cell death response phenotype, GTS+ cancer cells had impaired response to pro-inflammatory stimuli and FasL-mediated killing. However, we show that GTS imposes specific metabolic and proteotoxic vulnerabilities in cancer cells with potential implications for therapeutic targeting in the clinic. Overall, our integrated
computational and experimental analyses reveal a previously unknown clinically significant major phenotype of human cancers with clear implications in patient stratification for therapies, especially with immune modulatory agents, and therapeutic strategies to target this subset of cancers in the clinic.

1:50 PM - 2:05 PM

RG07: Three-dimensional analysis of regulatory features reveals functional enhancer-associated loops

Yao Wang¹, Junbai Wang², Yufan Zhou¹, Malaina Gaddis³, Rohit Jadhav¹, Xun Lan⁴, Tim Huang¹, Shili Lin⁵, Peggy Farnham⁶, Seth Frietze⁷, Victor Jin¹

¹ University of Texas Health Science Center at San Antonio, United States; ² Oslo University Hospital – Norwegian Radium Hospital, Norway; ³ University of Southern California, United States; ⁴ Stanford University, United States; ⁵ The Ohio State University, United States; ⁶ University of Southern California, United States; ⁷ University of Vermont, United States

Several critical gaps remain in our knowledge of the relationship of chromatin structure to gene regulation. These include 1) classifying different types of chromatin interactions (including promoter-enhancer contacts), 2) determining the relationships between classes of chromatin interactions and the epigenomic state, 3) deciphering the functional relevance of chromatin interactions, and 4) determining whether genes associated with different chromatin interaction classes are involved in disease.

To gain insight into the relationship between chromatin structure and gene expression, we conducted chromatin conformation analysis using PANC1 pancreatic cancer and MCF7 breast cancer cells. For PANC1, we carried out Tethered Chromatin Capture (TCC) on two biological replicates, and compared correlation between replicates to validate the data quality. For MCF7, we used both TCC and in situ Hi-C protocols on replicates and also performed correlation analysis. To analyze the 3D conformation in these two cancer cells, we first detected topologically associated domains (TADs) in each chromosome, then applied a novel Hi-C analysis algorithm and identified hundreds of thousands of Interacting Loci Pairs (ILPs) in each of the two cell types. We classified ILPs according to location with respect to gene structure, gene expression, different histone modifications, DNase hypersensitivity, and RNA polymerase II and CTCF binding. Interestingly, we found that a majority of ILPs are within a particular TAD, and only 5% of the ILPs are involved in promoter regions, with even fewer promoter-enhancer loops. To further explore the potential mechanism behind 3D conformation and gene expression, we conducted TCC on PANC1 treated with drug ICG001 known as a CBP inhibitor, and C646 known as a CBP/p300 competitor. We examined the changes of TADs and ILPs in the drug-treated PANC1, and the impact of pharmacological inhibition of histone acetylation on genes having promoter-enhancer loops in PANC1 cells. We find that genes associated with promoter-enhancer loops have cell-type-specific functional
annotations. We further demonstrated that genes with promoter-enhancer loops altered expression in response to drug treatment in PANC1, suggesting that the chromatin loops we identified are functional. Taken together, our study provides insights into the interdependence of three-dimensional chromatin looping and gene expression mediated by enhancer-promoter interactions.

**SB08: Network Maximal Correlation to Infer Nonlinear Gene Modules in Cancer**

Soheil Feizi\(^1\), Ali Makhdoumi\(^1\), Ken Duffy\(^2\), Manolis Kellis\(^1\), Muriel Medard\(^1\)

\(^1\)MIT, United States; \(^2\)Hamilton Institute, Maynooth University, Ireland

Identifying nonlinear relationships in large datasets is a daunting task particularly when the form of the nonlinearity is unknown. Here, we introduce Network Maximal Correlation (NMC) as a fundamental measure to capture nonlinear associations in networks without the knowledge of underlying nonlinearity shapes. NMC infers, possibly nonlinear, transformations of variables with zero means and unit variances by maximizing total nonlinear correlation over the underlying network. For the case of having two variables, NMC is equivalent to the standard Maximal Correlation metric. We characterize a solution of the NMC optimization using geometric properties of Hilbert spaces for both discrete and jointly Gaussian variables. For discrete random variables, we show that the NMC optimization is an instance of the Maximum Correlation Problem and provide necessary conditions for its global optimal solution. Moreover, we propose an efficient algorithm based on Alternating Conditional Expectation (ACE) which converges to a local NMC optimum. For this algorithm, we provide guidelines for choosing appropriate starting points to jump out of local maximizers. We also propose a distributed algorithm to compute a 1-epsilon approximation of the NMC value for large and dense graphs using graph partitioning. For jointly Gaussian variables, under some conditions, we show that the NMC optimization can be simplified to a Max-Cut problem, where we provide conditions under which an NMC solution can be computed exactly. Under some general conditions, we show that NMC can infer the underlying graphical model for functions of latent jointly Gaussian variables. These functions are unknown, bijective, and can be nonlinear. This result broadens the family of continuous distributions whose graphical models can be characterized efficiently. We illustrate the robustness of NMC in real world applications by showing its continuity with respect to small perturbations of joint distributions. We also show that sample NMC (NMC computed using empirical distributions) converges exponentially fast to the true NMC value. Finally, we apply NMC to different cancer datasets including breast, kidney and liver cancers, and show that NMC infers gene modules that are significantly associated with survival times of individuals while they are not detected using linear association measures.
RG08: An integrated model for detecting significant chromatin interactions from high-resolution Hi-C data

Mark Carty¹, Lee Zamparo¹, Alvaro Gonzalez¹, Raphael Pelossof¹, Olivier Elemento², Christina Leslie¹

¹ Memorial Sloan Kettering Cancer Center, United States; ² Weill Cornell Medical College, United States

We present an integrated model for estimating the significance of Hi-C interactions that accounts for systematic sources of variation in interaction read counts, including the dependence of random polymer ligation on genomic distance and GC content and mappability bias. Additionally, we explicitly model the zero-inflation and over-dispersion of counts in the contact matrix by using a generalized linear model approach based on hurdle regression. By learning a null model that incorporates all these statistical properties of Hi-C contact matrix counts, we show that our estimates of significance (P values) are less inflated than previous methods, yielding fewer false discoveries. We performed an analysis on a recent very high resolution in situ Hi-C data set for the lymphoblastoid cell line GM12878 and found that our method can identify significant interactions at the sub-topologically associating domain level, including DNA loops mediated by CTCF and/or cohesin and enhancer-promoter interactions. We also found that CTCF-associated interactions were more strongly enriched in the middle range of genomic distance (~700Kb-1.5Mb), while interactions involving actively marked DNase accessible elements were enriched both at short range (<500Kb) and longer range (>1.5Mb). Strikingly, we found a dramatic enrichment of longer range gene-gene interactions connecting the replication-dependent histone genes on chromosome 6, potentially representing the chromatin architecture at the histone locus body.

SB09: Integrating single cell transcriptional signatures and cell-cell interaction network models to study multiscale control of liver regeneration dynamics

Daniel Cook¹, Sirisha Achanta², Aalap Verma¹, Jan Hoek², Babatunde Ogunnaike¹, Rajanikanth Vadigepalli²

¹ University of Delaware, United States; ² Thomas Jefferson University, United States

Recent technological advances have allowed for transcriptional profiling of single cells from tissues. Challenges remain, however, in integrating single-cell scale transcriptional regulation with tissue-scale physiological response. We address these challenges by integrating experimentally identified transcriptional cell states into a cell-cell interaction model with physiological transitions between states. We applied this technique to the framework of liver resection and regeneration, a treatment for hepatocellular carcinoma and as a technique to enable live liver transplant.
Our approach gives unique insights into the multiscale control of liver regeneration dynamics. Furthermore, the availability of clinical data and pathology samples from regenerating and non-regenerating cases present an opportunity to employ modeling approaches that integrate data from different functional scales to better understand the mechanisms underlying clinical resection responses.

We used laser capture microdissection followed by high-throughput qPCR to acquire a new experimental data set on gene regulation in hundreds of single and pooled hepatocytes and hepatic stellate cells during liver regeneration in healthy and chronically diseased livers. We analyzed the high-dimensional, single-cell scale gene regulation data to identify distinct cell populations within each cell type. We then extrapolated from the single-cell scale to cell phenotype distributions at the whole-tissue scale. Our approach characterized the transcriptional state of individual hepatocytes and hepatic stellate cells in the in vivo regenerating liver at a level of detail not previously achievable. Characterization of transcriptional profiles revealed cell-type specific, functionally distinct states for hepatocytes as well as stellate cells. Surprisingly, we found the two liver cell types distributed among all identified functional states irrespective of the experimental condition (baseline or regenerating, healthy or chronic disease). However, relative proportions of the cells distributed among the identified functional states changed based on regeneration stage as well as healthy vs. chronic disease, suggesting a cell population balance signature of disease.

We developed a computational model that accounts for the multi-scale nature of liver regeneration by integrating physiological-scale interactions, transcriptional states of hepatocytes and liver non-parenchymal cells, and molecular signaling networks. We explored a range of model dynamics to identify parameter sets that account for experimentally observed regeneration profiles in health and disease. Model simulations suggest that aberrant activation dynamics of hepatocytes and hepatic stellate cells are key factors suppressing liver regeneration in multiple disease contexts.

Our integrated experimental and computational analysis points towards new avenues for therapeutic intervention based on renormalizing the cellular functional state balances to improve surgical outcomes.
elements that control the regulation of transcription and replication. Recent advancements in sequencing and analysis techniques – specifically Hi-C, or high-throughput chromosome conformation capture – have revealed these interactions genome-wide at unprecedented resolutions. Unfortunately, navigating the Hi-C data remains a daunting feat for many biologists, as its $O(n^2)$ complexity for the already big data intrinsic to mammalian genomes poses a challenge to its analysis (time and memory usage), storage and transfer.

Our laboratory has developed and extended the functionality of the 3D Genome Browser (http://3dgenome.org), a web-based, intuitive and accessible browser of Hi-C data. The browser adopts a genome feature-searchable approach and contextualizes the region with established University of California Santa Cruz (UCSC) Genomic Browser while retaining its flexibility to customize genome tracks and load personalized UCSC sessions. Furthermore, the browser allows the visualization of the Hi-C contact matrix as both rotated heatmap, as well as virtual 4C (linear plot detailing the number of interactions between a single genomic site of interest -bait or anchor locus- with other loci), which is further contextualized with DNase I Hypersensitive Site (DHS)-linkage, ChIA-PET and capture Hi-C tracks. The virtual 4C display is particularly valuable in the identification of cis-regulatory elements and their gene targets.

To ameliorate the $O(n^2)$ complexity of Hi-C data, we pioneered a new file format to represent contact matrices, called BUTLR (Binary Upper TrianguLar matRix). This format not only compresses the high-resolution Hi-C data by ~20, it permits random access such that interaction values among any regions could be queried fast and directly. Therefore, the 3D Genome Browser supports the visualization of Hi-C datasets in BUTLR format hosted on any HTTP accessible server without requiring the transfer of entire files, similar to bigWig/bigBed files for the UCSC Genome Browser.

Currently, our browser boasts ~3000 views every month and receives frequent improvements in its user interface. Several laboratories have successfully utilized the browser to visualize user-generated Hi-C interactions. With our gene-centric, binary-file browser approach, the 3D Genome Browser improves the accessibility in browsing Hi-C data. With the visualization of the spatial organization and long-range interactions of particular genomic regions along with their genetic and epigenetic context, our browser seeks to drive hypothesis-generation about and enrich the understanding of the intrinsic link between genomic organization and genetic regulation.
SB10: Understanding Breast Cancer Heterogeneity through Personalized Drosophila Models

Jennifer Long Diaz, Avi Ma’Ayan, Ross Cagan
Icahn School of Medicine at Mount Sinai, United States

Triple negative breast cancer (TNBC) is a molecularly heterogeneous disease characterized by poor therapeutic response, low survival rates, and few druggable molecular targets. We aim to study this heterogeneity by examining complex genetic patient-specific models of TNBC in Drosophila.

Building genetic models of TNBC requires identifying the genes responsible for tumor progression. TNBCs are largely driven by genes with altered copy number status. Our analysis of breast cancer copy number data from The Cancer Genome Atlas (TCGA) has identified prioritized likely putative drivers specific to TNBC from over 8000 genes in amplified and deleted regions.

We will functionally screen these putative driver genes for enhancement of cell migration and tissue expansion in transformed tissue in Drosophila. In preliminary studies, these phenotypes accurately identified several known driver genes. The newly identified driver genes, along with known driver genes harboring mutations, will be used to construct a set of complex, personalized Drosophila models for ten TNBC patients in TCGA.

To further guide the selection of key drivers for each model, we have developed novel Drosophila gene set enrichment tools, which identify key genes when applied to expression data from each patient. We will then use gene expression data to quantitatively track the accuracy of each model and select the best fit model for the patient’s actual tumor.

We are developing increasingly accurate personalized models through an iterative experimental-computational workflow. Each final personalized model has the potential to display unique properties, and we will use these models to study TNBC heterogeneity between patients. In each model, we will examine the extent of cell proliferation and migration, determine the signaling pathways responsible for these phenotypes by biochemical analysis and a chemical genetic screen, and measure response to standard-of-care chemotherapies. Our goal is to shed light on the molecular basis for patient-to-patient variability in survival and therapeutic response in TNBC.
RG10: Modeling methyl-sensitive transcription factor motifs with an expanded epigenetic alphabet

Coby Viner¹, James Johnson², Nicolas Walker³, Marcela Sjoberg⁴, David J. Adams⁴, Anne C. Ferguson-Smith³, Timothy L. Bailey², Michael M. Hoffman¹

1 University of Toronto and Princess Margaret Cancer Centre, Canada; 2 The University of Queensland, Australia; 3 University of Cambridge, United Kingdom; 4 Wellcome Trust Sanger Institute, United Kingdom

Many transcription factors (TFs) initiate transcription only in specific sequence contexts, providing the means for sequence specificity of transcriptional control. A four-letter DNA alphabet only partially describes the possible diversity of nucleobases a TF might encounter. Cytosine is often present in the modified forms: 5-methylcytosine (5mC) or 5-hydroxymethylcytosine (5hmC). TFs have been shown to distinguish unmodified from modified bases. Modification-sensitive TFs provide a mechanism through which widespread changes in DNA methylation and hydroxymethylation, found in many cancers, can dramatically shift active gene expression programs.

To understand the effect of modified nucleobases on gene regulation, we developed methods to discover motifs and identify TF binding sites in DNA with covalent modifications. Our models expand the standard A/C/G/T alphabet, adding m (5mC) and h (5hmC). We adapted the well-established position weight matrix formulation of TF binding affinity to this expanded alphabet.

We engineered several tools to work with expanded-alphabet sequence and position weight matrices. First, we developed a program, Cytomod, to create a modified sequence, using data from bisulfite and oxidative bisulfite sequencing experiments. Cytomod decides between multiple modifications at a single locus, using a configurable evidence model. Second, new versions of MEME (Multiple EM for Motif Elicitation), DREME (Discriminative Regular Expression Motif Elicitation), and MEME-ChIP enable de novo discovery of modification-sensitive motifs. A new version of CentriMo enables central motif enrichment analysis to infer direct DNA binding in an expanded-alphabet context. These versions permit users to specify new alphabets, anticipating future alphabet expansions.

We created an expanded-alphabet genome sequence using whole-genome maps of 5mC and 5hmC in naive ex vivo mouse T cells from BLUEPRINT. Using this sequence, expanded-alphabet position weight matrices, and ChIP-seq data from Mouse ENCODE and others, we identified cis-regulatory modules active only in the presence or absence of cytosine modifications. We reproduced various known methylation binding preferences, including the preference of ZFP57 and C/EBP for methylated motifs and the preference of c-Myc for unmethylated E-box motifs. Using these known binding preferences to tune model parameters enables discovery of novel modified motifs.
SB11: A linear time-invariant model of phenotype dynamics in breast cancer cell populations

Margaret P. Chapman1, Tyler Risom2, Roel Dobbe1, Rosalie Sears2, Claire Tomlin1
1 UC Berkeley, United States; 2 Oregon Health and Science University, United States

Background: Phenotypic heterogeneity, or cellular diversity on the phenotypic level, poses a fundamental challenge to effective treatment of certain cancers, including triple negative breast cancer (TNBC). Thus, the discovery of strategies to control heterogeneity is a priority in cancer biology. Understanding the phenomena that govern drug-induced changes in cell population phenotypes is needed to design therapeutic approaches that systematically control heterogeneous tumors. Here we present a mathematical framework to suggest why observed trends in phenotype dynamics occur and why they may change under therapeutic perturbation.

Methods: Phenotype and dead cell quantities were recorded from 4 replicate wells containing the TNBC cell line, HCC1143, at 6 time points every 12 hours under three distinct conditions. The conditions were application of a PI3K/mTOR inhibitor, a MEK inhibitor, or a control containing the drug vehicle DMSO but no medicinal agent. A linear time-invariant model of phenotype dynamics was derived to investigate the therapeutic effects on phenotypic heterogeneity. Cell division and death of each phenotype in addition to switching between phenotype pairs were represented explicitly. A convex optimization program was formulated and solved to estimate values of model parameters using measurements from two wells. Data from the two remaining wells were used for model validation. Parameter values were analyzed in the context of existing biological hypotheses. This methodology will be compared to an alternative approach, involving a stochastic process, from the literature.

Results: The model predicts monotonic trends well; predicting oscillatory behavior is reserved for future work. Analysis of model parameters agrees with existing knowledge on therapeutic effects and provides rationale for observed trends. For example, the model indicates that the PI3K/mTOR inhibitor reduces the prominence of basal phenotypes due to specific switching behavior and low cell division, rather than selective elimination. Further, the model suggests that the MEK inhibitor promotes the prominence of basal phenotypes via switching phenomena, instead of increased cell division or reduced death. Remarkably, the model results specify bidirectional switching between two phenotypes that precisely reflects the hypothesized effects of the MEK inhibitor.

Conclusion: Our framework motivates further experimentation to study model predictions and the design of precise metrics for therapeutic
effectiveness. Short-term future work includes: analyzing statistical significance, investigating measurement noise, and identifying gaps in the model's descriptive power. The long-term goals are to explore richer models to capture oscillatory trends (e.g., stochastic linear time-varying or non-linear) and incorporate additional sources of biological data into the methodology.

4:05 PM - 4:20 PM

RG11: Genome-wide mapping of histone marks at single-nucleosome resolution

Marcelo Rivas-Astroza¹, Shu Xiao¹, Xiaoyi Cao¹, Sheng Zhong¹
¹ University of California, San Diego, United States; ² University of Illinois, Urbana-Champaign, United States

Histone post-translational modifications (PTM) control much of the workings of the genome. However, it remains unclear how different histone PTM combine within individual nucleosomes to encode regulatory functions. Regular protocols of chromatin immunoprecipitation followed by sequencing (ChIP-seq) are unsuited to shed light on this matter as they are unable to trace histone PTM to individual nucleosomes. Alternative ChIP-seq protocols employing enzymatic digestion of DNA can overcome this limitation and produce histone PTM footprints at single-nucleosome resolution, but analytical obstacles still precludes determining whether a nucleosome is marked or not by a histone PTM. In particular, care has to be taken to avoid co-founding effects stemming from the relation between nucleosomes co-localization levels and nucleosomal enrichment of histone marks. To overcome these problems, here we devised computational and statistical methods to exploit the power of ChIP-seq protocols based on enzymatic digestion of DNA to trace histone PTM to individual nucleosomes. Using a ChIP-seq protocol where DNA was MNase-digested and our analytical methods we generated genome-wide maps where H3K4me3, H3K27Ac, H3K9me3, and H3K27me3 were traced to individual nucleosomes in mouse embryonic stem cells. A significant number of nucleosomes were marked by two or more of histone marks. Nucleosomes marked simultaneously by H3K4me3 and H3K27me3 were prevalent among bivalent domains compared to the genomic background. Nucleosomes having the repressive marks H3K27me3 and H3K9me3 were enriched at the transcription starting site of highly active genes only if they were also co-localized with the activating mark H3K27Ac. Inclusion of alternatively spliced exons on the final mRNA was correlated with nucleosomes marked by H3K4me3, H3K27Ac, or H3K27me3, but was largely unaffected by nucleosomes marked by H3K9me3. Together, these findings indicate that combinatorial patterns of histone PTM within individual nucleosomes are fundamental units of regulatory information.
Precision medicine is an emerging method for disease treatment and prevention that takes into consideration individual genetic and environmental variability for each person. However, the advance of precision medicine is hindered by a lack of mechanistic understanding of the energetics and dynamics of drug-target and genetic interactions in the context of the whole human genome and interactome. To address this challenge, we have developed a novel structural systems pharmacology approach to elucidate molecular basis and genetic biomarkers of drug action. Our approach combines big data analytics and mechanism-based modeling through integrating structural genomic, functional genomic, metabolomics, and interactomic data. By searching for all structurally-characterized human proteins and applying molecular modeling and machine learning, we are able to construct genome-scale high-resolution drug-target interaction models. Subsequently, we link the putative off-targets to genome-scale biological networks to identify drug modulation pathways and cryptic genetic factors. A novel algorithm k-DELADD is developed to model the impact of edge modification of biological network on network robustness, and is used to identify hidden genetic biomarkers of drug phenotypic response.

As proof-of-concept studies, we have applied our structural systems pharmacology approach to drug rescue and drug repurposing for precision medicine. We have identified cryptic genetic factors that account for the side effect of Torcetrapib, a cholesterol-lowering drug that failed in phase III clinical trial due to serious side effects. Recently, we have revealed molecular and genetic mechanisms of metformin, enabling us to repurpose metformin as a precision anti-cancer therapy. Metformin is a drug prescribed to treat type-2 diabetes, and exhibits anti-cancer effects in a portion of patients. However, the direct molecular and genetic interactions which produce this pleiotropic effect of metformin have not yet been fully described. Using our approach, we predict a set of potential molecular targets of metformin, and experimentally validate the interaction between metformin and a majority of our top-ranked off-targets. Notably, SGK1 was identified as a key target in mediating metformin’s effect. Moreover, key genetic network biomarkers are identified. They are largely consistent with existing experimental evidence. Our results shed new light on repurposing metformin as safe, effective, personalized therapies, and demonstrate that structural systems pharmacology is a potential powerful tool to facilitate the development of precision medicine.
RG12: Creating a library of genome-wide chromatin state patterns during B lymphopoiesis

Mark Maienschein-Cline¹, Pinal Kanabar¹, Neil Bahroos¹, Malay Mandal², Marcus Clark²
¹ University of Illinois at Chicago, United States; ² University of Chicago, United States

B lymphopoiesis proceeds through several stages, during which the cell undergoes rearrangements in its antibody gene content that change the specificity of its antigen recognition mechanisms. This process forms a crucial underpinning of our adaptive immune system. In addition to changes in the antibody genetic content, these stages are also defined by a number of important epigenetic modulations. However, only a fraction of these are currently understood, often only in the context of specific transcription factors in certain developmental stages. Obtaining more general patterns of chromatin state on a broader, lymphopoiesis-wide context would thus provide an invaluable resource underpinning our understanding of epigenomics changes in B lymphopoiesis.

To this end, we consider nucleosome positioning and chromatin accessibility, which play an important role in determining regions of the genome that regulatory factors can interact with. Methodologies like ATAC-seq, DNase-seq, and FAIRE-seq can be employed to detect loci with open chromatin. In particular, ATAC-seq is a newly developed and particularly powerful tool, as it involves a simpler protocol that can be applied to smaller populations of cells; ATAC-seq can also be used to infer nucleosome positioning (the converse of open chromatin) if paired-end sequencing is used. Additionally, the activity of specific transcription factors can be inferred using bioinformatic techniques like motif enrichment, making open chromatin measurements a valuable basis for additional epigenetic studies.

We have profiled eight B lymphopoietic stages using ATAC-seq and describe the exciting preliminary results and analysis strategy here. We first identified regions of interest genome-wide from open chromatin enrichment, and then used patterns of chromatin state changes to separate the loci into functionally differentiated groups. In particular, we used an unsupervised clustering approach to discover clusters of loci with concordant chromatin state changes, and employed consensus clustering to determine the number of distinct patterns that can be identified reliably. In our data, we discovered 11 distinct patterns that describe changes in chromatin state across >100,000 differentiated loci. In addition to revealing important changes to the regulatory landscape across B lymphopoiesis, we believe that these patterns and the associated loci can be used as a valuable reference library of chromatin state for future B lymphopoiesis studies.
SB13: Bringing big genomic data into focus for studying complex diseases in specific biological contexts

Arjun Krishnan¹, Ran Zhang¹, Victoria Yao¹, Chandra Theesfeld¹, Aaron Wong², Alicja Tadych¹, Natalia Volfovsky², Alan Packer², Alex Lash², Olga Troyanskaya²

¹ Princeton University, United States; ² Simons Foundation, United States

A big challenge in genomics is characterizing the genetic and functional dysregulation in complex diseases. Addressing this problem requires systematic computational approaches that can harness the explosion of data and bring ever-finer biological contexts into focus e.g. tissue, cell-type, sex and age. Towards this goal, we recently developed a Bayesian framework that integrates thousands of gene-expression, protein-interaction and regulatory-sequence datasets to predict tissue-specific functional relationships between genes in each of 144 specific human cell-types and tissues.

Here, using autism spectrum disorder (ASD) as an example, we demonstrate how tissue-specific networks provide a valuable apparatus for generating hypotheses about the molecular basis of human diseases. ASD has a strong genetic basis that remains poorly characterized by sequencing and quantitative genetics studies. Using an evidence-weighted machine learning approach that utilizes the human brain-specific functional gene network, we generated the first genome-wide prediction of autism-associated genes. These predictions were validated using an independent large case-control sequencing study. Leveraging these genome-wide predictions and the brain-specific network, our analyses demonstrate that the large set of ASD genes, including a host of novel candidates, converges on a smaller number of key cellular pathways and specific early developmental stages of the brain.

Manifesting in early development and being five times more common among boys than among girls, ASD is also an exemplar of diseases whose incidence or severity varies dramatically across the human lifespan and between the sexes. Therefore, our next goal lies in expanding our genomics toolkit to address age- and sex-specificity in addition to tissue/cell-type-specificity. we will conclude with preliminary results that demonstrate the promise of some of our approaches towards this goal.

