Pacific Symposium on Biocomputing 2014

Abstract Book

Poster Presenters: Poster space is assigned by abstract page number. Please find the page that your abstract is on and put your poster on the poster board with the corresponding number (e.g., if your abstract is on page 50, put your poster on board #50).

Proceedings papers with oral presentations on pages 10-41 are not assigned poster space.

Poster abstracts are organized by session with accepted proceedings papers with poster presentations listed first.

CANCER PANOMICS: COMPUTATIONAL METHODS AND INFRASTRUCTURE FOR INTEGRATIVE ANALYSIS OF CANCER HIGH-THROUGHPUT "OMICS" DATA ACCEPTED PROCEEDINGS PAPERS WITH ORAL PRESENTATIONS 9

Tumor haplotype assembly algorithms for cancer genomics 10
Derek Aguiar, Wendy S.W. Wong, Sorin Istrail

The Stream Algorithm: Computationally Efficient Ridge-Regression via Bayesian Model Averaging, and Applications to Pharmacogenomic Prediction of Cancer Cell Line Sensitivity 11
Elias Chaibub Neto, In Sock Jang, Stephen H. Friend, Adam A. Margolin

Sharing information to reconstruct patient-specific pathways in heterogeneous diseases 12
Anthony Gitter, Alfredo Braunstein, Andrea Pagnani, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Riccardo Zecchina, Ernest Fraenkel

Detecting Statistical Interaction Between Somatic Mutational Events and Germline Variation from Next-Generation Sequence Data 13
Hao Hu, Chad D. Huff

Systematic Assessment Of Analytical Methods for Drug Sensitivity Prediction from Cancer Cell Line Data 14
In Sock Jang, Elias Chaibub Neto, Justin Guinney, Stephen H. Friend, Adam A. Margolin

Integrative Analysis of Two Cell Lines Derived From a Non-Small-Lung Cancer Patient – a Panomics Approach 15
Oleg Mayba, Florian Gnad, Michael Peyton, Fan Zhang, Kimberly Walter, Pan Du, Melanie A. Huntley, Zhaoshi Jiang, Jinfeng Liu, Peter M. Haverty, Robert C. Gentleman, Ruiqiang Li, John D. Minna, Yingrui Li, David S. Shames, Zemin Zhang

An integrated approach to blood-based cancer diagnosis and biomarker discovery 16
Martin Renqiang Min, Salim Chowdhury, Yanjun Qi, Alex Stewart, Rachel Ostroff

Multiplex Meta-Analysis of Medulloblastoma Expression Studies with External Controls 17
Alexander A. Morgan, Matthew D. Li, Achal S. Achrol, Purvesh J. Khatri, Samuel H. Cheshier

COMPUTATIONAL APPROACHES TO DRUG REPURPOSING AND PHARMACOLOGY ACCEPTED PROCEEDINGS PAPERS WITH ORAL PRESENTATIONS 18

Anti-Infectious Drug Repurposing Using an Integrated Chemical Genomics and Structural Systems Biology Approach 19
Clara Ng, Ruth Hauptman, Yinliang Zhang, Philip E. Bourne, Lei Xie

Drug-Target Interaction Prediction by Integrating Chemical, Genomic, Functional and Pharmacological Data 20
Fan Yang, Jinbo Xu, Jianyang Zeng

Prediction of Off-Target Drug Effects through Data Fusion 21
Emmanuel R. Yera, Ann E. Cleves, Ajay N. Jain

Exploring the Pharmacogenomics Knowledge Base (PharmGKB) for Repositioning Breast Cancer Drugs by Leveraging Web Ontology Language (OWL) and Cheminformatics Approaches 22
Qian Zhu, Cui Tao, Feichen Shen, Christopher Chute

DETECTING AND CHARACTERIZING PLEIOTROPY: NEW METHODS FOR UNCOVERING THE CONNECTION BETWEEN THE COMPLEXITY OF GENOMIC ARCHITECTURE AND MULTIPLE PHENOTYPES ACCEPTED PROCEEDINGS PAPERS WITH ORAL PRESENTATIONS 23

Using the Bipartite Human Phenotype Network to Reveal Pleiotropy and Epistasis beyond the Gene **24**
Christian Darabos, Samantha H. Harmon, Jason H. Moore
Environment-Wide Association Study (EWAS) for Type 2 Diabetes in the Marshfield Personalized Medicine Research Project Biobank **25**
Molly A. Hall, Scott M. Dudek, Robert Goodloe, Dana C. Crawford, Sarah A. Pendergrass, Peggy Peissig, Murray Brilliant, Catherine A. McCarty, Marylyn D. Ritchie
Dissection of Complex Gene Expression Using the Combined Analysis of Pleiotropy and Epistasis **26**
Vivek M. Philip, Anna L. Tyler, Gregory W. Carter

PERSONALIZED MEDICINE: FROM GENOTYPES AND MOLECULAR PHENOTYPES TOWARDS THERAPY
ACCEPTED PROCEEDINGS PAPERS WITH ORAL PRESENTATIONS **27**

PATH-SCAN: A Reporting Tool For Identifying Clinically Actionable Variants **28**
Roxana Daneshjou, Zachary Zappala, Kim Kukurba, Sean M Boyle, Kelly E Ormond, Teri E Klein, Michael Snyder, Carlos D Bustamante, Russ B Altman, Stephen B Montgomery
Imputation-based Assessment of Next Generation Rare Exome Variant Arrays **29**
Alicia R. Martin, Gerard Tse, Carlos D. Bustamante, Eimear E. Kenny
Utilization of an EMR-Biorepository to Identify the Genetic Predictors of Calcineurin-Inhibitor Toxicity in Heart Transplant Recipients **30**
Matthew Oetjens, William S. Bush, Kelly A. Birdwell, Holli H. Dilks, Erica A. Bowton, Joshua C. Denny, Russell A. Wilke, Dan M. Roden, Dana C. Crawford
Robust Reverse Engineering of Dynamic Gene Networks under Sample Size Heterogeneity **31**
Ankur P. Parikh, Wei Wu, Eric P. Xing
Variant Priorization and Analysis Incorporating Problematic Regions of the Genome **32**
Anil Patwardhan, Michael Clark, Alex Morgan, Stephen Chervitz, Mark Pratt, Gabor Bartha, Gemma Chandratillake, Sarah Garcia, Nan Leng, Richard Chen
Joint Association Discovery and Diagnosis of Alzheimer's Disease by Supervised Heterogeneous Multiview Learning **33**
Shandian Zhe, Zenglin Xu, Yuan Qi, Peng Yu

TEXT AND DATA MINING FOR BIOMEDICAL DISCOVERY
ACCEPTED PROCEEDINGS PAPERS WITH ORAL PRESENTATIONS **34**

Vector Quantization Kernels for the Classification of Protein Sequences and Structures **35**
Wyatt T. Clark, Predrag Radivojac
Combining Heterogenous Data for Prediction of Disease Related and Pharmacogenes **36**
Christopher S. Funk, Lawrence E. Hunter, K. Bretonnel Cohen
A Novel Profile Biomarker Diagnosis for Mass Spectral Proteomics **37**
Henry Han
Towards Pathway Curation Through Literature Mining – A Case Study Using PharmGKB **38**
Ravikumar K.E., Kavishwar B. Wagholikar, Hongfang Liu
Sparse Generalized Functional Liner Model for Predicting Remission Status of Depression Patients **39**
Yashu Liu, Zhi Nie, Jiayu Zhou, Michael Farnum, Vaibhav A Narayan, Gayle Wiittenberg, Jieping Ye

Development of a Data-Mining Algorithm to Identify Ages at Reproductive Milestones in Electronic Medical Records **40**
Jennifer Malinowski, Eric Farber-Eger, Dana C. Crawford
An Efficient Algorithm to Integrate Network and Attribute Data for Gene Function Prediction **41**
Shankar Vembu, Quaid Morris

POSTER PRESENTATIONS **42**

CANCER PANOMICS: COMPUTATIONAL METHODS AND INFRASTRUCTURE FOR INTEGRATIVE ANALYSIS OF CANCER HIGH-THROUGHPUT "OMICS" DATA POSTER PRESENTATIONS **43**

Extracting Significant Sample-Specific Cancer Mutations Using Their Protein Interactions **44**
Liviu Badea
Combining GATK, SAMtools, Varscan2 and pibase to reduce false negatives in somatic SNV-calling **45**
Michael Forster, Andre Franke
Causes and Consequences of Cancer Transcriptome Variability **46**
Kjong-Van Lehmann, André Kahles, Oliver Stegle, William Lee, David Kuo, Rileen Sinha, Cancer Genome Atlas Research Network, Nikolaus Schultz, Robert Klein, Gunnar Rätsch
Bayesian Network (BN) Structure Learning Inference Application in Studying BRD4 and JMJD6 Function Cooperatively on Promoter-proximal Polymerase-II (Pol II) Pausing Release **47**
Qi Ma, Wen Liu, Michael. G. Rosenfeld
Integrated Proteomic, Phosphoproteomic, and Genomic Pathway Analysis for Patient-Derived Tumor Samples **48**
Jason E. McDermott, Vladislav A. Petyuk, Feng Yang, Marina A. Gritsenko, Matthew E. Monroe, Joshua T. Aldrich, Ronald J. Moore, Therese R. Clauss, Anil K. Shukla, Athena A. Schepmoes, Rosalie K. Chu, Samuel H. Payne, Tao Liu, Karin D. Rodland, Richard D. Smith

COMPUTATIONAL APPROACHES TO DRUG REPURPOSING AND PHARMACOLOGY POSTER PRESENTATIONS **49**

Challenges in Secondary Analysis of High Throughput Screening Data **50**
Aurora S. Blucher, Shannon K. McWeeney
Drug Intervention Response Predictions with Paradigm (DIRPP) Identifies Drug Resistant Cancer Cell Lines and Pathway Mechanisms of Resistance **51**
Douglas Brubaker, Analisa Difeo, Yanwen Chen, Taylor Pearl, Kaide Zhai, Gurkan Bebek, Mark Chance, Jill Barnholtz-Sloan
Structure-based Systems Biology: Identification of Off-Targets and Biological Pathways for Human Chymase Inhibitors **52**
Mahreen Arooj, Keun Woo Lee
Automated Discovery of Features that Determine Specificity and Affinity of Protein-Ligand Interactions Across Protein Families Using Machine Learning Analysis of Structures and Chemical Interaction Profiles **53**
Joseph Schoeniger, Peter Anderson
Predicting Drug Selectivity from Host-Pathogen Drug Networks **54**
Geoffrey Henry Siwo, Roger Smith, Asako Tan, Michael T. Ferdig

DETECTING AND CHARACTERIZING PLEIOTROPY: NEW METHODS FOR UNCOVERING THE CONNECTION BETWEEN THE COMPLEXITY OF GENOMIC ARCHITECTURE AND MULTIPLE PHENOTYPES
POSTER PRESENTATIONS 55

A SYSTEMS BIOLOGY APPROACH FOR UNDERSTANDING FLORAL TRANSITION IN PLANTS 56
Gitanjali Yadav

GENERAL POSTER PRESENTATIONS 57

Polygenic risk score analysis in Childhood Onset Schizophrenia study 58
Kwangmi Ahn, Steven S An, Judith L. Rapoport

Long-range chromatin contacts reveal a role for the pluripotency and Polycomb networks in genome organization 59
Giancarlo Bonora*, Matthew Denholtz*, Constantinos Chronis, Erik Splinter, Wouter de Laat, Jason Ernst, Matteo Pellegrini, and Kathrin Plath

A Survey of Coding Variation within 70 Pharmacogenes: The SPHINX Database 60
William S. Bush, Jonathan Boston, Jay Cowan, Jacqueline Kirby, John Wallace, Josh Denny, Marylyn D. Ritchie, Jonathan L. Haines

The Stream Algorithm: Computationally Efficient Ridge-Regression via Bayesian Model Averaging, and Applications to Pharmacogenomic Prediction of Cancer Cell Line Sensitivity 61
Elias Chaibub Neto, In Sock Jang, Stephen H. Friend, Adam A. Margolin

Detection of Unknown Bacterial Genomes with Clique Log-Linear Models 62
Adrian Dobra, Camillo Valdes, Bertrand Clarke, Jennifer Clarke

GeneSeer Aids Drug Discovery by Exploring Evolutionary Relationships Between Genes Across Genomes 63
Douglas D. Fenger, Matthew Shaw, Philip Cheung, Tim Tully

Methods for Investigating the Pleiotropic Effects of Mitochondrial Genetic Variation on Human Health and Disease 64
Sabrina Mitchell, Jacob Hall, Robert Goodloe, Jonathan Boston, Eric Farber-Eger, Sarah Pendergrass, William Bush, Dana Crawford

Molecular Predictors of Residual Disease After Cytoreductive Surgery in Patients with High-Grade Serous Ovarian Cancer 65
Shelley M. Herbrich, Susan L. Tucker, Kshipra Gharpure, Anna Unruh, Alpa M. Nick, Erin K. Crane, Robert L. Coleman, Charles W. Drescher, Sherry Wu, Gabriel Lopez-Berestein, Bulent Ozpolat, Christina Ivan, Keith A. Baggerly, Anil K. Sood

Pharmacogenomic Discovery with PharmGKB PGxplore 66
Darla Hewett, Michelle Whirl-Carrillo, Julia Barbarino, Ryan Whaley, Mark Woon, Russ B. Altman and Teri E. Klein

Modeling cell signaling network of neuronal differentiation 67
Tsuyoshi Iwasaki, Ryo Takiguchi, Takumi Hiraiwa, Tadamasa Kimura, Akira Funahashi, Noriko Hiroi, Kazuto Yamazaki

Protein Interactions as Drug Targets: A Combined Computational and Experimental Approach 68
Jouhyun Jeon, Joan Teyra, Satra Nim and Philip M. Kim

A Kernel Based L1-Norm Regularized Logistic Regression Method to Predict Drug-Target Interactions 69
Shinhyuk Kim, Daeyong Jin, Hyunju Lee

Integrated Immunotherapy: Implications for Personalized Cancer Treatment 70

Deborah H Lundgren, Veneta Qendro, Karim Rezaul, Sun-Il Hwang, Zanna Aristarova, Ardian Latifi, and David K Han
Predicting Combination Therapies Using DivRank on the Connectivity Map Data 71
J. Matthew Mahoney, Anna L. Tyler
An Integrated Approach To Blood-Based Cancer Diagnosis And Biomarker Discovery..... 72
Renqiang Min, Salim Chowdhury, Yanjun Qi, Alex Stewart, Rachel Ostroff
Methods for Investigating the Pleiotropic Effects of Mitochondrial Genetic Variation on Human Health and Disease 73
Sabrina Mitchell, Jacob Hall, Robert Goodloe, Jonathan Boston, Eric Farber-Eger, Sarah Pendergrass, William Bush, Dana Crawford
Kaleidoscopic Evolution of C2H2 Zinc Finger DNA Binding..... 74
Hamed S. Najafabadi, Sanie Mnaimneh, Frank W. Schmitges, Kathy N. Lam, Ally Yang, Mihai Albu, Matthew T. Weirauch, Ernest Radovani, Jack Greenblatt, Brendan J. Frey, and Timothy R. Hughes
Locating Hydrogen Atoms, the "Dark Matter" in Proteins..... 75
Ho Leung Ng, Matthew Bronstad, Jinny Ching
Machine Reading for Cancer Panomics..... 76
Hoifung Poon, Tony Gitter, Chris Quirk
Analysis of RNA-Binding Protein Dynamics with Elastic Network Models..... 77
Ann Quigley, Michael Terribilini
Comparison of RNA and DNA Single Nucleotide and Indel Variants in the NCI-60 cell line collection..... 78
Onur Sakarya, Jeremy Ku, Kunbin Qu, Thon de Boer, Bill Gibb, Kevin Kwei, Jennie Jeong, Mei-Lan Liu, Robert Pelham, Sam Levy, Ellen Beasley
Bayesian Network Reconstruction Using Systems Genetics Data: Comparison of MCMC Methods 79
Shinya Tasaki, Ben Sauerwine, Bruce Hoff, Hiroyoshi Toyoshiba, Chris Gaiteri, Elias Chaibub Neto
Cardiac Enhancers Harbor Undiscovered Genetic Variants Associated with Heart Contraction Traits..... 80
Xinchen Wang, Manolis Kellis, Laurie Boyer
Ancestry Informative Markers for Native Hawaiians 81
Hansong Wang, Laurence N. Kolonel, Loic Le Marchand
A Frequent Inactivating Mutation in RHOA GTPase in Angioimmunoblastic T-cell Lymphoma..... 83
Hae Yong Yoo, Min Kyung Sung, Seung Ho Lee, Sangok Kim, Haeseung Lee, Seongjin Park, Sang Cheol Kim, Byungwook Lee, Kyoohyoung Rho, Jong-Eun Lee, Kwang-Hwi Cho, Wankyu Kim, Hyunjung Ju, Jaesang Kim, Seok Jin Kim, Won Seog Kim, Sanghyuk Lee, and Young Hyeh Ko
Tribe: The Collaborative Platform for Mining Genomic Data in Biology..... 84
Rene A. Zelaya, Casey S. Greene

PERSONALIZED MEDICINE: FROM GENOTYPES AND MOLECULAR PHENOTYPES TOWARDS THERAPY POSTER PRESENTATIONS 85

Bags of Words Models of Epitope Sets: HIV Viral Load Regression with Counting Grids.... 86
Alessandro Perina, Pietro Lovato, Nebojsa Jojic
Findings from the third Critical Assessment of Genome Interpretation, CAGI 2013, a Community Experiment to Evaluate Phenotype Prediction..... 87
Steven E. Brenner, John Moult, CAGI Participants
The Effects of Electronic Medical Record Phenotyping Details on Genetic Association Studies: HDL-C as a Case Study 88

Logan Dumitrescu, Robert Goodloe, Eric Farber-Eger, Jonathan Boston, Dana C. Crawford
CLINVITAE: a freely available database of clinically observed variants 89
Reece K. Hart, Bruce Blyth, Tim Chu, John Garcia
Transcriptomic Analysis of Benign and Malignant Thyroid Nodules 90
Katayoon Kasaian, Karen L. Mungall, Jacquie Schein, Yongjun Zhao, Richard A. Moore, Martin Hirst, Marco A. Marra, Blair A. Walker, Sam M. Wiseman, Steven J.M. Jones
An Empirical Bayesian Framework for Somatic Mutation Detection from Cancer Genome Sequencing Data 91
Yuichi Shiraishi, Yusuke Sato, Kenichi Chiba, Yusuke Okuno, Yasunobu Nagata, Kenichi Yoshida, Norio Shiba, Yasuhide Hayashi, Haruki Kume, Yukio Homma, Masashi Sanada, Seishi Ogawa, Satoru Miyano
Genomic and Computational Approaches to Tuberculosis and Nontuberculous Mycobacterial Disease 92
Michael Strong, Rebecca Davidson, Gargi Datta, Ben Garcia, Nabeeh Hasan, Mary Jackson

TEXT AND DATA MINING FOR BIOMEDICAL DISCOVERY POSTER PRESENTATIONS 93

Matrix Factorization-Based Data Fusion for Gene Function Prediction in Baker's Yeast and Slime Mold 94
Marinka Zitnik, Blaz Zupan
Automating the Construction of Metabolic Pathways Using BRENDA, MetaCyc model Pathways and Literature Mining. 95
Jan Czarnecki, Irene Nobeli, Adrian M Smith and Adrian J Shepherd
Extracting Country-of-Origin from Electronic Medical Records for Gene-Environment Studies as Part of the Epidemiologic Architecture for Genes Linked to Environment (EAGLE) 96
Eric Farber-Eger, Robert Goodloe, Jonathan Boston, William S. Bush, Dana C. Crawford
Towards Knowledge Inference from Biomedical texts: Aligning Text-Mining Extractions to Pathway model semantics 97
Ravikumar Komandur Elayavilli, Kavishiwar Wagholikar, Hongfang Liu
Sparse Generalized Functional Liner Model for Predicting Remission Status of Depression Patients 98
Yashu Liu, Zhi Nie, Jiayu Zhou, Michael Farnum, Vaibhav A Narayan, Gayle Wittenberg, Jieping Ye
Unsupervised Discovery of Gene-Drug Interaction Patterns in Biomedical Text 99
Bethany Percha, Russ B. Altman

WORKSHOP: COMPUTATIONAL IDENTIFICATION AND FUNCTIONAL ANALYSIS OF NON-CODING RNAS INVITED SPEAKERS ABSTRACTS 100

An Integrative Approach for Identifying Functionally Important and/or Clinically Relevant Long Noncoding RNAs in Human Cancer 101
Yiwen Chen, Zhou Du, Teng Fei, Roel G W Verhaak, Zhen Su, Yong Zhang, Myles Brown, X Shirley Liu
Nuclear RNAi: Driving Selective Recognition of Noncoding RNAs to Control Transcription and Splicing 102
David R. Corey
Identification, Annotation, Classification and the Evolutionary History of Large Intergenic Non-Coding RNAs (lincRNAs) in Mammals 103
Manuel Garber
Lincing the Non-Coding World to the Mammalian Circadian Clock 104

John Hogenesch
Predicting RNA Secondary Structure Using Probabilistic Methods 105
David H. Mathews
Noncoding RNA Database and Functional Research 106
Yi Zhao, Tengfei Xiao, Jiao Yuan, Runsheng Chen
INDEX 107

CANCER PANOMICS: COMPUTATIONAL METHODS AND INFRASTRUCTURE FOR INTEGRATIVE ANALYSIS OF CANCER HIGH-THROUGHPUT "OMICS" DATA

ACCEPTED PROCEEDINGS PAPERS WITH ORAL PRESENTATIONS

TUMOR HAPLOTYPE ASSEMBLY ALGORITHMS FOR CANCER GENOMICS

Derek Aguiar, Wendy S.W. Wong, Sorin Istrail

Department of Computer Science and Center for Computational Molecular Biology
Brown University
Providence, RI 02912, USA

The growing availability of inexpensive high-throughput sequence data is enabling researchers to sequence tumor populations within a single individual at high coverage. But, cancer genome sequence evolution and mutational phenomena like driver mutations and gene fusions are difficult to investigate without first reconstructing tumor haplotype sequences. Haplotype assembly of single individual tumor populations is an exceedingly difficult task complicated by tumor haplotype heterogeneity, tumor or normal cell sequence contamination, polyploidy, and complex patterns of variation. While computational and experimental haplotype phasing of diploid genomes has seen much progress in recent years, haplotype assembly in cancer genomes remains uncharted territory.

In this work, we describe HapCompass-Tumor a computational modeling and algorithmic framework for haplotype assembly of copy number variable cancer genomes containing haplotypes at different frequencies and complex variation. We extend our polyploid haplotype assembly model and present novel algorithms for (1) complex variations, including copy number changes, as varying numbers of disjoint paths in an associated graph, (2) variable haplotype frequencies and contamination, and (3) computation of tumor haplotypes using simple cycles of the compass graph which constrain the space of haplotype assembly solutions. The model and algorithm are implemented in the software package HapCompass-Tumor which is available for download from http://www.brown.edu/Research/Istrail_Lab/.

THE STREAM ALGORITHM: COMPUTATIONALLY EFFICIENT RIDGE-REGRESSION VIA BAYESIAN MODEL AVERAGING, AND APPLICATIONS TO PHARMACOGENOMIC PREDICTION OF CANCER CELL LINE SENSITIVITY

Elias Chaibub Neto, In Sock Jang, Stephen H. Friend, Adam A. Margolin

Sage Bionetworks, 1100 Fairview Avenue North, Seattle, Washington 98109, USA

Computational efficiency is important for learning algorithms operating in the "large p, small n" setting. In computational biology, the analysis of data sets containing tens of thousands of features ("large p"), but only a few hundred samples ("small n"), is nowadays routine, and regularized regression approaches such as ridge-regression, lasso, and elastic-net are popular choices. In this paper we propose a novel and highly efficient Bayesian inference method for fitting ridge-regression. Our method is fully analytical, and bypasses the need for expensive tuning parameter optimization, via cross-validation, by employing Bayesian model averaging over the grid of tuning parameters. Additional computational efficiency is achieved by adopting the singular value decomposition re-parametrization of the ridge-regression model, replacing computationally expensive inversions of large p-by- p matrices by efficient inversions of small and diagonal n-by-n matrices. We show in simulation studies and in the analysis of two large cancer cell line data panels that our algorithm achieves slightly better predictive performance than cross-validated ridge-regression while requiring only a fraction of the computation time. Furthermore, in comparisons based on the cell line data sets, our algorithm systematically out-performs the lasso in both predictive performance and computation time, and shows equivalent predictive performance, but considerably smaller computation time, than the elastic-net.

SHARING INFORMATION TO RECONSTRUCT PATIENT-SPECIFIC PATHWAYS IN HETEROGENEOUS DISEASES

Anthony Gitter, Alfredo Braunstein, Andrea Pagnani, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Riccardo Zecchina, Ernest Fraenkel

Microsoft Research, Cambridge, MA, USA
Department of Biological Engineering, Massachusetts Institute of Technology Cambridge, MA, USA
DISAT and Center for Computational Sciences, Politecnico di Torino, Turin, Italy
Human Genetics Foundation, Turin, Italy

Advances in experimental techniques resulted in abundant genomic, transcriptomic, epigenomic, and proteomic data that have the potential to reveal critical drivers of human diseases. Complementary algorithmic developments enable researchers to map these data onto protein-protein interaction networks and infer which signaling pathways are perturbed by a disease. Despite this progress, integrating data across different biological samples or patients remains a substantial challenge because samples from the same disease can be extremely heterogeneous. Somatic mutations in cancer are an infamous example of this heterogeneity. Although the same signaling pathways may be disrupted in a cancer patient cohort, the distribution of mutations is long-tailed, and many driver mutations may only be detected in a small fraction of patients. We developed a computational approach to account for heterogeneous data when inferring signaling pathways by sharing information across the samples. Our technique builds upon the prize-collecting Steiner forest problem, a network optimization algorithm that extracts pathways from a protein-protein interaction network. We recover signaling pathways that are similar across all samples yet still reflect the unique characteristics of each biological sample. Leveraging data from related tumors improves our ability to recover the disrupted pathways and reveals patient-specific pathway perturbations in breast cancer.

DETECTING STATISTICAL INTERACTION BETWEEN SOMATIC MUTATIONAL EVENTS AND GERMLINE VARIATION FROM NEXT-GENERATION SEQUENCE DATA

Hao Hu, Chad D. Huff

Department of Epidemiology
The University of Texas MD Anderson Cancer Center
Houston, TX, 77030, USA

The two-hit model of carcinogenesis provides a valuable framework for understanding the role of DNA repair and tumor suppressor genes in cancer development and progression. Under this model, tumor development can initiate from a single somatic mutation in individuals that inherit an inactivating germline variant. Although the two-hit model can be an overgeneralization, the tendency for the pattern of somatic mutations to differ in cancer patients that inherit predisposition alleles is a signal that can be used to identify and validate germline susceptibility variants. Here, we present the Somatic-Germline Interaction (SGI) tool, which is designed to identify statistical interaction between germline variants and somatic mutational events from next-generation sequence data. SGI interfaces with rare-variant association tests and variant classifiers to identify candidate germline susceptibility variants from case-control sequencing data. SGI then analyzes tumor-normal pair next-generation sequence data to evaluate evidence for somatic-germline interaction in each gene or pathway using two tests: the Allelic Imbalance Rank Sum (AIRS) test and the Somatic Mutation Interaction Test (SMIT). AIRS tests for preferential allelic imbalance to evaluate whether somatic mutational events tend to amplify candidate germline variants. SMIT evaluates whether somatic point mutations and small indels occur more or less frequently than expected in the presence of candidate germline variants. Both AIRS and SMIT control for heterogeneity in the mutational process resulting from regional variation in mutation rates and inter-sample variation in background mutation rates. The SGI test combines AIRS and SMIT to provide a single, unified measure of statistical interaction between somatic mutational events and germline variation. We show that the tests implemented in SGI have high power with relatively modest sample sizes in a wide variety of scenarios. We demonstrate the utility of SGI to increase the power of rare variant association studies in cancer and to validate the potential role in cancer causation of germline susceptibility variants.

SYSTEMATIC ASSESSMENT OF ANALYTICAL METHODS FOR DRUG SENSITIVITY PREDICTION FROM CANCER CELL LINE DATA

In Sock Jang, Elias Chaibub Neto, Justin Guinney, Stephen H. Friend, Adam A. Margolin

Sage Bionetworks
1100 Fairview Ave. N Seattle, WA 98109, USA

Large-scale pharmacogenomic screens of cancer cell lines have emerged as an attractive pre-clinical system for identifying tumor genetic subtypes with selective sensitivity to targeted therapeutic strategies. Application of modern machine learning approaches to pharmacogenomic datasets have demonstrated the ability to infer genomic predictors of compound sensitivity. Such modeling approaches entail many analytical design choices; however, a systematic study evaluating the relative performance attributable to each design choice is not yet available. In this work, we evaluated over 110,000 different models, based on a multifactorial experimental design testing systematic combinations of modeling factors within several categories of modeling choices, including: type of algorithm, type of molecular feature data, compound being predicted, method of summarizing compound sensitivity values, and whether predictions are based on discretized or continuous response values. Our results suggest that model input data (type of molecular features and choice of compound) are the primary factors explaining model performance, followed by choice of algorithm. Our results also provide a statistically principled set of recommended modeling guidelines, including: using elastic net or ridge regression with input features from all genomic profiling platforms, most importantly, gene expression features, to predict continuous-valued sensitivity scores summarized using the area under the dose response curve, with pathway targeted compounds most likely to yield the most accurate predictors. In addition, our study provides a publicly available resource of all modeling results, an open source code base, and experimental design for researchers throughout the community to build on our results and assess novel methodologies or applications in related predictive modeling problems.

INTEGRATIVE ANALYSIS OF TWO CELL LINES DERIVED FROM A NON-SMALL-LUNG CANCER PATIENT – A PANOMICS APPROACH

Oleg Mayba, Florian Gnad, Michael Peyton, Fan Zhang, Kimberly Walter, Pan Du, Melanie A. Huntley, Zhaoshi Jiang, Jinfeng Liu, Peter M. Haverty, Robert C. Gentleman, Ruiqiang Li, John D. Minna, Yingrui Li, David S. Shames, Zemin Zhang

Departments of Bioinformatics and Computational Biology and
Development Oncology Diagnostics
Genentech, Inc., South San Francisco, CA 94080, USA

BGI-Shenzhen, Shenzhen 518083, China

Hamon Center for Therapeutic Oncology Research
UT-Southwestern Medical Center, Dallas, TX 75390, USA

Cancer cells derived from different stages of tumor progression may exhibit distinct biological properties, as exemplified by the paired lung cancer cell lines H1993 and H2073. While H1993 was derived from chemo-naive metastasized tumor, H2073 originated from the chemo-resistant primary tumor from the same patient and exhibits strikingly different drug response profile. To understand the underlying genetic and epigenetic bases for their biological properties, we investigated these cells using a wide range of large-scale methods including whole genome sequencing, RNA sequencing, SNP array, DNA methylation array, and de novo genome assembly. We conducted an integrative analysis of both cell lines to distinguish between potential driver and passenger alterations. Although many genes are mutated in these cell lines, the combination of DNA- and RNA-based variant information strongly implicates a small number of genes including TP53 and STK11 as likely drivers. Likewise, we found a diverse set of genes differentially expressed between these cell lines, but only a fraction can be attributed to changes in DNA copy number or methylation. This set included the ABC transporter ABCC4, implicated in drug resistance, and the metastasis associated MET oncogene. While the rich data content allowed us to reduce the space of hypotheses that could explain most of the observed biological properties, we also caution there is a lack of statistical power and inherent limitations in such single patient case studies.

AN INTEGRATED APPROACH TO BLOOD-BASED CANCER DIAGNOSIS AND BIOMARKER DISCOVERY

Martin Renqiang Min, Salim Chowdhury, Yanjun Qi, Alex Stewart, Rachel Ostroff

NEC Labs America, Princeton, NJ 08540, USA
Lane Center for Computational Biology, Carnegie Mellon University, Pittsburgh, PA 15213, USA
SomaLogic, Inc., Boulder, CO 80301, USA

Disrupted or abnormal biological processes responsible for cancers often quantitatively manifest as disrupted additive and multiplicative interactions of gene/protein expressions correlating with cancer progression. However, the examination of all possible combinatorial interactions between gene features in most case-control studies with limited training data is computationally infeasible. In this paper, we propose a practically feasible data integration approach, QUIRE (QUadratic Interactions among infoRmative fEatures), to identify discriminative complex interactions among informative gene features for cancer diagnosis and biomarker discovery directly based on patient blood samples. QUIRE works in two stages, where it first identifies functionally relevant gene groups for the disease with the help of gene functional annotations and available physical protein interactions, then it explores the combinatorial relationships among the genes from the selected informative groups. Based on our private experimentally generated data from patient blood samples using a novel SOMAmer (Slow Off-rate Modified Aptamer) technology, we apply QUIRE to cancer diagnosis and biomarker discovery for Renal Cell Carcinoma (RCC) and Ovarian Cancer (OVC). To further demonstrate the general applicability of our approach, we also apply QUIRE to a publicly available Colorectal Cancer (CRC) dataset that can be used to prioritize our SOMAmer design. Our experimental results show that QUIRE identifies gene-gene interactions that can better identify the different cancer stages of samples, as compared to other state-of-the-art feature selection methods. A literature survey shows that many of the interactions identified by QUIRE play important roles in the development of cancer.

MULTIPLEX META-ANALYSIS OF MEDULLOBLASTOMA EXPRESSION STUDIES WITH EXTERNAL CONTROLS

Alexander A. Morgan, Matthew D. Li, Achal S. Achrol, Purvesh J. Khatri, Samuel H. Cheshier

Stanford University School of Medicine
Stanford, CA 94305

We propose and discuss a method for doing gene expression meta-analysis (multiple datasets) across multiplex measurement modalities measuring the expression of many genes simultaneously (e.g. microarrays and RNAseq) using external control samples and a method of heterogeneity detection to identify and filter on comparable gene expression measurements. We demonstrate this approach on publicly available gene expression datasets from samples of medulloblastoma and normal cerebellar tissue and identify some potential new targets in the treatment of medulloblastoma.

COMPUTATIONAL APPROACHES TO DRUG REPURPOSING AND PHARMACOLOGY

ACCEPTED PROCEEDINGS PAPERS WITH ORAL PRESENTATIONS

ANTI-INFECTIOUS DRUG REPURPOSING USING AN INTEGRATED CHEMICAL GENOMICS AND STRUCTURAL SYSTEMS BIOLOGY APPROACH

Clara Ng, Ruth Hauptman, Yinliang Zhang, Philip E. Bourne, Lei Xie

Department of Computer Science, Hunter College, the City University of New York
695 Park Avenue, New York City, NY 10065, U. S. A.

Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego
9500 Gilman Drive, La Jolla, CA 92093, U. S. A.

The emergence of multi-drug and extensive drug resistance of microbes to antibiotics poses a great threat to human health. Although drug repurposing is a promising solution for accelerating the drug development process, its application to anti-infectious drug discovery is limited by the scope of existing phenotype-, ligand-, or target-based methods. In this paper we introduce a new computational strategy to determine the genome-wide molecular targets of bioactive compounds in both human and bacterial genomes. Our method is based on the use of a novel algorithm, ligand Enrichment of Network Topological Similarity (ligENTS), to map the chemical universe to its global pharmacological space. ligENTS outperforms the state-of-the-art algorithms in identifying novel drug-target relationships. Furthermore, we integrate ligENTS with our structural systems biology platform to identify drug repurposing opportunities via target similarity profiling. Using this integrated strategy, we have identified novel P. falciparum targets of drug-like active compounds from the Malaria Box, and suggest that a number of approved drugs may be active against malaria. This study demonstrates the potential of an integrative chemical genomics and structural systems biology approach to drug repurposing.

DRUG-TARGET INTERACTION PREDICTION BY INTEGRATING CHEMICAL, GENOMIC, FUNCTIONAL AND PHARMACOLOGICAL DATA

Fan Yang, Jinbo Xu, Jianyang Zeng

Department of Mathematical Sciences
Tsinghua University
Beijing, 100084, P. R. China

Toyota Technological Institute at Chicago
6045 S. Kenwood Ave.
Chicago, IL 60637, USA

Institute for Interdisciplinary Information Sciences
Tsinghua University
Beijing, 100084, P. R. China

In silico prediction of unknown drug-target interactions (DTIs) has become a popular tool for drug repositioning and drug development. A key challenge in DTI prediction lies in integrating multiple types of data for accurate DTI prediction. Although recent studies have demonstrated that genomic, chemical and pharmacological data can provide reliable information for DTI prediction, it remains unclear whether functional information on proteins can also contribute to this task. Little work has been developed to combine such information with other data to identify new interactions between drugs and targets. In this paper, we introduce functional data into DTI prediction and construct biological space for targets using the functional similarity measure. We present a probabilistic graphical model, called conditional random field (CRF), to systematically integrate genomic, chemical, functional and pharmacological data plus the topology of DTI networks into a unified framework to predict missing DTIs. Tests on two benchmark datasets show that our method can achieve excellent prediction performance with the area under the precision-recall curve (AUPR) up to 94.9. These results demonstrate that our CRF model can successfully exploit heterogeneous data to capture the latent correlations of DTIs, and thus will be practically useful for drug repositioning. Supplementary Material is available at http://iiis.tsinghua.edu.cn/~compbio/papers/psb2014/psb2014_sm.pdf.