RG13: AEF: A methodology based on Assortativity of Epigenetic Features in promoter centered chromatin interaction networks identifies Polycomb and RNA Polymerase as main players

Vera Pancaldi¹, Enrique Carrilo de Santa Pau¹, Peter Fraser², Mikhail Spivakov², Daniel Rico¹, Alfonso Valencia¹

¹ Spanish National Cancer Research Centre (CNIO), Spain; ² The Babraham Institute, United Kingdom of Great Britain and Northern Ireland
With the advent of genome-wide chromatin interaction mapping techniques, numerous methods have been proposed to analyze the resulting datasets. However, only few studies have focused on the comparison of different datasets and even fewer exploit the numerous tools of network theory in the analysis of these networks.

Assortativity is a widely used metric in the field of social networks, which characterizes the extent to which similar nodes are more likely to interact. We propose to analyze assortativity of a number of epigenetic features in chromatin interaction networks with the aim of identifying key features that associate with chromatin contacts.

We consider two very recent promoter-centered interaction maps for mouse Embryonic Stem Cells obtained with two variations of promoter-capture HiC (PCHiC and HiCap), a novel technique that detects chromatin contacts involving at least one promoter. To characterize the interacting chromatin fragments, the nodes of the network, we exploit a collection of 79 epigenetic features (cytosine and histone modifications and binding peaks of chromatin related proteins). We project the abundance of these features on the network as node attributes and calculate their assortativity.

Plotting assortativity versus abundance of each epigenetic feature we identify the Polycomb Group Proteins and associated histone marks as particularly important in the chromatin contacts. The results are reproduced in the two datasets, despite the technical differences in the experimental techniques and in the topology of the mapped chromatin networks. Moreover, we distinguish between two subnetworks: the network of contacts amongst promoters (PP) and the network of contacts between promoters and other genomics fragments (PO), presumably distal elements.

We observe substantial differences in the assortativity of certain features in these subnetworks, for example the active promoter mark H3K4me3 has negative assortativity in the PP contacts and positive assortativity in the PO contacts. Interestingly, we find a difference in assortativity of different forms of RNA Polymerase II in PO contacts, with the actively elongating form showing much higher assortativity compared to inactive or poised forms. These findings, reproduced in two independent datasets and validated on additional ChIA-PET networks, are inline with the model in which enhancers contact the elongating form of polymerase, which would be key in the regulation of transcription.

To conclude, we propose assortativity of features in general or in subnetworks as a topology-aware metric that can be used to compare multiple contact maps and to assess their biological relevance.
SB14: Integrating transcriptomic and proteomic data with predictive regulatory network models of host response to pathogens

Deborah Chasman\(^1\), Kevin B. Walters\(^2\), Tiago J.S. Lopes\(^2\), Dylan Pozorski\(^3\), Amie J. Eisfeld\(^2\), Yoshihiro Kawaoka\(^2\), Sushmita Roy\(^1\)

\(^1\) Wisconsin Institute for Discovery, University of Wisconsin-Madison, United States; \(^2\) School of Veterinary Medicine, Dept. of Pathobiological Sciences, Influenza Research Institute, University of Wisconsin-Madison, United States; \(^3\) Department of Computer Sciences, University of Wisconsin-Madison, United States

An emerging challenge in infectious disease research is to integrate diverse types of functional genomics datasets to identify important gene and pathway-level drivers of mammalian host response. Such datasets include genome wide measurements of host response to infectious agents at the mRNA, protein, and metabolite levels. We present a novel computational approach based on a predictive regulatory network model to integrate host mRNA levels, protein levels and protein-protein interactions and to identify important regulators of host response to pathogenic infections. We demonstrate our approach on transcriptomic and proteomic measurements of host response to multiple Influenza A viruses in two study systems: human Calu-3 cell line and mouse lung. Using our approach, we inferred integrated regulatory programs for modules that capture strain- and pathogenicity-specific patterns of mammalian immune response. Our regulatory networks and modules are significantly enriched for known pathways of immune response and additionally implicate a crosstalk between apoptosis, inflammation and interferon signaling. RNAi-based validation of our predicted regulators identified five novel regulators that had a significant impact on viral replication demonstrating the potential of our approach to discover novel regulators of host response to pathogenic infections. Our method is generally applicable to integrate multi-level omics data to gain gene, module and pathway level understanding of regulatory networks and to identify important regulators of host response to different types of infectious pathogens.

RG14: Nucleotide Sequence Composition Adjacent to Intronic Splice Sites Improves Splicing Efficiency and Reduces Translation Costs in Fungi

Zohar Zafrir and Tamir Tuller

Tel Aviv University, Israel

RNA splicing is the central process of intron removal in eukaryotes known to regulate various cellular functions. The canonical sequence elements which are essential for intron recognition are well-known.
However, the role of various sequence features affecting splicing efficiency, intronic retention, and translation regulation has yet to be thoroughly studied. Focusing on four fungi as model organisms (S. cerevisiae, S. pombe, A. nidulans, and C. albicans) we performed for the first time a comprehensive high resolution and large scale systems biology study, aimed at characterizing how splicing efficiency of introns and the crosstalk between gene splicing and translation are encoded in transcripts and affect their evolution. Our analysis suggests that pre-mRNA local folding strength at intronic boundaries is under selective pressure, as it directly affects splicing efficiency and improves recognition of intronic boundaries (Yofe* and Zafrir* et al., PLoS Genetic, 2014; Zafrir and Tuller, RNA, 2015). In addition, when considering the reading frame of exons upstream and adjacent to introns we find evidence of preference for intronic STOP codons close to the intronic 5'end and that the beginning of introns are selected for ÕcodonsÓ with higher translation efficiency, presumably to reduce translation and metabolic costs in cases of non-spliced introns. Ribosomal profiling data analysis in S. cerevisiae supports the conjecture that in this organism intron retention frequently occurs; thus, introns are partially translated, and their translation efficiency affects organismal fitness (Zafrir and Tuller, under revision, 2015). These new discoveries are contributory steps towards a broader understanding of splicing regulation, mRNA translation, intron evolution, and the effect of silent mutations on gene expression and organismal fitness.

10:05 AM - 10:20 AM
Independence C/D
SB15: Microbes are STICKY – Large-scale Inference and Topological Analysis of Microbial Interaction Networks
Christian L. Müller¹, Zachary D. Kurtz², Richard Bonneau²
¹ Simons Foundation, United States; ² New York University School of Medicine, United States

In recent years, 16S-rRNA and other environmental sequencing measurements of microbial communities have revealed phylogeny and the abundances of microbial populations across diverse ecosystems. While changes in microbial community composition are demonstrably associated with certain environmental conditions (from metabolic and immunological health in mammals to ecological stability in soils and oceans), the identification of underlying mechanisms and interaction patterns among microbes requires new statistical tools. Key challenges for 16S rRNA data analysis are that (i) the data are typically compositional, i.e., counts of Operational Taxonomic Units (OTUSs) are normalized to the total number of counts in the sample due to limits in sequencing capacity, and (ii) the data are underdetermined, i.e., the number of detected OTUs typically exceeds the number of available samples by an order of magnitude. Thus, traditional statistical measures for the detection of OTU-OTU relationships such as Pearson correlation can lead to spurious results. Our recently introduced SPIEC-EASI (SParse InversE...
Covariance Estimation for Ecological Association Inference) framework (http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004226) addresses both of these issues by combining compositional data transformations with algorithms for sparse neighborhood and inverse covariance selection that also work in the underdetermined regime.

Using the SPIEC-EASI framework we learned sparse interaction networks from a large collection of publicly available data sets across many ecological environments, ranging from gut and freshwater habitats to urban environments. Using concepts from graph theory, statistics, and theoretical ecology, we analyzed the topology of the learned interaction graphs. We identify a number of general network architecture principles that reproduce within and across habitats. We observe that all learned networks share the small-world property, common to many biological networks, and that the networks follow the May-Wigner criterion for ecological stability. In addition, we observe that the STICKY (or Chung-Lu) model is an excellent statistical network model that describes microbial networks accurately across multiple habitats. We believe that these findings shed a new light on the organization principles of microbial communities and present a promising step toward the unification of ecological theory and data-driven systems microbiology.

10:05 AM - 10:20 AM

**RG15: Lobular Scale Spatio-temporal Modeling of Calcium Signal Propagation in the Liver**

Aalap Verma¹,², Anil Antony¹, Joannes Hoek¹, Babatunde Ogunnaike², Rajanikanth Vadigepalli¹

¹ Thomas Jefferson University, United States; ² University of Delaware, United States

Regulation of Ca²⁺ signals in hepatocytes is crucial to normal liver function. Free Ca²⁺ in the intracellular domain regulates a wide range of hepatocyte functions such as metabolism, bile secretion, proliferation and apoptosis. Experimental observations suggest that, in response to extracellular stimuli, the oscillatory intracellular Ca²⁺ release manifests as a Ca²⁺ wave propagating across the liver lobule. Extracellular stimulants such as hormones elicit a cascade involving phospholipase C (PLC) cleavage, IP₃ generation, and subsequent Ca²⁺ release from intracellular stores. Gap-junction mediated transfer of IP₃ molecules between adjacent hepatocytes induces Ca²⁺ oscillations in neighboring hepatocytes which spreads through the tissue. This cell-cell interaction results in communication of extracellular stimulus levels between hepatocytes, leading to a coordinated response at the tissue level for efficient liver function. Disruption of such intercellular interaction may lead to pathological conditions. We stimulated intracellular Ca²⁺ release by treating isolated perfused mouse livers with vasopressin and captured Ca²⁺ wave transduction in liver lobules using confocal intact tissue
imaging techniques. We segmented the images to identify Ca2+ time series profiles in individual hepatocytes, and analyzed the resulting single cell Ca2+ intensity patterns. Our analysis revealed that the direction of signal propagation between adjacent hepatocytes lying along a sinusoid changed frequently during the course of measurement suggesting that a clear wave-like propagation of Ca2+ signals does not occur at the microscopic scale. We built a computational model based on our experimental data to further understand this non-canonical propagation of Ca2+ signals. Based on model simulations and analysis we predicted that differential expression of the hormone receptor in different regions of liver lobules explains the wave-like propagation of Ca2+ signals. Reversal of Ca2+ signal between adjacent hepatocytes at the microscopic scale could be explained by interference of waves starting at different loci and passing through the same hepatocyte. We further compared Ca2+ oscillations between control and chronic alcohol adapted mouse hepatocytes. Based on model simulations, we predict increased PLC activity and reduced intercellular communication between alcohol adapted hepatocytes. Our model thus provides a framework for generating hypotheses on lobular scale changes in intracellular Ca2+ oscillations parameters and intercellular Ca2+ signal propagation characteristics that potentially underlie normal and diseased states.

10:20 AM - 10:35 AM
Independence C/D
SB16: Multi-omics learning and optimal experimental design for microbial organisms
Minseung Kim, Navneet Rai, Violeta Zorraquino, Xiaokang Wang, Ilias Tagkopoulos
University of California at Davis, United States

Accurate prediction of cellular and molecular state in novel environments is one of the grand challenges in modern biology. Despite the availability of omics profiles, it remains unclear how and at what degree their integration can train a predictive model, or how current datasets can guide which new conditions should be investigated. To address these challenges, we developed a framework of omics data integration, predictive modeling and optimal experimental design. We constructed a comprehensive Escherichia coli compendium specifically structured for efficient machine learning. The compendium integrates 4,389 profiles in multiple layers ranging from transcriptome, proteome, metabolome, fluxome, and phenome with in-depth characterization of profiling conditions by 612 features of strain genotypes, chemical composition of medium used, stresses exposed, and genetic perturbations. The compendium was undergone in multi-step procedure of preprocessing to correct for gene-level noises, batch-effects, and platform-biases. We used this resource to train a multi-scale statistical model that integrates four omics layers to predict expression levels of 4096 transcripts, 1001 proteins, 2382 metabolic fluxes and 356 metabolite concentrations as well growth dynamics. To guide future
experimentation, we developed a methodology to identify experiments that optimally sample the experimental space and simultaneously decrease the uncertainty of the model. The proposed methodology takes into account two types of uncertainty in genome-scale prediction; prediction interval from bootstrapped RNNs and entropy estimated by Gaussian process. The genetic and environmental ontology that was reconstructed from the omics data is substantially different and complementary to the ontologies that are traditionally derived by using genetic and chemical information. Predictive performance (PCC) over novel conditions range from 0.54 to 0.87 for the various omics layers and their integration outperformed any single layer for growth rate prediction. Growth prediction of our model was particularly effective for novel wild type conditions (PCC=0.76). The efficacy of optimal experimental design was evaluated over 15 rounds of transcriptional profiling in novel conditions that resulted in a substantial improvement of the performance over alternative methods. The performance of genome-wide expression prediction for the condition space close to optimal conditions newly profiled was substantially improved after refinement (PCC from 0.41 to 0.61) and gradual decrease in uncertainty of the prediction model over the course of 15 rounds was significant than alternatives (P < 0.005). This work provides an integrative framework of omics-driven predictive modeling and experimentation that can be broadly applied to guide biological discovery.

**RG16: Dysregulated transcription factor networks and clusters in breast cancer subtypes, identified by coexpression and cistromic data integration**

Qian Zhu¹, Vessela Kristensen², Olga Troyanskaya¹

1 Princeton University, United States; 2 University of Oslo, Norway

Transcriptional regulation mediated by transcription factor binding provides a key mechanism for the dysregulation of genes, in that they influence gene expression via proximal element (promoter) and distal element (enhancer). Previously, several attempts have been made to construct a transcription factor (TF) network to dissect the genetic programs in breast cancer. However, these approaches have not been disease subtype specific or have been based on promoter sequence evidence alone. In this work, we sought to delineate subtype-specific TF networks in breast cancer by integrating transcriptomic data with experimental cistromic data (Chip-seq) from relevant MCF7, T47D cell lines, an epithelial (A549) and an embryonic stem cell cell-line (H1-HESC) deposited in ENCODE. Unlike previous approaches, our approach importantly is targeted by a coexpressed gene-set, which allows subtype signatures to be defined and subtype network to be constructed.

We leveraged over 170 expression datasets to first identify subtype-specific coexpression signatures, next prioritized 160 Chip-seq experiments
from cell lines to identify TFs which may govern coexpressed genes, and finally utilized these Chip-seq data to identify protein-DNA regulatory interactions between TFs in each subtype. We witnessed extensive clustering of in vitro TF motifs, and confirmed expected motifs enriched by coexpressed genes’ enhancer regions, which provide a sequence basis for protein-DNA interactions.

Importantly, the networks of TF illuminated the luminal, basal, and stem cell lineages involved in breast cancer subtypes. Transcriptional hierarchies were constructed and delineated the early molecular factors leading to the development of lineages. When we investigated the TFs’ deregulation potential, we identified 9 groups of TFs based on their shared profiles of deregulations across subtypes (i.e., SNPs, somatic mutations, DNA methylation and copy number aberrations), suggesting that these 9 TF clusters are consistently targeted by genetic and epigenetic alterations. Not only did the TF clusters illustrate deregulation differences among subtypes, but their patterns of deregulation were also highly reproducible in held-out datasets. The TF clusters included basal stem cell initiators SOX9, EN1, luminal progenitor GATA3, and luminal markers ESR1, BHLHE40, among others. Overall, we illustrate the informative value of a combined analysis involving coexpression and epigenetic, genetic alteration data in the study of subtype-specific transcriptional dysregulations.

11:50 AM - 12:05 PM
Independence C/D
RG17: Transcription and differential DNA repair underlies promoter mutation hotspots in cancer genomes
Dilmi Perera, Rebecca Poulos, Anushi Shah, Dominik Beck, John Pimanda, Jason Wong
University of New South Wales, Australia

Promoters are DNA sequences that play an essential role in controlling gene expression. While recent whole cancer genome analyses have identified numerous hotspots of somatic mutations within promoters, most do not appear to be functional as they do not perturb gene expression. As such, positive selection does not adequately explain the frequency of promoter mutations in cancer genomes. Our analysis indicates that increased mutation density at gene promoters is in fact linked to transcriptional activity and differential DNA repair.

By analyzing 1,163 cancer genomes, we found evidence for increased local density of somatic point mutations within the DNase I hypersensitive centre of gene promoters across 14 cancer types. Mutated promoters were strongly associated with transcriptional activity, with mutation density highest within transcription factor binding sites. By analysing genome-wide maps of nucleotide excision repair (NER), we found that NER is impaired within the DNase I hypersensitive centre of active gene promoters, inversely mirroring the increase in somatic mutation density.
Taken together, our analysis has uncovered the presence of a previously unknown mechanism linking transcription initiation and DNA repair, thereby implicating localised differential DNA repair as the underlying cause for the somatic mutation hotspots observed at gene promoters of cancer genomes.

11:50 AM - 12:05 PM

RG18: The regulation of distal enhancers and silencers from the Cebpa locus during hematopoiesis, inferred using a transcriptional model

Manu\textsuperscript{1}, John Reinitz\textsuperscript{2}, Eric Bertolino\textsuperscript{1}
\textsuperscript{1} University of North Dakota, United States; \textsuperscript{2} University of Chicago, United States

CCAAT/enhancer binding protein, alpha (C/EBP\textsubscript{a}) plays an instructive role in the macrophage-neutrophil cell-fate decision and its expression is necessary for neutrophil development. How Cebpa itself is regulated in the myeloid lineage is not known. We decoded the cis-regulatory logic of Cebpa, and two other myeloid transcription factors, Egr1 and Egr2, using a combined experimental-computational approach. We designed reporter genes capable of detecting both distal enhancers and silencers. We measured the activity of 46 putative cis-regulatory modules (CRMs) lying in 40-80 kb regions surrounding the three genes in cells representing myeloid progenitors, and derived early macrophages or neutrophils. This analysis revealed both enhancers and silencers in the vicinity of Cebpa, although the latter outnumber the former. The abundance of silencers is consistent with the downregulation of endogenous Cebpa in the differentiated cell types and in non-myeloid hematopoietic lineages. In a novel application of sequence-based transcriptional models, we determined the nas yet unknown regulatory roles of 15 potential transcriptional regulators of Cebpa CRMs. This was achieved by testing 32,768 alternative transcriptional models, representing all possible role combinations, against CRM activity data. Analysis of the models allowed us to infer the cis-regulatory logic of how the expression patterns of transcription factors (TFs) in the cell types control activity patterns of most of the CRMs. In particular, we found that silencer-mediated repression of Cebpa is effected mainly by TFs expressed in non-myeloid lineages. These results imply that cross-lineage antagonism is likely to be mediated by long-distance repression and, since silencers are yet to be characterized for most genes, is more widespread than previously believed. More generally, our results demonstrate that de novo cis-regulatory dissection is feasible on a large scale with the aid of transcriptional modeling.
RG19: Comparison of Methods to Predict Impact of Regulatory Variants

Felix Yu, Dongwon Lee, Michael Beer
Johns Hopkins University, United States

The vast majority of sequence variants associated with common human disease are intergenic, enriched in open chromatin regions, and likely regulatory. To identify functional variants within GWAS associated LD blocks, we have developed a sequence-based model based on our gapped k-mer SVM (gkm-SVM) (Lee et al., Nature Genetics, 2015; Ghandi et al., PLOS Comp Biol 2014). This approach uses cell-type specific epigenetic data to train a gkm-SVM whose scoring function encodes the relative regulatory importance of individual sequence features in the disease relevant cell-type. The change in sequence feature scores induced by a regulatory variant determines its predicted impact, a score we call deltaSVM. We have shown that deltaSVM is roughly 10x more accurate at predicting dsQTLs than other methods (Kircher et al., Nat Gen 2014; Ritchie et al., Nat Meth 2014) and our previous kmer-SVM (Lee et al., Gen Res 2011). We have also used deltaSVM to predict the expression change in massively parallel reporter assays, which shows good agreement with high throughput datasets in mouse liver (Patwardhan et al., Nat Biotech 2012), K562 cells, and HepG2 cells (Kheradpour et al., Gen Res 2013). Here, we compare the accuracy of deltaSVM to other computational approaches, including PWMs, other kmer-based approaches, and deep neural networks (Zhou and Troyanskaya, Nat Meth 2015).

RG20: Super-enhancers delineate disease-associated regulatory nodes in T cells

Golnaz Vahedi¹, Vittorio Sartorelli², John O’Shea²
1 University of Pennsylvania, United States; 2 NIH, United States

Enhancers regulate spatiotemporal gene expression and impart cell-specific transcriptional outputs that drive cell identity. Super-enhancers (SEs), also known as stretch-enhancers, are a subset of enhancers especially important for genes associated with cell identity and genetic risk of disease. CD4(+) T cells are critical for host defence and autoimmunity. Here we analysed maps of mouse T-cell SEs as a non-biased means of identifying key regulatory nodes involved in cell specification. We found that cytokines and cytokine receptors were the dominant class of genes exhibiting SE architecture in T cells. Nonetheless, the locus encoding Bach2, a key negative regulator of effector differentiation, emerged as the most prominent T-cell SE, revealing a network in which SE-associated genes critical for T-cell biology are repressed by BACH2. Disease-associated single-nucleotide polymorphisms for immune-mediated disorders, including rheumatoid arthritis, were highly enriched for T-cell SEs versus
typical enhancers or SEs in other cell lineages. Intriguingly, treatment of T cells with the Janus kinase (JAK) inhibitor tofacitinib disproportionately altered the expression of rheumatoid arthritis risk genes with SE structures. Together, these results indicate that genes with SE architecture in T cells encompass a variety of cytokines and cytokine receptors but are controlled by a 'guardian' transcription factor, itself endowed with an SE. Thus, enumeration of SEs allows the unbiased determination of key regulatory nodes in T cells, which are preferentially modulated by pharmacological intervention. Super-enhancers delineate disease-associated regulatory nodes in T cells.


12:20 PM - 12:35 PM

RG21: Evaluating Genetic Variation Impact on Transcription Factor Binding Sites

Wenqiang Shi, Oriol Fornes, Wyeth Wasserman
University of British Columbia, Canada

Current clinical sequence analysis focuses on exomes that highlight protein coding regions, despite awareness that cis-regulatory variations can cause human genetic disorders. Genome-wide association studies have identified thousands of disease-related variations most of which fall within cis-regulatory regions. Whole genome sequencing is now widely used in clinical genetics research, but the bioinformatics methods for the identification of functional regulatory changes are inadequate. The need to interpret and prioritize regulatory variations is becoming urgent for clinical genome analysis.

In this project, we focus on prioritizing variations likely to disrupt transcription factor binding sites in cis-regulatory elements. In developing the methods for cis-regulatory sequence analysis, we focus on differential transcription factor (TF) binding between two alleles distinguished by single nucleotide alterations. In ChIP-seq data, allele-specific binding (ASB) events, which indicate a TF selectively binds to one of two alleles at heterozygous positions, directly reveal the impact of cis-regulatory variation on TFBS within the same cellular context. We extracted ASB events from ENCODE ChIP-Seq data coupled with available WGS data in the corresponding cells. This key ASB reference collection exhibits a strong relationship between the predicted strength of TF-DNA interactions (as scored with position weight matrices (PWM)) and observed TF binding in vivo. DNase I accessibility differences between two alleles are also strongly associated with TF binding difference across multiple TFs and cells. In a TF-specific manner, cofactors can be quantitatively identified based on the differential overlap of cofactor ChIP-seq peaks between ASB and non-ASB events. Combining the available feature data, a classifier model trained to distinguish between ASB and non-ASB events achieves good accuracy (e.g. 78% for CTCF).
RG22: Identifying condition specific transcription factor binding with ATAC-seq

Roger Pique-Regi, Donovan Watza, Molly Estill, Sophia Chaudhry, Francesca Luca
Wayne State University, United States

Specific regulatory sequences control gene transcription response when a cell is exposed to changes in the cellular environment (e.g. drug treatment). Recent technical advances in functional genomics have facilitated the profiling of regulatory sequences across many cell-types and tissues, yet we are still very far from mapping the sequences that control cell transcriptional response to many external stimuli. Profiling across different environmental conditions the binding activity of these TFs can be quickly accomplished at a genome-wide scale with the recently developed technique ATAC-seq, which utilizes the Tn5 transposase to fragment and tag accessible DNA. When coupled with a computational method such as CENTIPEDE, footprint models for TFs with known motifs can be generated across the genome to detect binding. To date, there are no methods that efficiently incorporate the information provided by paired-end sequencing which allows both the identification of the library fragment length as well as the two cleavage locations that generated the fragment. We have extended CENTIPEDE to utilize fragment length information to exploit the joint statistics of cleavage pairs. Our results indicate that paired-end sequencing provides a more informative footprint model for ATAC-seq libraries which leads to greater accuracy in predicting TF binding. These results were validated with ChIP-seq data (ENCODExProject) for multiple factors including CTCF, NRSF, NRF-1, and NFkB. We then assayed TF activity in lymphoblastoid cell-lines (LCLs) across multiple treatments (selenium, copper, retinoic acid and glucocorticoids) for which we previously determined significant differences in gene expression levels. From our initial sequencing results we were able to resolve 383 actively bound motifs across all conditions. We were also able to characterize 5236 regions that have significantly changed chromatin accessibility (FDR < 10%) in response to both copper and selenium. We have extended the CENTIPEDE model hierarchical prior to detect motifs that have differences in footprint activity in treatment vs. control experiments. For both metal ions we have detected a significant increase of binding for ETS and CRE motifs. Our results demonstrate that ATAC-seq together with an improved footprint model are excellent tools for rapid profiling of transcription binding factor activity to study cellular regulatory response to the environment.
SB17: tRNA-derived fragments in Drosophila and their potential targets

Spyros Karaiskos, Ammar Naqvi, Karl Swanson, Andrey Grigoriev
Rutgers University, United States

In our paper, “Age-driven modulation of tRNA-derived fragments in Drosophila and their potential targets” (1), we presented rather unusual suspects in the world of gene regulation – transfer RNA fragments (tRFs). Such fragments have previously escaped detection or have been generally ignored as noise due to low count numbers in of small RNA libraries. While the focus in the analysis of such libraries has been primarily on microRNAs (miRNAs), recent studies have reported findings of fragments of transfer RNAs (tRFs) across a range of organisms.