PREDICTION OF OFF-TARGET DRUG EFFECTS THROUGH DATA FUSION

Emmanuel R. Yera, Ann E. Cleves, Ajay N. Jain

Bioengineering and Therapeutic Sciences, University of California, San Francisco, San Francisco, CA 94143, USA

We present a probabilistic data fusion framework that combines multiple computational approaches for drawing relationships between drugs and targets. The approach has special relevance to identifying surprising unintended biological targets of drugs. Comparisons between molecules are made based on 2D topological structural considerations, based on 3D surface characteristics, and based on English descriptions of clinical effects. Similarity computations within each modality were transformed into probability scores. Given a new molecule along with a set of molecules sharing some biological effect, a single score based on comparison to the known set is produced, reflecting either 2D similarity, 3D similarity, clinical effects similarity or their combination. The methods were validated within a curated structural pharmacology database (SPDB) and further tested by blind application to data derived from the ChEMBL database. For prediction of off-target effects, 3D-similarity performed best as a single modality, but combining all methods produced performance gains. Striking examples of structurally surprising off-target predictions are presented.

EXPLORING THE PHARMACOGENOMICS KNOWLEDGE BASE (PHARMGKB) FOR REPOSITIONING BREAST CANCER DRUGS BY LEVERAGING WEB ONTOLOGY LANGUAGE (OWL) AND CHEMINFORMATICS APPROACHES

Qian Zhu, Cui Tao, Feichen Shen, Christopher Chute

Department of Health Sciences Research, Mayo Clinic, Rochester, MN 55905, USA
School of Biomedical Informatics, University of Texas Health Science Center at Houston, TX 77030, USA
School of Computing and Engineering, University of Missouri-Kansas City, Kansas City, MO 64110, USA

Computational drug repositioning leverages computational technology and high volume of biomedical data to identify new indications for existing drugs. Since it does not require costly experiments that have a high risk of failure, it has attracted increasing interest from diverse fields such as biomedical, pharmaceutical, and informatics areas. In this study, we used pharmacogenomics data generated from pharmacogenomics studies, applied informatics and Semantic Web technologies to address the drug repositioning problem. Specifically, we explored PharmGKB to identify pharmacogenomics related associations as pharmacogenomics profiles for US Food and Drug Administration (FDA) approved breast cancer drugs. We then converted and represented these profiles in Semantic Web notations, which support automated semantic inference. We successfully evaluated the performance and efficacy of the breast cancer drug pharmacogenomics profiles by case studies. Our results demonstrate that combination of pharmacogenomics data and Semantic Web technology/Cheminformatics approaches yields better performance of new indication and possible adverse effects prediction for breast cancer drugs.

DETECTING AND CHARACTERIZING PLEIOTROPY: NEW METHODS FOR UNCOVERING THE CONNECTION BETWEEN THE COMPLEXITY OF GENOMIC ARCHITECTURE AND MULTIPLE PHENOTYPES

ACCEPTED PROCEEDINGS PAPERS WITH ORAL PRESENTATIONS

USING THE BIPARTITE HUMAN PHENOTYPE NETWORK TO REVEAL PLEIOTROPY AND EPISTASIS BEYOND THE GENE

Christian Darabos, Samantha H. Harmon, Jason H. Moore

Institute for the Quantitative Biomedical Sciences
The Geisel Medical School at Dartmouth College
Lebanon, NH 03756, U.S.A.

With the rapid increase in the quality and quantity of data generated by modern high-throughput sequencing techniques, there has been a need for innovative methods able to convert this tremendous amount of data into more accessible forms. Networks have been a corner stone of this movement, as they are an intuitive way of representing interaction data, yet they offer a full set of sophisticated statistical tools to analyze the phenomena they model. We propose a novel approach to reveal and analyze pleiotropic and epistatic effects at the genome-wide scale using a bipartite network composed of human diseases, phenotypic traits, and several types of predictive elements (i.e. SNPs, genes, or pathways). We take advantage of publicly available GWAS data, gene and pathway databases, and more to construct networks different levels of granularity, from common genetic variants to entire biological pathways. We use the connections between the layers of the network to approximate the pleiotropy and epistasis effects taking place between the traits and the predictive elements. The global graph-theory based quantitative methods reveal that the levels of pleiotropy and epistasis are comparable for all types of predictive element. The results of the magnified "glaucoma" region of the network demonstrate the existence of well documented interactions, supported by overlapping genes and biological pathway, and more obscure associations. As the amount and complexity of genetic data increases, bipartite, and more generally multipartite networks that combine human diseases and other physical attributes with layers of genetic information, have the potential to become ubiquitous tools in the study of complex genetic and phenotypic interactions.

ENVIRONMENT-WIDE ASSOCIATION STUDY (EWAS) FOR TYPE 2 DIABETES IN THE MARSHFIELD PERSONALIZED MEDICINE RESEARCH PROJECT BIOBANK

Molly A. Hall, Scott M. Dudek, Robert Goodloe, Dana C. Crawford, Sarah A. Pendergrass, Peggy Peissig, Murray Brilliant, Catherine A. McCarty, Marylyn D. Ritchie

The Pennsylvania State University, University Park, PA 16802, USA
Vanderbilt University, Nashville TN, 37232, USA
The Marshfield Clinic, Marshfield, WI, USA
Essentia Institute of Rural Health, Duluth, MN, USA

Environment-wide association studies (EWAS) provide a way to uncover the environmental mechanisms involved in complex traits in a high-throughput manner. Genome-wide association studies have led to the discovery of genetic variants associated with many common diseases but do not take into account the environmental component of complex phenotypes. This EWAS assesses the comprehensive association between environmental variables and the outcome of type 2 diabetes (T2D) in the Marshfield Personalized Medicine Research Project Biobank (Marshfield PMRP). We sought replication in two National Health and Nutrition Examination Surveys (NHANES). The Marshfield PMRP currently uses four tools for measuring environmental exposures and outcome traits: 1) the PhenX Toolkit includes standardized exposure and phenotypic measures across several domains, 2) the Diet History Questionnaire (DHQ) is a food frequency questionnaire, 3) the Measurement of a Person's Habitual Physical Activity scores the level of an individual's physical activity, and 4) electronic health records (EHR) employs validated algorithms to establish T2D case-control status . Using PLATO software, 314 environmental variables were tested for association with T2D using logistic regression, adjusting for sex, age, and BMI in over 2,200 European Americans. When available, similar variables were tested with the same methods and adjustment in samples from NHANES III and NHANES 1999-2002. Twelve and 31 associations were identified in the Marshfield samples at $p<0.01$ and $p<0.05$, respectively. Seven and 13 measures replicated in at least one of the NHANES at $p<0.01$ and $p<0.05$, respectively, with the same direction of effect. The most significant environmental exposures associated with T2D status included decreased alcohol use as well as increased smoking exposure in childhood and adulthood. The results demonstrate the utility of the EWAS method and survey tools for identifying environmental components of complex diseases like type 2 diabetes. These high-throughput and comprehensive investigation methods can easily be applied to investigate the relation between environmental exposures and multiple phenotypes in future analyses.

DISSECTION OF COMPLEX GENE EXPRESSION USING THE COMBINED ANALYSIS OF PLEIOTROPY AND EPISTASIS

Vivek M. Philip, Anna L. Tyler, Gregory W. Carter

The Jackson Laboratory,
Bar Harbor, ME, 04609, USA

Global transcript expression experiments are commonly used to investigate the biological processes that underlie complex traits. These studies can exhibit complex patterns of pleiotropy when trans-acting genetic factors influence overlapping sets of multiple transcripts. Dissecting these patterns into biological modules with distinct genetic etiology can provide models of how genetic variants affect specific processes that contribute to a trait. Here we identify transcript modules associated with pleiotropic genetic factors and apply genetic interaction analysis to disentangle the regulatory architecture in a mouse intercross study of kidney function. The method, called the combined analysis of pleiotropy and epistasis (CAPE), has been previously used to model genetic networks for multiple physiological traits. It simultaneously models multiple phenotypes to identify direct genetic influences as well as influences mediated through genetic inte ractions. We first identified candidate trans expression quantitative trait loci (eQTL) and the transcripts potentially affected. We then clustered the transcripts into modules of co-expressed genes, from which we compute summary module phenotypes. Finally, we applied CAPE to map the network of interacting module QTL (modQTL) affecting the gene modules. The resulting network mapped how multiple modQTL both directly and indirectly affect modules associated with metabolic functions and biosynthetic processes. This work demonstrates how the integration of pleiotropic signals in gene expression data can be used to infer a complex hypothesis of how multiple loci interact to co-regulate transcription programs, thereby providing additional constraints to prioritize validation experiments.

PERSONALIZED MEDICINE: FROM GENOTYPES AND MOLECULAR PHENOTYPES TOWARDS THERAPY

ACCEPTED PROCEEDINGS PAPERS WITH ORAL PRESENTATIONS

PATH-SCAN: A REPORTING TOOL FOR IDENTIFYING CLINICALLY ACTIONABLE VARIANTS

Roxana Daneshjou, Zachary Zappala, Kim Kukurba, Sean M Boyle, Kelly E Ormond, Teri E Klein, Michael Snyder, Carlos D Bustamante, Russ B Altman, Stephen B Montgomery

Stanford University
Stanford, CA 94305

The American College of Medical Genetics and Genomics (ACMG) recently released guidelines regarding the reporting of incidental findings in sequencing data. Given the availability of Direct to Consumer (DTC) genetic testing and the falling cost of whole exome and genome sequencing, individuals will increasingly have the opportunity to analyze their own genomic data. We have developed a web-based tool, PATH-SCAN, which annotates individual genomes and exomes for ClinVar designated pathogenic variants found within the genes from the ACMG guidelines. Because mutations in these genes predispose individuals to conditions with actionable outcomes, our tool will allow individuals or researchers to identify potential risk variants in order to consult physicians or genetic counselors for further evaluation. Moreover, our tool allows individuals to anonymously submit their pathogenic burden, so that we can crowd source the collection of quantitative information regarding the frequency of these variants. We tested our tool on 1092 publicly available genomes from the 1000 Genomes project, 163 genomes from the Personal Genome Project, and 15 genomes from a clinical genome sequencing research project. Excluding the most commonly seen variant in 1000 Genomes, about 20% of all genomes analyzed had a ClinVar designated pathogenic variant that required further evaluation.

IMPUTATION-BASED ASSESSMENT OF NEXT GENERATION RARE EXOME VARIANT ARRAYS

Alicia R. Martin, Gerard Tse, Carlos D. Bustamante, Eimear E. Kenny

Stanford University, Stanford, CA 94305
Icahn School of Medicine at Mount Sinai, New York, NY 10029

A striking finding from recent large-scale sequencing efforts is that the vast majority of variants in the human genome are rare and found within single populations or lineages. These observations hold important implications for the design of the next round of disease variant discovery efforts—if genetic variants that influence disease risk follow the same trend, then we expect to see population-specific disease associations that require large samples sizes for detection. To address this challenge, and due to the still prohibitive cost of sequencing large cohorts, researchers have developed a new generation of low-cost genotyping arrays that assay rare variation previously identified from large exome sequencing studies. Genotyping approaches rely not only on directly observing variants, but also on phasing and imputation methods that use publicly available reference panels to infer unobserved variants in a study cohort. Rare variant exome arrays are intentionally enriched for variants likely to be disease causing, and here we assay the ability of the first commercially available rare exome variant array (the Illumina Infinium HumanExome BeadChip) to also tag other potentially damaging variants not molecularly assayed. Using full sequence data from chromosome 22 from the phase I 1000 Genomes Project, we evaluate three methods for imputation (BEAGLE, MaCH-Admix, and SHAPEIT2/IMPUTE2) with the rare exome variant array under varied study panel sizes, reference panel sizes, and LD structures via population differences. We find that imputation is more accurate across both the genome and exome for common variant arrays than the next generation array for all allele frequencies, including rare alleles. We also find that imputation is the least accurate in African populations, and accuracy is substantially improved for rare variants when the same population is included in the reference panel. Depending on the goals of GWAS researchers, our results will aid budget decisions by helping determine whether money is best spent sequencing the genomes of smaller sample sizes, genotyping larger sample sizes with rare and/or common variant arrays and imputing SNPs, or some combination of the two.

UTILIZATION OF AN EMR-BIOREPOSITORY TO IDENTIFY THE GENETIC PREDICTORS OF CALCINEURIN-INHIBITOR TOXICITY IN HEART TRANSPLANT RECIPIENTS

Matthew Oetjens, William S. Bush, Kelly A. Birdwell, Holli H. Dilks, Erica A. Bowton, Joshua C. Denny, Russell A. Wilke, Dan M. Roden, Dana C. Crawford

Vanderbilt University
Nashville, TN 37212, USA

Calcineurin-inhibitors CI are immunosuppressive agents prescribed to patients after solid organ transplant to prevent rejection. Although these drugs have been transformative for allograft survival, long-term use is complicated by side effects including nephrotoxicity. Given the narrow therapeutic index of CI, therapeutic drug monitoring is used to prevent acute rejection from underdosing and acute toxicity from overdosing, but drug monitoring does not alleviate long-term side effects. Patients on calcineurin-inhibitors for long periods almost universally experience declines in renal function, and a subpopulation of transplant recipients ultimately develop chronic kidney disease that may progress to end stage renal disease attributable to calcineurin inhibitor toxicity (CNIT). Pharmacogenomics has the potential to identify patients who are at high risk for developing advanced chronic kidney disease caused by CNIT and providing them with existing alternate immunosuppressive therapy. In this study we utilized BioVU, Vanderbilt University Medical Center's DNA biorepository linked to de-identified electronic medical records to identify a cohort of 115 heart transplant recipients prescribed calcineurin-inhibitors to identify genetic risk factors for CNIT We identified 37 cases of nephrotoxicity in our cohort, defining nephrotoxicity as a monthly median estimated glomerular filtration rate (eGFR) <30 mL/min/1.73m2 at least six months post-transplant for at least three consecutive months. All heart transplant patients were genotyped on the Illumina ADME Core Panel, a pharmacogenomic genotyping platform that assays 184 variants across 34 genes. In Cox regression analysis adjusting for age at transplant, pre-transplant chronic kidney disease, pre-transplant diabetes, and the three most significant principal components (PCAs), we did not identify any markers that met our multiple-testing threshold. As a secondary analysis we also modeled post-transplant eGFR directly with linear mixed models adjusted for age at transplant, cyclosporine use, median BMI, and the three most significant principal components. While no SNPs met our threshold for significance, a SNP previously identified in genetic studies of the dosing of tacrolimus CYP34A rs776746, replicated in an adjusted analysis at an uncorrected p-value of 0.02 (coeff(S.E.) = 14.60(6.41)). While larger independent studies will be required to further validate this finding, this study underscores the EMRs usefulness as a resource for longitudinal pharmacogenetic study designs.

ROBUST REVERSE ENGINEERING OF DYNAMIC GENE NETWORKS UNDER SAMPLE SIZE HETEROGENEITY

Ankur P. Parikh, Wei Wu, Eric P. Xing

School of Computer Science, Carnegie Mellon University
Pittsburgh, PA 15213, USA

Simultaneously reverse engineering a collection of condition-specific gene networks from gene expression microarray data to uncover dynamic mechanisms is a key challenge in systems biology. However, existing methods for this task are very sensitive to variations in the size of the microarray samples across different biological conditions (which we term sample size heterogeneity in network reconstruction), and can potentially produce misleading results that can lead to incorrect biological interpretation. In this work, we develop a more robust framework that addresses this novel problem. Just like microarray measurements across conditions must undergo proper normalization on their magnitudes before entering subsequent analysis, we argue that networks across conditions also need to be "normalized" on their density when they are constructed, and we provide an algorithm that allows such normalization to be facilitated while estimating the networks. We show the quantitative advantages of our approach on synthetic and real data. Our analysis of a hematopoietic stem cell dataset reveals interesting results, some of which are confirmed by previously validated results.

VARIANT PRIORIZATION AND ANALYSIS INCORPORATING PROBLEMATIC REGIONS OF THE GENOME

Anil Patwardhan, Michael Clark, Alex Morgan, Stephen Chervitz, Mark Pratt, Gabor Bartha, Gemma Chandratillake, Sarah Garcia, Nan Leng, Richard Chen

Personalis Inc.
1350 Willow Road, Suite 202, Menlo Park, CA, 94025, USA

In case-control studies of rare Mendelian disorders and complex diseases, the power to detect variant and gene-level associations of a given effect size is limited by the size of the study sample. Paradoxically, low statistical power may increase the likelihood that a statistically significant finding is also a false positive. The prioritization of variants based on call quality, putative effects on protein function, the predicted degree of deleteriousness, and allele frequency is often used as a mechanism for reducing the occurrence of false positives, while preserving the set of variants most likely to contain true disease associations. We propose that specificity can be further improved by considering errors that are specific to the regions of the genome being sequenced. These problematic regions (PRs) are identified a-priori and are used to down-weight constitutive variants in a case-control analysis. Using samples drawn from 1000-Genomes, we illustrate the utility of PRs in identifying true variant and gene associations using a case-control study on a known Mendelian disease, cystic fibrosis(CF).

JOINT ASSOCIATION DISCOVERY AND DIAGNOSIS OF ALZHEIMER'S DISEASE BY SUPERVISED HETEROGENEOUS MULTIVIEW LEARNING

Shandian Zhe, Zenglin Xu, Yuan Qi, Peng Yu

Purdue University
West Lafayette, IN 47907, USA

A key step for Alzheimer's disease (AD) study is to identify associations between genetic variations and intermediate phenotypes (e.g., brain structures). At the same time, it is crucial to develop a noninvasive means for AD diagnosis. Although these two tasks--association discovery and disease diagnosis--have been treated separately by a variety of approaches, they are tightly coupled due to their common biological basis. We hypothesize that the two tasks can potentially benet each other by a joint analysis, because (i) the association study discovers correlated biomarkers from different data sources, which may help improve diagnosis accuracy, and (ii) the disease status may help identify disease-sensitive associations between genetic variations and MRI features. Based on this hypothesis, we present a new sparse Bayesian approach for joint association study and disease diagnosis. In this approach, common latent features are extracted from different data sources based on sparse projection matrices and used to predict multiple disease severity levels based on Gaussian process ordinal regression; in return, the disease status is used to guide the discovery of relationships between the data sources. The sparse projection matrices not only reveal the associations but also select groups of biomarkers related to AD. To learn the model from data, we develop an efficient variational expectation maximization algorithm. Simulation results demonstrate that our approach achieves higher accuracy in both predicting ordinal labels and discovering associations between data sources than alternative methods. We apply our approach to an imaging genetics dataset of AD. Our joint analysis approach not only identies meaningful and interesting associations between genetic variations, brain structures, and AD status, but also achieves significantly higher accuracy for predicting ordinal AD stages than the competing methods.

TEXT AND DATA MINING FOR BIOMEDICAL DISCOVERY

ACCEPTED PROCEEDINGS PAPERS WITH ORAL PRESENTATIONS

VECTOR QUANTIZATION KERNELS FOR THE CLASSIFICATION OF PROTEIN SEQUENCES AND STRUCTURES

Wyatt T. Clark, Predrag Radivojac

Department of Computer Science and Informatics, Indiana University
Bloomington, Indiana 47405, U.S.A.

We propose a new kernel-based method for the classification of protein sequences and structures. We first represent each protein as a set of time series data using several structural, physicochemical, and predicted properties such as a sequence of consecutive dihedral angles, hydrophobicity indices, or predictions of disordered regions. A kernel function is then computed for pairs of proteins, exploiting the principles of vector quantization and subsequently used with support vector machines for protein classification. Although our method requires a significant pre-processing step, it is fast in the training and prediction stages owing to the linear complexity of kernel computation with the length of protein sequences. We evaluate our approach on two protein classification tasks involving the prediction of SCOP structural classes and catalytic activity according to the Gene Ontology. We provide evidence that the method is competitive when compared to string kernels, and useful for a range of protein classification tasks. Furthermore, the applicability of our approach extends beyond computational biology to any classification of time series data.

COMBINING HETEROGENOUS DATA FOR PREDICTION OF DISEASE RELATED AND PHARMACOGENES

Christopher S. Funk, Lawrence E. Hunter, K. Bretonnel Cohen

Computational Bioscience Program, University of Colorado School of Medicine
Aurora, CO 80045, USA

Identifying genetic variants that affect drug response or play a role in disease is an important task for clinicians and researchers. Before individual variants can be explored efficiently for effect on drug response or disease relationships, specific candidate genes must be identified. While many methods rank candidate genes through the use of sequence features and network topology, only a few exploit the information contained in the biomedical literature. In this work, we train and test a classifier on known pharmacogenes from PharmGKB and present a classifier that predicts pharmacogenes on a genome-wide scale using only Gene Ontology annotations and simple features mined from the biomedical literature. Performance of F=0.86, AUC=0.860 is achieved. The top 10 predicted genes are analyzed. Additionally, a set of enriched pharmacogenic Gene Ontology concepts is produced.

A NOVEL PROFILE BIOMARKER DIAGNOSIS FOR MASS SPECTRAL PROTEOMICS

Henry Han

Department of Computer and Information Science
Fordham University, New York NY 10023 USA

Quantitative Proteomics Center, Columbia University, New York 10027 USA

Mass spectrometry based proteomics technologies have allowed for a great progress in identifying disease biomarkers for clinical diagnosis and prognosis. However, they face acute challenges from a data reproducibility standpoint, in that no two independent studies have been found to produce the same proteomic patterns. Such reproducibility issues cause the identified biomarker patterns to lose repeatability and prevent real clinical usage. In this work, we propose a profile biomarker approach to overcome this problem from a machine-learning viewpoint by developing a novel derivative component analysis (DCA). As an implicit feature selection algorithm, derivative component analysis enables the separation of true signals from red herrings by capturing subtle data behaviors and removing system noises from a proteomic profile. We further demonstrate its advantages in disease diagnosis by viewing input data as a profile biomarker. The results from our profile biomarker diagnosis suggest an effective solution to overcoming proteomics data's reproducibility problem, present an alternative method for biomarker discovery in proteomics, and provide a good candidate for clinical proteomic diagnosis

TOWARDS PATHWAY CURATION THROUGH LITERATURE MINING – A CASE STUDY USING PHARMGKB

Ravikumar K.E., Kavishwar B. Wagholikar, Hongfang Liu

Department of Health Sciences Research, College of Medicine, Mayo Clinic, Rochester, MN, 55905

The creation of biological pathway knowledge bases is largely driven by manual effort to curate based on evidences from the scientific literature. It is highly challenging for the curators to keep up with the literature. Text mining applications have been developed in the last decade to assist human curators to speed up the curation pace where majority of them aim to identify the most relevant papers for curation with little attempt to directly extract the pathway information from text. In this paper, we describe a rule-based literature mining system to extract pathway information from text. We evaluated the system using curated pharmacokinetic (PK) and pharmacodynamic (PD) pathways in PharmGKB. The system achieved an F-measure of 63.11% and 34.99% for entity extraction and event extraction respectively against all PubMed abstracts cited in PharmGKB. It may be possible to improve the system performance by incorporating using statistical machine learning approaches. This study also helped us gain insights into the barriers towards automated event extraction from text for pathway curation.

SPARSE GENERALIZED FUNCTIONAL LINER MODEL FOR PREDICTING REMISSION STATUS OF DEPRESSION PATIENTS

Yashu Liu, Zhi Nie, Jiayu Zhou, Michael Farnum, Vaibhav A Narayan, Gayle Wiittenberg, Jieping Ye

Department of Computer Science and Engineering
Center for Evolutionary Medicine and Informatics, The Biodesign Institute
Arizona State University, Tempe, AZ 85287, USA

Johnson & Johnson Pharmaceutical Research & Development, LLC,
Titusville, NJ, USA

Complex diseases such as major depression affect people over time in complicated patterns. Longitudinal data analysis is thus crucial for understanding and prognosis of such diseases and has received considerable attention in the biomedical research community. Traditional classification and regression methods have been commonly applied in a simple (controlled) clinical setting with a small number of time points. However, these methods cannot be easily extended to the more general setting for longitudinal analysis, as they are not inherently built for time-dependent data. Functional regression, in contrast, is capable of identifying the relationship between features and outcomes along with time information by assuming features and/or outcomes as random functions over time rather than independent random variables. In this paper, we propose a novel sparse generalized functional linear model for the prediction of treatment remission status of the depression participants with longitudinal features. Compared to traditional functional regression models, our model enables high-dimensional learning, smoothness of functional coefficients, longitudinal feature selection and interpretable estimation of functional coefficients. Extensive experiments have been conducted on the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) data set and the results show that the proposed sparse functional regression method achieves significantly higher prediction power than existing approaches.

DEVELOPMENT OF A DATA-MINING ALGORITHM TO IDENTIFY AGES AT REPRODUCTIVE MILESTONES IN ELECTRONIC MEDICAL RECORDS

Jennifer Malinowski, Eric Farber-Eger, Dana C. Crawford

Vanderbilt University
Nashville, TN 37232, USA

Electronic medical records (EMRs) are becoming more widely implemented following directives from the federal government and incentives for supplemental reimbursements for Medicare and Medicaid claims. Replete with rich phenotypic data, EMRs offer a unique opportunity for clinicians and researchers to identify potential research cohorts and perform epidemiologic studies. Notable limitations to the traditional epidemiologic study include cost, time to complete the study, and limited ancestral diversity; EMR-based epidemiologic studies offer an alternative. The Epidemiologic Architecture for Genes Linked to Environment (EAGLE) Study, as part of the Population Architecture using Genomics and Epidemiology (PAGE) I Study, has genotyped more than 15,000 patients of diverse ancestry in BioVU, the Vanderbilt University Medical Center's biorepository linked to the EMR (EAGLE BioVU). We report here the development and performance of data-mining techniques used to identify the age at menarche (AM) and age at menopause (AAM), important milestones in the reproductive lifespan, in women from EAGLE BioVU for genetic association studies. In addition, we demonstrate the ability to discriminate age at naturally-occurring menopause (ANM) from medically-induced menopause. Unusual timing of these events may indicate underlying pathologies and increased risk for some complex diseases and cancer; however, they are not consistently recorded in the EMR. Our algorithm offers a mechanism by which to extract these data for clinical and research goals.

AN EFFICIENT ALGORITHM TO INTEGRATE NETWORK AND ATTRIBUTE DATA FOR GENE FUNCTION PREDICTION

Shankar Vembu, Quaid Morris

University of Toronto
Toronto, ON, Canada

Label propagation methods are extremely well-suited for a variety of biomedical prediction tasks based on network data. However, these algorithms cannot be used to integrate feature-based data sources with networks. We propose an efficient learning algorithm to integrate these two types of heterogeneous data sources to perform binary prediction tasks on node features (e.g., gene prioritization, disease gene prediction). Our method, LMGraph, consists of two steps. In the first step, we extract a small set of "network features" from the nodes of networks that represent connectivity with labeled nodes in the prediction tasks. In the second step, we apply a simple weighting scheme in conjunction with linear classifiers to combine these network features with other feature data. This two step procedure allows us to (i) learn highly scalable and computationally efficient linear classifiers, (ii) and seamlessly combine feature-based data sources with networks. Our method is much faster than label propagation which is already known to be computationally efficient on large-scale prediction problems. Experiments on multiple functional interaction networks from three species (mouse, fly, C.elegans) with tens of thousands of nodes and hundreds of binary prediction tasks demonstrate the efficacy of our method.

POSTER PRESENTATIONS

CANCER PANOMICS: COMPUTATIONAL METHODS AND INFRASTRUCTURE FOR INTEGRATIVE ANALYSIS OF CANCER HIGH-THROUGHPUT "OMICS" DATA

POSTER PRESENTATIONS

ACCEPTED PROCEEDINGS PAPER WITH POSTER PRESENTATION.

EXTRACTING SIGNIFICANT SAMPLE-SPECIFIC CANCER MUTATIONS USING THEIR PROTEIN INTERACTIONS

Liviu Badea

University Politehnica Bucharest and Bioinformatics Group, ICI
8-10 Averescu Blvd, Bucharest, Romania
Email: badea.liviu@gmail.com

We present a joint analysis method for mutation and gene expression data employing information about proteins that are highly interconnected at the level of protein to protein (pp) interactions, which we apply to the TCGA Acute Myeloid Leukemia (AML) dataset. Given the low incidence of most mutations in virtually all cancer types, as well as the significant inter-patient heterogeneity of the mutation landscape, determining the true causal mutations in each individual patient remains one of the most important challenges for personalized cancer diagnostics and therapy. More automated methods are needed for determining these "driver" mutations in each individual patient. For this purpose, we are exploiting two types of contextual information: (1) the pp interactions of the mutated genes, as well as (2) their potential correlations with gene expression clusters. The use of pp interactions is based on our surprising finding that most AML mutations tend to affect nontrivial protein to protein interaction cliques.

COMBINING GATK, SAMTOOLS, VARSCAN2 AND PIBASE TO REDUCE FALSE NEGATIVES IN SOMATIC SNV-CALLING

Michael Forster, Andre Franke

Somatic variant calling in paired tumor/normal samples is challenging due to tumor heterogeneity on the one hand and the random DNA fragment sampling nature of next-generation sequencing on the other hand, and potentially also the additional biases introduced by exome capture or other targeted sequencing methods. We therefore use GATK and Samtools with the lowest possible stringency to detect potential novel mutations, VarScan2 to detect potential novel somatic mutations, and pibase to test each thus detected novel SNV site as well as all known dbSNP sites. We annotate the resulting candidate somatic variants and then „biologically" filter the large number of candidates to a reduced set of interesting candidates that can be more easily validated. In Pan-Omic analyses, we use pibase for high-throughput cross-validations of somatic SNVs between e.g. genomes, exomes, and transcriptomes. We here release a new version of pibase with improved low-coverage sensitivity. The originally published version uses the two-tailed 2x4 Fisher's exact test in combination with a minor allele noise threshold if the major alleles are identical in the paired samples. At low coverages, e.g. below 20x, Fisher's exact test may not yield significant p-values. For this reason the new release includes an additional test for allelic imbalance between the paired samples. The new release also introduces the new pibase_fisherdiff_mergevalidate tool for cross-validations between technical replicates or different Omics. pibase can be downloaded from http://www.ikmb-uni-kiel.de/pibase

CAUSES AND CONSEQUENCES OF CANCER TRANSCRIPTOME VARIABILITY

Kjong-Van Lehmann, André Kahles, Oliver Stegle, William Lee, David Kuo, Rileen Sinha, Cancer Genome Atlas Research Network, Nikolaus Schultz, Robert Klein, Gunnar Rätsch

The Cancer Genome Atlas (TCGA) provides a comprehensive survey of the molecular characteristics of 20 different tumor types and enables large scale analysis across multiple cancers. The genetic heterogeneity of the samples, differences in sequencing quality and protocols, population structure, and drastic expression differences in various tumor types require sophisticated tools for the joint analysis of the 4000 samples across multiple tumor types. We have re-aligned all RNA-seq and WXS-seq data in a uniform manner, applying the same mapping pipeline consistently across all samples. Gene expression, alternative splicing isoforms and splicing efficiency have been estimated on these new alignments. We further performed unified SNP calling on exome sequencing data from tumor tissue across all samples using the Genome Analysis Toolkit. A common variant association study (CVAS) using mixed models to account for the heterogeneity of the samples has been undertaken in order to improve our understanding of the relationship between common germline variants and alternative splicing patterns allowing us to find determinants of transcriptome variation. We investigate the effect of rare somatic variants known to have a potentially significant effect on transcriptional regulation. A rare variant association study (RVAS) using somatic variations from exon sequencing data is being applied to investigate the basis of somatic changes on transcriptional regulation. A decomposition of genomic covariances into trans and cis effects elucidates the importance of such factors across different cancer types which will not only improve our understanding of the molecular basis of cancer but also provide potential targets for the development of new treatment options.

BAYESIAN NETWORK (BN) STRUCTURE LEARNING INFERENCE APPLICATION IN STUDYING BRD4 AND JMJD6 FUNCTION COOPERATIVELY ON PROMOTER-PROXIMAL POLYMERASE-II (POL II) PAUSING RELEASE

Qi Ma, Wen Liu, Michael. G. Rosenfeld

Jumonji C domain-containing protein 6 (JMJD6), which is revealed to have demethylase activity toward H3R2me2/1 and H4R3me2/1, functions together with Bromodomain-containing protein 4 (BRD4), which is a positive regulatory component of P-TEFb complex, on promoter-proximal Pol II pausing release and transcriptional elongation. Based on observation of their physical interaction and TR (traveling ratio) tests by analyzing Pol II chromatin immunoprecipitation coupled with high throughput sequencing (ChIP-seq) and global nuclear run-on coupled with sequencing (Gro-seq) in cells transfected with control siRNA or siRNAs specifically targeting on JMJD6 or BRD4. We revealed that knockdown of JMJD6 or BRD4 caused Pol II pausing on promoter regions of genes positively regulated by both proteins. But the regulatory mechanism model is still not revealed. Genome-wide profiling of JMJD6 and BRD4 binding sites through ChIP-seq by mHMM model revealed that these two proteins regulate Pol II pausing release through their binding on distal-regulatory elements, but the regulatory relationship have not been studied in detail. The Bayesian network (BN) structure-learning model is an ideal probabilistic model of inferring regulatory relationships/interactions between a set of biological factors, it can infer the direct and indirect interaction and de novo identify the potential causal relationship, characterizing biological functions by integrating large-scale sequencing datasets. In our study, we apply this model to study and predict the regulatory patterns correlated regions existing in the observed combination of target factors and histone marks that might effects on the Pol II pausing release mechanism. And we found that BRD4 & JMJD6 are characterized to co-localized together in enhancer regions and co-regulated genes' promoter regions with association with different transcriptional factors, which suggests they function together on Pol-II promoter-proximal pausing release by combinatory effects from promoter and enhancers both. And which further suggests the existence of enhancer-promoter looping in this system.

INTEGRATED PROTEOMIC, PHOSPHOPROTEOMIC, AND GENOMIC PATHWAY ANALYSIS FOR PATIENT-DERIVED TUMOR SAMPLES

Jason E. McDermott, Vladislav A. Petyuk, Feng Yang, Marina A. Gritsenko, Matthew E. Monroe, Joshua T. Aldrich, Ronald J. Moore, Therese R. Clauss, Anil K. Shukla, Athena A. Schepmoes, Rosalie K. Chu, Samuel H. Payne, Tao Liu, Karin D. Rodland, Richard D. Smith

Cancer is, in general, a complex disease; it operates on pathways and systems, not solely on the individual components of those systems (genes or proteins). We have gathered global proteomic and phosphoproteomic data from a set of tumors with existing genomic data (sequencing, methylation, miRNA and mRNA expression) associated with a range of survival phenotypes. We report on our recent progress in analyzing proteomic, phosphoproteomic, and accompanying genomic data in terms of pathways that are significantly associated with breast and ovarian cancer progression and mortality. We show that proteomics and phosphoproteomics data both highlight a pathway set that is partly distinct and that partly overlaps with more traditional gene expression data. Further we use novel analysis approaches to integrate proteomic and genomic data to identify pathway activation and show that data integration improves resolution of activated pathways in short- and long- term survivors and in different clinical and molecularly-defined subtypes. We employ data integration and network analysis approaches to identify novel potential drivers of cancer progression. Our results provide a number of intriguing predictions of pathway and protein drivers of breast and ovarian cancer progression.

COMPUTATIONAL APPROACHES TO DRUG REPURPOSING AND PHARMACOLOGY

POSTER PRESENTATIONS

ACCEPTED PROCEEDINGS PAPER WITH POSTER PRESENTATION.

CHALLENGES IN SECONDARY ANALYSIS OF HIGH THROUGHPUT SCREENING DATA

Aurora S. Blucher, Shannon K. McWeeney

Division of Bioinformatics and Computational Biology, Oregon Health & Science University
Portland, OR 97203 USA
Emails: blucher@ohsu.edu, mcweeney@ohsu.edu

Repurposing an existing drug for an alternative use is not only a cost effective method of development, but also a faster process due to the drug's previous clinical testing and established pharmokinetic profiles. A potentially rich resource for computational drug repositioning approaches is publically available high throughput screening data, available in databases such as PubChem Bioassay and ChemBank. We examine statistical and computational considerations for secondary analysis of publicly available high throughput screening (HTS) data with respect to metadata, data quality, and completeness. We discuss developing methods and best practices that can help to ameliorate these issues.

ACCEPTED PROCEEDINGS PAPER WITH POSTER PRESENTATION.