We had originally investigated the age-related dynamics of miRNA loading into different RNA-induced silencing complexes (RISC) of Drosophila melanogaster using small RNA-Seq (2). Here we describe tRFs found in the same transcriptome libraries, and focus on their structural and functional features that make these fragments similar to miRNAs. Similar to miRNAs, tRFs have distinct isoforms with precise ends preferentially originating from 5' or 3' end of a precursor molecule (tRNA). Analogously to the seed sequences in miRNAs, we observe that tRF ends possess short 7-mer sequences matching conserved regions across 12 Drosophila genomes, preferentially in 3' UTRs but also in introns and exons. Like miRNAs, tRFs display specific isoform loading into Ago1 and Ago2 and thus likely function in RISC complexes. And finally, as is the case with miRNAs, we observe the levels of tRF loading into Ago1 and Ago2 to differ considerably and both tRF expression and loading appear to be age-dependent, indicating potential regulatory changes from young to adult organisms.

We found that Drosophila tRF reads mapped to tRNA genes for all 20 amino acids, while previous studies have usually reported fragments from only a few tRNAs. Moreover, we detected fragments of both nuclear and mitochondrial tRNAs, while only the former have been described. Following the similarities with miRNAs and based on complementarity with conserved Drosophila genome regions we described seed sequence found in the most abundant tRFs. Further, we identified their possible targets with matches in the Drosophila melanogaster 3'UTR regions. Strikingly, these potential target genes of the most abundant tRFs show significant Gene Ontology enrichment in development and neuronal function. This observation suggests that involvement of tRFs in the RNA interfering pathway may play a role in brain activity or brain changes with age.

RG23: ReMiCs: Regularized Regression Model for Inference of Condition-specific Gene Regulatory Networks from Multiple Data Sets

Nooshin Omranian and Zoran Nikoloski
Max-Planck Institute of Molecular Plant Physiology, Germany

Reconstruction of gene regulatory networks (GRNs) from gene expression levels alone often yields consensus regulatory interactions over multiple data sets. Yet, insights in the molecular basis of an organism’s response to developmental and environmental changes can only be obtained by identifying condition-specific GRNs. Here we propose a two-step approach, termed ReMiCs, based on regularized regression to infer condition-specific GRNs: In the first step, we determine the consensus regulatory interactions by employing least absolute shrinkage and selection operator (LASSO) with fusion penalty. The consensus GRN is reconstructed from multiple data sets by imposing three constraints modeled as fused LASSO: (1) sparsity of GRNs, (2) similarity of GRNs inferred from individual data sets, and (3) concordance of differential behavior between regulatory gene and its targets. In the second step, we determine the residual of each condition-specific data set upon removing the consensus effects found in the first step. We then infer the condition-specific interactions by employing weighted LASSO regression on the respective residual data set. We tested the performance of ReMiCs on two transcriptomics time-series data from Mycobacterium tuberculosis under hypoxia and reaeration as well as on transcriptomics time-series data from Arabidopsis thaliana under eight combinations of light and temperature perturbations. The comparative analysis against condition-specific regulatory interactions extracted from literature demonstrated the advantages of the ReMiCs over the existing state-of-the-art approaches applied on condition-specific data sets. The approach can readily be extended to include ChIP-Seq data used for pre-selecting likely interactions. The study indicates that the combination of sparse regression techniques with other biologically meaningful constraints is a promising framework for reconstructing consensus as well as condition-specific GRNs solely based on transcriptomics data.
SB18: High-throughput allele-specific expression across 250 environmental conditions

Gregory Moyerbrailean1, Chris Harvey1, Omar Davis1, Adnan Alazizi1, Donovan Watza1, Yoram Sorokin1, Karoline Pruder1, Nancy Hauff1, Xiaoquan Wen2, Roger Pique-Regi1, Francesca Luca1
1 Wayne State University, United States; 2 University of Michigan, United States

Adaptations to local environments have played major roles in shaping allele frequency distributions in human populations. Yet, a mismatch between genotype and environment may be responsible for higher disease risk. Recent studies have shown that GxE interactions can be detected when studying molecular phenotypes that are relevant for complex traits (e.g. infection response eQTLs in immune cells). Despite these relevant examples, the extent to which the environment can modulate genetic effects on quantitative phenotypes is still to be defined. Here we have taken a high-throughput approach to achieve a comprehensive characterization of GxE interactions in humans through allele-specific expression (ASE) analysis. To this end we have investigated the transcriptional response to 50 treatments in 5 different cell types (for a total of 250 cellular environments and 3 individuals per cell type). Across 56 cellular environments (cell type/treatment with large changes in gene expression) we discovered 6073 instances of ASE (FDR<10%), corresponding to 4310 unique genes. We found that in an individual sample, on average, 0.5% of genes with heterozygous SNPs are ASE genes. We observe that the majority of ASE is consistent across conditions (“shared” ASE), confirming previous conditional eQTL analyses. Overall, we find 248 loci with evidence for GxE interaction (conditional ASE), 120 with control-only ASE and 128 with treatment-only ASE genes. We used a multinomial generalized linear model with elastic net regularization (glmnet) to assess which factors influence the likelihood of conditional ASE. This model allows us to control for factors that may influence ASE and potential confounders (e.g., gene expression, cell type, treatment). Cell type seems to be an important factor for shared ASE: Melanocytes show a 30% increase in the probably of ASE, while LCLs show an 18% reduction. When we focus on treatment-only ASE, there are significant differences across treatments but these are largely explained by the changes in gene expression. For genes that are differentially expressed, each 2-fold increase in gene expression response corresponds to a 2.22-fold increase in the probability of treatment-only ASE. Finally, integrating our results with data from 18 traits from GWAS meta-analysis revealed enrichments for genes differentially expressed in specific treatments. For example, variants associated with Crohn’s disease are enriched in genes that respond to aspirin in PBMCs and HUVECs, thus identifying candidate genes for aspirin aggravating effects on Crohn’s symptoms.
RG24: Transcriptional regulatory network inference for rare immune cell populations from gene expression and chromatin accessibility measurements

Emily Miraldi1, Maria Pokrovskii1, Jason A. Hall1, Nicholas Carriero2, Dan R. Littman1, Richard Bonneau2
1 NYU School of Medicine, United States; 2 Simons Center for Data Analysis, United States

Innate lymphoid cells (ILCs) compose a newly discovered and relatively rare population of immune cell lineages (ILC1, ILC2, ILC3) that play many important roles, including early defense against infections, tissue homeostasis and repair, and autoimmune disease. These diverse ILC physiological functions require coordination of complex gene expression patterns, involving thousands of genes. Our goal here is the inference of transcriptional regulatory networks (TRNs) to model these gene expression responses as multivariate functions of transcription factor (TF) activities. Although TF mRNA can serve as a proxy for TF activity, this assumption is imperfect, given the importance of downstream processing (e.g., translation, chemical modification) to TF. Other groups have proposed using prior knowledge of TF target gene expression to estimate TF activity, and the Bonneau lab has recently integrated this TFA estimation procedure into our parsimonious TRN inference method, the Inferelator, yielding great performance advantages in the model organism B. subtilis, for which a database of known TF-target interactions exists (Ortiz, Hafemeister, et al., submitted). Unfortunately, very few transcriptional regulatory interactions are known for ILCs, and, thus, we developed a method to derive a OpriorÓ of putative TF-gene interactions from ATAC-seq chromatin accessibility measurements and gene expression in ILCs. The ATAC-seq data, when integrated with TF motif information, can provide a list of TF candidate regulators that bind cis to target genes. However, given that motifs among TF families can be degenerate and often more than one TF in a given family is expressed per condition, a key challenge with ATAC-seq data is determining which transcription factors (TFs) contribute to accessible regions (e.g., peaks) for a given experimental condition. Here, I develop methods to harness the quantitative signal in ATAC-seq datasets across conditions to best ascribe ATAC-seq signal to the activity of a single TF. These ATAC-seq-derived priors then serve as input to the Inferelator, which uses context likelihood of relatedness (CLR) and Bayesian best subset regression (BBSR) to infer TRNs from gene expression and transcription factor activity estimates.

Importantly, we validate our integrative method for TRN inference from ATAC-seq and gene expression data in the more well-studied context of in vitro Th17 cells, for which we also generated ATAC-seq data and for which a gold-standard network based on TF knockout and ChIP-seq data exists. We then apply our methods to ILCs and use the resulting TRNs to understand ILC regulation in host health and pathophysiology.
SB19: DREISS: dynamics of gene expression driven by external and internal regulatory networks based on state space model

Daifeng Wang¹, Fei He², Sergei Maslov², Mark Gerstein¹
1 Yale University, United States; 2 Brookhaven National Laboratory, United States

Gene expression is controlled by combinatorial effects of regulatory factors from different biological subsystems driving specific regulatory functions such as general transcription factors, cellular growth factors and microRNAs. A subsystem’s gene expression may be controlled by its internal regulatory factors, exclusively, or by other external subsystems, or by both. It is thus useful to distinguish the degree to which a subsystem is regulated internally or externally – e.g., how species-specific regulatory factors affect the expression of conserved genes during evolution.

We developed a computational method (DREISS) for dynamics of gene expression driven by external and internal regulatory modules based on state space model to help dissect the effects of different regulatory subsystems on gene expression (dreiss.gersteinlab.org) [1]. Given a subsystem, the “state” and “control” in the model refer to its own (internal) and another subsystem’s (external) gene expression levels. The state at a time is determined by the state and control at previous time. Because typical time-series data do not have enough samples to estimate the model’s parameters, DREISS uses dimensionality reduction, and identifies canonical temporal expression trajectories (e.g., degradation, growth, oscillation) representing the regulatory effects coming from various subsystems.

To demonstrate capabilities of DREISS, we study the regulatory effects of evolutionary conserved vs. divergent transcription factors across distant species. In particular, we applied it to the time-series gene expression datasets of C. elegans and D. melanogaster during their embryonic development. We analyzed the expression dynamics of the conserved, orthologous genes (orthologs), seeing the degree to which these can be accounted for by orthologous (internal) versus species-specific (external) transcription factors (TFs). We found that between two species, the orthologs have matched internally driven expression dynamic patterns but very different externally driven patterns. This is particularly true for genes with evolutionarily ancient functions (e.g. the ribosomal proteins), in contrast to those with more recently evolved functions (e.g., cell-cell communication). This suggests that despite striking morphological differences, some fundamental embryonic-developmental processes are still controlled by ancient regulatory systems.

RG25: An Experimentally Supported Model of the Bacillus subtilis Global Transcriptional Regulatory Network

Mario L. Arrieta1, Christoph Hafemeister1, Ashley Rose Bate1, Timothy Chu1, Alex Greenfield1, Bentley Shuster1, Samantha N. Barry1, Matthew Gallitto1, Brian Liu1, Thadeous Kacmarczyk1, Francis Santoriello1, Jie Chen1, Christopher D.A. Rodrigues2, Tsutomu Sato3, David Z. Rudner2, Adam Driks4, Richard Bonneau1, Patrick Eichenberger1
1 New York University, United States; 2 Harvard Medical School, United States; 3 Hosei University, Japan; 4 Loyola University Chicago, United States

Organisms from all domains of life use gene regulation networks to control cell growth, identity, function, and responses to environmental challenges. Although accurate global regulatory models would provide critical evolutionary and functional insights, they remain incomplete, even for the best studied organisms. Efforts to build comprehensive networks are confounded by challenges including network scale, degree of connectivity, complexity of organism-environment interactions, and difficulty of estimating the activity of regulatory factors. Taking advantage of the large number of known regulatory interactions in Bacillus subtilis and two transcriptomics datasets (including one with 38 separate experiments collected specifically for this study), we use a new combination of network component analysis and model selection to simultaneously estimate transcription factor activities and learn a substantially expanded transcriptional regulatory network for this bacterium. In total, we predict 2258 novel regulatory interactions and recall 74% of the previously known interactions. We obtained experimental support for 391 (out of 635 evaluated) novel regulatory edges (62% accuracy), thus significantly increasing our understanding of various cell processes, such as spore formation.

SB20: SplashRNA: accurate prediction of potent shRNAs with a sequential classification strategy

Raphael Pelossof1, Lauren Fairchild1, Christina Widmer2, Chun-Hao Huang1, Vipin Sreedharan1, Qing Xiang1, Ralph Garippa1, Gunnar Ratsch1, Scott Lowe1, Christof Fellmann3, Christina Leslie1
1 MSKCC, United States; 2 Activision, United States; 3 UC Berkeley, United States

Short hairpin RNAs (shRNAs) are a ubiquitously used experimental tool for knocking down genes, including now in CRISPR-shRNA combination models. Therefore, prediction of potent shRNA sequences is a critical computational problem. We have developed a novel machine learning approach to accurately predict potent shRNAs by training on massively parallel sensor assay data, pooled screens of RNAi reporters that identify shRNAs with potent knock-down. Our method, SplashRNA, achieves more than double the accuracy of DSIR, a standard algorithm in the field.
in a blind sensor screen that tested the potency of 250,000 shRNAs, and 20% improvement in accuracy against DSIR and Sherwood, a commercial algorithm, on blind cell viability screens for 2,000 shRNAs.

shRNA pooled screens are designed as sequential enrichment screens of RNAi reporters, where each reporter consists of an shRNA (under control of doxycycline) and a sensor targeted by the shRNA (in the 3′UTR of a reporter gene) expressed together in a single cell, so that on-dox fluorescence enables sorting for potent shRNAs. The first published sensor assay screen started with an unbiased input library of shRNAs, and analysis of shRNAs that passed the screen established potency rules. Due to the high costs of such screens, later screens are designed to test biased sets of shRNA that are likely to be potent given the learned potency rules. We have generated such a screen for the training of our algorithms, testing approximately 20,000 shRNAs that were split into 64 shRNAs per 300 genes.

Our final classifier sequentially aggregates classifiers in a soft cascade scheme. In this learning methodology each classifier has the option either to pass the shRNA to the next classifier or to stop and assign a weaker potency score to the shRNA. This approach is beneficial when each screen only tests a subset of the shRNAs, and therefore we would design the first classifier to screen unbiased sequences, and the second consecutive classifier to classify more potent shRNA sequences that it receives from the first. Each one of the classifiers of the cascade is trained with an SVM and uses a spectrum kernel and a weighted-degree kernel as feature representations for shRNA sequences.

Validation of SplashRNA on PTEN in 3T3 cells by western blot, shows that we are able achieve perfect classification ranking, implying that SplashRNA may have fully learned sequence specific potency rules of cellular shRNA processing. An implementation of our algorithm is publicly available at splashrna.mskcc.org.

**RG26: Network model of normal gene expression predicts gene perturbation fold changes**

**Sudhir Varma**

*HiThru Analytics, United States of America*

Gene expression exhibits a network effect whereby perturbations of some genes (in the form of siRNA knockdown or drug treatment in vitro or mutations, changes in methylation or aneuploidy in disease) influence the expression of downstream genes. Thus the downstream genes are simply responding to the dysregulation of the root-cause genes and are themselves not the source of the perturbation. If the expression (or fold change) of a gene is explained by the expression (or change in expression) of other genes, it becomes less likely to be the source of the perturbation.
Conversely, genes with a large positive or negative difference between the predicted and actual expression show evidence of being the main drivers in the experimental or disease condition.

Using a set of 4277 normal samples from various organs compiled from public datasets, we have built a network where the expression of each gene (target gene) is modeled as a linear combination of a small number of other genes (source genes). A single model for each gene was fitted to samples from all organs. For each organ we tested the fit of the model using the correlation between the predicted and actual values. We used the network model to predict the fold changes using a set of 658 siRNA knockdown samples.

The network model predicts the expression of a median of 42% (20%-71%) of variable-expression genes (log expression range>0.5) in all of the organs with a correlation>0.80. On the siRNA knockdown samples, the model predicted the resulting fold changes of all genes with a median correlation of 0.31 across the samples (0.13-0.68).

We demonstrate that a single linear regression model (per gene) is sufficient to predict the expression of most genes for multiple organs. The relationships between source genes and target gene defines a network which is capable of quantitatively predicting the downstream effects of a perturbation. Conversely, the difference between the predicted and true expression in a disease sample points to possible root causes of the disease.

We have implemented a web tool for exploring the network predictions on a variety of disease samples (http://www.explainbio.com).
Dream Challenges Posters

DR01: From Shape to Smell: Predicting Olfactory Perceptual Descriptors using Molecular Structural Information

Richard C Gerkin
Arizona State University, United States

This poster is also presented as a talk. The abstract is on page 29.

DR02: Using Big Data and Crowdsourcing for Catalyzing Breakthroughs in ALS

Neta Zach1, Robert Küffner2, Hagit Alon1, Nazem Atassi3, Barbara di Camillo4, Merit Cudkowicz5, Javier Garcia-Garcia5, Orla Hardiman6, Guang Li7, Lara Mangravite8, Raquel Norel9, Thea Norman8, Alexander Sherman9, Liuxia Wang7, Gustavo Stolovitzky9
1 Prize4Life, Israel; 2 Helmholtz Center, Germany; 3 Massachusetts General Hospital, United States; 4 University of Padova, Italy; 5 Universitat Pompeu Fabra, Spain; 6 Beaumont Hospital and Trinity College Dublin, Ireland; 7 Origent Data Solutions, United States; 8 Sage Bionetworks, United States; 9 IBM, United States

This poster is also presented as a talk. The abstract is on page 30.

DR03: DREAMTools: a Python Package for scoring collaborative challenges

Thomas Cokelaer1, Mukesh Bansal2, Christopher Bare3, Erhan Bilal4, Brian M. Bot5, Elias Chaibub Neto5, Federica Eduati1, Mehmet Gönen5, Steven Hill6, Bruce Hoff3, Jonathan R. Karr7, Robert Küffner8, Michael Menden1, Pablo Meyer4, Raquel Norel4, Abhishek Pratap5, Robert J. Prill9, Matthew T. Weirauch10, James C. Costello11, Gustavo Stolovitzky4, Julio Saez-Rodriguez12
1 EMBL-EBI, United Kingdom; 2 Department of Systems Biology, Columbia University, United States; 3 Sage Bionetworks, United States; 4 IBM, TJ Watson, Computational Biology Center, United States; 5 Oregon Health & Science University, United States; 6 MRC Biostatistics Unit, Cambridge Institute of Public Health, United Kingdom; 7 Department of Genetics & Genomic Sciences, Icahn School of Medicine at Mount Sinai, United States; 8 Institute of Bioinformatics and Systems Biology, German Research Center for Environmental Health, Germany; 9 IBM Almaden Research Center, San Jose, United States; 10 Center for Autoimmune Genomics and Etiology and Divisions of Biomedical Informatics and Developmental Biology, Cincinnati Children’s Hospital, United States; 11 Department of Pharmacology, University of Colorado Anschutz Medical Campus United States; 12 RWTH Aachen University Medical Hospital, Germany

ABSTRACT: DREAM challenges are community competitions designed to advance computational methods and address fundamental questions in system biology and translational medicine. Each challenge asks participants to develop and apply computational methods to either predict unobserved outcomes or to identify unknown model parameters given a set of training data. Computational methods are evaluated using an automated scoring metric, scores are posted to a public leaderboard, and methods are published to facilitate community discussions on
how to build improved methods. By engaging participants from a wide range of science and engineering backgrounds, DREAM challenges can comparatively evaluate a wide range of statistical, machine learning, and biophysical methods. Here, we describe DREAMTools, a Python package for evaluating DREAM challenge scoring metrics. DREAMTools allows one to reproduce results from past DREAM challenges. The software also provides a command line interface that enables researchers to test new methods on past challenges, as well as a framework for scoring new challenges. As of September 2015, DREAMTools includes more than 80\% of completed DREAM challenges. DREAMTools complements the data, metadata, and software tools available at the DREAM website (http://dreamchallenges.org) and on the Synapse platform (https://www.synapse.org). In the poster, we will give an overview of the past and present challenges and how the DREAMTools package can be used to reproduce scores from previous competitions. We will also describe the scoring functions that are currently available within the package and how new challenges can be included into the package.

**DR04: A Two-layer Predictor for DREAM 9.5 Olfaction Prediction Challenge**

Ping-Han Hsieh¹, Bor-Wei Cherng², Yu-Chuan Chang¹, Ming-Yi Hong¹, Yi-An Tung³, Yen-Jen Oyang⁴, Chien-Yu Chen⁵

¹ National Taiwan University, Taipei, Taiwan; ² National Taiwan University and Academia sinica, Taipei, Taiwan; ³ Genome and System biology program, National Taiwan University and Academia sinica, Taipei, Taiwan; ⁴ Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, Taipei, Taiwan; ⁵ Department of Bio-Industrial Mechatronics Engineering, National Taiwan University, Taipei, Taiwan

**ABSTRACT:** Olfaction is one of the most important sensibilities in animal behavior. Understanding the olfactory precept from the aspect of molecular properties may broaden our understanding of sensory cognition and create more odorant applications for industries. Though there have been many published methodologies in predicting human’s olfactory flavor, the prediction accuracy is expected to be further improved by additional training data sets. This challenge was designed to predict personal olfactory flavor based on chemical features of compounds. Based on the concept of the one-neuron-receptor role in olfactory sensory studies, odor molecules conjugate with specific types of olfactory receptors, and transmit neuron signals to multiple brain regions to generate the olfactory perception. As the result of neuronal circuitry conservation, it is reasonable to hypothesize that there are hidden relations between olfactory perception and the molecular properties, making us precisely process smell perception. To tackle this olfaction prediction challenge, we built up a machine learning-based pipeline to generate individual-specialized ensemble-based linear models, modified from the PCA-based baseline model delivered by the challenge organizers. The adopted features include molecular physicochemical properties from
Dragon molecular descriptors, while the target information came from perceptual data collected in Rockefeller University Smell Study. Due to the sensory varieties between individuals, we, therefore, hypothesized that the predicting models for different persons may slightly differ from each other, though it may still have some common factors embedded in the models. In this regard, we built predicting models for each single person with respect to each chemical compound. It is expected that the individual-specific features were carefully selected by the ensemble approach. The proposed method also employed a second-layer predicting framework to predict the target value: ‘valence/pleasantness’. The results revealed that the two-layer approach performed better than the conventional design, the single-layer framework, suggesting that some of the olfactory senses are highly related.

**DR05: Attractor Metafeatures Discover Molecular Signatures for Odor Perception Prediction**

**Andrew Matteson**  
*Applied BioMath, United States*

**ABSTRACT:** Chemoinformatics predictors are generally more numerous than the number of samples in a target variable. The key challenge in working with these imbalanced data sets is avoiding over-fitting. My approach sought to preserve as much information about the target variable while reducing the dimensionality of the data. My methods focused on the use of mutual information to either select features, or project all the features onto a space that was a good predictor of the target variables. For selection, my methods are connected to the technique of maximum-relevance-minimum-redundancy (MRMR) feature selection.

I engineered “metapredictors” from the chemoinformatics predictors using the “Attractor Metagene” algorithm (1). The algorithm engineers features in an unsupervised way by weighted averaging over the original features. Weights are chosen as a function of the mutual information between the engineered feature, and the original features.

Several of the metapredictors map onto known chemical structures associated with particular smells. Other metapredictors correspond to chemical structures not yet identified as corresponding to odors. These metapredictors enable connections to be made between molecular structure and perception that are directly interpretable.

The learning methods I developed fall into an enrich-project-predict framework. I discuss opportunities to extend these methods to other chemoinformatics and bioinformatics machine learning problems.

DR06: Reflecting on the Prostate Cancer Dream Challenge: Lessons Learned

Team Jayhawks: Devin C. Koestler¹, Joseph Usset¹, Stefan Graw¹, Richard Meier¹, Rama Raghavan¹, Junqiang (Eric) Dai¹, Prabhakar Chalise¹, Shellie Ellis², Brooke L. Fridley¹

¹ Department of Biostatistics, University of Kansas Medical Center, Kansas City, United States; ² Department of Health Policy and Management, University of Kansas Medical Center, Kansas City, United States

From March through August 2015, nearly 60 teams from around the world participated in the Prostate Cancer Dream Challenge, cosponsored in part by: the Prostate Cancer Foundation, National Cancer Institute (NCI), and the American Joint Committee on Cancer (AJCC). Participating teams were faced with the task of developing prediction models for patient survival and treatment toxicity using clinical variables collected from the comparator arms of four phase III clinical trials, including over 2,000 metastatic castrate resistant prostate cancer patients treated with first-line docetaxel. In this poster presentation, we describe: (a) the 3 sub challenges comprising the Prostate Cancer Dream Challenge, (b) the statistical metrics used by the challenge organizers for benchmarking the performance of prediction models for each sub-challenge, and (c) our team’s overall analytic strategy for addressing each of the challenge objectives. Specifically, we discuss our approach for identifying clinically important risk-predictors (i.e., feature selection and dimension reduction), the methodological framework(s) considered by our team for model development and validation, including the ensemble-based Cox proportional hazards regression model representing our final submission, and the adaptation of our modeling framework based on the results from the intermittent leaderboard rounds. As the aftermath of the Prostate Cancer Dream Challenge has prompted our team to reflect on the lessons learned throughout challenge, we also provide our perspectives on the importance of delegation, collaboration, data cleaning, and organization in challenges such as the Prostate Cancer Dream Challenge.

DR07: Feature Selection & Random Forest for ALS prediction

Witold R. Rudnicki¹, Wojciech Lesiński¹, Aneta Polewko-Klim¹, Krzysztof Mnich², Agnieszka Golińska¹

¹ Department of Bioinformatics, University of Białystok, Poland; ² Computational Centre, University of Białystok, Poland

The main goal of the challenge was to find the clustering of patients that would improve prediction of the progression of the ALS disease. The success of the clustering was measured by comparing prediction of the progress of the disease with actual data. We provided answers to all questions of the challenge, namely predicting disease progress and eventual death of patients for two data sets.
The modeling performed in three steps — feature construction, selection and model building. Features describing the time series data were constructed following the approach proposed by winners of the ALS DREAM 7 Challenge[1].

We have attempted clustering using informative features, however without success, hence the final clusters were based only on the availability of data for given object.

Feature selection was based on the information entropy. Information gain was computed all variables and all pairs of variables. Informative variables and pairs were selected and redundant features were removed.

Final models were built using random forest classifier [2], using six original features. All possible combinations of variables were tested using cross-validation.