DRUG INTERVENTION RESPONSE PREDICTIONS WITH PARADIGM (DIRPP) IDENTIFIES DRUG RESISTANT CANCER CELL LINES AND PATHWAY MECHANISMS OF RESISTANCE

Douglas Brubaker, Analisa Difeo, Yanwen Chen, Taylor Pearl, Kaide Zhai, Gurkan Bebek, Mark Chance, Jill Barnholtz-Sloan

Case Center for Proteomics and Bioinformatics, Case Western Reserve University School of Medicine, BRB 932, 10900 Euclid Avenue, Cleveland, Ohio 44106, USA

Case Comprehensive Cancer Center, Case Western Reserve University School of Medicine, 11100 Euclid Avenue, Cleveland, Ohio 44106, USA

Genomic Medicine Institute, Cleveland Clinic Lerner
Research Institute, 9500 Euclid Avenue, Cleveland, Ohio 44195, USA

Massachusetts Institute of Technology
77 Massachusetts Avenue, Cambridge, MA 02139, USA

The revolution in sequencing techniques in the past decade has provided an extensive picture of the molecular mechanisms behind complex diseases such as cancer. The Cancer Cell Line Encyclopedia (CCLE) and The Cancer Genome Project (CGP) have provided an unprecedented opportunity to examine copy number, gene expression, and mutational information for over 1000 cell lines of multiple tumor types alongside IC50 values for over 150 different drugs and drug related compounds. We present a novel pipeline called DIRPP, Drug Intervention Response Predictions with PARADIGM7, which predicts a cell line's response to a drug intervention from molecular data. PARADIGM (Pathway Recognition Algorithm using Data Integration on Genomic Models) is a probabilistic graphical model used to infer patient specific genetic activity by integrating copy number and gene expression data into a factor graph model of a cellular network. We evaluated the performance of DIRPP on endometrial, ovarian, and breast cancer related cell lines from the CCLE and CGP for nine drugs. The pipeline is sensitive enough to predict the response of a cell line with accuracy and precision across datasets as high as 80 and 88% respectively. We then classify drugs by the specific pathway mechanisms governing drug response. This classification allows us to compare drugs by cellular response mechanisms rather than simply by their specific gene targets. This pipeline represents a novel approach for predicting clinical drug response and generating novel candidates for drug repurposing and repositioning.

STRUCTURE-BASED SYSTEMS BIOLOGY: IDENTIFICATION OF OFF-TARGETS AND BIOLOGICAL PATHWAYS FOR HUMAN CHYMASE INHIBITORS

Mahreen Arooj, Keun Woo Lee

Objective: To better understand the mechanisms of chymase inhibitors and also for drug repurposing exercises to find novel uses for these inhibitors. Background: Off-target binding connotes the binding of a small molecule of therapeutic significance to a protein target in addition to the primary target for which it was proposed. Chymase is an enzyme of the hydrolase class that catalyzes the hydrolysis of peptide bonds. A link between heart failure and chymase has also been ascribed, and a chymase inhibitor is in clinical phase II for treatment of heart failure. However, the underlying mechanisms of the off-target effects of human chymase inhibitors are still unclear. Methodology: Putative off-targets for chymase inhibitors were identified through various structural and functional similarity analyses along with molecular docking studies. Finally, literature survey was carried out to incorporate these off-targets into various biological pathways and to establish links between pathways and particular adverse effects. Results and Conclusion: Off-targets of chymase inhibitors are linked to various biological pathways such as classical and lectin pathways of complement system, intrinsic and extrinsic pathways of coagulation cascade, and fibrinolytic system. Tissue kallikreins, granzyme M, neutrophil elastase, and mesotrypsin are also identified as off-targets. These off-targets and their associated pathways are elucidated for the affects of inflammation, cancer, hemorrhage, thrombosis and central nervous system diseases (Alzheimer's disease). Here, we have developed a robust computational strategy which is valuable for applications such as multi-target drug design and drug combinations.

AUTOMATED DISCOVERY OF FEATURES THAT DETERMINE SPECIFICITY AND AFFINITY OF PROTEIN-LIGAND INTERACTIONS ACROSS PROTEIN FAMILIES USING MACHINE LEARNING ANALYSIS OF STRUCTURES AND CHEMICAL INTERACTION PROFILES

Joseph Schoeniger, Peter Anderson

A fundamental issue in molecular recognition is how small molecules–from naturally occurring metabolites to exogenous drug compounds–discriminate between their protein targets and alternative protein structures and how protein receptors discriminate between small molecules. Ligand binding is largely driven by the three-dimensional arrangement of binding site structural features on both protein and ligand that participate in intermolecular interactions, rather than by amino acid sequence of the protein alone. Likewise, ligand-binding specificity cannot be understood by analyzing only ligand structure and chemistry. The ligand-protein interface is, however, a geometrically and chemically complex dynamic surface which is difficult to analyze. We present a new approach for identifying features of ligand-protein binding interfaces that predict binding selectivity and demonstrate its effectiveness for predicting kinase inhibitor specificity [1]. We analyzed a large set of human kinases and kinase inhibitors using clustering of experimentally determined inhibition constants (to define specificity classes of kinases and inhibitors) and virtual ligand docking (to extract structural and chemical features of the ligand-protein binding interfaces). We then used statistical methods to identify features characteristic of each class. Machine learning methods (random forests) were employed to determine which combinations of characteristic features were predictive of class membership and to predict binding specificities and affinities of new compounds. Experiments showed predictions were 70% accurate. These results show that our method can automatically pinpoint, on the three-dimensional binding interfaces, pharmacophore-like features that act as "selectivity filters". In recent work, we have also found that it is possible, using the interaction features harvested, to generate QSAR models and identify for each specificity class affinity-determining features that may function independent of the specificity-determining features for that class. These features are suitable for screening libraries of compound structures in order to enrich for ligands that, in principle, reject off-target binding and have tailored specificity and affinity. The method is not restricted to kinases, requires no prior hypotheses about specific interactions, and can be applied to any protein families for which sets of structures and ligand binding data are available. Examples will be given of preliminary analysis of viral protease inhibitor interactions.

[1] Anderson, P. C.; De Sapio, V.; Turner, K. B.; Elmer, S. P.; Roe, D. C.; Schoeniger, J. S. J. Med. Chem. 2012, 55, 1926.

PREDICTING DRUG SELECTIVITY FROM HOST-PATHOGEN DRUG NETWORKS

Geoffrey Henry Siwo, Roger Smith, Asako Tan, Michael T. Ferdig

Thousands of small molecules possess cytotoxic activity against infectious agents in vitro, but the successful development of therapeutic agents also demands that those molecules be safe for use in the human host. Host-pathogen drug selectivity remains a key barrier to the effective deployment of new therapeutic agents. Computational methods that predict selectivity could greatly enhance the pace of drug discovery. Here, we test the idea that relationships among drugs within and between the human host and an infectious agent can be used to predict selectivity of individual drugs and their combinations. To determine these drug relationships between the human host and the malaria parasite Plasmodium falciparum, we computed pairwise correlations among 10 drugs in both host cell lines and in parasite cultures using data from the connectivity map and our own transcriptional perturbation studies, respectively. Comparing drug pairs across the 2 species, we find strong positive correlations between a drug's set of relationships in the host and pathogen, demonstrating conservation in drug response patterns and mechanisms of action. For example, methotrexate's correlations to other drugs in the host is highly positively correlated to its relationships in the parasite ($r = 0.74$ compared to $r = 0.01$ in randomized data), an observation consistent with its high cytotoxicity in both species. However, we also find that some drugs have distinct patterns in the host and parasite, highlighting potential drug selectivity. We further explore the idea of applying pairwise drug correlations to predict selective drug combinations. The anti-malarial drug artemisinin is highly negatively correlated to chloramphenicol in the parasite ($r = -0.83$) compared to the human host ($r = -0.04$). We propose that differential drug correlations in host versus parasite could indicate the value of precisely deployed drug combinations against a pathogen at concentrations that are safe for host cells but cytotoxic to the pathogen (selective synergism). Our study demonstrates high conservation of transcriptional responses to drugs across 2 species and challenges the dogma that the P. falciparum transcriptome is hard-wired. In addition, we demonstrate the value of integrating data from infectious agents with the connectivity map of host transcriptional responses to facilitate the in silico discovery of selective chemical agents.

DETECTING AND CHARACTERIZING PLEIOTROPY: NEW METHODS FOR UNCOVERING THE CONNECTION BETWEEN THE COMPLEXITY OF GENOMIC ARCHITECTURE AND MULTIPLE PHENOTYPES

POSTER PRESENTATIONS

A SYSTEMS BIOLOGY APPROACH FOR UNDERSTANDING FLORAL TRANSITION IN PLANTS

Gitanjali Yadav

The 'Floral Transition' event in plants refers to the progression from vegetative to reproductive growth, and is known to be determined not by a single gene but by a highly complex gene network. In this network, the LEAFY (LFY) gene, a transcription factor with protein binding function that expresses widely in both vegetative and reproductive tissues, plays an important role as a switch that triggers flower formation by interacting and coordinating between several other genes. A critical level of LFY expression is essential for regulation of flowering-time and floral transition, and LFY plays a role in inflorescence and floral organ development as well. In this work, we investigate the process of floral transition at a genome-wide network level in Arabidopsis thaliana with LFY as the master regulator. In an effort to reduce data dimensionality, graph theoretical approach was used to superimpose interactomes, gene ontologies and expression profiles of the known floral genes, and this facilitated a reconstruction of the Arabidopsis floral network. The network enabled us to understand how LFY, situated at the center of the network, positively or negatively regulates the level or activities of other genes in the general physiology of flowering. A dynamic analysis of network topology led to the identification of relationships among genes in different modules and assisted in predicting regulators and downstream effectors of LFY. The study revealed that LFY directly affects the action of about 2000 other genes. Indirectly, it impacts about 18000 genes, covering over 70% of the Arabidopsis genome. Randomly selected other genes were unable to capture any significant portion of the genome. A low topological diameter ensures strong communication between the genes comprising the floral network, so that any perturbation to the system is likely to transmit very fast across all nodes. Since LFY stands at the very center of the network of flower development, even minute modifications in its expression/interactions may contribute to the appearance of floral structures in evolution. Floral meristems in Arabidopsis arise in continuous succession directly on the flanks of the inflorescence meristem and hence it is presumable that the pathways that regulate inflorescence and floral meristem identity would operate simultaneously and in close spatial proximity. Accordingly, we have identified several gene modules in the floral network distinct from floral development. Some of these 'high-confidence' genes are involved in unrelated plant processes such as Stress Reponses, Plant Immunity and Vegetative growth. It has been suggested earlier that defense responses are ancestral LFY roles, and our results confirm the same, apart from expanding the known floral network of Arabidopsis thaliana, providing new insights into the molecular basis of diverse floral morphologies.

GENERAL POSTER PRESENTATIONS

POLYGENIC RISK SCORE ANALYSIS IN CHILDHOOD ONSET SCHIZOPHRENIA STUDY

Kwangmi Ahn, Steven S An, Judith L. Rapoport

Childhood onset schizophrenia (COS) is a rare and severe form of the disorder, with more salient brain developmental and genetic abnormalities. Recent GWAS studies have provided strong support for a substantial common polygenic contribution to genetic susceptibility for schizophrenia. In this analysis, we examined the association between a direct measure of genetic risk of schizophrenia in126 COS probands, 95 their healthy siblings and 104 non - schizophrenia psychiatric patients (AD). Using data from the Psychiatric GWAS Consortia analysis of schizophrenia, we selected 135 SNPs in order to construct polygenic rick scores. Probans had higher generic risk score of schizophrenia than their health siblings ($P<0.0001$), but AD patients had similar scores to COS patients ($P=0.879$). While these initial results confirm the presence of a genetic risk for schizophrenia, we did not find evidence for the genetic risk differentiating schizophrenia from other psychiatric disorders.

LONG-RANGE CHROMATIN CONTACTS REVEAL A ROLE FOR THE PLURIPOTENCY AND POLYCOMB NETWORKS IN GENOME ORGANIZATION

Giancarlo Bonora*, Matthew Denholtz*, Constantinos Chronis, Erik Splinter, Wouter de Laat, Jason Ernst, Matteo Pellegrini, and Kathrin Plath

We mapped long-range chromatin interactions in mouse embryonic stem cells (ESCs), induced pluripotent stem cells (iPSCs), pre-iPSCs, and fibroblasts by means of 4C-seq, and uncovered an ESC-specific genome organization that is gradually re-established during reprogramming. Confirming previous results, we show that open/accessible versus closed chromatin character is the primary determinant of long-range chromatin interaction preferences. Importantly, we find that in ESCs, genomic regions extensively occupied by the pluripotency factors Oct4, Sox2, and Nanog preferentially co-localize. Similarly, regions strongly enriched for Polycomb-proteins and H3K27me3 frequently interact, and loss of the Polycomb-protein Eed diminishes these interactions without dramatically changing overall chromosome-conformation. These data reveal that transcriptional networks that govern ESC-identity play a role in determining genome-organization.

A SURVEY OF CODING VARIATION WITHIN 70 PHARMACOGENES: THE SPHINX DATABASE

William S. Bush, Jonathan Boston, Jay Cowan, Jacqueline Kirby, John Wallace, Josh Denny, Marylyn D. Ritchie, Jonathan L. Haines

The eMERGE Network is a national consortium combining DNA biorepositories with phenotype information extracted from the electronic medical record (EMR). With the unique resources available in eMERGE for pharmacogenomics discovery, the eMERGE-PGx project was initiated, applying the PGRN-Seq platform developed by the Pharmacogenomics Research Network (PGRN) to capture rare genetic variation within a set of 83 known pharmacogenes over ~ 10,000 individuals. Variants identified in this exploratory study will be made publicly available through an online database and portal called SPHINX. Through both a web interface and API, users can search identified variants by a variety of criteria, including basic attributes such as gene symbol, variant location, or RS number. More advanced searches are made possible using data from PharmGKB and a variety of other online databases. These resources allow searches by drug, and metabolic pathway, allowing higher-level hypotheses to be investigated. For eMERGE Network members, a restricted site is available to link this information to demographic data, ICD-9 and CPT codes, and medications mapped to RxNorm qualifiers. This new resource will provide investigators with detailed information for hypothesis generation in the field of pharmacogenomics.

THE STREAM ALGORITHM: COMPUTATIONALLY EFFICIENT RIDGE-REGRESSION VIA BAYESIAN MODEL AVERAGING, AND APPLICATIONS TO PHARMACOGENOMIC PREDICTION OF CANCER CELL LINE SENSITIVITY

Elias Chaibub Neto, In Sock Jang, Stephen H. Friend, Adam A. Margolin

Computational efficiency is important for learning algorithms operating in the "large p, small n" setting. In computational biology, the analysis of data sets containing tens of thousands of features ("large p"), but only a few hundred samples ("small n"), is nowadays routine, and regularized regression approaches such as ridge-regression, lasso, and elastic-net are popular choices. In this paper we propose a novel and highly efficient Bayesian inference method for fitting ridge-regression. Our method is fully analytical, and bypasses the need for expensive tuning parameter optimization, via cross-validation, by employing Bayesian model averaging over the grid of tuning parameters. Additional computational efficiency is achieved by adopting the singular value decomposition re-parametrization of the ridge-regression model, replacing computationally expensive inversions of large p-by-p matrices by efficient inversions of small and diagonal n-by-n matrices. We show in simulation studies and in the analysis of two large cancer cell line data panels that our algorithm achieves slightly better predictive performance than cross-validated ridge-regression while requiring only a fraction of the computation time. Furthermore, in comparisons based on the cell line data sets, our algorithm systematically out-performs the lasso in both predictive performance and computation time, and shows equivalent predictive performance, but considerably smaller computation time, than the elastic-net.

DETECTION OF UNKNOWN BACTERIAL GENOMES WITH CLIQUE LOG-LINEAR MODELS

Adrian Dobra, Camillo Valdes, Bertrand Clarke, Jennifer Clarke

We use a mutation-based technique on next generation sequencing (NGS) data to test for the presence of various bacterial strains in multiple metagenomic samples. Our technique is novel in the we use mutations at the nucleotide level to define sparse multidimensional contingency tables associated with the genomes in a given reference list. We focus on a specific class of hierarchical log-linear models which we call clique log-linear models. The models in this class have key desirable features that makes them quite suitable for the analysis of ultra high-dimensional categorical data. We devise efficient Markov chain Monte Carlo (MCMC) techniques for model determination. We apply our framework to data from the Human Microbiome Project (HMP).

GENESEER AIDS DRUG DISCOVERY BY EXPLORING EVOLUTIONARY RELATIONSHIPS BETWEEN GENES ACROSS GENOMES

Douglas D. Fenger, Matthew Shaw, Philip Cheung, Tim Tully

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

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

METHODS FOR INVESTIGATING THE PLEIOTROPIC EFFECTS OF MITOCHONDRIAL GENETIC VARIATION ON HUMAN HEALTH AND DISEASE

Sabrina Mitchell*, Jacob Hall*, Robert Goodloe, Jonathan Boston, Eric Farber-Eger, Sarah Pendergrass, William Bush, Dana Crawford (*co-first author)

Mitochondria play a critical role in the cell, and have DNA independent of the nuclear genome. There is evidence for a role of mitochondrial DNA (mtDNA) variation in human health and disease; however this area of investigation has lagged behind research into the role of the nuclear genetic variation on outcomes/traits. Phenome-wide association studies (PheWAS) investigate the association between a wide range of traits and genetic variation. To date, this approach has not been used to investigate the relationship between mtDNA variants and phenotypic variation. Herein, we describe the development of an analysis pipeline for PheWAS with mtDNA variants (mt-PheWAS). Using the Metabochip custom genotyping array, nuclear and mitochondrial DNA variants were genotyped in 11,519 African Americans from the Vanderbilt University biorepository, BioVU. We employed polygenic modeling to explore the relationship between mtDNA variants and a group of eight cardiovascular-related traits obtained from de-identified electronic medical records within BioVU. Results from polygenic analysis were prioritized for further analysis via single mtDNA variant tests of association. We found evidence for an effect of mtDNA variation on total cholesterol and type 2 diabetes, and identified an association between the mt16189 variant and type 2 diabetes in African Americans, an association previously reported in European-descent populations. Results from this initial study suggest that our analysis pipeline is a valid approach for investigating the relationship between mitochondrial genome variation and a range of phenotypes providing a framework for future mt-PheWAS.

MOLECULAR PREDICTORS OF RESIDUAL DISEASE AFTER CYTOREDUCTIVE SURGERY IN PATIENTS WITH HIGH-GRADE SEROUS OVARIAN CANCER

Shelley M. Herbrich, Susan L. Tucker, Kshipra Gharpure, Anna Unruh, Alpa M. Nick, Erin K. Crane, Robert L. Coleman, Charles W. Drescher, Sherry Wu, Gabriel Lopez-Berestein, Bulent Ozpolat, Christina Ivan, Keith A. Baggerly, Anil K. Sood

Background: Growing data suggest that high-grade serous ovarian cancer (HGSOC) patients most likely to benefit from upfront cytoreductive surgery are those with no gross residual disease (RD). Earlier attempts to find markers contrasting “optimal” and “suboptimal” surgical outcomes have not been reliable. We clarify the endpoint by contrasting no versus any RD. Methods: We used publicly available microarray datasets with clinical annotation for filtration, contrasting RD and No-RD cohorts in each and requiring consistent changes significant at a false discovery rate (FDR) of 10% in both. We selected a small number of filtered genes with a wide dynamic range (>16-fold) for further development. We assayed expression of these genes in new patient samples with qRT-PCR, with outcomes blinded to those performing the assays and those making predictions. All samples used in final validation were from the ovary itself, and were obtained prior to initiation of chemotherapy. We used a one-sided Fisher’s exact test to evaluate our performance. Results: We identified 47 probesets significant in both datasets at 10% FDRs. These included probesets for FABP4 and ADH1B, which tracked tightly and showed dynamic ranges >16-fold. Using the top quartile of FABP4 PCR values as a pre-specified cutoff, we found 30/35 RD cases in the high expression group, and 54/104 in the low group (p=0.0002). Conclusions: High FABP4 and ADH1B expression are associated with significantly higher risk of RD in HGSOC patients. Patients with high tumoral levels of FABP4 may be better candidates for neoadjuvant chemotherapy.