Eleven informative features for question 1 are (variables used by best model denoted by *, second model by ^): onset_delta*, hands^, Q1_Speech^, Q9_Climbing_Stairs^, Q5_Cutting*, fvc*, ALSFRS_Total^, Q3_Swallowing^, Q6_Dressing, Creatinine, Q4_Handwriting.

The cross-validated correlation of these models was 0.43, 0.42 respectively and RMSD was 0.549 and 0.568.

Eight informative variables for second question are (marked as previously): onset_delta^, ALSFRS_Total^, Creatinine^, weight^, fvc*, fvc_percent^ Chloride^, Gender^.

Cross-validated classification error for the first model was 0.22 for 12 months and 0.18 for the 18 and 24 months.

There were six informative variables both for the question 3: (trunk, Q4_Handwriting, Q3_Swallowing, Q8_Walking, hands, ALSFRS_R), and for the question 4: (ALSFRS_R_Total, ALSFRS_Total, hands, MEDHx_Thyroid, Q6_Dressing_and_Hygiene, R3_Respiratory_Insufficiency).

The quality of data was much lower for the second data set and while we obtained sets of informative variables, but the results were on the border of random. The correlation coefficient for third question was 0.15, classification error for the fourth question was 0.35 — both results were estimated by cross-validation.

**DR08: Predicting Discontinuation of Docetaxel Treatment for Metastatic Castration-Resistant Prostate Cancer (mCRPC) with hill climbing and random forest**

*Team Yoda: Daniel Kristiyanto, Kevin Anderson, Seyed Sina Khankhajeh, Kaiyuan Shi, Seth West, Ling Hong Hung, Azu Lee, Qi Wei, Migao Wu, Yunhong Yin, Ka Yee Yeung*

Institute of Technology, University of Washington, Tacoma, United States

* Corresponding author: kayee@uw.edu

**Motivation:** Prostate cancer patients may develop resistance to androgen deprivation therapy (ADT) [1,2]. In the DREAM 9.5 Prostate Cancer Challenge sub challenge 2 [2], we developed predictive models to predict patient outcomes in metastatic castrate-resistant prostate cancer (mCRPC) with subsequent discontinuation of docetaxel therapy.

**9 CREDITS,** The input data consist of 131 variables measured across clinical data from three clinical trials, namely, Memorial Sloan Kettering (MSK, with 476 patients), Celgene (with 526 patients), Sanofi (with 598 patients). The goal is to predict which patients in a fourth clinical trial (test data), AstraZeneca (AZ, with 470 patients), would discontinue treatment due to adverse events within 3 months.

**Data & Methods:** Data cleansing and pre-processing. The data cleansing were done separately within each clinical trial and later merged back together. Our data cleansing and pre-processing procedures include imputation of missing data [4], and removal of clinical variables with a high percentage of missing data. Data augmentation were also performed by converting selected multi-label variables into binary variables. Feature selection. We observed that univariate feature selection methods did not perform well. Hence, we adopted a hill-climbing [4] approach that optimized the AUC within 10-fold cross validation of the training data. We also addressed the issue of imbalanced data (total of 1292 negative samples and 197 positive samples) by randomly removing negative samples to meet a ratio roughly of 60% negative and 40% positive samples.

**Classification:** We applied random forest [5] using Sanofi as the hold-out, setting the parameters “mtry” to 25% of the number of features and number of trees to 100 times of the number of features.

**Assessment:** For validation, we repeated the training step 10 times, with average AUC as the assessment criteria. We also conducted additional assessment by holding out one of the three clinical trials. Our predictive model using MSK and Celgene data as the training set and Sanofi data as the test set yielded AUC = 0.165, accuracy = 0.9, precision = 0.21, F1 = 0.092, and recall = 0.06.

**Results:** Our final submission in predicting the discontinuation of docetaxel in the AstraZeneca clinical trial (using MSK, Celgene and Sanofi as training data) resulted in AUC of 0.13. Across the 470 in AstraZeneca clinical trial, 8 patients are predicted to discontinue the treatment within 3 months.
Acknowledgement: Ling Hong Hung and Ka Yee Yeung are supported by NIH grant U54-HL127624. This project used computing resources provided by Microsoft Azure. We would like to thank all students in TCSS 588 Bioinformatics in Spring 2015 at University of Washington Tacoma who contributed to this project.

REFERENCES

DR09: Predicting olfaction response for each individual
Yuanfang Guan
Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor
This poster is also presented as a talk. The abstract is on page 28.

DR10: Predicting olfactory perception from chemical structure: a gradient boosting model
Chung Wen Yu1, Yusuke Ihara1,2, and Joel D. Mainland1,3
1 Monell Chemical Senses Center, Philadelphia, United States; 2 Institute for Innovation, Ajinomoto Co., Inc., Kawasaki, Japan; 3 Department of Neuroscience, University of Pennsylvania School of Medicine, Philadelphia, United States
A fundamental problem in olfaction is to understand how the physical properties of a stimulus relate to perceptual characteristics. In vision, wavelength translates into color; in audition, frequency translates into pitch. By contrast, the mapping from chemical structure to olfactory percept is unknown. In other words, there is not a scientist or perfumer in the world who can view a novel molecular structure and predict how it will smell. Here we used an unpublished dataset where 49 subjects rated 476 odors to develop a model that uses physicochemical descriptors to predict 21 perceptual features. Previously published models for predicting pleasantness have used principal components of physicochemical descriptors or molecular complexity to predict pleasantness; both performed poorly on this dataset (Khan et al. 2007 r = 0.25, p < 0.001; Kermen et al., 2011 r = 0.26, p < 0.0001). Our model outperformed these previously published models on a validation set (r=0.61, p < 0.001).


**DR11: Predicting patient survival in the DREAM 9.5 mCRPC challenge**

Team FIMM-UTU: Teemu D. Laajala¹,², Suleiman Khan², Antti Airola³, Tuomas Mirtti²,⁴, Tapio Pahikkala³, Peddinti Gopalacharyulu², Tero Aittokallio¹,²

¹ Department of Mathematics and Statistics, University of Turku, Finland; ² Institute for Molecular Medicine Finland, University of Helsinki, Finland; ³ Department of Information Technology, University of Turku, Finland; ⁴ Department of Pathology, HUSLAB, Helsinki University Hospital, Finland

This poster is also presented as a talk. The abstract is on page 25.

**DR12: GENESIS: A variation discovery framework for clinical cancer genomic profiling**

Allen Chi-Shing YU¹,²*, Aldrin Kay-Yuen YIM¹,³*, Marco Jing-Weoi LI¹,²*

¹ Codex Genetics Limited, Hong Kong; ² School of Life Sciences, The Chinese University of Hong Kong, Hong Kong; ³ Computational & System Biology Program, Washington University School of Medicine; * Co-first authors

Large scale cancer genomics projects carried out by consortia such as The Cancer Genome Atlas (TCGA) and International Cancer Genome Consortium (ICGC) have enormous impact on the understanding, diagnosis and treatment of cancer. Data generated in these projects provides important clinical insights for the advancement of bioinformatics algorithms, for instance somatic variant callers with increasing accuracy, the identification of very rare malignant clones as well as the DNA/RNA structural alterations. To participate in the ICGC-TCGA SMC-DNA DREAM Challenge (Intel-10 SNV real-tumor Sub-Challenge), we have developed a machine learning-based ensemble somatic variant calling pipeline – Genesis. Genesis was trained with the public cancer genomic dataset with validated mutation calls, and over thirty informative metrics were gathered from the four individual variant callers (MuTect, Strelka, Vardict and VarScan2) and genomic sequence features such as regional GC level, entropy, mappability and mutational hotspots. Variants were then classified into true or false positives using Support Vector Machine (SVM) models that are optimized for the tumor type and variant type (SNP or indel). Based on the 88 published hepatocellular carcinoma dataset and in-house clinical cancer samples, we demonstrated that our machine learning approach can achieve higher sensitivity and specificity when comparing to the individual call set. It is therefore expected that Genesis is well-applicable for clinical cancer genomic profiling.
**DR13: Predicting odor perception from molecular structure using a “nearest neighbor” approach**

Aharon Ravia, Lavi Secundo, Kobi Snitz, Noam Sobel  
*Department of Neurobiology, Weizmann Institute of Science*

The DREAM olfaction challenge was to predict the perceptual qualities of novel molecules according to their structural properties. The data consisted of 476 molecules rated by 49 subjects across 21 different descriptors.

Predicting odor perception from odor structure is a major goal in olfaction research. We and others have made initial steps in this direction such that we can now predict aspects like odorant pleasantness (Khan et al., 2007; Zarzo 2007; Koulakov et al., 2011) and pairwise odorant similarity (Snitz et al., 2013) from odorant structure alone. These abilities rest in part on the observation that principal component analysis (PCA) of molecular descriptors has a predictive power whereby the first principal component of the physio-chemical space is related to perceived odor pleasantness.

Here we set out to apply PCA analysis and “nearest neighbors approach” to the challenge data using a variation of the Rotation Forest method (Rodríguez, Kuncheva, & Alonso, 2006). This method uses PCA on random splits of the features in order to create weak predictors and then combine them as an Ensemble. In the Rotation Forest method classification is done on a rotated subspace. To answer the challenge we needed to create a continuous estimator instead of a classifier, and used nearest neighbor approach for this purpose. Moreover, we executed the estimation on one subspace of features each time.

This method yielded better results than other methods we tried, such as linear regression. For example, when we correlated predicted vs. actual pleasantness ratings of the leaderboard data we got results of $R = 0.52$. We think that although others were able to get similar and better predictions using other methods, this kind of analysis may reveal facts about the physio-chemical space, and extract characteristics such as proximity between molecules. Finally, we examine the possibility that the olfactory system solves olfactory space using similar strategy.

**DR14: Ensemble Approaches To Prostate Cancer Dream Challenge**

*Team The Data Wizard: Wen-Chieh Fang¹, Li-Min Tu², Huan-Jui Chang³, Chia-Tse Chang¹, Yu-Fu Wang¹, Mu-Hung Tsai⁴, Alexey Yu. Lupatov⁵, Konstantin N. Yarygin⁵, Hsih-Te Yang¹,⁴, Chiang Jung-Hsien¹,⁴*

¹ Department of Computer Science and Information Engineering, National Cheng Kung University, Taiwan; ² Department of Computer Science and Engineering, New York University, United States; ³ Department of Economics, National Cheng Kung University, Taiwan; ⁴ Institute of Medical Informatics, National Cheng Kung University, Taiwan; ⁵ Orekhovich Institute of Biomedical Chemistry of the Russian Academy of Medical Sciences, Moscow, Russia
Prostate cancer is the most common cancer diagnosed in many countries and the third most common cause of cancer death. However, prognostic models for overall survival for patients are dated. Based on preliminary exploration of the three sets of raw trial data, we model the data using ensemble learning in the subchallenge 1a. We provide a novel boosting approach to tackle the problem in subchallenge 1b. For the second subchallenge, we combine three different prediction models to derive the final prediction.

We first preprocess data to handle missing values and use filter method for feature selection. For predicting overall survival, we derive different feature sets for the three training sets (ASCENT2, CELGENE, EFC6546), respectively and then apply the Cox model with maximum penalized likelihood on the selected features to train a specific model for each set. We combine the ranking results from the different models. To predict the exact time to event (death of a patient), we develop a novel adaboost-like regression algorithm called “adaboost-s” for survival problem, especially to predict time to event. In the training phase, if the predicted time to event of a censored data point is smaller than the lower bound, the data point is considered incorrectly predicted and its weight is increased. To predict treatment discontinuation for patients treated with docetaxel due to adverse events at early time points, we apply ensemble technique to combine the ranking results of several methods. The methods includes two random forest classifiers and one gradient boosting classifier.

The main advantage of ensembles of different models is that it is unlikely that all models will make the same mistake. Ensembles tend to reduce the variance of models. Therefore, we apply ensemble approaches to deal with the prediction problems in this challenge. In the leaderboard, our approaches outperform the baseline and the methods of many teams.

**DR15: Supervised ensembles to boost the predictive power of DREAM challenges**

Gaurav Pandey1, Gustavo Stolovitzky1,2, Sean Whalen3, Lara Mangravite4, Solveig Siebert5, Abhishek Pratap5 and Om Prakash Pandey1

1 Department of Genetics and Genomic Sciences and Icahn Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, United States; 2 IBM Research, New York, United States; 3 Gladstone Institutes, University of California at San Francisco, United States; 4 Sage Bionetworks, Seattle, United States

Prediction problems in biomedical sciences, such as those posed in DREAM challenges, are well-known to be quite difficult to solve convincingly. This is due in part to incomplete knowledge of the biomedical phenomenon of interest, the appropriateness and data quality of the variables and measurements used for prediction, as well as a lack of consensus regarding the ideal predictor(s) for specific problems. These issues are reflected in the diversity of prediction techniques, datasets, domain knowledge and other ingredients used to develop submissions to
DREAM challenges. In such scenarios, a powerful approach to improving prediction performance is to construct ensemble predictors that combine the output of complimentary individual predictors derived from diverse techniques and/or datasets. Traditional ensemble methods like boosting, bagging and random forest are insufficient for this task as they (generally) assume that the individual/base predictors are of the same type. They also expect the ensemble process to have control over the generation of these predictors from (generally) a single training set. Both these important assumptions do not hold for the challenge setting. Thus, in this work, we propose the use of heterogeneous ensemble methods, such as stacking and ensemble selection, for building effective ensembles for DREAM challenges as well as other biomedical prediction problems. First, using several protein function and genetic interaction prediction datasets, we illustrated how such heterogeneous ensembles can provide statistically significant gains over individual predictors, including those based on boosting and random forests (Whalen et al., Methods, 2015). Deeper analysis shows that the superior predictive ability of these methods, especially stacking, can be attributed to their attention to the following aspects of the ensemble learning process: (i) better balance of diversity and performance, (ii) more effective calibration of outputs and (iii) more robust incorporation of additional individual predictors. Motivated by these results, we built stacking-based ensembles of subchallenge 2 of the Rheumatoid Arthritis anti-TNF drug response challenge. Using only six of the individual predictors, these ensembles (AUPR=0.5228) again provided prediction gains over the two best individual predictors (AUPR = 0.5099 and 0.5071). In current work, we are trying to realize such gains for other DREAM challenges as well, especially by systematically addressing the theoretical and implementation issues associated with this task.

**DR16: A Boosting approach and Cox model for Predicting Slope and Survival of ALS**

Wen-Chieh Fang¹, Chen Yang¹, Huan-Jui Chang², Hsih-Te Yang¹,³, Jung-Hsien Chiang¹,³

1 Department of Computer Science and Information Engineering, National Cheng Kung University, Taiwan; 2 Department of Economics, National Cheng Kung University, Taiwan; 3 Institute of Medical Informatics, National Cheng Kung University, Taiwan

Amyotrophic lateral sclerosis (ALS) is a progressive neurological disease that leads muscle weakness and gradually impacts on the functioning of the body, leading to eventual death. It greatly reduces an individual’s life expectancy. Currently, experts do not know precisely what causes ALS. There is no known cure for ALS. The DREAM ALS Stratification Prize4Life Challenge is held with the goal to enable better understanding of patient profiles and application of personalized ALS treatments.

In our approach to Dream ALS Challenge, we first disregard those features which 90 percent of values are missing. For the remaining, we replace
any missing value with the mean of that variable for all other cases. For all the four subchallenges, we apply equal frequency binning that divides the response variable into three groups such that each group contains approximately same number of values. There are two kinds of features: static features and ‘time-resolved’ features (those with different values when time varies). For the latter, we try two designated measurements, the minimum and the maximum as features. Then for both two kinds of features, we apply feature selection based on information gain to select the top-six features. In order to select optimal features, we run cross validation on the feature candidates. We apply Gradient Boosted Regression Trees (GBRT) to predict the ALSFRS slopes. GBRT computes a sequence of simple decision trees, where each successive tree is built for the prediction residuals of the preceding tree. We apply Cox model with maximum penalized likelihood to predict the survival probability.

In this challenge, we think that the feature selection is one of the most important steps and we believe that the most appropriate features dominate the performance of the model. In the last round of leaderboard, our team was the second best team and our approaches outperformed the methods of most teams.

**DR17: Using aggregated weights along paths across random forest to select important features and predict ALS progression**

_Jinfeng Xiao_ and Jian Peng

_University of Illinois at Urbana-Champaign, United States_

This poster is also presented as a talk. The abstract is on page 32.
Systems Biology and Regulatory Genomics Posters

P01: Distinct specificities of the androgen and glucocorticoid receptors revealed using feature-based recognition model analysis of SELEX data

Liyang Zhang1, Gabriella Martini2, H. Tomas Rube2, Vincent D. Fitzpatrick2, Harmen J. Bussemaker2, Miles A. Pufall1
1 University of Iowa, United States; 2 Columbia University, United States

This poster is also presented as a talk. The abstract is on page 42.

P02: Quantitative modeling of gene expression from sequence, using DNA shape-based model of binding sites

Pei-Chen Peng1, Saurabh Sinha1
University of Illinois at Urbana Champaign, United States

This poster is also presented as a talk. The abstract is on page 45.

P03: Presentation Title: Estimating the Number and Diversity of Cancer Mutations In the Overall Population from 5,319 Complete Cancer Genomes

Prathik Naidu1, Joseph Kaplinsky2, Ramy Arnaout2,3
1 Thomas Jefferson High School for Science and Technology, United States; 2 Beth Israel Deaconess Medical Center, United States; 3 Harvard Medical School, United States

This poster is also presented as a talk. The abstract is on page 41.

P04: MiSL: a method for mining synthetic lethal partners of recurrent cancer mutations uncovers novel mutation-specific therapeutic targets

Subarna Sinha1, Daniel Thomas1, Yang Gao2, Steven Chan1, Diede Brunen3, Rene Bernards3, Ravindra Majeti1, David L. Dill1
1 Stanford University, United States; 2 University of California at Berkeley, United States; 3 Netherlands Cancer Institute, Netherlands

This poster is also presented as a talk. The abstract is on page 38.

P05: Tracking the Evolution of 3D Gene Organization

Alon Diament and Tamir Tuller
Tel Aviv University, Israel

This poster is also presented as a talk. The abstract is on page 47.
P06: From phenotypic to molecular synergy: A transcriptional study of the dynamics of drug combinations based on single drug responses

Mehmet Eren Ahsen$^{1,2}$, Jennifer E. L. Diaz$^{1,2}$, Xintong Chen$^2$, Bojan Losic$^2$, Gustavo Stolovitzky$^1$

1 IBM Research; 2 Icahn School of Medicine at Mount Sinai, United States

This poster is also presented as a talk. The abstract is on page 44.

P07: Three-dimensional analysis of regulatory features reveals functional enhancer-associated loops

Yao Wang$^1$, Junbai Wang$^6$, Yufan Zhou$^1$, Malaina Gaddis$^8$, Rohit Jadhav$^1$, Xun Lan$^4$, Tim Huang$^1$, Shili Lin$^5$, Peggy Farnham$^3$, Seth Frietze$^6$, Victor Jin$^1$

1 University of Texas Health Science Center at San Antonio, United States; 2 Oslo University Hospital – Norwegian Radium Hospital, Norway; 3 University of Southern California, United States; 4 Stanford University, United States; 5 The Ohio State University, United States; 6 University of Vermont, United States

This poster is also presented as a talk. The abstract is on page 49.

P08: Understanding Breast Cancer Heterogeneity through Personalized Drosophila Models

Jennifer Long Diaz, Avi Ma’Ayan, Ross Cagan
Icahn School of Medicine at Mount Sinai, United States

This poster is also presented as a talk. The abstract is on page 54.

P09: Presentation Title: Creating a library of genome-wide chromatin state patterns during B lymphopoiesis

Mark Maienschein-Cline, Pinal Kanabar, Neil Bahroos, Malay Mandal, Marcus Clark
University of Chicago, United States

This poster is also presented as a talk. The abstract is on page 59.

P10: Bringing big genomic data into focus for studying complex diseases in specific biological contexts

Arjun Krishnan$^1$, Ran Zhang$^1$, Victoria Yao$^1$, Chandra Theesfeld$^1$, Aaron Wong$^2$, Alicja Tadych$^1$, Natalia Volfovsky$^2$, Alan Packer$^2$, Alex Lash$^2$, Olga Troyanskaya$^2$

1 Princeton University, United States; 2 Simons Foundation, United States

This poster is also presented as a talk. The abstract is on page 60.

P11: Nucleotide Sequence Composition Adjacent to Intronic Splice Sites Improves Splicing Efficiency and Reduces Translation Costs in Fungi

Zohar Zafir and Tamir Tuller
Tel Aviv University, Israel

This poster is also presented as a talk. The abstract is on page 62.
P12: Multi-omics learning and optimal experimental design for microbial organisms
Minseung Kim, Navneet Rai, Violeta Zorraquino, Xiaokang Wang, Ilias Tagkopoulos
University of California at Davis, United States
This poster is also presented as a talk. The abstract is on page 65.

P13: Comparison of Methods to Predict Impact of Regulatory Variants
Felix Yu, Dongwon Lee, Michael Beer
Johns Hopkins University, United States
This poster is also presented as a talk. The abstract is on page 69.

P14: High-throughput allele-specific expression across 250 environmental conditions
Gregory Moyerbrailean¹, Chris Harvey¹, Omar Davis¹, Adnan Alazizi¹, Donovan Watza¹, Yoram Sorokin¹, Karoline Pruder¹, Nancy Hauff¹, Xiaoquan Wen², Roger Pique-Regi¹, Francesca Luca¹
¹ Wayne State University, United States; ² University of Michigan, United States
This poster is also presented as a talk. The abstract is on page 74.

P15: Evaluating Genetic Variation Impact on Transcription Factor Binding Sites
Wenqiang Shi, Oriol Fornes, Wyeth Wasserman
University of British Columbia, Canada
This poster is also presented as a talk. The abstract is on page 70.

P16: Network model of normal gene expression predicts gene perturbation fold changes
Sudhir Varma
HiThru Analytics, United States
This poster is also presented as a talk. The abstract is on page 78.

P17: Identifying condition specific transcription factor binding with ATAC-seq
Roger Pique-Regi, Donovan Watza, Molly Estill, Sophia Chaudhry, Francesca Luca
Wayne State University, United States
This poster is also presented as a talk. The abstract is on page 71.
P18: An integrative and applicable phylogenetic footprinting framework for cis-regulatory motifs identification in prokaryotic genomes

Bingqiang Liu¹, Qin Ma²
1 Shandong University, China; 2 South Dakota State University, United States

ABSTRACT: Phylogenetic footprinting is an important computational technique for identifying cis-regulatory motifs (motifs for short) in orthologous regulatory regions of query genes from multiple genomes, based on the viewpoint that motifs tend to evolve slower than their surrounding non-functional sequences. However, the real power of this strategy is yet to be fully realized, as people still have obstacles in how to optimize the selection of orthologous data and how to effectively reduce the false positives in motif prediction. Here we present an integrative phylogenetic footprinting framework, named MP3, for prokaryotic genomes based on a new orthologous data preparation procedure and a novel promoter scoring and pruning method, in support of accurate motif predictions. Specifically, we collected orthology broadly from all prokaryotic genuses and building the orthologous regulatory regions based on sequence similarity of promoter regions, which not only fully made use of the large-scale genomic data and taxonomy information but also filtered the promoters with limited contribution out, thus can produce the high quality reference set. On the other hand, the promoter scoring and pruning is implemented through motif voting by a set of complementary predicting tools, which mine motif candidates as many as possible and eliminate the effect of random noise simultaneously. We have applied the framework to Escherichia coli k12 genome and get the prediction performance evaluated by comparing with seven existing programs. This evaluation is carried out in a systematic way both in nucleotide level and binding site level by adopting the benchmark method proposed by Tompa, along with additional statistical measurement. The results showed that MP3 performs better with 98% and 88% improvement in Performance Coefficient and Correlation Coefficient in nucleotide level over MDscan, which is the best one of other tools. In binding sites level, MP3 outperforms MDscan by 60% in F-Score and 46% in Average Site Performance. Most importantly, we have integrated this phylogenetic footprinting framework into our motif identification and analysis server DMINDA, through which the users can efficiently identify and analyze motifs for any prokaryotic genes.
P19: The CoGAPS matrix factorization algorithm infers feedback mechanisms from therapeutic inhibition of EGFR that increases expression of growth factor receptors

Elana J. Fertig1, Hiroyuki Ozawa2, Manjusha Thakar1, Jason Howard1, Gabriel Krigsfield1, Alexander V. Favorov1, Daria A. Gaykalova1, Michael F. Ochs3, Christine H. Chung4

1 Johns Hopkins University, United States; 2 Department of Otorhinolaryngology-Head and Neck Surgery, Keio University School of Medicine, Japan; 3 The College of New Jersey, United States; 4 Moffitt Cancer Center & Research Institute, United States

ABSTRACT: Next generation sequencing technologies open a door for a precise personalized medicine. Thus, patients with oncogene driven tumors are currently treated with targeted therapeutics such as EGFR inhibitors. However, drug interactions with other activated signaling pathways in treated tumors often alter predicted therapeutic response. Therefore, bioinformatics algorithms are needed to infer unanticipated molecular interactions from anticipated molecular response to targeted therapeutics in diverse genetic backgrounds. To model heterogeneous genetic backgrounds in HNSCC, we use HaCaT cells with forced overexpression of EGFR, HRAS, and PIK3CA. Previously, the CoGAPS matrix factorization algorithm was shown to infer the specific signaling pathways that were activated in these HaCaT knock-in constructs from gene expression data. In this study, we evaluated whether CoGAPS could also delineate unanticipated signaling changes from anticipated cellular signaling response caused by targeted therapeutic in diverse genetic backgrounds. To test this hypothesis, we measured gene expression after treating the modified HaCaT cells with three EGFR targeted agents (gefitinib, cetuximab and afatinib) for 24 hours. The CoGAPS matrix factorization algorithm distinguished a gene expression signature associated with the anticipated silencing of the EGFR network and a signature associated with unanticipated transcriptional feedback in HaCaT constructs that were sensitive to EGFR inhibitors. Notably, the feedback signature showed that EGFR gene expression itself increased in cells that were responsive to EGFR inhibitors. The CoGAPS algorithm further associated such feedback with increased expression of several growth factor receptors by the AP-2 family of transcription factors. Once transcribed, these growth factor receptors may ultimately compensate for EGFR inhibition in these sensitive cells. Our data suggest, that CoGAPS gene expression signatures delineate on and off target effects of drugs related to therapeutic sensitivity in diverse genetic backgrounds.
P20: A network approach to monitor progression of treatment in tuberculosis

Awanti Sambarey1, Abhinandan Devaprasad1, Nagasuma Chandra1
Indian Institute of Science, India

ABSTRACT: Tuberculosis remains one of the leading causes of mortality due to an infectious agent, affecting ~9 million people each year and resulting in about 1.5 million deaths annually. Delayed diagnosis, presence of co-morbidities and emerging drug resistance further compound the problem and underscore the need to gain mechanistic insights into the host response to infection and treatment. Unraveling correlates of successful host response and outcome of therapy are essential for the development of improved therapeutic strategies. The host response to tuberculosis is multifaceted, complex and dynamic, involving intricate cross-talk among several processes occurring simultaneously across different host compartments, and it becomes important to address the underlying complexity in these interdependent molecular networks in order to elucidate the relationship between molecular origins of disease and the manifested phenotype, thereby necessitating a systems approach.

Despite the increased availability of host genome-scale omics data in infection and over the course of treatment in tuberculosis, the molecular end points of therapy have not been clearly elucidated. While transcriptomics data illuminates the differential regulation of genes in individual patients, the cause and consequences of such differential regulation are still not understood. In this study, we have used network-based approaches, together with gene expression data to capture inter- and intra-cellular communication in the host and identify markers that can predict treatment prognosis. We first construct a comprehensive genome-scale network of host processes comprising 11,017 proteins and 1,51,645 interactions. We then integrate genome-wide gene expression data to generate dynamic response networks that monitor progress of therapy in tuberculosis over multiple weeks. Through weighted shortest path analysis we have identified molecular processes that are differentially regulated over the course of treatment, highlighting the importance of host signaling processes and lipid metabolism in governing outcome of therapy. By shifting the focus from individual genes to pathway-based analysis, network-based studies help illuminate the effects of local changes on the global system, and can aid in modifying therapeutic design for effective tuberculosis treatment.
P21: Epigenetic dysregulation of human myogenesis affects time regulated eRNA and associated transposable element expression

Loqmane Seridi¹, Yanal Ghosheh¹, Beatrice Bodega², Gregorio Alanis-Lobato¹, Timothy Ravasi¹, Valerio Orlando¹

¹ King Abdullah University of Science and Technology, Saudi Arabia; ² Istituto Nazionale Genetica Molecolare ‘Romeo ed Enrica Invernizzi’, Italy

ABSTRACT: Transcriptional regulation is a complex process that involves the interaction of transcription factors, promoters, enhancers, noncoding RNAs, transposable elements and chromatin states. To understand the transcriptional regulome, spatiotemporal measurements of its components is necessary. Myogenesis is a model system to study transcriptional regulation because factors driving the process are well known and evolutionary conserved. However, most time course studies of myogenesis are limited to few time points and cell lines. Here, using RNA-Seq and CAGE, we deep sequenced a high-resolution time-course of myogenesis transcriptome from human primary cells of healthy donors and donors affected by Duchenne Muscular Dystrophy (DMD). We compiled a full catalog of coding and non-coding RNAs, promoters, enhancers, and active transposable elements. Comparative analysis of the two time-courses suggests a major change in epigenetic landscape in DMD leading to global dysregulation of coding and non-coding genes, enhancers, and full-length transposable elements. It also indicates a high correlation between enhancers and transposable elements activities.

P22: Characterizing the dynamics of enzyme localization

Pablo Meyer, Stacey Gifford
IBM T.J. Watson Research Center, United States

ABSTRACT: To better understand how enzyme localization affects enzyme activity we studied using timelapse microscopy the cellular localization of necessary enzymes for cell wall synthesis (MurA and MurG) in the bacteria Bacillus subtilis. Enzymes localize during exponential growth and measuring the diffusion coefficient of their complex shows that it diffuses actively around the cell in an antibiotic-dependent manner. Point mutations in the helical domain of one of the proteins, disrupts its localization to the membrane caused severe sporulation defects, but did not affect localization nor caused detectable defects during exponential growth. We found a lipid-dependent mechanism for MurG localization, as in strains where the cardiolipin-synthesizing genes were deleted, MurG levels were diminished at the forespore. These results support localization as a critical factor in the regulation of proper enzyme function and catalysis.
P23: The Systems Toxicology Computational Challenge: Identification of Exposure Response Markers

Vincenzo Belcastro, Carine Poussin, Stephanie Boue, Florian Martin, Alain Sewer, Bjoern Titz, Manuel C Peitsch, Julia Hoeng
PMI, Switzerland

ABSTRACT: Risk assessment in the context of 21st century toxicology relies on the identification of specific exposure response markers and the elucidation of mechanisms of toxicity, which can lead to adverse events. As a foundation for this future predictive risk assessment, diverse set of chemicals or mixtures are tested in different biological systems, and datasets are generated using high-throughput technologies. However, the development of effective computational approaches for the analysis and integration of these data sets remains challenging. The sbv IMPROVER (Industrial Methodology for Process Verification in Research; http://sbvimprover.com/) project aims to verify methods and concepts in systems biology research via challenges posed to the scientific community. In fall 2015, the 4th sbv IMPROVER computational challenge will be launched which is aimed at evaluating algorithms for the identification of specific markers of chemical mixture exposure response in blood of humans or rodents. The blood is an easily accessible matrix, however remains a complex biofluid to analyze. This computational challenge will address questions related to the classification of samples based on transcriptomics profiles from well-defined sample cohorts. Moreover, it will address whether gene expression data derived from human or rodent whole blood are sufficiently informative to identify human-specific or species-independent blood gene signatures predictive of the exposure status of a subject to chemical mixtures (current/former/non-exposure). Participants will be provided with high quality datasets to develop predictive models/classifiers and the predictions will be scored by an independent scoring panel. The results and post-challenge analyses will be shared with the scientific community, and will open new avenues in the field of systems toxicology.

P24: E-Flux2 and SPOT: Validated methods for inferring intracellular metabolic flux distributions from transcriptomic data

Min Kyung Kim, Anatoliy Lane, James Kelly, Desmond Lun
Rutgers University, United States

ABSTRACT: Several methods have been developed to predict system-wide intracellular metabolic fluxes by integrating transcriptomic data with genome-scale metabolic models. While powerful in many ways, existing methods have several shortcomings, and because of limited validation against experimentally measured intracellular fluxes, it is unclear which method has the best accuracy in general.
We present a general strategy for inferring intracellular metabolic flux distributions using transcriptomic data coupled with genome-scale metabolic reconstructions. It consists of two different template models called DC (determined carbon source model) and AC (all possible carbon sources model) and two different new methods called E-Flux2 (E-Flux method combined with minimization of l2 norm) and SPOT (Simplified Pearson correlation with Transcriptomic data), which can be chosen and combined depending on the availability of knowledge on carbon source or objective function. This enables our strategy to be applied to a broad range of experimental conditions. We examined E. coli and S. cerevisiae as representative prokaryotic and eukaryotic microorganisms respectively. The predictive accuracy of our algorithm was validated by calculating the uncentered Pearson correlation between predicted fluxes and measured fluxes. To this end, we compiled 20 experimental conditions (11 in E. coli and 9 in S. cerevisiae), of transcriptome measurements coupled with corresponding central carbon metabolism intracellular flux measurements determined by 13C metabolic flux analysis (13C-MFA), which is largest dataset assembled to date for the purpose of validating inference methods for predicting intracellular fluxes. In both organisms, our method achieves an average correlation coefficient ranging from 0.59 to 0.87, outperforming a representative sample of competing methods. Easy-to-use implementations of E-Flux2 and SPOT are available as part of the open-source package MOST (http://most.ccib.rutgers.edu/).

Our method represents a significant advance over existing methods for inferring intracellular metabolic flux from transcriptomic data. It not only achieves higher accuracy, but it also combines into a single method a number of other desirable characteristics including applicability to a wide range of experimental conditions, production of a unique solution, fast running time, and the availability of a user-friendly implementation.

P25: SetRank: A highly specific tool for pathway analysis

Cedric Simillion¹, Robin Liechti², Heidi Lischer³, Vassilios Ioannidis², Rémy Bruggmann¹

¹ Bern University, Switzerland; ² SIB Swiss Institute of Bioinformatics, Switzerland; ³ University of Zurich, Switzerland

ABSTRACT: The purpose of gene set enrichment analysis (GSEA) is to find general trends in the huge lists of genes or proteins generated by many functional genomics techniques and bioinformatics analyses. We present SetRank, an advanced GSEA algorithm which is able to eliminate many false positive hits. The key principle of the algorithm is that it discards gene sets that have initially been flagged as significant, if their significance is only due to the overlap with another gene set. The algorithm is explained in detail and its performance is compared to that of other methods using objective benchmarking criteria. The benchmarking results show that SetRank is a highly specific and accurate tool for GSEA. Furthermore, we show that the reliability of results can be improved by taking sample
source bias into account. SetRank and the accompanying visualization tools are available both as R/Bioconductor packages and through an online web interface.

**P26: The role of genome accessibility in transcription factor binding in bacteria**

Antonio Gomes, Harris Wang  
*Columbia University, United States*

**ABSTRACT:** ChIP-seq enables the identification of regulatory regions that govern gene expression at genome-scale. However, the biological insights generated from ChIP-seq analysis have been limited to predictions of binding sites and cooperative interactions. Furthermore, ChIP-seq data often poorly correlate with in vitro measurements or predicted motifs, highlighting that binding affinity alone is insufficient to explain transcription factor (TF)-binding in vivo. A more comprehensive biophysical representation of TF-binding will improve our ability to understand, predict, and alter gene expression. Here, we show that genome accessibility is a key parameter that impacts TF-binding in bacteria. We developed a thermodynamic model that parameterizes ChIP-seq coverage in terms of genome accessibility and binding affinity. The role of genome accessibility is validated using a large-scale ChIP-seq dataset of the *M. tuberculosis* regulatory network. We find that accounting for genome accessibility led to a model that explains 69% of the ChIP-seq profile variance, while a model based in motif conservation alone explains only 46% of the variance. Moreover, our framework enables de novo ChIP-seq peaks prediction and is useful for inferring TF-binding peaks in new experimental conditions by reducing the need for additional experiments. We observe that the genome is more accessible in intergenic regions, and that increased accessibility is positively correlated with gene expression and anti-correlated with distance to the origin of replication. Our biophysical model provides a more comprehensive description of TF-binding in vivo from first principles towards a better representation of gene regulation in silico, with promising applications in systems biology.

**P27: Does the overall shape of gene networks differ between cancer and normal states? Towards a comprehensive understanding of cancer system biology by meta-analysis of various cancer transcriptomes**

Pegah Khosravi¹, Esmaeil Ebrahimie²  
¹ School of Biological Sciences, Institute for Research in Fundamental Sciences (IPM), Tebran, Iran (Islamic Republic of); ² Department of Genetics and Evolution, School of Biological Sciences, The University of Adelaide, Australia

**ABSTRACT:** Recent advances in computational biology have provided the possibility of formulating the characteristics of gene networks in terms of network topology statistics. The aim of the present study is to find the
possible network topology rules which can distinguish different types of cancer from normal state. To this end, meta-analysis is employed to analyse the gene regulatory networks of 8 different types of cancer (breast, cervical, esophageal, head and neck, leukemia, prostate, rectal, lung and two subtypes of lung cancer (small cell lung and non-small cell lung)) in comparison to normal state. Microarray data were downloaded from the GEO database, NCBI. Gene regulatory networks were constructed using the ARACNE algorithm through the Cyni toolbox; consequently, 20 network statistics were calculated using NetworkAnalyzer plugin for Cytoscape. These statistics mainly describe number of edges, clustering coefficient, connected components, network diameter, network centralization, characteristics path length, average number of neighbors, number of nodes, network density, and heterogeneity in networks. Discriminant function analysis show that number of edges, network diameter, and average number of neighbors are the main network topology statistics which discriminate cancer networks from normal ones. Cancer networks have lower number of edges with shorter diameter, and fewer number of neighbors that confirms the extensive networks rewiring during cancer progression. Discriminant function analysis is able to predict gene network of cancer from normal with 70% accuracy according to cross-validation test. PCA analysis demonstrates the similarity in network statistics between cervical cancer and breast cancer. Lung cancer have a distinguished different network pattern with low network centralization and diameter. This study demonstrates the possibility of finding universal pattern in different types of cancers based on network topological statistics. It also shows that decision tree models (pattern recognition) are successful in finding the pattern of cancer induction based on the important network statistics.

P28: Comparative Assessment Suite for Transcription Factor Binding Motifs

Caleb Kipkurui Kibet and Philip Machanick
Department of Computer Science and Research Unit in Bioinformatics (RUBi), Rhodes University, Grahamstown, South Africa

ABSTRACT: Predicting transcription factor (TF) binding sites remains an active challenge due to degeneracy and multiple potential binding sites in the genome. The advent of high throughput sequencing has seen several experimental approaches, including ChIP-seq, DNase-seq and ChIP-exo, and dozens of algorithms developed to address the challenge. An increasing number of motif models has been published and those in databases have more than doubled in the last two years. However, there is no standardized means of motif assessment let alone a computational tool to rank the available motifs for a given TF. This makes it hard to choose the best models and for algorithm developers to benchmark, test, quantify and improve on their tools. We introduce a web server hosting a suite of tools that assesses PWM-based motif models using scoring, comparison and
enrichment approaches. Given that there is no agreed standard for motif quality assessment, we present a range of measures so users can apply their own judgement. An assess-by-scoring approach uses motif models to score benchmark data partitioned into positive and background sets, then uses AUC, Pearson, MNCP and Spearman’s rank statistics to quantify their performance – scoring functions are energy, GOMER, sum occupancy and sum log-odds. An assess-by-comparison approach seeks to rank, for a given TF, motifs based on similarity to all available motifs in the database using TOMTOM’s Euclidean distance function and FISim. It assumes the best model should be representative of information in the others, provided a variety of data and algorithms is used. This is a quick data-independent approach that has proved to be powerful, reproducing assessment-by-score ranks with over 0.7 average correlation. A web interface to the tools uses the Django framework with a MySQL back end. The database contains 6,530 human and mouse motif models and benchmark data derived from available databases and publications. A user-entered test motif for a given TF is ranked against motifs for the same TF in the database using the available benchmark data as well as user-supplied data in BED or FASTA format. Results are returned in interactive visuals providing further information on motif clustering, similarity and ranks, with options to download publication-ready figures and ranked motif data. We have demonstrated the benefit of our web server in motif choice and ranking as well as in motif discovery. Web server and command-line versions are available (link to be added once available, estimated mid-October 2015).

P29: Loregic: A method to characterize the cooperative logic of regulatory factors

Daifeng Wang¹, Koon-Kiu Yan¹, Cristina Sisu¹, Chao Cheng², Joel Rozowsky¹, William Meyerson¹, Mark Gerstein¹

¹ Yale University, United States; ² Dartmouth Medical School, United States

ABSTRACT: Gene expression is controlled by various gene regulatory factors. Those factors work cooperatively forming a complex regulatory circuit on genome wide. Corruptions of regulatory cooperativity may lead to abnormal gene expression activities such as cancer. Traditional experimental methods, however, can only identify small-scale regulatory activities. Thus, to systematically understand the cooperativity between and among different types of regulatory factors, we need the efficient and systematic computational methods. Regulatory circuits have been found to behavior very analogous to the electronic circuits in which a wide variety of electronic elements work coordinately to function correctly. Recently, an increasing amount of next generation sequencing data provides great resources to study regulatory activity, so it is possible to go beyond this and systematically study regulatory circuits in terms of logic elements. To this end, we develop Loregic, a computational method integrating gene expression and regulatory network data, to characterize the cooperativity of regulatory factors for the first time in cancers such as acute myeloid
leukemia, which provided unprecedented insights into the gene regulatory
logics in complex biological systems [1]. Loregic uses all 16 possible
two-input-one-output logic gates (e.g. AND or XOR) to describe triplets
of two factors regulating a common target. We attempt to find the gate
that best matches each triplet’s observed gene expression pattern across
many conditions. We make Loregic available as a general-purpose tool
(loregic.gersteinlab.org). We validate it with known yeast transcription-
factor knockout experiments. Next, using human ENCODE ChIP-Seq
and TCGA RNA-Seq data, we are able to demonstrate how Loregic
characterizes complex circuits involving both proximally and distally
regulating transcription factors (TFs) and also miRNAs. Furthermore,
we show that MYC, a well-known oncogenic driving TF, can be modeled
as acting independently from other TFs (e.g., using OR gates) but
antagonistically with repressing miRNAs. Finally, we inter-relate Loregic’s
gate logic with other aspects of regulation, such as indirect binding
via protein-protein interactions, feed-forward loop motifs and global
regulatory hierarchy.

[1] Daifeng Wang, Koon-Kiu Yan, Cristina Sisu, Chao Cheng, Joel Rozowsky, William
Meyerson, Mark Gerstein, “Loregic: A method to characterize the cooperative logic of regulatory

P30: Affymetrix Probesets as Proxies for Mature
MicroRNAs

Rebecca Tagett and Sorin Draghici
Wayne State University, United States

ABSTRACT: Motivation: MicroRNAs are small non-coding RNAs
that regulate mRNA abundance post-transcriptionally, and have been
implicated in many contexts, from development to disease. Primary
microRNAs (pri-miRNAs) from intergenic regions, which represent
around half of the known population, are independently transcribed,
5-prime capped and poly-adenylated. Like precursor messenger RNAs,
they can be kilobases in length, and must undergo extensive processing.
Previous studies have suggested that the expression of some intergenic pri-
miRNAs can be used as surrogates for expression of mature microRNAs.
Little is known about which pri-miRNAs have this property, and sequence
annotation is un-available for the majority of pri-miRNAs.

The Affymetrix HG U133 Plus 2.0 array includes probesets for 19,612
protein coding genes and 15,943 non-coding, mostly un-annotated
transcripts. With over 3000 public data series, these experiments cover a
large variety of human tissues, conditions and disease states. We identify
pri-miRNAs among the non-protein-coding probesets, and show that
some of them can be used as proxies for mature microRNAs, opening
up the possibility of combined microRNA - mRNA studies of mRNA
abundance in relation to the presence of microRNAs.
Methods: Using sequence similarity, we leverage the Unigene database to connect microRNA stem-loop sequences to target sequences from U133 Plus 2.0, identifying over 250 probe-set-microRNA pairs. Fifty of these are one-to-one matches between a Unigene cluster and a probe-set, with one or several associated stem-loop precursors. Having compiled a set of public datasets from experiments where samples are run both on a microRNA platform and U133 Plus 2.0, we identify pri-miRNA probe sets that correlate with mature microRNA. We select two well-studied microRNAs from these results and screen all public U133 plus 2.0 series to find which sets express the pri-miRNA of interest. We then validate the selected pri-miRNAs based on published literature, and perform functional analyses on the genesets that are co-expressed and anti-co-expressed with these microRNAs.

Results: We present the set of probesets from U133 Plus 2.0 which target pri-microRNA transcripts, highlighting which are acceptable surrogates for mature microRNA abundance. Those that do not correlate to mature microRNA abundance may be useful to study microRNA processing regulation, tissue specificity, co-transcription and transcription factor activity. Those that are proxies for mature microRNAs can be studied within the pool of all coding mRNAs, over the vast repository of U133 Plus 2.0 chips, to generate new, testable hypotheses regarding microRNA function. We present some intriguing results for the two pri-miRNAs selected for in-depth analysis.

P31: Non-coding isoforms of coding genes in B cell development and malignancies
Irtisha Singh, Shih-Han Lee, Christina Leslie, Christine Mayr
Memorial Sloan Kettering Cancer Institute, United States

ABSTRACT: Alternative cleavage and polyadenylation (ApA) is most often viewed as the selection of alternative pA signals in the 3'UTR, generating 3'UTR isoforms that code for the same protein. However, ApA events can also occur in introns, generating either non-coding transcripts or truncated protein-coding isoforms due to the loss of C-terminal protein domains, leading to diversification of the proteome. Since previous studies have demonstrated the cell type and condition specific expression of 3'UTR isoforms, we decided to investigate the cell type specificity and potential functional consequences of isoforms generated by intronic ApA. We therefore carried out an analysis of 3'-seq and RNA-seq profiles from chronic lymphocytic leukemia (CLL) and multiple myeloma (MM) samples as compared to mature human B cells (naïve and CD5+) and plasma cells, respectively, together with our previous 3'-seq atlas generated from a wide variety of tissues and cell lines. This analysis showed that intronic ApA is a normal and regulated process, most widely used in immune cells, with intronic ApA events enriched near the start of the transcription unit, yielding non-coding transcripts or messages with minimal coding sequence (CDS). These early intronic ApA events...
preferentially occur in transcription factors, chromatin regulators, and ubiquitin pathway genes. De novo assembly of RNA-seq data supports ~60% of the intronic ApA events from plasma cells and MM samples, leading to >2000 candidate alternative transcripts arising from intronic ApA, with ~900 transcripts ending near the start of the transcription unit, retaining less than 25% of the coding sequence. Our analysis showed that two thirds of these intronic ApA isoforms have minimal coding potential, likely generating non-coding isoforms from protein coding genes. CLL cells increase the expression of early intronic ApA events relative to mature B cells, while MM cells decrease the expression of these events relative to plasma cells. For a fraction of genes, increased expression of isoforms generated by intronic ApA coincides with reduced expression of the full length mRNA in CLLs compared to mature B cells; conversely, lower expression of intronic ApA events coincides with higher full length mRNA expression for some genes in MM samples compared to plasma cells. In these genes, expression of the intronic event may function as a switch to alter full-length mRNA expression. The other fraction of these non-coding isoforms may potentially act as scaffolds for recruiting regulatory factors to the locus.

P32: Predicting Metabolic Networks through Pairwise Rational Kernels

Abiel Roche-Lima
Medical Science Campus, University of Puerto Rico, Puerto Rico

ABSTRACT: Metabolic networks are represented by the set of metabolic pathways. Metabolic pathways are a series of chemical reactions, in which the product from one reaction serves as the input to another reaction. Many pathways remain incompletely characterized, and in some of them not all enzyme components have been identified. One of the major challenges of computational biology is to obtain better models of metabolic pathways. Existing models are dependent on the annotation of the genes. This propagates error accumulation when the pathways are predicted by incorrectly annotated genes.

Pairwise kernel frameworks have been used in supervised learning approaches, e.g., Pairwise Support Vector Machines (SVMs), to predict relationships among two pairs of entities. Pairwise kernel methods are computationally expensive in terms of processing, especially when used to manipulate pairs of sequences, for example to predict metabolic networks. Rational kernels are based on transducers to manipulate sequence data, computing similarity measures between sequences or automata. Rational kernels take advantage of the smaller and faster representation and algorithms of weighted finite-state transducers. They have been effectively used in problems that handle large amount of sequence information such as protein essentiality, natural language processing and machine translations.
We propose a new framework, Pairwise Rational Kernels (PRKs), to manipulate pairs of sequence data, as pairwise combinations of rational kernels. We develop experiments using SVM with PRKs applied to metabolic pathway predictions in order to validate our methods. As a result, we obtain faster execution times with PRKs than similar pairwise kernels, while maintaining accurate predictions. Because raw sequence data can be used, the predictor model avoids the errors introduced by incorrect gene annotations. We also obtain a new type of Pairwise Rational Kernels based on automaton and transducer operations. In this case, we define new operations over two pairs of automata to obtain new rational kernels. We also develop experiments to validate these new PRKs to predict metabolic networks. As a result, we obtain the best execution times when we compare them with pairwise kernels and the previous PRKs.

**P33: Measuring and interpreting similarity between scale-free biological networks**

Qian Peng, Bingqing Xie, Gady Agam  
*Department of Computer Science, Illinois Institute of Technology, United States*

**ABSTRACT:** Biological networks such as metabolism networks, protein-protein interaction networks, and gene regulation networks, are used in numerous applications to reveal the functions of genes, proteins, and molecules. Measuring the similarity between such networks is important for both clustering algorithms and the validation of algorithms. In clustering algorithms the similarity measure is used to determine networks that can be grouped in the same cluster. In the validation of algorithms the similarity measure is used to determine the similarity between a synthetic network and an actual one. Synthetic networks are useful for algorithm validation because they can be synthesized in a manner where the ground truth is known (e.g. A network where the clusters are known).

Existing methods for measuring network similarity, such as NetSimile, do not target biological networks specifically and lack absolute interpretation of their measurements. In this paper we propose a principled metric using machine learning which consistently measures the similarity between biological networks and apply it to measure the similarity between actual networks and synthesized ones. In addition to improved performance, similarity in our approach has a meaning of edge rewiring percentage which makes interpreting absolute similarity results easier.

Our similarity classifier uses several network features such as: maximum degree, local and global clustering coefficients, degree exponent, degree distribution, and some other geometric features. To train our model we used a set of 140 actual biological networks for which we generated perturbed versions at various levels by randomly rewiring the edges in them. We use Random forest regression with 10 fold cross validation. Our cross validation results show that we can accurately estimate the known percentage of edge rewire with an average accuracy of roughly 5.5%.
After training the model we measured its performance on predicting the similarity between synthesized networks. In this evaluation we synthesized 100 networks using the Barabasi scale-free network synthesis algorithm. The objective was to measure whether our model can accurately estimate the difference between networks generated with various synthesis parameters (number of vertices and degree distribution). We compared the proposed approach to a standard implementation of NetSimile. We observed that in the proposed approach the produced measure monotonically increased as the difference in synthesis parameters increased whereas in NetSimilie this was not always the case. The paper provides the full details of this evaluation.