PHARMACOGENOMIC DISCOVERY WITH PHARMGKB PGXPLORE

Darla Hewett, Michelle Whirl-Carrillo, Julia Barbarino, Ryan Whaley, Mark Woon, Russ B. Altman and Teri E. Klein

We recognize that the pharmacogenomic, biomedical, and scientific research communities are undergoing a transformation where decision-making, including clinical decision-making, is based on information extracted from large-scale, and diverse data sets, including PharmGKB. Additionally data mining and knowledge visualization are central to all modes of scientific exploration and discovery, including hypothesis generation, hypothesis testing, hypothesis acceptance, and discovery integration. By giving scientists and clinicians the ability to quickly create their own data mining and visualization algorithms, without writing code, and the ability to include their own data sources, we have provided a new way for scientists and clinicians to extract, visualize, and contribute useful information in a timely fashion.

MODELING CELL SIGNALING NETWORK OF NEURONAL DIFFERENTIATION

Tsuyoshi Iwasaki, Ryo Takiguchi, Takumi Hiraiwa, Tadamasa Kimura, Akira Funahashi, Noriko Hiroi, Kazuto Yamazaki

After the establishment of human iPS cells, it is more significant to apply stem cells to medical treatment. However, it is difficult to control differentiation process of stem cells stably. Our research purpose is to predict the state changes of stem cells via simulation for helping to overcome these difficulties. We integrated binary relationships of genes or molecules from public knowledge to got whole signaling network about neuronal differentiation. As a result of making cascade of differentiation process on this network, it was suggested that neuronal differentiation process could be consist of almost four steps which include both positive and negative feedback. To validate these four steps process, we got experimental data about neural differentiation from Gene Expression Omnibus (GEO). The time-series microarray data of rat neural stem cell differentiation showed qualitative state changes that we assumed in the four steps. The up-down directions of gene expressions which were estimated from cascade of the process and feedback regulations were accord with the microarray data. To validate the process quantitatively, we built the four steps mathematical model by CellDesigner which is the one of the standard modeling tool of biochemical networks, and these equations were automatically given by SBML squeezer. We fitted parameters of the model to the microarray data with COPASI. The simulation result showed that the four steps process could express the state changes of some genes quantitatively. These components in the process might be important marker or regulator during neuronal differentiation. This model will be helpful in R&D of regenerative medicine, we suggest this process could be the basic framework to simulate neuronal differentiation.

PROTEIN INTERACTIONS AS DRUG TARGETS: A COMBINED COMPUTATIONAL AND EXPERIMENTAL APPROACH

Jouhyun Jeon, Joan Teyra, Satra Nim and Philip M. Kim

Protein interactions and their networks have been at the focus of recent biomedical science. In particular, there is growing interest in targeting protein interactions with future therapeutic agents. I will outline our efforts to combine advances in structure modeling, machine learning and state-of-the-art combinatorial chemistry to target protein interactions using synthetic peptides. Our integrated pipeline covers everything from the identification of particular protein interactions important in different cancer types to the validation of candidate compounds in vivo. Specifically, we first utilize modern machine learning techniques to integrate a large variety of cancer genomic datasets to identify suitable drug targets. We then use structural modeling to find the subset of targets that can be inhibited using peptides. Using peptide phage display, we obtain high-affinity binders that inhibit the protein-protein interaction in question. Finally, we use lenti-viral delivery to verify the efficacy of our peptides in cell lines. Thus far, we have obtained peptide binders for about a hundred high-value protein domains.

A KERNEL BASED L1-NORM REGULARIZED LOGISTIC REGRESSION METHOD TO PREDICT DRUG-TARGET INTERACTIONS

Shinhyuk Kim, Daeyong Jin, Hyunju Lee

Computational methods for predicting drug-target interactions have become one of the most important approaches in drug research because they help to reduce the time, cost, and failure rates for developing new drugs. Recently, with the accumulation of drug-related data sets such as drug's side effects and pharmacological data, it has became possible to predict drug-target interactions based on these data sets. In this study, we focus on drug-drug interactions (DDI), the adverse effects (DDIAE) and pharmacological information (DDIPharm), and investigate the relationship among chemical structures, side effects, and DDIs from several data sources. First, DDIPharm data from the STITCH database, DDIAE from drugs.com, and drug-target pairs from ChEMBL and SIDER were collected. By applying two machine learning approaches, a support vector machine (SVM) and a kernel based L1-norm regularized logistic regression (KL1LR), we showed that DDI is a promising feature in predicting drug-target interactions. Next, the accuracies of predicting drug-target interactions using DDI were compared to those obtained using the chemical structure and side effects based on SVM and KL1LR approaches, showing that DDI was the most contributing data source for predicting drug-target interactions. We also showed that KL1LR was comparable to SVM in predicting drug-target interactions.

INTEGRATED IMMUNOTHERAPY: IMPLICATIONS FOR PERSONALIZED CANCER TREATMENT

Deborah H Lundgren, Veneta Qendro, Karim Rezaul, Sun-Il Hwang, Zanna Aristarova, Ardian Latifi, and David K Han

ABSTRACT: Anti-tumor immunity was proposed as a novel concept over 50 years ago, but its effective utilization has been hampered by a series of challenges. One critical challenge to its success is the identification of tumor-specific antigen targets capable of eliciting an effective immune response. To address this issue we have devised a multi-step strategy integrating genomics, proteomics and immunology. As a proof of principle, we are applying this strategy to breast cancer analysis. First, exome sequencing is performed on a patient's tumor samples, focusing on a limited set of known or putative breast cancer driver genes in order to increase read depth and sensitivity. Second, based on mutations identified through exome sequencing, a searchable mutant protein database is created. Third, the patient's tumor samples are analyzed via shotgun proteomics to identify which genetic mutations are actually translated into mutant proteins in sufficient quantities to be detected by high-throughput mass spectrometry. Fourth, mutant peptides predicted by exome sequencing and verified by proteomic analysis are spotted in wild type/mutant pairs on a peptide chip. Fifth, the chip is then incubated with patient serum, and a comparison of wild type and mutant spot intensities reveals which mutant peptides exhibit a strong affinity to antibodies in the serum. Based on these results, multiple patient-specific immunogenic mutant peptides can be experimentally selected as targets for personalized cancer vaccines and therapeutic monoclonal antibodies.

PREDICTING COMBINATION THERAPIES USING DIVRANK ON THE CONNECTIVITY MAP DATA

J. Matthew Mahoney, Anna L. Tyler

The Connectivity Map (CMap) gene expression database contains gene expression response profiles of human cells in response to perturbations from thousands of FDA approved compounds. CMap allows for in silico prediction of drugs that target disease-derived gene expression signatures. The CMap user supplies a pair of gene lists, one up- and the other down-regulated, and CMap converts this query into a ranked list of drugs with high-ranking compounds likely to modify the disease gene expression and possibly provide therapeutic benefit to a patient. The CMap rankings, however, suffer from the high correlation between compounds with similar indications. For example, anti-inflammatory medications induce simliar gene expression profiles indicating their similar function. Thus, similar compounds populate the top of the list and compete for attention from the user. Furthermore, the typical CMap user is working on a disease for which there is only a weak (or nonexistent) standard-of-care treatment. These situations likely arise from the heterogeneity of the underlying molecular biology of the disease relative to existing therapies. Thus, we require a ranking system that can automatically detect promising combination therapies by giving high rank to diverse compounds that have distinct actions and cover as much of the target gene expression profile as possible. Here we discuss the use of the DivRank algorithm for generating such diverse rankings. DivRank is a random walk based ranking method that achieves diversity through a competitive process by which nodes compete locally for high scores. Thus, high scores are not smoothed across the neighborhoods of highly central nodes as in PageRank, but rather concentrated at distinct hubs. The CMap data yields a natural, bipartite gene-drug network. We propose using DivRank on a drug-drug projection of this network which incorporates the user's query signature. The interface for the user is identical to that of CMap, but the output should indicate a plausible combination therapy that gives maximal coverage of the user's signature with minimal redundancy of drug action.

AN INTEGRATED APPROACH TO BLOOD-BASED CANCER DIAGNOSIS AND BIOMARKER DISCOVERY

Renqiang Min, Salim Chowdhury, Yanjun Qi, Alex Stewart, Rachel Ostroff

Disrupted or abnormal biological processes responsible for cancers often quantitatively manifest as disrupted additive and multiplicative interactions of gene/protein expressions correlating with cancer progression. However, the examination of all possible combinatorial interactions between gene features in most case-control studies with limited training data is computationally infeasible. In this paper, we propose a practically feasible data integration approach, QUIRE, to identify discriminative complex interactions among informative gene features for cancer diagnosis and biomarker discovery directly based on patient blood samples. QUIRE works in two stages, where it first identifies functionally relevant gene groups for the disease with the help of gene functional annotations and available physical protein interactions, then it explores the combinatorial relationships among the genes from the selected informative groups. Based on our private experimentally generated data from patient blood samples using a novel SOMAmer (Slow Off-rate Modified Aptamer) technology, we apply QUIRE to cancer diagnosis and biomarker discovery for Renal Cell Carcinoma (RCC) and Ovarian Cancer (OVC). To further demonstrate the general applicability of our approach, we also apply QUIRE to a publicly available Colorectal Cancer (CRC) dataset that can be used to prioritize our SOMAmer design. Our experimental results show that QUIRE identifies gene-gene interactions that can better identify the different cancer stages of samples, as compared to other state-of-the-art feature selection methods. A literature survey shows that many of the interactions identified by QUIRE play important roles in the development of cancer.

METHODS FOR INVESTIGATING THE PLEIOTROPIC EFFECTS OF MITOCHONDRIAL GENETIC VARIATION ON HUMAN HEALTH AND DISEASE

Sabrina Mitchell*, Jacob Hall*, Robert Goodloe, Jonathan Boston, Eric Farber-Eger, Sarah Pendergrass, William Bush, Dana Crawford (*co-first author)

Mitochondria play a critical role in the cell, and have DNA independent of the nuclear genome. There is evidence for a role of mitochondrial DNA (mtDNA) variation in human health and disease; however this area of investigation has lagged behind research into the role of the nuclear genetic variation on outcomes/traits. Phenome-wide association studies (PheWAS) investigate the association between a wide range of traits and genetic variation. To date, this approach has not been used to investigate the relationship between mtDNA variants and phenotypic variation. Herein, we describe the development of an analysis pipeline for PheWAS with mtDNA variants (mt-PheWAS). Using the Metabochip custom genotyping array, nuclear and mitochondrial DNA variants were genotyped in 11,519 African Americans from the Vanderbilt University biorepository, BioVU. We employed polygenic modeling to explore the relationship between mtDNA variants and a group of eight cardiovascular-related traits obtained from de-identified electronic medical records within BioVU. Results from polygenic analysis were prioritized for further analysis via single mtDNA variant tests of association. We found evidence for an effect of mtDNA variation on total cholesterol and type 2 diabetes, and identified an association between the mt16189 variant and type 2 diabetes in African Americans, an association previously reported in European-descent populations. Results from this initial study suggest that our analysis pipeline is a valid approach for investigating the relationship between mitochondrial genome variation and a range of phenotypes providing a framework for future mt-PheWAS.

KALEIDOSCOPIC EVOLUTION OF C2H2 ZINC FINGER DNA BINDING

Hamed S. Najafabadi, Sanie Mnaimneh, Frank W. Schmitges, Kathy N. Lam, Ally Yang, Mihai Albu, Matthew T. Weirauch, Ernest Radovani, Jack Greenblatt, Brendan J. Frey, and Timothy R. Hughes

The largest and most diverse class of eukaryotic transcription factors contain Cys2-His2 zinc fingers (C2H2-ZFs), each of which typically binds a DNA nucleotide triplet within a larger binding site. Frequent recombination and diversification of their DNA-contacting residues suggests that C2H2-ZFs play a prevalent role in adaptive evolution. Very little is known about the function and evolution of the vast majority of C2H2-ZFs, including whether they even bind DNA. We determined DNA-binding motifs for thousands of individual natural C2H2-ZFs, correlated them with the specificity residues, and examined the evolution of their sequence preferences. We conclude that most extant C2H2-ZFs are DNA-binding and many have evolved from ancestors with completely different motifs, including recent and much more ancient events. Nonetheless, C2H2-ZFs that recognize most DNA triplets can be traced to the last common ancestor of eutherians. A popular belief is that C2H2-ZFs silence endogenous retroviral elements; we found that human ZNF528 primarily binds L1 subclasses, but for most C2H2-ZFs, binding data in human cells predominantly support other roles. These data show that animals contain a rapidly evolving and largely unstudied adaptive C2H2-ZF regulatory network, in parallel to well-characterized and highly conserved basic developmental programs.

LOCATING HYDROGEN ATOMS, THE "DARK MATTER" IN PROTEINS

Ho Leung Ng, Matthew Bronstad, Jinny Ching

Hydrogen constitutes half of the atoms in proteins and plays critical roles in enzyme function and receptor binding. The roles of the hydrogen atoms are incompletely understood as few experimental methods can precisely determine their precise locations. We present results using a new computational approach called HyPO (Hydrogen Prediction and Observation) on X-ray and neutron diffraction data to identify and locate hydrogen atoms in protein structures. Traditionally, crystallographers determine whether an atom is present in a crystal structure by visual comparison with a contoured map: can one see the atom in the map? HyPO takes a fundamentally different, quantitative approach: What arrangement of atoms is most probable given a map? HyPO is sensitive at resolution ranges up to 1.6 angstroms, which includes 12% of all protein structures in the PDB. In a data set of structures up to 1.0 angstroms, HyPO detects electron density peaks at rotameric positions for 60% of 428 alanine residues, consistent with hydrogen atoms. We demonstrate validation of HyPO results against gold standard neutron crystallography data.

MACHINE READING FOR CANCER PANOMICS

Hoifung Poon, Tony Gitter, Chris Quirk

Advances in sequencing technology have made available a plethora of panomics data for cancer research, yet it remains an outstanding challenge to integrate such heterogeneous data for identifying disease genes and drug targets. Biological knowledge such as pathways can be used to construct powerful probabilistic models (e.g., graphical models) for effective integration of panomics data. The majority of knowledge resides in text (e.g., journal publications), which has been undergoing its own exponential growth, making it mandatory to develop machine reading methods for automatic knowledge extraction. In this presentation, we will review progress and challenges in machine reading for pathway extraction, and analyze the impact of extracted pathways on a pathway-based graphical model for cancer driver inference.

ANALYSIS OF RNA-BINDING PROTEIN DYNAMICS WITH ELASTIC NETWORK MODELS

Ann Quigley, Michael Terribilini

RNA binding proteins play crucial roles in many cellular processes especially gene expression. The induced fit model has been used to describe the conformational change undergone by many proteins upon binding to RNA. X-ray crystallographic structures illustrate the conformational adjustments that proteins undergo in the process of binding, with transitions in motion inferred in the progression of the unbound state of the protein to the bound state in complex with the partner molecule. Though crystallographic data can be used to understand the apex of motion, it cannot examine the dynamic changes that are induced upon binding. Computational models that correspond to the real-time motion of the protein allow for exploration of the dynamic motion of binding sites of specific proteins instead of relying on static representations of the molecule. The elastic network model (ENM) is one such computational dynamic model that uses a simplified representation of the protein structure and intramolecular interactions to estimate large-scale motions of a protein. Previous work with ENMs has demonstrated the agreement between the model and experimental data on protein dynamics and has been effective in analyzing protein-protein interactions. The dynamic motion of RNA binding proteins was explored to observe whether or not binding sites exhibited consistent and distinct behaviors from nonbinding sites present on the same protein. Dynamics data was simulated with an ENM of known RNA-protein complexes and the motion of binding and non-binding residues was analyzed. This study used two datasets of RNA-binding proteins; first a small paired set where both the bound and unbound structures of the protein were available, and second, a larger set where only the unbound structure was available. The paired dataset was used to determine the correlation between the ENM data when starting with the unbound conformation compared to the bound conformation because the ENM is known to be influenced by the starting conformation. However, the data indicated a very high correlation indicating the ability to start with either the bound or unbound conformation of the protein. Using the larger dataset of proteins in complex with RNA, it was determined that RNA binding residues are more dynamic than nonbinding surface residues. This observed trend was applicable regardless of the amino acid type involved. These findings suggest that the intrinsic dynamic motion of the protein may play a pivotal role in protein-RNA recognition.

COMPARISON OF RNA AND DNA SINGLE NUCLEOTIDE AND INDEL VARIANTS IN THE NCI-60 CELL LINE COLLECTION

Onur Sakarya, Jeremy Ku, Kunbin Qu, Thon de Boer, Bill Gibb, Kevin Kwei, Jennie Jeong, Mei-Lan Liu, Robert Pelham, Sam Levy, Ellen Beasley

RNA sequencing (RNA-seq) can be employed to measure gene expression to detect fusion transcripts and non-coding RNAs. RNA-Seq can also be used to detect nucleotide-scale variations in expressed transcripts. However, compared to DNA, RNA variant detection has additional technical challenges. These challenges include uneven read coverage introduced by gene and allele specific expression, ambiguous read mapping due to alternative splicing and artifacts introduced by reverse transcriptase. In this study, we addressed to what extent sequence variation can be accurately detected from RNA molecules. We sequenced both ribosomal RNA depleted total RNA and exome enriched DNA from the NCI-60 cancer cell lines using paired 100 base pair reads on the Illumina HiSeq 2000. We mapped the reads from RNA and DNA using GSNAP and BWA respectively. Sequence variants on both datasets were detected with FreeBayes and GATK and passed through a filtration cascade to remove low quality variants. We built a "consensus" variant set from DNA and RNA variants that were identified by both FreeBayes and GATK regardless on whether the call was made from DNA or RNA. The precision and sensitivity of variants detected in RNA was determined by comparing against the consensus calls. On average, we detected 11,092 consensus variants on sites that have sufficient coverage from both DNA and RNA. This represents approximately 25% of exonic variants that would be detected by DNA alone. RNA data alone identifies 10,237 variants while DNA alone finds 10,494, of which 9,763 (precision: 95%, sensitivity: 88%) and 10,085 (precision: 96%, sensitivity: 91%), respectively, overlap with the consensus calls. Through a manual investigation of alignments, half of the RNA only variants were attributed to bioinformatics technical artifacts such as mapping differences between GMAP and BWA, low frequency of alternative alleles in DNA for heterozygous sites and variants at the 5' edge of the reads. For the remaining half (estimated 237 coding variants per cell line), variant calls appeared valid. We continue to investigate systematic ways to filter out false positive variants and the possibility of RNA-editing versus reverse-transcription substitutions. Although just one fourth of exonic DNA variant sites were expressed in RNA, we would expect these to be enriched for changes that influence phenotype the most. This result is compatible with a recent study in GM12878 cells that identified 33.4% of exome variants on coding sites by RNA-Seq (Piskol et al., AJHG, 2013). Within the expressed sites, our RNA variant calling was reliable with an average 95% precision and 88% sensitivity. To the best of our knowledge, these results comprise a first catalogue of RNA mutations for the NCI-60 cell line collection.