P34: Global Functional Annotation and Visualization of the 2015 Yeast Genetic Interaction Network

Anastasia Baryshnikova1, Michael Costanzo2, Chad Myers3, Brenda Andrews2, Charlie Boone2
1 Princeton University, United States; 2 University of Toronto, Canada; 3 University of Minnesota, United States

ABSTRACT: Large-scale biological networks map functional relationships between most genes in the genome and can potentially uncover high level organizing principles governing cellular functions. Despite the availability of an incredible wealth of network data, our current understanding of their functional organization is very limited and nearly inaccessible for biologists. To facilitate the discovery of functional structure and advance its biological interpretation, we developed a systematic quantitative approach to determine which functions are represented in a network, which parts of the network they are associated with and how they are related to one another. Our method, named Spatial Analysis of Functional Enrichment (SAFE), detects network regions that are statistically overrepresented for a functional group or a quantitative phenotype of interest, and provides an intuitive visual representation of their relative positioning within the network. Using SAFE, we examined the most recent genetic interaction network from budding yeast Saccharomyces cerevisiae, which was derived from the quantitative growth analysis of over 20 million double mutants. By annotating the genetic interaction network with GO biological process, protein localization and protein complex membership data, SAFE showed that the network is structured hierarchically and reflects the functional organization of the yeast cell at many different levels of resolution. In addition, we analyzed the network using a large-scale chemical genomics dataset and generated a global view of the yeast cellular response to chemical treatment. This view recapitulated the known modes-of-action of chemical compounds and identified a potentially novel mechanism of resistance to the anti-cancer drug bortezomib. Our results demonstrate that SAFE is a powerful tool for annotating biological networks and a unique framework for understanding the global wiring diagram of the cell.
P35: Memory of Inflammation in Regulatory T Cells
Joris van der Veeken, Alvaro J. González, Hyunwoo Cho, Aaron Arvey, Christina S. Leslie, Alexander Y. Rudensky
Memorial Sloan Kettering Cancer Center, United States

ABSTRACT: Regulatory T (Treg) cells are a specialized lineage of suppressive CD4 T cells that act as critical negative regulators of inflammation in various biological contexts. Treg cells exposed to an inflammatory environment undergo numerous transcriptional and epigenomic changes, acquire highly enhanced suppressive capacity, and show altered tissue homing potential. Whether these changes represent stable differentiation akin to memory T cells, or a transient adaptation to the inflammatory environment, is currently unclear.

We used an inducible lineage tracing system to analyze the long-term stability of inflammation-induced transcriptional, epigenomic, and functional changes in inflammation-experienced Treg cells. To this end, we performed an integrative computational analysis of ATAC-seq, histone modification (H3K27ac, H3K27me3, H3K4me1) ChIP-seq, and RNA-seq profiles of Treg cells before, during, and two months after exposure to an acute inflammatory environment. We found that Treg cells, in contrast to memory T cells, showed a striking ability to revert activation-induced transcriptional and epigenomic changes and maintained only a selective and specific memory of inflammation. Genes undergoing stable expression changes underwent qualitatively similar but more dramatic chromatin remodeling than genes undergoing transient changes. Stable gene expression changes were further reinforced during secondary Treg cell activation, while genes undergoing transient expression changes were similarly regulated during primary and secondary responses. Moreover, transiently expressed genes did not maintain stable chromatin modifications that would facilitate their reactivation. Importantly, while the activation-induced increase in Treg cell suppressive function was transient, inflammation-experienced Treg cells acquired a stable non-lymphoid tissue preference characterized by differential expression of tissue homing molecules. These data suggest that memory of inflammation allows Treg cells to preferentially localize to non-lymphoid organs to dampen ongoing tissue inflammation, without becoming stably hyperactive and causing an immunosuppressed state.

P36: Enhancing the detection of genomic rearrangements to better understand cancer pathology
Francesca Cordero¹, Marco Beccuti¹, Maddalena Arigoni², Raffaele Calogero²
¹ Department of Computer Science, University of Torino, Italy; ² University of Torino, Italy

ABSTRACT: Among the genome structural variants, the genomic rearrangements are one of the major sources of genetic diversity in human cancer. The chimera (or fusion) genes are derived by recombination event
formed by the breakage and re-joining of two DNA sequences. There are some intrinsic difficulties in the detection of these rearrangements due to both the experimental protocol used and the computational methodologies implemented to detect the fusion events. Fusion events occurring in a specific cell type are usually detected at transcription-level and results generated in different laboratories are only partially overlapping. A prototypical example is the case of MCF7 analysis for fusion detection Edgren et al. (Genome Biology, 2011), Kangaspeska et al. (PLoS One, 2012), Sakarya et al. (PLoS Computational Biology, 2012), Maher et al. (PNAS, 2009), and Inaki et al. (Genome Research 2011) used different tools and also sequencing was done using both Illumina and Solid sequencing technologies and starting from polyA selected RNA or totalRNA. To understand the effect of library preparation on fusion detection, we compare different sequencing protocols: polyA selection, ACCESS protocol and ribosomal depleted total RNA.

Our data indicated that sequencing polyA selected RNAs is the least effective method to detect MCF7 known fusions, while ribosomal depleted total is the most efficient.

Taken together our data and those previously published, MCF7 cell line represents an ideal model to evaluate the presence of specific genomic roles to define those sites involved in aberrant translocations. However, RNAseq does not provide information on the effective region in which the translocation is located. Thus, we have sequenced at the MCF7 genome at 35X and we have detected the breakpoint region of the MCF7 know fusions. Preliminary data indicates the presence of some patterns that are associated to these events. We are actually evaluating if we could identify genomic roles that could be also observed in translocations annotated in COSMIC database.

**P37: In-silico Analysis of Circular RNA as Regulators of miRNA**

**Nicholas Akers, Xintong Chen, Eric Schadt, Bojan Losic**

*Icahn School of Medicine at Mount Sinai, United States*

**ABSTRACT:** MicroRNA (miRNA) are well characterized as important non-coding regulators of cellular gene expression. Less well understood are the mechanisms that regulate miRNA. Recently, circular RNA (cRNA), have been described as a well expressed, non-coding, tissue specific RNA product with an ambiguous cellular function. One plausible hypothesis for the function of cRNA is to specifically bind cellular regulators such as miRNA. This hypothesis was informed by the discovery that cRNA ciRS-7 contains numerous binding sites for miRNA miR-7, allowing the cRNA to attenuate the effects of the miRNA. This finding has led to speculation that the cellular role of many cRNA is to ‘sponge’ miRNA as a tier of control over gene expression. To investigate this hypothesis we have aligned with BLAST the sequences of all known miRNA with
all reported cRNA to look for enrichment of miRNA binding sites. As negative controls we have also created a random nucleotide database of miRNA and aligned this to our cRNA database as well. Our results considered ~236M cRNA/miRNA pairs, and indicate that published cRNA are 41% more likely to contain a binding site for known miRNA than randomly generated miRNA (p<2.2x10^-16). In addition to this, ciRS-7 and miR-7 are among the strongest cRNA/miRNA pairs, residing in the top 0.001% of all combinations in binding sites per base pair. The next step in this experiment is to examine the effects of cRNA expression on miRNA targets. We are analyzing the effect of expression of the most prolific miRNA binding cRNA on mRNA targets of these miRNA in a public access RNA-Seq dataset. If the hypothesis that cRNA attenuate the effects of miRNA is true, we expect to find that mRNA targets of miRNA degradation will be increased with greater cRNA expression. The results of this analysis will be presented in order to provide clarity on the role of cRNA in regulating mRNA within the cell.

P38: A parallel negative feedback motif exhibits bidirectional control based on differential kinetics in cytokine regulatory networks

Warren Anderson¹, Hirenkumar Makadia¹, Andrew Greenhalgh², James Schwaber¹, Samuel David², Rajanikanth Vadigepalli¹

¹ Thomas Jefferson University, United States; ² McGill University, Canada

ABSTRACT: Negative feedback is critical for maintaining homeostasis within and between cells. Inflammatory immune diseases aberrant negative feedback interactions within cytokine regulatory networks. We developed a computational model of a macrophage cytokine interaction network to study the regulatory mechanisms of macrophage-mediated inflammation. We established a literature-based cytokine network, including TNF, TGF, and IL-10, and fitted a mathematical model to published data from LPS-treated microglia (brain macrophage). We evaluated the validity of our model by testing whether it could recapitulate the experimentally determined “tolerance” response to the endotoxin LPS. We applied two doses of LPS and determined the gain of the peak TNF responses. Our results were consistent with published experimental data demonstrating tolerance to LPS. Global sensitivity analysis revealed that TGF - and IL-10-mediated inhibition of TNF was critical for regulating network behavior. Further analysis revealed that TNF exhibited adaptation to sustained LPS stimulation. We simulated the effects of functionally inhibiting TGF and IL-10 on TNF adaptation. Our analysis showed that TGF and IL-10 knockouts (TGF KO and IL-10 KO) exert divergent effects on adaptation. TGF KO attenuated TNF adaptation whereas IL-10 KO enhanced TNF adaptation. We experimentally tested the hypothesis that IL-10 KO enhances TNF adaptation in murine macrophages and found supporting evidence. Next, we tested the effect of IL-10 and TGF KO on tolerance using our computational model. Surprisingly, we found...
that IL-10 KO enhanced tolerance of the TNF response to sequentially applied LPS doses. In contrast, TGF KO repressed LPS tolerance. These opposing effects could be explained by differential kinetics of negative feedback. Inhibition of IL-10 reduced early negative feedback that results in enhanced TNF-mediated TGF expression. To further assess whether the relative effects of IL-10 and TGF could be explained by their differential kinetics, we adapted our macrophage model to a 3-node system. We found that the 3-node parallel negative feedback topology supported robust adaptation and tolerance. Inhibition of relatively fast negative feedback enhanced adaptation and tolerance. In contrast, inhibition of relatively slow negative feedback attenuated adaptation and tolerance. We propose that differential kinetics in parallel negative feedback loops constitute a novel mechanism underlying the complex and non-intuitive pro-versus anti-inflammatory effects of individual cytokine perturbations. Based on the data from our reduced 3-node network, we posit that parallel negative feedback motifs with differential kinetics can be tuned for bi-directional control (i.e., negative and positive influences) in contexts ranging from intracellular biochemical signaling to inter-cellular interactions.

**P39: Sequence biases in CLIP experimental data are incorporated in protein RNA-binding models**

**Yaron Orenstein and Bonnie Berger**

*MIT, United States*

**ABSTRACT:** Protein-RNA interactions play important roles in many processes in the cell. CLIP-based methods measure protein RNA-binding in vivo in a high-throughput manner on a genome-wide scale. In these technologies, the protein is cross-linked to the RNA and pulled down. The protein is then removed following the cleaving of the RNA by a restriction enzyme. Later, bound RNA segments are sequenced and mapped back to the genome to be called as peaks.

Here, we present a newly-identified bias in CLIP peaks, which we call the ‘terminating G’. Most called peaks terminate in a G, since RNase T1 cleaves at accessible G’s much more strongly than at other nucleotides. The fact that most raw sequences do not terminate at a G implies that this bias is introduced in the peak calling process. Unfortunately, protein RNA-binding preferences are not easily disentangled from enzyme specificities. Thus, we call for an appropriate experimental control to measure the cleaving enzyme specificities. These should later be incorporated as co-variants in the peak calling process to identify unbiased binding sites. Then, better algorithms may be developed to predict more accurate binding sites.
P40: A novel study of the scope and limitations of baker’s yeast as a model organism for human tissue-specific pathways

Shahin Mohammadi¹, Baharak Saberidokht¹, Shankar Subramaniam², Ananth Grama¹

¹ Purdue University, United States; ² University of California, San Diego, United States

ABSTRACT: Budding yeast, S. cerevisiae, has been used extensively as a model organism for studying cellular processes in evolutionarily distant species, including humans. However, different human tissues, while inheriting a similar genetic code, exhibit distinct anatomical and physiological properties. Driving biochemical processes and associated biomolecules that mediate the differentiation of various tissues are not completely understood, neither is the extent to which a unicellular organism, such as yeast, can be used to model these processes within each tissue.

We propose a novel computational framework coupled with the corresponding statistical model to assess the suitability of yeast as a model organism for different human tissues. Using our method, we dissect the functional space of human tissue-specific networks according to their conservation both across species and among different tissues.

Using a case study of GNF Gene Atlas dataset, we classify different tissues based on their similarity to yeast. In cases where suitability of yeast can be established, through conservation of tissue-specific pathways in yeast, it can serve as an experimental model for further investigations of new biomarkers, as well as an unbiased phenotypic screen to assay pharmacological and genetic interventions. On the other hand, for tissues with missing functionality in yeast, we provide molecular constructs (gene insertions) for creating more appropriate, tissue-engineered humanized yeast models.

P41: Deciphering single-cell transcriptional heterogeneity to understand principles of neuronal phenotype organization and plasticity

James Park¹, Babatunde Ogunnaike¹, James Schwaber², Rajanikanth Vadigepalli²

¹ University of Delaware, United States; ² Thomas Jefferson University, United States

ABSTRACT: Reconciling a cell’s transcriptional state and its phenotypic function is confounded by the transcriptional heterogeneity observed at the single-cell scale. This transcriptional heterogeneity conflicts with the traditional expectation that a neuronal phenotype consists of functionally identical neurons that respond uniformly to synaptic and neuromodulatory inputs. Moreover, this transcriptional heterogeneity is prominent within and across post-mitotic neuronal populations.
throughout the brain, where neurons interact to form circuits that regulate physiological function. High-throughput “-omic” level analysis, however, suggests that a more complex molecular organization potentially underlies neuronal phenotypic function and emergent systems-level behavior that occurs in the brain. In order to understand the functional relevance of this transcriptional heterogeneity, we examined two distinct brain nuclei by analyzing the transcriptional responses of individual neurons responding to specific physiological perturbations. In the first case, we generated a high-dimensional gene expression data set from individual blood pressure-regulating neurons within the nucleus tractus solitarius (NTS) that were collected from rats undergoing an acute hypertensive challenge. In the second case, we analyzed the transcriptional states of hundreds of single neurons within the suprachiasmatic nucleus (SCN) from mice responding to a light-induced phase shift in circadian rhythms. Using a combination of multivariate analytical techniques, graph network theory, and a novel fuzzy logic-based regulatory network modeling methodology, we identified molecular organizational structures in which individual neurons from both brain nuclei form distinct transcriptional states that align with synaptic/neuromodulatory inputs. Concomitantly, our quantitative regulatory network models and simulations of NTS neurons suggest that distinct networks correspond to these subtypes and drive heterogeneous gene expression behavior in a continuous fashion. Within the SCN, the presence of transcriptionally distinct neuronal subtypes provides insight into the organization and intercellular interactions underlying SCN regulation of circadian function. Having identified these SCN neuronal subtypes, we are now able to postulate a cellular interaction network in which specific neuronal subtypes fulfill specific functional roles in regulating circadian phase-shift behavior.

P42: microRNA-mediated feed forward disinhibition of multiple functional pathways amplifies prohypertensive signaling

Danielle Decicco, James Schwaber, Rajanikanth Vadigepalli
Thomas Jefferson University, United States

ABSTRACT: microRNAs have emerged as novel post-transcriptional regulators of many cellular disease processes. However, in essential hypertension, there has been no characterization of the microRNA expression landscape in key neuroanatomical blood pressure control regions during hypertension development. Using a global analysis of microRNA expression levels in these regions, we quantified 419 well-annotated microRNAs in the brainstem, and we identified 24 microRNAs showing stage-dependent differential expression in hypertensive rats compared to controls. We constructed microRNA regulatory networks based on predicted targets from bioinformatic databases including RNA22 and miRWALK. Our microRNA regulatory networks indicated that predicted targets primarily fell into functional pathways previously
associated with hypertension such as inflammation and Angiotensin II signaling. We measured the putative targets using high-throughput qPCR to evaluate correlations between microRNAs and their predicted gene targets. Our analysis revealed a similar extent of positive and negative correlations between the microRNA and predicted target transcript patterns suggesting regulatory relationships. We discovered a pair of microRNAs, previously shown to be enriched in different cells types: miR-135a in astrocytes and miR-376a in neurons, which demonstrated stronger anti-correlational relationships with their putative targets in the hypertensive state compared to controls. These microRNAs demonstrate expression levels which are negatively correlated with key target expression levels in the inflammation and Angiotensin II pathways. Interestingly, the key putative targets are known inhibitors of these functional pathways that show increased activity in hypertension. Such feed forward disinhibition by microRNA-135a and microRNA-376a of the inflammatory and Angiotensin II pathways occurred at the onset of hypertension suggesting a mechanistic role for this regulatory network. Given that both pathways are hyperactive in the chronic hypertensive stage, microRNA regulatory network-mediated disinhibition of those pathways at the onset stage is likely to have a causal effect of amplifying those pathways, contributing to the development of hypertension. This feed-forward disinhibition by miR-135a and miR-376a suggests synergistic network activity contributing to the development of hypertension.

P43: Prioritizing animal models for human diseases using genome-wide functional networks

Max Homilius¹, Arjun Krishnan¹, Calum MacRae², Olga Troyanskaya¹
1 Princeton University, United States; 2 Brigham and Women’s Hospital, Harvard Medical School, United States

ABSTRACT: Model organisms are key to studying the molecular basis of human traits and diseases. Therefore, for a rare or common disease defined by a group of implicated genes, it is valuable to identify relevant model organism phenotypes to transfer knowledge and propel further investigation. Yet, we lack tools to seamlessly search across organisms to identify the model phenotype equivalent to a human disease (or the human disease corresponding to a model phenotype of interest). The most straightforward approach – mapping disease to phenotype based on overlapping homologous genes – is severely limiting because, 1) our knowledge of associated genes for most diseases and phenotypes is largely incomplete, thus leaving many actual disease-phenotype pairs with little to no ‘common’ genes; 2) treating diseases and phenotypes as bags of genes ignores the underlying complex organism-specific biology. Here we present a framework for systematically matching diseases and phenotypes that overcomes both of these limitations. By jointly using genome-scale functional gene interaction networks in both human and the model organism, we create and match genome-wide representations of human...
diseases and model phenotypes, and further filter nonspecific matches to arrive at highly resolved disease-phenotype mappings. Further, for each disease-phenotype pair, in addition to known genes, we report the novel homologous genes most associated with the disease/phenotype, which are prime candidates for experimental follow-up. We have made our approach available through a dynamic web-interface that allows researchers to easily use their own gene set (or a previously known disease/phenotype) to query a large collection of resources containing disease-gene and phenotype-gene associations in human and five model organisms (mouse, zebrafish, fly, worm and yeast). Users can readily see prioritized diseases/phenotypes, list candidate genes, explore them in the context of the underlying networks, and export all results.

**P44: Discovery of bruchid resistance-related variations in regulatory regions by genome-wide sequence comparison**

Dung-Chi Wu¹, Mao-Sen Liu¹, Tony Chien-Yen Kuo¹, Kuan-Yi Li¹, Roland Schafleitner², Hsiao-Feng Lo³, Long-Fang O. Chen¹, Chien-Yu Chen⁴, Chia-Yun Ko¹, Huei Mei Chen²

¹ Institute of Plant and Microbial Biology, Academia Sinica, Taiwan; ² AVRDC-the World Vegetable Center, Taiwan; ³ Department of Horticulture and Landscape Architecture, National Taiwan University, Taiwan; ⁴ Dept. of BIME, National Taiwan University, Taiwan

**ABSTRACT:** Mungbean, Vigna radiata [L.] R. Wilczek, is one of the most important legume crops with valuable nutritional and medical value. The bruchid beetle (Callobruchus maculates), known as bean weevils, would attack mungbeans both in the field and in storage, resulting in great losses in the stored grains. A wild mungbean, V. radiata var sublobata (TC1966) from Madagascar, was resistant to many bean weevils and with the ability of crossing with V. radiata. Though little knowledge has been uncovered for weevil resistance, breeding for bruchid resistance is still a major goal in mungbean studies. In this study, we first de novo assembled the genome of a bruchid-resistant recombinant inbreeding line 59 (RIL59) which derived from TC1966 and a bruchid-susceptible variety NM92. The primitive genomic data was combined with additional genome and transcriptome analysis for different levels of bruchid-resistance mungbean lines, including the two parent, TC1966 and NM92, and the other 12-inbred-generation progenies, to investigate where might the major distinct loci between bruchid-resistant and bruchid-susceptible lines. The ab initio predicted gene models of RIL59 consist of 44,317 genes, representing 49,952 transcripts. The genome-wide variation analysis performed on NM92, TC1966 and RIL59 revealed that 3,162 genes have sequence variants, including non-synonymous substitutions and INDELs, on exons to cause protein sequence changes. These genes were suspected to be related to the bruchid-resistance. On the other hand, a draft bruchid-susceptible mungbean (Vigna radiata var. radiata VC1973A) genome was previously published. We mapped the above-mentioned putative bruchid-
resistance-related genes to this bruchid-susceptible draft genome and found a hot region on Vr05. This result was consistent with the genotype-by-sequencing (GBS) data which also suggested that the region from 5M bps to 12M bps of Vr05 is strongly related to bruchid-resistance. These two draft genomes were aligned to identify 127 scaffolds of RIL59 that together correspond to the Vr05 of VC1973A. Among them, about 50 scaffolds were considered associated with this region. In total, 508 genes were identified in these scaffolds. If considering the upstream 2,000 bps of each gene model as the promoter, there were 544 promoters falling in the suspected resistance-related region. Comparison on the promoters between the bruchid-resistant and bruchid-susceptible mungbean lines discovered some large structure variations, suggesting the gain or loss of regulatory elements might play key roles in bruchid resistance. In summary, the comparison of promoters of the two draft genomes reveals the potential impact of regulatory regions in affecting resistant phenotypes of mungbeans.

**P45: Empirical Evidence Supporting a Systematic Approach to Gene Network Identification**

Sweta Sharma and Desmond Lun
*Rutgers University, United States*

**ABSTRACT:** A major cellular systems biology challenge of the past decade has been the development of a comprehensive model for gene regulatory networks (GRNs). Particularly, there is growing impetus for the extraction of regulatory information from expression data as it becomes increasingly available and accurate. Identifying networks from such information requires deciphering direct interactions from indirect ones. For instance, if gene A regulates gene B and B regulates gene C, then changing A’s expression will directly affect B’s expression while indirectly affecting C’s.

Recently, Birget et al proposed a systematic approach for network identification. They consider a binary model that captures the non-linear dependencies of GRNs and reverse-engineer the network using assignments (perturbations to the expression level of a single gene) and whole transcriptome steady-state expression measurements. Under this model, their approach achieves identification of acyclic networks with worst-case complexity costs in terms of assignments and measurements that scale quadratically with the size of the network. For networks with cycles, the worst-case complexity cost scales cubically.

We conduct a proof-of-concept experiment for this approach by reverse-engineering a five-gene sub-network of the outer-membrane protein regulator (ompR) in E. coli. Through assignments achieved by gene deletions and expression measurements from qPCR, we successfully identify the regulatory relationships and discern direct from indirect
interactions. We also performed computational experiments on in silico networks derived from known regulatory relationships in E. coli and S. cerevisiae, where gene regulation is thermodynamically modeled using the system of ODEs that was used to generate data for previous DREAM challenges. We achieve 100% identification for noiseless acyclic networks of size ranging from 100 to 4,000 genes. For noisy acyclic E. coli networks of size 100, we obtain an AUPR of .95. This is significantly improved from the .71 AUPR obtained by the top performer in the DREAM3 inference challenge for acyclic in silico networks. Furthermore, we achieve this using ten-fold fewer assignments and measurements. For noisy cyclic E. coli networks of size 100, we obtain an AUPR of .75, compared to .45 for the top performer in the DREAM4 InSilico_Size100 sub-challenge containing cyclic networks. We achieve this using roughly the same number of assignments and half as many measurements.

Taken together, our results imply that the reverse engineering method of Birget et al is not only experimentally feasible but uses reasonable resources. It can therefore serve as the basis for systematic, accurate reverse engineering of large-scale gene regulatory networks.

**P46: Dysregulation of co-regulatory microRNA networks by chronic ethanol consumption leads to impaired liver regeneration**

**Austin Parrish, Egle Juskeviciute, Jan Hoek, Rajanikanth Vadigepalli**

*Thomas Jefferson University, United States*

**ABSTRACT:** microRNAs are a class of small, non-coding RNAs ~21 nucleotides long that regulate numerous cellular processes in a post-transcriptional manner. Previous research has identified several microRNAs of interest involved in liver regeneration and hepatocellular carcinoma, including miR-21, which has been shown by our lab to be significantly upregulated following liver damage by 70% partial hepatectomy, along with chronic ethanol consumption. Given that microRNAs often exert their effects in regulatory networks that display both positive and negative cooperation, we sought to identify additional microRNAs involved in liver regeneration alongside miR-21. In order to accomplish this, we performed in vivo knockdown of miR-21 using a locked nucleic acid (LNA) probe containing a complementary sequence to miR-21. Whole liver tissue samples were collected from both control and ethanol-fed Sprague-Dawley rats at baseline conditions and 24 hours post-partial hepatectomy. These samples were analyzed for microRNA expression using the NanoString microRNA microarray platform. Analysis of the expression data reveals twelve microRNAs that show differential expression in response to miR-21 knockdown. Of these genes, three show positive correlations with miR-21 expression while eight are negatively correlated. Using target prediction software, we developed a network of putative microRNA-gene interactions and compared the
predicted targets to genes identified as differentially expressed based on Affymetrix microarray analysis. This network of putative targets identifies a number of genes that are potentially regulated by these differentially expressed microRNAs. Gene ontology and pathway analysis reveals that multiple predicted targets are involved in processes relating to cell cycle progression. In conclusion, these studies identified a set of co-regulatory microRNAs whose dysregulation by chronic ethanol consumption may lead to impaired liver regeneration.

**P47: Furthering understanding of Parkinson’s Disease through integrative analysis in C. elegans**

Victoria Yao, Rachel Kaletsky, Coleen Murphy, Olga Troyanskaya  
*Princeton University, United States*

**ABSTRACT:** The etiology of complex human diseases, especially in the context of aging, such as Parkinson’s disease, is likely a combination of many environmental and genetic factors. Elucidating the molecular basis of pathophysiologies of such diseases requires a combination of systems-level studies in human and model systems. The nematode *C. elegans* is an effective and efficient model for human disease due to its sufficient complexity and high genetic conservation with humans, combined with short lifespan and the abundance of genetic tools and assays. In particular, the complexity of *C. elegans* at the tissue level allows for in depth investigations of relevant diseases in a tissue-specific manner. To this end, we developed a novel semi-supervised regularized Bayesian integration method to integrate a large compendium of heterogenous datasets for the construction of 203 tissue- and cell-type specific networks in *C. elegans*. We demonstrate the accuracy of these networks in detecting tissue-specific functional signal, even for very small and specific tissues and cell types. We then use the dopaminergic neuron worm network combined with Parkinson’s disease genes identified in quantitative genetic studies in human to predict new genes implicated in Parkinson’s disease. A subset of these predictions has been experimentally confirmed to have Parkinson’s disease endophenotypes in *C. elegans* and are conserved in human, providing potential therapeutic targets.

**P48: Transcription Network Inference using RNA Expression and Degradation Rate Data in S. cerevisiae**

Konstantine Tchourine\(^1\), Christian L. Mueller\(^2\), Chirstine Vogel\(^1\), Richard Bonneau\(^1\)  
\(^1\) NYU - Center for Genomics and Systems Biology, United States; \(^2\) Simons Center for Data Analysis, United States

**ABSTRACT:** Despite many years of research and the availability of large-scale datasets, modeling RNA transcription and predicting transcriptional regulatory interactions on a systems level in eukaryotes remains a challenging problem and requires modeling changes in RNA
abundance due to both the regulation of synthesis and degradation. Even Saccharomyces cerevisiae has several hundred putative TFs and ~6,000 potential targets, rendering the theoretical regulatory interaction space enormous. Further, eukaryotes are marked by extensive promoter regions, many response pathways, and additional regulatory layers, e.g. RNA decay, which further confound gene expression regulation. For these reasons, even the best network inference algorithms have so far performed very poorly in yeast. To address this challenge, we are taking several steps towards constructing the first high-quality, high-coverage yeast regulatory network. I am developing an expanded version of an existing gene regulatory inference framework, Inferelator-BBSR, that incorporates RNA decay rates to predict new regulatory interactions, estimate each interaction's contribution to the dynamics of the transcription process, and estimate gene-dependent RNA decay rates. Incorporation of RNA decay rates can be done either computationally by finding optimal decay rates for different modes of regulation in yeast, or empirically by directly incorporating RNA decay rate data into the inference procedure. In this presentation, I will show that both ways of incorporating RNA decay rates into the inference framework improve regulatory network inference. Furthermore, I will show that the inferred regulatory network can help identify different modes of stress adaptation which require different average RNA decay rates.

P49: The optimized high-throughput siRNA screening : Applications in cancer target discovery

Nayoung Kim and Sukjoon Yoon
Sookmyung Women's University, The Republic of Korea

ABSTRACT: RNA interference (RNAi) has become a powerful tool for drug target discovery, and the systematic loss-of-function screens using RNAi libraries can now be performed to identify the biological functions of specific genes or pathways in various diseases. Cancer target discovery studies on clinically relevant drug applications and their mode of actions can be accelerated by integrating multi-level omics data such as genome, transcriptome, proteome and phosphatome data together with siRNA screening data.

We introduce the siRNA screening platform composed of the image-based assay optimization, primary screening, data analysis and hit selection criteria using some studies to investigate novel therapeutic targets in cancer. We applied two different samples to siRNA screening. One example is a study using a specific gene-knockdown cell line. In this study, in order to identify novel therapeutic targets in STK11-deficient lung cancer cells, we utilized a large-scale siRNA screening to identify genes that would sensitize STK11-deficient lung cancer cells (A549) with or without AMPK. And another example is a genome-wide siRNA screening using a sphere-forming (3D) culture system similar to in vivo. 3D growths of cancer cells in vitro are more reflective of in situ cancer cell growth than growth in
monolayer (2D). This study is designed to identify genes reducing sphere size on 3D as compared to 2D.

In the study using a stable knockdown cell line, the perturbation of several genes exhibited significant inhibitory effect on the growth of AMPK-knockdown cells. And we identified that specific hits inducing inhibition of cell growth with AMPK knockdown were related to metabolism and signal transduction among various functional categories. These results highlight the potential of synthetic lethal siRNA screens with AMPK inhibitors to define new determinants of potential therapeutic targets. And in another screening using 3D culture system, we found specific genes reducing sphere formation. These hits were related to lipid metabolism. From these results, we can find new therapeutic target-related drugs for inhibition of tumor progression and metastasis.

This screening platform can be provided as a valuable tool to find novel therapeutic targets and drugs for cancer therapy. We now provide this platform service to academic and industrial organizations.

**P50: MACE: a web-based application for analyzing mutation-specific drug response and gene expression in cancers**

**Yourae Hong** and **Sukjoon Yoon**
*Sookmyung Women’s University, The Republic of Korea*

**ABSTRACT:** Systematic understanding of mutation-oriented drug sensitivity on cancer cell lines will provides therapeutic benefits on the cancer therapy. Here, we present the MACE database as a web-based interactive tool for interpreting drug response and gene expression in the genotypic classification of cancer cell lines. Chemical screening and DNA microarray data on NCI60 cell lines were organized to identify mutation- or lineage-specific chemicals and gene expression signatures. In this system, users can perform the individual and combined analysis to find potential associations of chemicals and genes with major gene mutations of cancers. The present MACE database can be used to understand how gene mutation is interconnected with the drug response and gene expression in cancer subtypes. This database provides a valuable tool to predict and optimize the therapeutic window for anticancer agents and related gene targets. The MACE web database is available at http://mace.sookmyung.ac.kr/.

**P51: Reverse engineering gene regulatory networks from structural and epigenetic datasets**

**Brittany Baur** and **Serdar Bozdag**
*Marquette University, United States*

**ABSTRACT:** One of the major challenges in computational biology is the identification of “driver” copy number changes that promote cancer cell progression. The goal of this study is to identify genes within regions
with aberrant copy number and DNA methylation changes that have widespread downstream effects, and their associated targets. We first identified these aberrant regions by integrating DNA methylation or copy number datasets with gene expression datasets in luminal A breast cancer patients. We then identified candidate genes within the aberrated regions which could act as regulators of downstream targets by integrating the expression levels of the regulators and potential targets with pathway analysis. Based on gene ontology, we established that genes associated with aberrant copy number and DNA methylation changes are enriched in terms associated with the regulation of various biological processes. This indicates that these genes are potentially regulators of other genes. We identified several candidate genes within these regions that are likely regulators strongly affected by copy number or DNA methylation aberrations. By identifying causal genes within the aberrant regions, this study could aid in the discovery of therapeutic targets of cancer drugs.

**P52: Alternative Splicing During Heat Stress in Arabidopsis thaliana**

Gaurav Kandoi and Julie A Dickerson

*Iowa State University, United States*

ABSTRACT: Alternative splicing (AS) which produces multiple messenger RNAs by different combination of various regions of the precursor transcript is a major cause of diversity in gene products. Recent estimates suggest the rate of AS to be as high as 95% in humans and 60% in plants. Despite the prevalence of AS events, their functional consequences are largely unknown. Although the impact of abiotic stresses (temperature, salt, light etc.) on AS events in Arabidopsis thaliana has been widely studied, not much is known about how differential splicing affects the metabolic pathways under such stress conditions. High-throughput RNA sequencing (RNA-seq) data from a heat stress experiment in A. thaliana, was used to find regions which undergo differential splicing. Even though heat stress leads to an increase in the number of AS events, only ~90 of such alternatively spliced genes are also differentially spliced (DS) between the two conditions. Most of these are nuclear genes and have been annotated with biological processes such as response to stress, response to abiotic or biotic stimulus and cell organization, and biogenesis. A significant portion of these differentially spliced genes are also linked with molecular functions like binding (DNA or RNA, nucleotide, protein and nucleic acid) and enzymatic activity (transferase, hydrolase and kinase). For the most part, the novel spliced isoforms are predicted to be more abundant than the normal transcript in the heat stress condition. Conserved domain analyses indicate that novel spliced isoforms share similar domain architecture with the normal transcripts more often than not. By studying the effect of such alternative splicing events on protein function, we can identify important metabolic networks. Combination of these differential networks across the spectrum of stress conditions generates metabolic models with a high-level regulatory framework.
P53: An integrated computational pipeline for analysing genetic, molecular, and functional variations in complex diseases

Bajuna Salehe, Chris Jones, Giuseppe Di Fatta, Liam McGuffin
University of Reading, United Kingdom

ABSTRACT: The ongoing advancement of the technologies used for generating ‘omic data has led to the flood in biological data. This ‘big data’ phenomenon has increased the challenge for biologists and biomedical experts of finding a better analytical strategies that are capable of integrating variation data from different omic states, using integrated computational approaches to further understanding phenotypes (complex diseases/polygenic traits). Here, we have proposed an ‘omic variation framework focusing initially on single nucleotide polymorphisms (SNPs), which is one of the key ‘omic variation types that are studied in order to understand the relationships underpinning complex traits. However, this framework should also be adaptable to other ‘omic variation data, such as methylomic, transcriptomic and copy number variation (CNVs). Furthermore, we have designed a pipeline for an integrated computational approach to implement this framework, which we have applied to study platelet proteomic data sets. In this case study the aim is to understand the association of SNPs at different levels with the adenosine diphosphate (ADP) activated platelet response. Platelets play key roles in the thrombus formation, which is one of the major risks for cardiovascular diseases (CVDs), and ADP activated platelet response is highly involved during the thrombus formation, as well as being variable among individuals. Using the initial implementation of this pipeline we have been able to identify key genetic variants (SNPs) such as rs6141803 and rs7007145 in PTK2B and COMMD7 genes respectively that are significantly associated with platelet aggregation. Many of our identified SNPs were previously unidentified, and have been independently reported to be associated with the risk of CVDs.

P54: Exploration of Breast Cancer Genes and Bioinformatics Analyses

Shahrzad Eslamian and Jonathan P. Leidig
Grand Valley State University, United States

ABSTRACT: Information visualization may be applied to bioinformatics research tools to assist in understanding the complex (often textual) datasets. The main goal of this work was to design an interactive visualization tools to detail the genes potentially responsible for breast cancer as they are discovered through bioinformatics analysis. The dataset is derived from the publically shared research as maintained by the bioinformatics research community. The visualization aims to detail the explicit relationships and existing analyses of these target genes and their related micro RNA, considering the distributed nature of this field of research and disaggregation of the underlying datasets.
P55: Spectral coherence classification of uORF translation in a neuroblastoma cell model of differentiation

Sang Chun, Caitlin Rodriguez, Peter Todd, Ryan Mills
University of Michigan, United States

ABSTRACT: Upstream open reading frames (uORFs) are prevalent in the human transcriptome and may negatively regulate the abundance of canonically encoding proteins through the promotion of mRNA decay and competitive expression, among other mechanisms. uORFs are conserved across species and have been annotated to genes with diverse biological functions, including but not limited to oncogenes, cell cycle control and differentiation, and stress response. As such, the aberrant expression of certain uORFs has been implicated in the development and progression of various diseases. Therefore, the positive identification and validation of uORFs as translational products is critical for understanding their role in complex biological processes and disease etiology. Where mRNA-Seq has been used to approximate the transcriptomic content of a cell, or group of cells, the recently developed method of sequencing ribosome-protected fragments aims to profile the translational landscape of a sample. In concert, various algorithms have been developed to differentiate coding transcripts from non-coding transcripts based on the alignment of ribosome-protected fragments to a reference transcriptome. We have developed a classification algorithm based on the magnitude of coherence between the aligned ribosome profiling reads and tri-nucleotide periodic signal inherent to protein-coding sequences. In this study, we compare our spectral coherence-based classification algorithm (SPECtre) against existing methods and apply our approach to positively identify variably translated uORFs related to differentiation of SH-SY5Y neuroblastoma cells.

P56: Transcriptional Regulatory Networks During the Endothelial-to-Hematopoietic Transition in the Mouse Embryo

Long Gao1, Joanna Tober2, Peng Gao1, Jianshu Zhang1, Changya Chen1, Nancy A. Speck2, Kai Tan1
1 The University of Iowa, United States; 2 University of Pennsylvania, United States

ABSTRACT: Hematopoietic stem cells (HSCs) in the embryo are derived from hemogenic endothelia (HE) of the arterial wall from the aorta/gonad/mesonephros (AGM) region and yolk sac (YS). HE from AGM and YS has different developmental potentials. HE from YS primarily produces committed erythroid/myeloid progenitor and HE from AGM can produce lymphoid progenitors and HSCs. The transcriptional regulatory networks (TRN) that control the endothelial-to-hemogenic transition in AGM and YS are poorly understood. Here we compared the transcriptomes of
endothelium and hemogenic endothelium from embryonic (E) day 9.5 and E10.5 AGM and YS by RNA-Seq. We developed a novel computational method for constructing condition-specific transcriptional regulatory networks (TRNs) by sample elimination and network comparison with limited number of samples. By modeling developmental-stage-specific TRNs, we identified 73 gene modules (1429 genes) with differential activities between E and HE and between AGM HE and YS HE. We further identified a number of transcription factors that regulate the endothelial-to-hemogenic transitions, including Runx1, Sox7, Hoxa7, and Hoxd9. Long intergenic noncoding RNAs (lincRNAs) have been shown to regulate the development of various lineages. However, nothing is known about the role of lincRNAs during embryonic hematopoiesis. We identified 18 and 41 novel lincRNAs that are specifically expressed in E and HE, respectively. Among them, 10 lincRNAs were differentially expressed between E and HE, suggesting a role in regulating the development of hemogenic endothelium. In summary, our systematic analysis of the transcriptomes during endothelial-to-hemogenic transition has uncovered a number of novel regulators and gene pathways of this critical developmental transition.

P57: Dynamic organization and activation of enhancers and super-enhancers dictate effector and memory CD8+ T cell responses

Bing He, Haihui Xue, Kai Tan
University of Iowa, United States

ABSTRACT: CD8 T cells are critical in controlling infection by intracellular pathogens including viruses and intracellular bacteria. Differentiation of naïve CD8 T cells (TN) to effector (TE) and memory CD8 T cells (TCM) is accompanied by dynamic gene expression and epigenetic modification changes at promoters as revealed by previous analyses. However, there is virtually no information regarding the dynamics of enhancers during CD8 T cells responses to date. Here, we have mapped four histone modification marks in TN, TE, and TCM cells after viral infection. Our results suggest that the chromatin environment at regulatory DNA sequences in TCM is more permissive than in TN and TE. We further predicted the enhancers, super enhancers, and their targets, and constructed condition-specific transcriptional regulatory networks (TRNs) in three T cell stages. We have identified a highly dynamic repertoire of the enhancers and their targets during CD8 T cell responses, as 77% of the enhancers and 82% of the enhancer-promoter interactions are stage-specific. Our results suggest the dynamic change of enhancer activity during cell stage transition leads to TRN rewiring, which explains the expression change of the key factors of T cell function.
P58: Data and Computing Platform to facilitate NCER-PD (National Centre of Excellence in Research on Parkinson’s Disease) Project
Venkata Satagopam, Peter Banda, Jan Martens, Joachim Kutzera, Kirsten Roomp, Wei Gu, Patrick May, Reinhard Schneider
LCSB, University of Luxembourg

ABSTRACT: The Data and Computing Platform provides key infrastructure for the integration, curation and interrogation of anonymized clinical and experimental data. The platform manages multidimensional data associated with clinical research, including patient data, sample-associated information, and high-throughput molecular readouts from these samples. These different data flows are integrated at their source with the help of advanced data capture and transfer approaches. Clinical data can be entered remotely, via electronic forms at the time of collection, assuring their integrity and standardization. To attain this goal, REDCap[1], a state-of-the-art clinical research data management system has been implemented. All entered data will be immediately anonymized and sample-associated data will be accessed directly at their storage location, the IBBL, via secure communication with the LIMS of the biobank. High-throughput experimental data will be uploaded directly to the database service provided by LCSB for handling large, heterogeneous biomedical datasets: the tranSMART system[2]. tranSMART enables sharing, integration, standardization and analysis of heterogeneous data from collaborative translational research. It is used in pharmaceutical industry and in Innovative Medicine Initiative projects (e.g. eTRIKS[3], AETIONOMY[4]) to store and share curated phenotypic data such as clinical observations and adverse events; omics data like transcriptomics, proteomics, metabolomics and genotyping. Well-grounded machine learning and computational modeling approaches will enable data analysis and interpretation.


P59: Emergent Topological and Statistical Properties of Gene Regulatory Grids
Wilberforce Ouma, Mohammadmahdi Yousefi, Andrea Doseff, Erich Grotewold
Ohio State University, United States

ABSTRACT: Gene regulatory grids (GRGs) are static representations of gene regulatory networks (GRNs) encompassing all possible regulator-target gene interactions that provide a system-wide view of transcriptional gene regulation. To understand their architectural organization, we constructed and investigated emergent topological and statistical properties of GRGs of the following model organisms: Caenorhabditis elegans, Drosophila melanogaster, Saccharomyces cerevisiae and Arabidopsis
thaliana. We implemented a formal statistical approach for fitting a power-law function to the empirical degree distribution of the grid and observed that the out-degree, and not the in-degree, follows a power-law distribution, suggesting a scale-free property of GRGs that have transcription factors as hubs. The four GRGs however exhibit different power-law exponents. A computational sub-sampling of sub-grids from the original grids showed that for D. melanogaster, the exponent was invariant for a large number of sub-grids. With this invariant property of the exponent, we derived a mathematical formulation that estimates the number of interactions in a fully-connected grid. We hypothesize that a consequence of the scale-free property in cellular networks is reduction of the average path-length in the grid, resulting in faster signal propagation.

P60: Integrated Feature Detection From Chromatin State Measurements

Anastasia Shcherban, Matti Nykter, Juha Kesseli
Institute of Biosciences and Medical Technology (BioMediTech), University of Tampere, Finland

ABSTRACT: Identification and extraction of biologically relevant features, such as active regulatory regions from high throughput sequencing (HTS) data is a major challenge. A number of algorithms have been developed to perform feature detection on HTS data. In most cases, features are detected from one track of HTS data at a time, followed by further analysis and integration of peak detection results obtained from multiple sources. Thus, these methods fail to efficiently discover and utilize the complementary information found in multiple signals.

In this work we present our approach extending an existing algorithm called ZINBA (Zero-inflated negative binomial algorithm) to analyze multiple data tracks simultaneously. Our goal is to shed light on the relationships between data tracks using our statistical model while improving detection results for features that can be found from a selected pair of tracks. The statistical model is built by incorporating a correlation term for HTS data, so that the algorithm can be run for two tracks in parallel with improved results. To estimate the parameters of this model, the iterative algorithm (of EM-type) is extended from the original by including correlation estimation based on the results of logistic regression and generalized linear model fitting steps. As an output, our algorithm provides feature detection results for both tracks separately, a correlation model describing the interaction between the two tracks, and an optional consensus track output showing feature calls based on data from both of the two tracks used.

Here, we consider three examples of integrating HTS data across multiple data types and evaluate the performance of our algorithm with Receiver Operating Characteristics (ROC) curves and Area Under Curve (AUC) values. The example cases have been selected as pairs of tracks from
ENCODE datasets that share promoter and enhancer activity information. First, we combine ChIP-seq histone modification marks H3k4me3 and H3k27ac (K562 cell line). Second, we apply our algorithm to ChIP-seq TFs P300 and FAIRE-seq open chromatin data (K562). Finally, we apply our algorithm to a pair of FAIRE-seq replicates (K562) and study the resulting consensus track. In all the cases the analysis of algorithm performance shows a significant improvement in comparison to single track analysis.

P61: Identifying sequence-dependent regulators of gene expression from a novel massively parallel reporter assay

Vincent Fitzpatrick¹, Joris van Arensbergen², Marcel de Haas², J Omar Yáñez-Cuna², Ludo Pagie², Bas van Steensel², Harmen Bussemaker¹

1 Columbia University, Department of Biological Sciences, United States; 2 Netherlands Cancer Institute, Netherlands

ABSTRACT: DNA-binding proteins regulate expression through sequence-specific interactions with gene promoters. These interactions are further mediated by local chromatin context extrinsic to the promoter sequence, making it difficult to separate sequence-dependent regulatory mechanisms from other contextual factors. To this end, our collaborators in the Van Steensel lab (NKI) have developed SuRE-seq, a high-throughput reporter assay that screens for genomic fragments capable of driving expression of a uniform plasmid reporter. SuRE-seq quantifies the relative expression rate of millions of genomic elements in parallel, providing insight into genome-wide mechanisms of transcription regulation. Using a regression-based approach, we have discovered sequence-specific, spatially-dependent mechanisms of gene regulation in Drosophila and human cell lines, including motifs attributable to known transcription factors and low-complexity sequence patterns with strand-dependent contributions to expression. These results allow us to separate sequence-intrinsic regulatory properties of gene promoters and enhancers that are independent of endogenous chromatin context.

P62: Characterization of phased, secondary, small interfering RNAs (phasiRNAs) using Machine Learning

Parth Patel, Sandra Mathioni, Atul Kakrana, Hagit Shatkay, Blake Meyers

University of Delaware, United States

ABSTRACT: Small RNAs (sRNAs) in plant range in size from 21 to 24 nucleotides, and play important roles in biological processes such as development, epigenetics modification, and plant defense. They can be partitioned into three major classes: microRNAs (miRNAs); heterochromatic small interfering RNAs (hc-siRNAs); and phased, secondary, small interfering RNAs (phasiRNAs) (Fei et al., 2013). Our study focuses on phasiRNAs, for which the knowledge about functionality is still limited. We (Zhai et al. 2015) and others have shown that maize anthers (male reproductive organs), express two classes of phasiRNAs
(21-nt and 24-nt) during different developmental time points (pre-meiotic and meiosis). Other data suggest these phasiRNAs are required for fertility.

Given the important role grasses such as maize and rice play as a prime food-source in many countries and as influential factors in the global economy, we aim to identify and understand the function of grass-specific phasiRNAs in maize and rice development. To this end, we use the framework of hidden Markov models (HMMs) in order to model both phasiRNA and non-phasiRNA sequences, and to distinguish between the two types of these small RNAs. We performed ANOVA with Dunnett’s method, demonstrating that the probability assigned by the resulting HMMs to phasiRNAs (21/24-nt) from rice and maize is significantly different from that assigned to other genomic sequences of similar length. Future work will include classification to distinguish phasiRNA sequences from non-phasiRNA sequences using other machine learning classifier(s), aiming to extract patterns (i.e., motifs, GC content) occurring in phasiRNAs to provide further insight into their biological function.

**P63: The transcription factor GABP selectively binds and activates the mutant TERT promoter in cancer**

Robert J.A. Bell¹, **H. Tomas Rube**², Alex Kreig³, Andrew Mancini¹, Shaun F. Fouse¹, Raman P. Nagarajan¹, Serah Choi¹, Chibo Hong¹, Daniel He¹, Melike Pekmezci¹, John K. Wiencke¹, Margaret R. Wrensch¹, Susan M. Chang¹, Kyle M. Walsh¹, Sua Myong³, Jun S. Song³, Joseph F. Costello¹

¹ University of California San Francisco, United States; 2 Columbia University, United States; 3 University of Illinois Urbana-Champaign, United States

**ABSTRACT:** Reactivation of telomerase reverse transcriptase (TERT) expression enables cells to overcome replicative senescence and escape apoptosis, which are fundamental steps in the initiation of human cancer. Multiple cancer types, including up to 83% of glioblastomas (GBMs), harbor highly recurrent TERT promoter mutations of unknown function but specific to two nucleotide positions. We identified the functional consequence of these mutations in GBMs to be recruitment of the multimeric GA-binding protein (GABP) transcription factor specifically to the mutant promoter. Allelic recruitment of GABP is consistently observed across four cancer types, highlighting a shared mechanism underlying TERT reactivation. Tandem flanking native E26 transformation-specific motifs critically cooperate with these mutations to activate TERT, probably by facilitating GABP heterotetramer binding. GABP thus directly links TERT promoter mutations to aberrant expression in multiple cancers.

**P64: Charting the human genome’s regulatory landscape with transcription factor binding site predictions**

Xi Chen¹ and Richard Bonneau¹²

¹ New York University, United States; 2 Simons Foundation, United States

**ABSTRACT:** Transcription factor (TF) binding is an essential step in the regulation of gene expression. Differential binding of multiple TFs at key
cis-regulatory loci allows the specification of progenitor cells into various cell types, tissues and organs. ChIP-Seq is a technique that can reveal genome-wide patterns of TF binding. However, it lacks the scalability to cover the range of factors, cell types and dynamic conditions a multicellular eukaryotic organism sees. So charting the regulatory landscape spanning multi-lineage differentiation requires computational methods to predict TF binding sites (TFBS) in an efficient and scalable manner.

We develop a method to predict binding sites for over 800 human TFs using a rich collection of DNA binding motifs. We integrate genomic features, including chromatin accessibility, motif scores, TF footprints, CpG/GC content, evolutionary conservation and the proximity of TF motifs to transcription start sites in sparse logistic regression classifiers. We label candidate motif sites with ChIP-Seq data and apply correlation-based filter and L1 regularization to select relevant features for each trained TF. The resulted logistic regression classifiers accurately predict TFBS and perform favorably in comparison to the current best TFBS prediction methods. Further, we map TFs based on feature distance to a nearest trained TF neighbor. Cross-TF predictions allow us to scale and expand the repertoire of putative TFBS to any TFs where motif data is available and to any cell types where accessibility data is obtainable. Our method has the potential to be applied in previously intractable domains, such as the identification of cell type-specific cis-regulatory modules, and reveal key properties underlying the regulatory complexity of multicellular eukaryotes.

P65: Deconvolving discriminative sequence features in overlapping categories of TF binding sites

Akshay Kakumanu¹, Silvia Velasco², Esteban Mazzoni², Shaun Mahony¹
1 Penn State, United States; 2 New York University, United States

ABSTRACT: Given multiple ChIP-seq experiments, we often seek to define clusters of binding sites that describe site properties across experiments. For example, we may categorize a given transcription factor’s binding sites as condition-independent or condition-specific across multiple condition ChIP-seq experiments. Similarly, we may categorize a transcription factor’s binding sites as being located in active enhancers or not based on overlaps with appropriate histone modification ChIP-seq experiments. Given such binding site categories, it is natural to ask what sequence features are associated with a category label. However, discovering such label-specific sequence features is often confounded by overlaps between binding site categories. For example, if condition-independent transcription factor binding sites are also more likely to be located within promoter regions, any sequence features specific to condition-independent binding behavior will be convolved with sequence features specific to promoters. Therefore, in order to identify sequence signals specifically associated with a given binding label it is necessary to deconvolve discriminative sequence signals from overlapping labels.
In order to meet this challenge, we developed SeqUnwinder, a principled approach to identifying interpretable discriminative sequence features for overlapping categories of transcription factor binding sites. SeqUnwinder uses local k-mer frequencies as predictors for a multiclass logistic classifier. Class label relationships between clusters are incorporated through an L-2 norm regularizer that encourages clusters sharing a label to have similar predictor weights. Our approach yields an integrated framework that identifies discriminative sequence signals for individual TF binding class labels and all combinations of labels, making it easy to gain more insights into TF binding preferences in a given in vivo system.