BAYESIAN NETWORK RECONSTRUCTION USING SYSTEMS GENETICS DATA: COMPARISON OF MCMC METHODS

Shinya Tasaki, Ben Sauerwine, Bruce Hoff, Hiroyoshi Toyoshiba, Chris Gaiteri, Elias Chaibub Neto

Reconstruction of biological networks using high-throughput technology has the potential to produce condition-specific interactomes. But are these reconstructed networks a reliable source of biological interactions? Do some network inference methods offer dramatically improved performance on certain types of networks? Here, we report a large-scale simulation study to compare the performance of Markov chain Monte Carlo (MCMC) samplers for reverse engineering of Bayesian networks. MCMC samplers investigated include traditional and state of the art Metropolis-Hastings and Gibbs sampling approaches, as well as novel samplers we have designed. By designing our simulation as a multi-factorial experiment with crossed factors we can determine the effect of any simulation parameter on the relative performance of any two methods. In this study, simulation parameters, such as sample size, network size, average edge density, network topology, amount of signal, amount of noise, and integration of genetics and gene expression data, play the role of factors, and the difference in area under the curve between different samplers play the role of a response variable. Our simulations reveal that network size, edge density, and amount of gene-to-gene signal are major parameters that define the performance of samplers and each sampler shows distinct behaviors to the parameters. Specifically, state of the art samplers outperform traditional samplers for highly interconnected large networks with strong gene-to-gene signal, which is one of the most difficult tasks. Our newly developed samplers show comparable or superior performance over the existing state of the art samplers for this task. Furthermore, this performance gain is strongest in networks with biologically-oriented topology, which indicates that our samplers have suitable characteristics for biological networks. Examination of the performance of MCMC samplers can help guide our choice of the more appropriate method for network reconstruction using systems genetics data.

CARDIAC ENHANCERS HARBOR UNDISCOVERED GENETIC VARIANTS ASSOCIATED WITH HEART CONTRACTION TRAITS

Xinchen Wang, Manolis Kellis, Laurie Boyer

Genome wide association studies (GWAS) have emerged as a powerful approach for identifying single nucleotide polymorphisms (SNPs) that mark complex trait loci. Most significant SNPs fall outside protein-coding regions, however, making it difficult to understand disease mechanism, and current GWAS hits collectively often explain only a small proportion of the total heritability for a disease. We hypothesized that SNPs within GWAS loci alter the activity of enhancer elements and result in changes in the expression of nearby genes. We used genome-wide enhancer maps of 127 human tissues and demonstrate that GWAS loci associated with ventricular contractility traits are specifically enriched within enhancers that are active in both fetal and adult hearts. Our data indicate that lead SNPs reported by GWAS are often unlikely to be the putative causal variants. Further analysis revealed that these enhancers share similar sequence properties including transcription factor motifs that may reflect common regulatory mechanisms. We demonstrate that SNPs within these loci likely disrupt transcription factor binding and alter expression of nearby genes. Finally, we show that enhancer characteristics learned from GWAS loci can be used to identify many additional, nominally significant loci that do not reach genome-wide significance, suggesting that enhancer annotations can facilitate the discovery of trait-associated loci. Collectively, our work demonstrates that GWAS loci can affect the activity of tissue-specific enhancers, and support the relevance of epigenomics in understanding the mechanistic underpinnings of complex disease.

ANCESTRY INFORMATIVE MARKERS FOR NATIVE HAWAIIANS

Hansong Wang, Laurence N. Kolonel, Loic Le Marchand

Admixture mapping is a powerful approach that specifically examines admixed populations to find disease susceptibility loci that differ both in gene frequencies and in disease prevalence between ancestral populations. Self-reported Native Hawaiians are admixed with Hawaiian/Polynesian, East Asian and European ancestries. There is considerable interest in applying this method in the study of obesity, tobacco-derived biomarkers in relation to lung cancer and breast cancer in women in Native Hawaiians. From one million genotyped markers in Native Hawaiians, we identified Native Hawaiians with almost 100% Hawaiian ancestry with principal component analysis. With reference to the 1000 Genomes Project sequencing data of Europeans and East Asians, multiple panels of Ancestry Informative Markers (AMIs) of different sizes was derived by 1) maximizing the informativeness measure In (Rosenberg et al. 2003) across the ancestral populations (East Asians, Europeans and 100% Hawaiians), or 2) choosing from markers with small frequency differences between Europeans and East Asians and large frequency differences between Hawaiian and East Asian/Europeans. We evaluated the performance of these AIMs panels in admixture mapping and in controlling for population structure in association studies in Native Hawaiians with simulation. We also applied admixture mapping to search for loci associated with BMI, breast cancer survival, and a few tobacco derived carcinogens.

Wittig

In solution HLA capture and high-resolution NGS-based typing method and an automated, integrated analysis framework

Michael Wittig, Jarl Andreas Anmarkrud, Michael Forster, Eva Ellinghaus, Kristian Holm, Lars Wienbrandt, Sascha Sauer, Manfred Schimmler, Malte Ziemann, Siegfried Görg, Tom Hemming Karlsen, Andre Franke

The human leukocyte antigen (HLA) locus contains the most polymorphic genes in the human genome. These genes play an important role in immune response and much is already known about their role in autoimmunity and infectious disease. The classical characterization of these genes is based on Sanger- or next-generation sequencing (NGS) of a limited amplicon repertoire or labeled oligonucleotides, which identify allele-specific sequences. Using these traditional methods, the rate of possible ambiguities is high and requires manual evaluation of the results, which is also an error-prone process. Here, we introduce a highly automated method, which employs comprehensive in- solution targeted capturing of the complete classical class I and class II loci in combination with NGS. Our implemented fully automated analysis allowed for the accurate characterization of HLA-A (0.99 allele calling rate), HLA-B (0.99), HLA-C (0.99), HLA-DRB1 (0.98), HLA-DQA1 (0.99), HLA-DQB1 (0.99), HLA-DPA1 (0.98) and HLA-DPB1 (0.96). Including possible ambiguities and manual verification allowed for the exact HLA typing of all our reference samples. The reference sample set comprises 261 samples so far and were derived from the International Histocompatibility Working Group (IHWG) biobank and from another German center. The allelic diversity of the reference sample was maximized before enrichment and NGS. For HLA-A we identified 66 different alleles, for HLA-B 106, HLA-C 49, HLA-DRB1 71, HLA-DQA1 20, HLA-DQB1 17, HLA-DPA1 5 and HLA-DPB1 39, respectively. In summary, our method provides a straight-forward workflow, which is mainly due to the use of in-solution targeted capturing rather than traditional amplicon-based tecniques. The fully automated allele calling delivers high confident allele calls and the number of possible ambiguities is drastically reduced compared to traditional typing (e.g. class I with on avg. 2.5 possible alternatives per sample). At this very early stage of development, one technician can characterize 182 samples in one week with high confidence and high resolution (6-8 digits).

A FREQUENT INACTIVATING MUTATION IN RHOA GTPASE IN ANGIOIMMUNOBLASTIC T-CELL LYMPHOMA

Hae Yong Yoo, Min Kyung Sung, Seung Ho Lee, Sangok Kim, Haeseung Lee, Seongjin Park, Sang Cheol Kim, Byungwook Lee, Kyoohyoung Rho, Jong-Eun Lee, Kwang-Hwi Cho, Wankyu Kim, Hyunjung Ju, Jaesang Kim, Seok Jin Kim, Won Seog Kim, Sanghyuk Lee, and Young Hyeh Ko

The molecular mechanisms underlying angioimmunoblastic T-cell lymphoma (AITL), a common type of mature T-cell lymphoma of poor prognosis, are largely unknown. Here, we report a frequent somatic mutation in RHOA (G17V) using exome and transcriptome sequencing of samples from AITL patients. Further examination of the RHOA G17V mutation in 239 lymphoma samples revealed that the mutation was specific to T-cell lymphoma but absent in B-cell lymphoma. We demonstrate that RHOA G17V mutation, which was found in 46% (13 out of 28 patients) of the AITL cases examined, is oncogenic in nature using multiple molecular assays. Molecular modeling and docking simulations provided a structural basis for the loss of GTPase activity in the RHOA G17V mutant. Our experimental data and modeling results suggest that RHOA G17V is a driver mutation in AITL and that RHOA GTPase activity is disrupted by the missense mutation. Based on these data and through integrated pathway analysis, we build a comprehensive signaling network for AITL oncogenesis.

TRIBE: THE COLLABORATIVE PLATFORM FOR MINING GENOMIC DATA IN BIOLOGY

Rene A. Zelaya, Casey S. Greene

Tribe is a biological knowledge storage system that allows users to connect literature and experimental results to data mining webservers and services. Gene identifiers are mapped to a common format, and selected genes are maintained over time as gene symbols change. Tribe implements version control for user collections providing provenance for each change to promote reproducible science. Tribe allows collaborators to jointly access and build collections, and to use those collections to mine large scale data compendia through Tribe-enabled webservers. Our Tribe server provides an open API allowing other webservers and analytical platforms to interact with user's Tribe resources. To connect with Tribe, developers register for an API key, which requires only an e-mail address and client name. Collections from both users and curated resources are made available as JSON objects, a format for which native parsers are available for many programming languages. Servers and analytical platforms can interact with users' private collections by authenticating users through the standard OAuth2 protocol. For developers, we provide an open source demonstration client server capable of authenticating users and using private collections. Tribe provides a platform that new and existing webservers can leverage to provide a community of users with advanced bioinformatics capabilities while preserving analytical reproducibility.

PERSONALIZED MEDICINE: FROM GENOTYPES AND MOLECULAR PHENOTYPES TOWARDS THERAPY

POSTER PRESENTATIONS

ACCEPTED PROCEEDINGS PAPER WITH POSTER PRESENTATION.

BAGS OF WORDS MODELS OF EPITOPE SETS: HIV VIRAL LOAD REGRESSION WITH COUNTING GRIDS

Alessandro Perina, Pietro Lovato, Nebojsa Jojic

The immune system gathers evidence of the execution of various molecular processes, both foreign and the cells' own, as time- and space-varying sets of epitopes, small linear or conformational segments of the proteins involved in these processes. Epitopes do not have any obvious ordering in this scheme: The immune system simply sees these epitope sets as disordered "bags" of simple signatures based on whose contents the actions need to be decided. The immense landscape of possible bags of epitopes is shaped by the cellular pathways in various cells, as well as the characteristics of the internal sampling process that chooses and brings epitopes to cellular surface. As a consequence, upon the infection by the same pathogen, different individuals' cells present very different epitope sets. Modeling this landscape should thus be a key step in computational immunology. We show that among possible bag-of-words models, the counting grid is most fit for modeling cellular presentation. We describe each patient by a bag-of-peptides they are likely to present on the cellular surface. In regression tests, we found that compared to the state-of-the-art, counting grids explain more than twice as much of the log viral load variance in these patients. This is potentially a significant advancement in the field, given that a large part of the log viral load variance also depends on the infecting HIV strain, and that HIV polymorphisms themselves are known to strongly associate with HLA types, both effects beyond what is modeled here.

FINDINGS FROM THE THIRD CRITICAL ASSESSMENT OF GENOME INTERPRETATION, CAGI 2013, A COMMUNITY EXPERIMENT TO EVALUATE PHENOTYPE PREDICTION

Steven E. Brenner, John Moult, CAGI Participants

The Critical Assessment of Genome Interpretation (CAGI, 'kā-jē) is a community experiment to objectively assess computational methods for predicting the phenotypic impacts of genomic variation. In the experiment, participants are provided genetic variants and make predictions of resulting phenotype. These predictions are evaluated against experimental characterizations by independent assessors. A long-term goal for CAGI is to improve the accuracy of phenotype and disease predictions in clinical settings.

The third CAGI experiment (concluded in July 2013) consisted of ten diverse challenges. CAGI deliberately extends challenges from previous years, with the continuity allowing measurement of progress. For example, in the second CAGI, in a challenge to predict Crohn's disease from exomes, one group was able to identify 80% of affected individuals before the first false positive healthy person. In the third CAGI experiment, this challenge used an improved dataset, and several groups performed remarkably well, with one group achieving a ROC AUC of 0.94. The experiment also revealed important population structure to Crohn's disease in Germany.

For three years, CAGI has posed a challenge with Personal Genome Project (PGP) genome data. This year, two groups were able to successfully map a significant number of complete genomes to their corresponding trait profiles submitted by PGP participants. In the expanded challenge to predict benign versus deleterious variants in DNA double-strand break repair MRN genes—Rad50 (from last year), Mre11, and Nbs1—as determined by those that appear in a breast cancer case versus healthy control, predictions show how methods differ sharply in their effectiveness even amongst proteins in the same complex.

A new challenge this year was to use exomes from families with lipid metabolism disorders. In the case of hypoalphalipoproteinemia (HA), a company made predictions which showed how understanding the problem structure and employing an extensive knowledgebase led to remarkably good results. Another related challenge revealed a twist wherein real-world data differed sharply from theoretical models.

The other challenges were to predict which variants of BRCA1 and BRCA2 are associated with increased risk of breast cancer; to predict how variants in p53 gene exons affect mRNA splicing; to predict how well variants of a p16 tumor suppressor protein inhibit cell proliferation; and to identify potential causative SNPs in disease-associated loci.

Overall, CAGI revealed that the phenotype prediction methods embody a rich and diverse representation of biological knowledge, and they are able to make predictions that are highly statistically significant. However, we also found the accuracy of prediction on the phenotypic impact of any specific variant was unsatisfactory and of questionable clinical utility. The most effective predictions came from methods honed to the precise challenge, including the specific genes of interest as well as the problem context. Prediction methods are clearly growing in sophistication, yet there are extensive opportunities for further progress. Complete information about CAGI may be found at http://genomeinterpretation.org.

THE EFFECTS OF ELECTRONIC MEDICAL RECORD PHENOTYPING DETAILS ON GENETIC ASSOCIATION STUDIES: HDL-C AS A CASE STUDY

Logan Dumitrescu, Robert Goodloe, Eric Farber-Eger, Jonathan Boston, Dana C. Crawford

Biorepositories linked to de-identified electronic medical records (EMRs) have the potential to complement traditional epidemiologic studies in genotype-phenotype studies of complex human diseases and traits. A major challenge in meeting this potential is the use of EMR-derived data to extract phenotypes and covariates for genetic association studies. Unlike traditional epidemiologic data, EMR-derived data are collected for clinical care and are therefore highly variable across patients. The variability of clinical data coupled with the challenges associated with searching unstructured clinical notes requires the development of algorithms to extract phenotypes for analysis. Given the number of possible algorithms that could be developed for any one EMR-derived phenotype, we explored here the impact algorithm decision logic has on genetic associations study results for a single quantitative trait, high density lipoprotein cholesterol (HDL-C). We used five different algorithms to extract HDL-C from African American subjects genotyped on the Illumina Metabochip (n=11,519) as part of Epidemiologic Architecture for Genes Linked to Environment (EAGLE). Tests of association between HDL-C and genetic risk scores for HDL-C associated variants suggest that the genetic effect size does not vary substantially across the five HDL-C definitions. These data collectively suggest that, at least for this quantitative trait, algorithm decision logic and phenotyping details do not appreciably impact genetics association study tests statistics.

CLINVITAE: A FREELY AVAILABLE DATABASE OF CLINICALLY OBSERVED VARIANTS

Reece K. Hart, Bruce Blyth, Tim Chu, John Garcia

The frequent discovery of variants of uncertain significance (VUS) in sequencing projects creates a time-consuming analysis burden for clinical geneticists and genetic counselors. Many practitioners have recognized that a comprehensive, easy-to-use shared resource of clinical variants would facilitate interpretation and improve consistency. While ClinVar is a compelling resource for submitted clinical variants, numerous additional sources of clinically observed variants are not represented there. CLINVITAE aims to aggregate these unrepresented sources and provide an easy-to-use web interface for searching variants and comparing reported classifications. CLINVITAE is free to use and download.

For the first iteration of CLIVITAE, we aggregated variants from ClinVar (http://www.ncbi.nlm.nih.gov/clinvar/), Emory Genetics Laboratory Variant Classification Catalog (http://genetics.emory.edu/egl/emvclass/emvclass.php), Carver Mutation Database (https://www.carverlab.org/database) and the Kathleen Cunningham Foundation Consortium for Research into Familial Breast Cancer (http://www.kconfab.org/index.shtml). We also used the ARUP Mutation Databases (http://www.arup.utah.edu/database/index.php), the variants of which were deposited into ClinVar during the course of our project and are currently listed as discrete entries. Each record within CLINVITAE has two components: the verbatim source record, and a reconciled record that facilitates comparisons across sources. For example, because different data sources used different terminology to describe their classification (e.g., variant of uncertain significance, variant of unknown significance, VOUS, etc.), each discrete classification was mapped to one of the five ACMG variant classification categories when such mapping was reasonably clear. Similarly, we also performed “liftover” from the reported transcript to a canonical transcript using the genomic coordinate system as a reference. As of November 2013, CLINVITAE contains 79,907 variants in 9,189 genes.

The web interface supports searching and filtering by gene, transcript, source lab, and inferred classification, allowing a user to see only the variants relevant to their investigation. Queries may be bookmarked for sharing or repeated analyses. All entries have a link to the original source for attribution and show both the original variant description and remapped variant to enable manual verification. The entire database, and the active query, can be exported as a tab-separated value (TSV) file.

As part of InVitae's commitment to the Free the Data campaign (http://free-the-data.org/), we have developed a tool that will facilitate the sharing of clinical variants and their pathogenicity, thereby catalyzing the community's efforts to reach consensus about variant interpretation. We expect that increased sharing of genomic data, with appropriate consent, is in the best interest of patients. We expect tools like CLINVITAE to play important roles in classifying variants of uncertain significance and lessening the analysis burden for clinicians. The most recent version of our database can be found at http://clinvitae.invitae.com.