We demonstrate SeqUnwinder by using it to characterize TF binding during direct motor neuron programming. In our system, over-expression of Ngn2, Isl1, and Lhx3 (NIL) induces rapid and highly efficient conversion of mouse embryonic stem (ES) cells to spinal motor neurons. However, little is known about how the NIL factor combination achieves direct programming. We used ChIP-seq to profile NIL binding at three intermediate time points during the direct programming process. We then formed overlapping clusters of binding sites according to two criteria: dynamics over the course of programming, and the chromatin context in the initial ES cells. SeqUnwinder enables us to identify several meaningful sequence features associated with each cluster label, and thereby allows us to formulate hypotheses about the mechanisms through which over-expression of NIL can alter the fate of ES cells into induced motor neurons.

**P66: Implementation of a Deep Learning Framework to Predict De Novo Anticancer Drug Activity**

**Jose Zamalloa** and **Mona Singh**  
*Princeton University, United States*

**ABSTRACT:** Cancer treatment can greatly benefit from highly accurate drug prediction models. Current methods aim to identify key features in genomic data in order to predict known efficacies of a particular drug across cancer cell lines. The Cancer Cell Line Encyclopedia (CCLE) and Cancer Genome Project (CGP) provide the cancer drug panel data to build predictive models based on known drug compounds and genomic backgrounds of cancer cells. However, given that known compounds are not sufficient to efficiently treat cancer at the moment, there is a pressing need to develop methods that can accurately predict drug activity of compounds for which we have no prior information. The present method aims to solve this problem by combining chemical information across compounds and cancer genetic backgrounds to predict an unknown drug activity using a Deep Learning framework. We incorporate structural information of endogenous metabolites to describe chemical features of drug compounds and integrated them as features into our predictor along with selected genomic information.
We applied our approach to the CCLE dataset. We train our model on all the dataset but the compound of interest and test it on such compound in order to simulate the prediction of an unknown drug. Our preliminary results show that our accuracy is on par or better than current methods suggesting its potential use in predicting untested cancer drug candidates.

**P67: Computational Discovery of Transcription Factors Associated with Drug Response**

**Casey Hanson1, Junmei Cairns2, Liewei Wang2, Saurabh Sinha1**

1 University of Illinois at Urbana - Champaign, United States; 2 Mayo Clinic, United States

**ABSTRACT:** Genome wide association studies in pharmacogenomics generally involve associating drug-induced response with biomarkers. While GWAS suffers from sensitivity issues after correction, even signals that survive face problems of functional interpretation. Our study ameliorates this issue by posing the statistical test in the context of gene regulation. Rather than identifying SNPs or genes associated with drug response, we integrate biomarkers with genome-wide transcription factor (TF) binding data to elucidate whether a TF’s regulatory influence is associated with the drug. Our approach (GENMi) integrates gene expression, genotype, and drug-response data with ENCODE TF tracks to quantify the association between TFs and drugs via cis-regulatory eQTLs.

**Methods:** The GENMi method for testing a (TF, drug) combination consists of the following procedure. First, SNPs located outside of the TF’s ENCODE peak are discarded. Considering the 50kb upstream region of a gene as a putative cis-regulatory region, the gene is scored by the most significant eQTL under the TF’s peak. The top 400 eQTL genes are then tested for overlap with all genes correlated with the drug’s-induced cytotoxicity, using Gene Set Enrichment Analysis.

**Results:** We analyzed 114 TFs and 24 treatments using GENMi, yielding 334 significantly associated (TF, drug) pairs. The top 20 sparse (TF, drug) pairs yielded literature support for 13 associations, often from studies where perturbation of the TF’s expression changes drug response. We demonstrate the advantage of our approach by contrasting it with a baseline without using gene expression data. Our method reports more associations than the baseline approach at identical false positive rates (FPR). We further tested 14 TFs GENMi associated with either anthracycline (doxorubicin or epirubicin) and 21 TFs associated with either taxanes (paclitaxel or docetaxel). MTS cytotoxicity assays after TF knockdowns in two triple negative breast cancer cell lines, BT549 and MDA-MB231, yielded 6 TFs that significantly de-sensitized the cell to taxane induced apoptosis and 4 TFs that significantly de-sensitized the cell to anthracycline induced apoptosis.
P68: Pervasive variation of transcription factor orthologs contributes to regulatory network divergence

Shilpa Nadimpalli, Anton V. Persikov, Mona Singh

Princeton University, United States

ABSTRACT: Differences in transcriptional regulatory networks underlie much of the phenotypic variation observed across organisms. Changes to cis-regulatory elements are widely believed to be the predominant means by which regulatory networks evolve, yet examples of regulatory network divergence due to transcription factor (TF) variation have also been observed. To systematically ascertain the extent to which TFs contribute to regulatory divergence, we analyzed the evolution of the largest class of metazoan TFs, Cys2-His2 zinc finger (C2H2-ZF) TFs, across 12 Drosophila species spanning ~45 million years of evolution. Remarkably, we uncovered that a significant fraction of all C2H2-ZF 1-to-1 orthologs in flies exhibit variations that can affect their DNA-binding specificities. In addition to loss and recruitment of C2H2-ZF domains, we found diverging DNA-contacting residues in ~44% of domains shared between D. melanogaster and the other fly species. These diverging DNA-contacting residues, present in ~70% of the D. melanogaster C2H2-ZF genes in our analysis and corresponding to ~26% of all annotated D. melanogaster TFs, show evidence of functional constraint: they tend to be conserved across phylogenetic clades and evolve more slowly than other diverging residues. These same variations were rarely found as polymorphisms within a population of D. melanogaster flies, indicating their rapid fixation. The predicted specificities of these dynamic domains gradually change across phylogenetic distances, suggesting stepwise evolutionary trajectories for TF divergence. Further, whereas proteins with conserved C2H2-ZF domains are enriched in developmental functions, those with varying domains exhibit no functional enrichments. Our work suggests that a set of highly dynamic and largely unstudied TFs are a likely source of regulatory variation in Drosophila and other metazoans.

P69: A Novel Experimental Model Sheds Light on the Mechanism of Host-Gut Microbiome Interactions

Allison Richard1, Michael Burns2, Adnan Alazizi1, Roger Pique-Regi1, Ran Blekhman2, Francesca Luca1

1 Wayne State University, United States; 2 University of Minnesota, United States

ABSTRACT: The components of the human gut microbiome vary between physiological and pathological states. It has been shown that the gut microbiome differs in individuals with certain diseases, such as diabetes. However, the question of the cause and effect of the differences in both gut microbiome and host state remains unsolved. In order to delve into the relationship between gut microbiome and host, we treated human primary colonic epithelial cells (colonocytes) with different concentrations of gut microbiome from a healthy donor for varying time
intervals. Each experiment was performed in triplicate. We found that a gut microbiome to host ratio of 10:1 is best to simulate the symbiotic environment of the colon. Under these conditions, we performed RNA-seq to determine changes in the host gene expression that comprise a response to microbiome exposure. RNA-sequencing reads were aligned using BWA-mem and differentially expressed genes were identified using DESeq2. We found 2,111 genes and 1,110 genes that change expression in the host colonocytes following exposure to the microbiome for 4 and 6 hours, respectively (FDR = 1%). These genes are enriched for a variety of pathways involved in the interaction of host cells and gut microbiome. Specifically, we found enrichment in pathways involved in cell adhesion and cell surface receptor signaling. In addition, we performed 16S sequencing on bacterial DNA derived from the same co-cultures as the host colonocytes in order to study changes in the composition of the gut microbiome following exposure to the host. We found that after 4 hours of co-culturing, there was a decrease in the proportional abundance of the phylum Firmicutes and a corresponding increase in the phylum Proteobacteria. Furthermore, there was a decrease in overall diversity of the gut microbiome following exposure to the host colonocytes. Together, these results help us to identify which pathways are involved in the host response following microbiome exposure and in turn, how the microbiome is changed by host exposure. Study of both of these responses will help us to understand the cause of the differences in gut microbiome composition that have been seen in various pathological states.

P70: Experimentally identified gene-environment interactions contribute to heritability of complex traits

Cynthia Kalita¹, Gregory Moyerbrabalean¹, Omar Davis¹, Chris Harvey¹, Adnan Alizizi², Donovan Watza¹, Xiaoquan Wen², Xiang Zhou², Roger Pique-Regi¹, Francesca Luca¹

¹Wayne State University, United States; ²University of Michigan, United States

ABSTRACT: Genome wide association studies (GWAS) have identified thousands of common genetic variants associated with complex traits, including normal traits and common diseases. However, the significant SNPs found in these association studies explain only a small proportion of disease heritability. One possible explanation for this missing heritability is that the effect of the variant on the trait can be detected only under the right environmental conditions.

To test this hypothesis, we used GEMMA to jointly analyze summary statistics from 18 GWAS meta-analysis studies with annotations of regulatory variation. Our annotations are derived from: SNPs with allele specific expression (ASE) in 48 cellular environments, eQTLs (Wen et al 2014), and SNPs with conditional allele specific expression (cASE).
GEMMA (Genome-wide Efficient Mixed Model Association) tests for the proportion of variance in phenotypes explained (PVE) by typed genotypes, for example, “chip heritability”. At the same time, it estimates enrichment of a set of annotations within a GWAS trait. We find a range of enrichments for SNPs in genes with ASE, up to 7.84 for mean platelet volume. In comparison, for this same trait, SNPs in genic regions without ASE show an enrichment value of 1.03. When we consider SNPs in genes with cASE, we observe an enrichment of 5.10 as compared to 3.93 (SNPs in genes with ASE) and 1.04 (SNPs in genic regions). This approach, which integrates regulatory variation and gene-environment interactions into GWAS signals, can provide a much better understanding of the molecular mechanisms underlying inter-individual variation in complex traits.

**P71: A systematic survey of the Cys2His2 zinc finger DNA-binding landscape**

Joshua Wetzel¹, Anton Persikov¹, Mona Singh¹, Marcus Noyes²

1 Princeton University, United States; 2 NYU Institute for Systems Genetics, United States

**ABSTRACT:** Cys2His2 zinc fingers (C2H2-ZFs) comprise the largest class of metazoan DNA-binding domains. Despite this domain’s well-defined DNA-recognition interface, and its successful use in the design of chimeric proteins capable of targeting genomic regions of interest, much remains unknown about its DNA-binding landscape. To help bridge this gap in fundamental knowledge and to provide a resource for design-oriented applications, we screened large synthetic protein libraries to select binding C2H2-ZF domains for each possible three base pair target. The resulting data consist of >160 000 unique domain–DNA interactions and comprise the most comprehensive investigation of C2H2-ZF DNA-binding interactions to date. An integrated analysis of these independent screens yielded DNA-binding profiles for tens of thousands of domains and led to the successful design and prediction of C2H2-ZF DNA-binding specificities. Computational analyses uncovered important aspects of C2H2-ZF domain–DNA interactions, including the roles of within-finger context and domain position on base recognition. We observed the existence of numerous distinct binding strategies for each possible three base pair target and an apparent balance between affinity and specificity of binding. In sum, our comprehensive data help elucidate the complex binding landscape of C2H2-ZF domains and provide a foundation for efforts to determine, predict and engineer their DNA-binding specificities.
P72: Statistical Algorithms for Motif Discovery on SELEX Data

Chaitanya Rastogi, Harmen Bussemaker
Department of Biological Sciences, Columbia University, United States

ABSTRACT: SELEX-seq is an experimental and computational platform that combines biophysical modeling and deep sequencing in order to determine the DNA binding specificity of a transcription factor complexes [1]. Recent work has demonstrated the protocol’s ability to elucidate novel recognition properties of the eight Drosophila Hox proteins [2]. SELEX-seq analyses require detailed oligomer count information to infer affinities, a challenging computational task given the size of the data. Efficient implementations of the computational pipeline are required as the adoption of SELEX-seq increases. Following the methodology set out in [1,2], we have developed a suite of R/Bioconductor functions, named “SELEX,” to facilitate the analysis of SELEX-seq data. Thanks to efficient algorithms, this software can run on a standard laptop computer. Our package includes functionality for kmer counting, Markov model construction, and information gain (Kullback-Leibler divergence) calculations, along with integrated solutions for painless annotation and management of SELEX-seq experiments. Significantly, the package forms the basis for advanced feature-based modeling of TF binding sites. These novel statistical models directly infer G values for nucleotide, dinucleotide, and DNA shape features without any prior information about the binding factor in question.


Sponsors

Gold Sponsor

SANOFI

Sanofi, an integrated global healthcare leader, discovers, develops and distributes therapeutic solutions focused on patients’ needs. Sanofi has core strengths in the field of healthcare with seven growth platforms: diabetes solutions, human vaccines, innovative drugs, consumer healthcare, emerging markets, animal health and the new Genzyme. Sanofi is listed in Paris (EURONEXT: SAN) and in New York (NYSE: SNY).

Sanofi is the holding company of a consolidated group of subsidiaries and operates in the United States as Sanofi US, also referred to as Sanofi-aventis U.S. LLC. For more information on Sanofi US, please visit http://www.sanofi.us or call 1-800-981-2491.
Bronze Sponsor

ICAHN INSTITUTE FOR GENOMICS AND MULTISCALE BIOLOGY AT MOUNT SINAI

General Sponsors

ASTRAZENECA

IBM RESEARCH
http://www.research.ibm.com/

IBM Research

For more than sixty years, IBM Research, has been the innovation engine of IBM. We have been the location of some of the most notable technological and scientific business breakthroughs of the 20th and 21st centuries. From helping the Apollo missions land on the moon to the discovery of fractals; from the technology behind laser eye surgery to a question answering computer called Watson now being applied to health care, IBM Research continues to define the future of technology.
# Presenting Authors Index

<table>
<thead>
<tr>
<th>Name</th>
<th>Initials</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antti Airola</td>
<td></td>
<td>25</td>
</tr>
<tr>
<td>Tero Aittokallio</td>
<td></td>
<td>25</td>
</tr>
<tr>
<td>Gabriela Alexe</td>
<td></td>
<td>43</td>
</tr>
<tr>
<td>Nicholas Akers</td>
<td></td>
<td>110</td>
</tr>
<tr>
<td>Warren Anderson</td>
<td></td>
<td>111</td>
</tr>
<tr>
<td>Ramy Arnaout</td>
<td></td>
<td>41</td>
</tr>
<tr>
<td>Mario L. Arrieta</td>
<td></td>
<td>77</td>
</tr>
<tr>
<td>Peter Banda</td>
<td></td>
<td>126</td>
</tr>
<tr>
<td>Anastasia Baryshnikova</td>
<td></td>
<td>108</td>
</tr>
<tr>
<td>Britanny Baur</td>
<td></td>
<td>121</td>
</tr>
<tr>
<td>Michael Beer</td>
<td>69, 94</td>
<td></td>
</tr>
<tr>
<td>Vincenzo Belcastro</td>
<td></td>
<td>99</td>
</tr>
<tr>
<td>Bonnie Berger</td>
<td></td>
<td>16</td>
</tr>
<tr>
<td>Paul Boutros</td>
<td></td>
<td>34</td>
</tr>
<tr>
<td>James Bradner</td>
<td></td>
<td>17</td>
</tr>
<tr>
<td>Harmen J. Bussemaker</td>
<td></td>
<td>42, 92</td>
</tr>
<tr>
<td>Mark Carty</td>
<td></td>
<td>51</td>
</tr>
<tr>
<td>Elias Chaibub Neto</td>
<td></td>
<td>36</td>
</tr>
<tr>
<td>Huan-Jui Chang</td>
<td></td>
<td>30</td>
</tr>
<tr>
<td>Margaret P. Chapman</td>
<td></td>
<td>56</td>
</tr>
<tr>
<td>Deborah Chasman</td>
<td></td>
<td>62</td>
</tr>
<tr>
<td>Chien-Yu Chen</td>
<td>81, 116</td>
<td></td>
</tr>
<tr>
<td>Xi Chen</td>
<td></td>
<td>129</td>
</tr>
<tr>
<td>Jung-Hsien Chiang</td>
<td></td>
<td>30</td>
</tr>
<tr>
<td>Zechen Chong</td>
<td></td>
<td>34</td>
</tr>
<tr>
<td>Sang Chun</td>
<td></td>
<td>124</td>
</tr>
<tr>
<td>Daniel Cook</td>
<td></td>
<td>51</td>
</tr>
<tr>
<td>Francesca Cordero</td>
<td></td>
<td>109</td>
</tr>
<tr>
<td>James C. Costello</td>
<td></td>
<td>24</td>
</tr>
<tr>
<td>Danielle Decicco</td>
<td></td>
<td>114</td>
</tr>
<tr>
<td>Amit Dhurandhar</td>
<td></td>
<td>29</td>
</tr>
<tr>
<td>Alon Diament</td>
<td></td>
<td>47, 92</td>
</tr>
<tr>
<td>Laura L. Elo</td>
<td></td>
<td>26</td>
</tr>
<tr>
<td>Mehmet Eren Ahsen</td>
<td></td>
<td>44, 93</td>
</tr>
<tr>
<td>Shahrazad Eslamian</td>
<td></td>
<td>123</td>
</tr>
<tr>
<td>Lauren Fairchild</td>
<td></td>
<td>77</td>
</tr>
<tr>
<td>Wen-Chieh Fang</td>
<td></td>
<td>30, 90</td>
</tr>
<tr>
<td>Soheil Feizi</td>
<td></td>
<td>50</td>
</tr>
<tr>
<td>Elana J. Fertig</td>
<td></td>
<td>96</td>
</tr>
<tr>
<td>Vincent Fitzpatrick</td>
<td></td>
<td>128</td>
</tr>
<tr>
<td>Long Gao</td>
<td></td>
<td>124</td>
</tr>
<tr>
<td>Richard C Gerkin</td>
<td>29, 80</td>
<td></td>
</tr>
<tr>
<td>Yuanfang Guan</td>
<td>27, 31, 86</td>
<td></td>
</tr>
<tr>
<td>Antonio Gomes</td>
<td></td>
<td>101</td>
</tr>
<tr>
<td>Peddinti Gopalacharyulu</td>
<td></td>
<td>25</td>
</tr>
<tr>
<td>Raluca Gordan</td>
<td></td>
<td>43</td>
</tr>
<tr>
<td>Andrey Grigoriev</td>
<td></td>
<td>72</td>
</tr>
<tr>
<td>Casey Hanson</td>
<td></td>
<td>132</td>
</tr>
<tr>
<td>Bing He</td>
<td></td>
<td>125</td>
</tr>
<tr>
<td>Outi Hirvonen</td>
<td></td>
<td>26</td>
</tr>
<tr>
<td>Max Homilius</td>
<td></td>
<td>115</td>
</tr>
<tr>
<td>Yourae Hong</td>
<td></td>
<td>121</td>
</tr>
<tr>
<td>Sirkku Jyrkkö</td>
<td></td>
<td>26</td>
</tr>
<tr>
<td>Akshay Kakumanu</td>
<td></td>
<td>130</td>
</tr>
<tr>
<td>Cynthia Kalita</td>
<td></td>
<td>134</td>
</tr>
<tr>
<td>Gaurav Kandoo</td>
<td></td>
<td>122</td>
</tr>
<tr>
<td>Andrea Levchencko</td>
<td></td>
<td>21</td>
</tr>
<tr>
<td>Jennifer Long Diaz</td>
<td>54, 93</td>
<td></td>
</tr>
<tr>
<td>Qiu Ma</td>
<td></td>
<td>95</td>
</tr>
<tr>
<td>Michael Menden</td>
<td>35, 40, 80</td>
<td></td>
</tr>
<tr>
<td>Pablo Meyer</td>
<td></td>
<td>98</td>
</tr>
<tr>
<td>Emily Miraldi</td>
<td></td>
<td>75</td>
</tr>
<tr>
<td>Tuomas Mirtti</td>
<td></td>
<td>25</td>
</tr>
<tr>
<td>Shahin Mohammadi</td>
<td></td>
<td>113</td>
</tr>
<tr>
<td>Gregory Moyerbrailean</td>
<td></td>
<td>74, 94</td>
</tr>
<tr>
<td>Christian L. Müller</td>
<td></td>
<td>63</td>
</tr>
<tr>
<td>Silpa Nadimpalli</td>
<td></td>
<td>133</td>
</tr>
<tr>
<td>Prathik Naidu</td>
<td></td>
<td>92</td>
</tr>
<tr>
<td>Raquel Norel</td>
<td></td>
<td>80</td>
</tr>
<tr>
<td>Nooshin Omranian</td>
<td></td>
<td>73</td>
</tr>
<tr>
<td>Yaron Orenstein</td>
<td></td>
<td>112</td>
</tr>
<tr>
<td>Tapio Pahikkala</td>
<td></td>
<td>25</td>
</tr>
<tr>
<td>Bajaj Nairidas</td>
<td></td>
<td>60</td>
</tr>
<tr>
<td>Gaurav Pandey</td>
<td></td>
<td>89</td>
</tr>
<tr>
<td>James Park</td>
<td></td>
<td>113</td>
</tr>
<tr>
<td>Aparajith Sridhar</td>
<td></td>
<td>118</td>
</tr>
<tr>
<td>Parth Patel</td>
<td></td>
<td>128</td>
</tr>
<tr>
<td>Pei-Chen Peng</td>
<td></td>
<td>45, 92</td>
</tr>
<tr>
<td>Dimi Perera</td>
<td></td>
<td>67</td>
</tr>
<tr>
<td>Roger Pique-Regi</td>
<td></td>
<td>71, 94</td>
</tr>
<tr>
<td>Steve Quake</td>
<td></td>
<td>14</td>
</tr>
<tr>
<td>Chaitanya Rastogi</td>
<td></td>
<td>136</td>
</tr>
<tr>
<td>Aharon Ravia</td>
<td></td>
<td>88</td>
</tr>
<tr>
<td>Allison Richard</td>
<td></td>
<td>133</td>
</tr>
<tr>
<td>Marcelo Rivas-Atronza</td>
<td></td>
<td>57</td>
</tr>
<tr>
<td>Abiel Roche-Lima</td>
<td></td>
<td>106</td>
</tr>
<tr>
<td>H. Tomas Rube</td>
<td></td>
<td>129</td>
</tr>
<tr>
<td>Alexander Y. Rudensky</td>
<td></td>
<td>109</td>
</tr>
<tr>
<td>Anastasia Scherban</td>
<td></td>
<td>127</td>
</tr>
<tr>
<td>Michael Shen</td>
<td></td>
<td>20</td>
</tr>
<tr>
<td>Wengjiang Shi</td>
<td></td>
<td>70, 94</td>
</tr>
<tr>
<td>Cedric Similion</td>
<td></td>
<td>100</td>
</tr>
<tr>
<td>Irtisha Singh</td>
<td></td>
<td>105</td>
</tr>
<tr>
<td>Subarna Sinha</td>
<td></td>
<td>38, 92</td>
</tr>
<tr>
<td>William Stafford Noble</td>
<td></td>
<td>18</td>
</tr>
<tr>
<td>Joshua M. Stuart</td>
<td></td>
<td>33</td>
</tr>
<tr>
<td>Rebecca Tagett</td>
<td></td>
<td>104</td>
</tr>
<tr>
<td>Konstantine Tchourine</td>
<td></td>
<td>119</td>
</tr>
<tr>
<td>Olga Troyanskaya</td>
<td></td>
<td>22</td>
</tr>
<tr>
<td>Li-Min Tu</td>
<td></td>
<td>88</td>
</tr>
<tr>
<td>Golnaz Vahedi</td>
<td></td>
<td>69</td>
</tr>
<tr>
<td>Sudhir Varma</td>
<td></td>
<td>78, 94</td>
</tr>
<tr>
<td>Aapal Verma</td>
<td></td>
<td>64</td>
</tr>
<tr>
<td>Coby Viner</td>
<td></td>
<td>55</td>
</tr>
<tr>
<td>Leslie Vossshall</td>
<td></td>
<td>14</td>
</tr>
<tr>
<td>Daifeng Wang</td>
<td>76, 103</td>
<td></td>
</tr>
<tr>
<td>Yahli Wang</td>
<td></td>
<td>52</td>
</tr>
<tr>
<td>Yao Wang</td>
<td>49, 93</td>
<td></td>
</tr>
<tr>
<td>Chung Wen Yu</td>
<td></td>
<td>86</td>
</tr>
<tr>
<td>Joshua Wetzel</td>
<td></td>
<td>135</td>
</tr>
<tr>
<td>Jinfeng Xiao</td>
<td></td>
<td>32, 91</td>
</tr>
<tr>
<td>Bingqing Xie</td>
<td></td>
<td>107</td>
</tr>
<tr>
<td>Lei Xie</td>
<td></td>
<td>58</td>
</tr>
<tr>
<td>Chen Yang</td>
<td></td>
<td>30</td>
</tr>
<tr>
<td>Hsin-Te Yang</td>
<td></td>
<td>30</td>
</tr>
<tr>
<td>Victoria Yao</td>
<td></td>
<td>119</td>
</tr>
<tr>
<td>Neta Zach</td>
<td></td>
<td>30</td>
</tr>
<tr>
<td>Zohar Zafir</td>
<td></td>
<td>62, 93</td>
</tr>
<tr>
<td>Jose Zamalloa</td>
<td></td>
<td>131</td>
</tr>
<tr>
<td>Julia Zeitlinger</td>
<td></td>
<td>20</td>
</tr>
<tr>
<td>Liz Zhou</td>
<td></td>
<td>24</td>
</tr>
<tr>
<td>Qian Zhu</td>
<td></td>
<td>66</td>
</tr>
</tbody>
</table>
2016 RECOMB/ISCB CONFERENCE on REGULATORY and SYSTEMS GENOMICS with DREAM CHALLENGES

November 6 - 11, 2016
Pointe Hilton Squaw Peak Resort
Phoenix, Arizona, USA

www.iscb.org/recomb-regsysgen2016
THANK YOU TO OUR SPONSORS

GOLD SPONSOR
SANOFI

BRONZE SPONSOR
Mount Sinai
Icahn Institute for Genomics and Multiscale Biology

GENERAL SPONSORS
IBM Research
AstraZeneca
Sage