TRANSCRIPTOMIC ANALYSIS OF BENIGN AND MALIGNANT THYROID NODULES

Katayoon Kasaian, Karen L. Mungall, Jacquie Schein, Yongjun Zhao, Richard A. Moore, Martin Hirst, Marco A. Marra, Blair A. Walker, Sam M. Wiseman, Steven J.M. Jones

Thyroid cancer is the fastest growing cancer with an estimated 4% to 7% of the population developing a clinically significant thyroid nodule during their lifetime. In up to a quarter of cases, preoperative diagnosis by needle biopsy is inconclusive and so a large proportion of individuals undergo thyroidectomy as a diagnostic procedure for cancer. Limited understanding of the molecular mechanisms driving thyroid tumorigenesis and progression has hindered the identification of efficient diagnostics and administration of targeted treatments. We have undertaken transcriptomic analysis of thyroid malignant and benign samples using massively parallel sequencing technologies in order to characterize the molecular changes underlying these phenotypes.

Whole transcriptome sequencing of 9 papillary thyroid carcinomas and 15 benign thyroid nodules were performed using Illumina HiSeq2000 technologies. Sequence reads were aligned to the hg19 human reference genome using TopHat and de novo assembled using Trans-ABySS. Unsupervised clustering of the normalized sequence read counts identified distinct sub-groups in both the malignant and benign cohorts, demonstrating the heterogeneity of these tumors. Differential gene expression and exon usage analyses enabled the identification of distinct pathways and processes that are altered in these subgroups. Those differentially expressed exons coding for functionally vital protein domains can potentially serve as the needed diagnostics in the clinic.

Thyroid tumors have a low mutation rate; the only expressed mutations identified in the transcriptome data included V600E BRAF mutation in the malignant tumors and RAS hotspot mutations in a subset of benign nodules. Similarly, there were no recurrent gene fusions detected through de novo assembly of the data. However, the analysis of the assembled contigs has led to the identification of novel splice isoforms which have never been defined before. These can not only be used as diagnostic markers but also as therapeutic targets. One novel splicing event, coding for a cell surface protein, shows a deletion of 30 amino acids from the extracellular region of the protein. Such an event can be used for designing specific antibody-drug conjugates, targeting the malignant cells while sparing the normal healthy cells.

AN EMPIRICAL BAYESIAN FRAMEWORK FOR SOMATIC MUTATION DETECTION FROM CANCER GENOME SEQUENCING DATA

Yuichi Shiraishi, Yusuke Sato, Kenichi Chiba, Yusuke Okuno, Yasunobu Nagata, Kenichi Yoshida, Norio Shiba, Yasuhide Hayashi, Haruki Kume, Yukio Homma, Masashi Sanada, Seishi Ogawa, Satoru Miyano

Thanks to advances in high-throughput sequencing technologies, a comprehensive dissection of the cancer genome has become possible and a large number of somatic mutations have been detected in a wide variety of cancer types. However, due to the ambiguity of short read alignment and sequencing errors, a detection of somatic mutations is not a straightforward task. A number of statistical methods have been proposed for mutation calling, most of which evaluate the statistical difference in the observed allele frequencies of possible variants between tumors and paired normal samples. However, in case of low sequencing depths or tumor contents, an accurate detection of somatic mutations is still difficult.

We have proposed a novel method, EBCall (Empirical Bayesian mutation Calling, https://github.com/friend1ws/EBCall), for detecting somatic mutations. Unlike previous methods, EBCall uses sequencing data of multiple non-paired normal samples to estimate possible systematic sequencing errors at each genomic site, and the allele frequencies of the observed variants in the tumor DNA are then compared to the inferred sequencing error distribution at the corresponding genomic positions to discriminate genuine mutation from sequencing errors.

With whole exome sequencing data, we show that our method outperforms several existing methods in calling mutations with moderate allele frequencies. Furthermore our method enables accurate calling of somatic mutations with low allele frequencies (≤10%) harbored within a minor subpopulation, which lead to the deciphering of fine substructures within a tumor specimen.

GENOMIC AND COMPUTATIONAL APPROACHES TO TUBERCULOSIS AND NONTUBERCULOUS MYCOBACTERIAL DISEASE

Michael Strong, Rebecca Davidson, Gargi Datta, Ben Garcia, Nabeeh Hasan, Mary Jackson

The past decade has marked a significant shift in the biological sciences, increasingly moving toward larger and larger datasets ranging from genomic to proteomic scale datasets. As a result, our challenge in many cases has transitioned from that of data generation to data analysis and integration. To address this need, our group has put much effort into genome, transcriptome, and protein modeling projects in order to better understand and make use of omic scale datasets in a more robust, rapid, and meaningful manner. Most of our projects have focused on mycobacterial pathogens that cause severe respiratory diseases ranging from tuberculosis to Nontuberculous mycobacterial (NTM) infections. Through these efforts, we have been able implement genomic and bioinformatic pipelines to gain a better understanding of the diversity of TB and NTM pathogens at the genomic level. These studies range from comparisons of clinically derived M. abscessus strains to MDR TB strains. We have also implemented improved functional annotation and enrichment pipelines for the annotation of previously uncharacterized mycobacterial genes and the analysis of gene sets, which we hope will facilitate research in diverse mycobacterial fields. We also are actively working on translational projects to develop improved molecular diagnostic tests for drug resistant TB and NTM speciation. Together we hope these projects move us forward toward the collective goal of identifying better ways to prevent, limit, and combat deadly mycobacterial diseases.

TEXT AND DATA MINING FOR BIOMEDICAL DISCOVERY

POSTER PRESENTATIONS

ACCEPTED PROCEEDINGS PAPER WITH POSTER PRESENTATION.

MATRIX FACTORIZATION-BASED DATA FUSION FOR GENE FUNCTION PREDICTION IN BAKER'S YEAST AND SLIME MOLD

Marinka Zitnik, Blaz Zupan

The development of effective methods for the characterization of gene functions that are able to combine diverse data sources in a sound and easily-extendible way is an important goal in computational biology. We have previously developed a general matrix factorization-based data fusion approach for gene function prediction. In this manuscript, we show that this data fusion approach can be applied to gene function prediction and that it can fuse various heterogeneous data sources, such as gene expression profiles, known protein annotations, interaction and literature data. The fusion is achieved by simultaneous matrix tri-factorization that shares matrix factors between sources. We demonstrate the effectiveness of the approach by evaluating its performance on predicting ontological annotations in slime mold D. discoideum and on recognizing proteins of baker's yeast S. cerevisiae that participate in the ribosome or are located in the cell membrane. Our approach achieves predictive performance comparable to that of the state-of-the-art kernel-based data fusion, but requires fewer data preprocessing steps.

AUTOMATING THE CONSTRUCTION OF METABOLIC PATHWAYS USING BRENDA, METACYC MODEL PATHWAYS AND LITERATURE MINING.

Jan Czarnecki, Irene Nobeli, Adrian M Smith and Adrian J Shepherd

There are over 20 million research articles referenced in PubMed. This figure is increasing so rapidly that it is becoming increasingly difficult for large organisations, let alone individual scientists, to keep up with new research. The effect of this information overload can be seen within the metabolic pathways field, where most of the data in resources such as BioCyc is predicted from full genome sequences and the experimentally verified pathway data associated with a small number of well-studied model organisms; much relevant experimental data has yet to be incorporated.

Here we present a tool that combines literature mining with information from existing data resources to construct metabolic pathways for organisms of interest. The tool accepts a seed pathway from the user which is used to obtain relevant article abstracts (and full text articles if available in PubMed Central). The tool extracts metabolic reactions from the retrieved text using a combination of open source named entity recognition tools and a rule based algorithm. The reactions are then scored for relevance and accuracy based a) on the properties of the texts from which they are extracted, b) on what is known about plausible reactions given the set of curated pathways in MetaCyc and evidence from the BRENDA enzyme database, and c) the structure of the returned metabolic network.

The tool is designed to address different sub-tasks, including the extraction of: reactions that fill gaps in existing pathway data; evidence for variants of a given pathway not currently documented in the BioCyc or KEGG databases; and cross-links between existing pathways. Output from the tool enables immediate visualization of putative pathways using Cytoscape.

The tool was evaluated on its ability to extract MetaCyc alternative pathways from the literature, achieving promising results within certain domains. For instance, the 'allantoin degradation to glyoxylate' pathways in MetaCyc show the different routes between two small molecules in a range of different organisms. Pathway I shows a 3-step pathway constructed from Saccharomyces cerevisiae experimental evidence, while pathways II and III show a 4-step pathway found in a number of organisms, including Arabidopsis thaliana. Pathways II and III were provided to the tool and with instructions to find the corresponding pathway within S. cerevisiae. The tool was able to correctly reconstruct pathway I, with the appropriate reactions marked as both accurate and relevant. While certain areas are problematic (such as lipid metabolism), the tool has the potential to provide a useful line of research complementary to current metabolic pathway prediction methods.

EXTRACTING COUNTRY-OF-ORIGIN FROM ELECTRONIC MEDICAL RECORDS FOR GENE-ENVIRONMENT STUDIES AS PART OF THE EPIDEMIOLOGIC ARCHITECTURE FOR GENES LINKED TO ENVIRONMENT (EAGLE)

Eric Farber-Eger, Robert Goodloe, Jonathan Boston, William S. Bush, Dana C. Crawford

Biorepositories linked to electronic medical records are important resources for genetic association studies for diverse populations. Much effort has been expended to define and harmonize phenotypes within and across EMRs, and this effort is now being extended to include environmental exposures important for gene-environment studies. We describe here the extraction of country origin, an acculturation variable relevant for datasets representing diverse populations, in 15,863 DNA samples from non-European descent subjects linked to de-identified EMRs and genotyped using the Illumina Metabochip as part of the Epidemiologic Architecture for Genes Linked to Environment (EAGLE), a study site of the Population Architecture using Genomics and Epidemiology (PAGE) network. Overall, through a combination of automated searches and manual review, we were able to assign country of origin outside of the United States to 1,911 subjects in EAGLE BioVU. Among these, the distribution of the countries assigned followed expectations based on known migration patterns to the United States with an emphasis on the southeastern region. In addition to providing acculturation data, country of origin also proved a useful variable for quality control and evaluation of population stratification.

TOWARDS KNOWLEDGE INFERENCE FROM BIOMEDICAL TEXTS: ALIGNING TEXT-MINING EXTRACTIONS TO PATHWAY MODEL SEMANTICS

Ravikumar Komandur Elayavilli, Kavishiwar Wagholikar, Hongfang Liu

In the past decade, literature-mining systems have played critical role in annotation and curation of biological databases. However text-mining systems play very little role in the annotation of complex bio-molecular events in biological pathways. The curation of such biological pathway databases has been largely manual and takes significant human time and effort. The primary reason is the gap between the requirements for pathway curation and the task definitions for literature mining. In an ongoing work, our analysis reveals that there are fundamental differences between the annotation framework adopted by the text mining community and the pathway curation community. The semantic annotation of the text-mining world does not adequately capture the underlying semantics conveyed in the biological cascades. Consider the example phrase “ATR/FRP-1 also phosphorylated p53 in Ser 15”. While the text mining semantics consider p53 (molecule) as the theme, Ser 15 (molecular site) as the site modified in p53 (molecule) and ATR/FRP-1 (molecule) as the cause of the event phosphorylation. Phosphorylation is enzyme catalysis where the enzyme (ATR/FRP-1) catalyzes a state transition of p53 from un-phosphorylated form to phosphorylated form at serine residue 15. The text mining annotation representation fails to represent the end product “phosphorylated-p53”). For other reactions such as binding, transport etc. the end product complex and the sub-cellular location of the molecule has no separate representation in text mining annotations. The BioNLP 2013 PC shared task [1]is an ideal start in this direction. Allowing annotations that are not tied to textual mentions will not facilitate in translating the textual annotations to pathway model representations such as BioPAX, SBGN and SBML but will be highly useful in knowledge deduction and inference as well. Our preliminary work in rendering biomedical text-based extractions to pathway model standards such as SBGN, BioPAX reinforce that there is a critical need to align the goals of text mining deliverables to the demands of the pathway model semantics. References 1. Ohta, T., et al. Overview of the pathway curation (PC) task of bioNLP shared task 2013. in Proceedings of BioNLP Shared Task 2013 Workshop. 2013.

SPARSE GENERALIZED FUNCTIONAL LINER MODEL FOR PREDICTING REMISSION STATUS OF DEPRESSION PATIENTS

Yashu Liu, Zhi Nie, Jiayu Zhou, Michael Farnum, Vaibhav A Narayan, Gayle Wittenberg, Jieping Ye

Complex diseases such as major depression affect people over time in complicated patterns. Longitudinal data analysis is thus crucial for understanding and prognosis of such diseases and has received considerable attention in the biomedical research community. Traditional classification and regression methods have been commonly applied in a simple (controlled) clinical setting with a small number of time points. However, these methods cannot be easily extended to the more general setting for longitudinal analysis, as they are not inherently built for time-dependent data. Functional regression, in contrast, is capable of identifying the relationship between features and outcomes along with time information by assuming features and/or outcomes as random functions over time rather than independent random variables. In this paper, we propose a novel sparse generalized functional linear model for the prediction of treatment remission status of the depression participants with longitudinal features. Compared to traditional functional regression models, our model enables high-dimensional learning, smoothness of functional coefficients, longitudinal feature selection and interpretable estimation of functional coefficients. Extensive experiments have been conducted on the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) data set and the results show that the proposed sparse functional regression method achieves significantly higher prediction power than existing approaches.

UNSUPERVISED DISCOVERY OF GENE-DRUG INTERACTION PATTERNS IN BIOMEDICAL TEXT

Bethany Percha, Russ B. Altman

Pharmacogenomics is the study of how individual genomic variation influences drug response phenotypes. Therefore, a full understanding of the relationships among drugs, genes and phenotypes is of paramount importance to the field. However, extracting relevant pharmacogenomic relations from the biomedical literature is a very challenging proposition, mostly because of data sparsity - there are simply too many different ways to say the same thing. Here we apply an unsupervised biclustering technique, called information-theoretic co-clustering, to the problem of biomedical relation extraction. We show how this method can be used to semantically group gene-drug relations, effectively normalizing textual patterns into semantically-equivalent groups without the use of annotated training data. The method also "fills in" unobserved textual patterns connecting drugs and genes, reducing sparsity by approximately 70-fold. We show how it can be used to augment the set of manually-curated pharmacogenomic relationships currently contained in PharmGKB, generating new drug-gene relationship candidates for the database and providing evidence for its choices. This unsupervised approach could prove highly useful in natural language processing domains, like biomedicine, where little or no human-annotated training data is available.

WORKSHOP: COMPUTATIONAL IDENTIFICATION AND FUNCTIONAL ANALYSIS OF NON-CODING RNAS

INVITED SPEAKERS ABSTRACTS

AN INTEGRATIVE APPROACH FOR IDENTIFYING FUNCTIONALLY IMPORTANT AND/OR CLINICALLY RELEVANT LONG NONCODING RNAS IN HUMAN CANCER

Yiwen Chen, Zhou Du, Teng Fei, Roel G W Verhaak, Zhen Su, Yong Zhang, Myles Brown, X Shirley Liu

Despite growing appreciations of the number and importance of long non-coding RNA (lncRNA) in normal physiology and disease, our knowledge of cancer-related lncRNA remains limited. By repurposing microarray probes, we constructed the expression profile of 10,207 lncRNA genes in approximately 1,300 tumors over four different cancer types. Through integrative analysis of the lncRNA expression profiles with clinical outcome and somatic copy number alteration (SCNA), we identified lncRNA that are associated with cancer subtypes and clinical prognosis, and predicted those that are potential drivers of cancer progression. We validated our computational predictions by experimentally confirming prostate cancer cell growth dependence on two novel lncRNA. Our analysis provided a useful resource of clinically relevant lncRNA for future development of lncRNA biomarkers and identification of lncRNA therapeutic targets in cancer. It also demonstrated the power of integration of publically available genomic datasets and clinical information for discovering disease associated lncRNA.

NUCLEAR RNAI: DRIVING SELECTIVE RECOGNITION OF NONCODING RNAS TO CONTROL TRANSCRIPTION AND SPLICING

David R. Corey

Although many long noncoding RNAs (lncRNAs) have been discovered, their function, and their association with RNAi factors in the nucleus has remained obscure. Here, we identify RNA transcripts that overlap the cyclooxygenase-2 (COX-2) promoter and contain two adjacent binding sites for an endogenous miRNA, miR-589. We find that miR-589 binds the promoter RNA and activates COX-2 transcription. In addition to miR-589, fully complementary duplex RNAs that target the COX-2 promoter transcript activate COX-2 transcription. Activation by small RNA requires RNAi factors argonaute-2 (AGO2) and GW182, but does not require AGO-2 mediated cleavage of the promoter RNA. Instead, the promoter RNA functions as a scaffold. Binding of AGO2 protein/small RNA complexes to the promoter RNA triggers gene activation. Gene looping allows interactions between the promoters of COX-2 and phospholipase A2 (PLA2G4A), an adjacent pro-inflammatory pathway gene that produces arachidonic acid, the substrate for COX-2 protein. miR-589 and fully-complementary small RNAs regulate both COX-2 and PLA2G4A gene expression, revealing an unexpected connection between key steps of the eicosanoid signaling pathway. The work demonstrates the potential for RNA to coordinate locus-dependent assembly of related genes to form functional operons through cis looping.

IDENTIFICATION, ANNOTATION, CLASSIFICATION AND THE EVOLUTIONARY HISTORY OF LARGE INTERGENIC NON-CODING RNAS (LINCRNAS) IN MAMMALS

Manuel Garber

My presentation will cover computational and experimental methods to annotate lincRNAs with special consideration to their lower expression profile. I will then discuss phylogenetic footprint of these transcripts including both their sequence and expression profiles. We find that up to 65% of lincRNAs expressed in human are not orthologous to mouse, however the expression of Mammalian-conserved lincRNAs shows remarkably strong conservation of tissue specificity, suggesting it is essential and selectively maintained. Even in Mammalian-conserved lincRNAs we find abundant splice site turnover, suggesting that exact splice sites are not critical to their function. Hominid-specific lincRNAs are more tissue-specific, enriched for testis, and faster-evolving within the lineage.

LINCING THE NON-CODING WORLD TO THE MAMMALIAN CIRCADIAN CLOCK

John Hogenesch

I have a longstanding interest in noncoding RNA function (e.g. Hogenesch et al., Cell, 2001). In addition to several projects aimed at understanding function through global gene expression analysis (the Gene Atlas), my lab has also pioneered using functional screening to gain insight into the potential cellular functions of Linc RNAs (e.g. Willingham et al., Science, 2005). Recently, we have extended this work to functional screens of microRNAs (e.g. Olarerin et al., Genome Biology, 2013), and using microRNAs as markers of human disease (e.g. Anton, Olarerin et al., American Journal of Pathology, 2013). Most recently, dovetailing with the major interest of my lab, timing, we have developed a map of time and space of the expression of Linc RNAs in the mouse (Zhang et al., in preparation).

PREDICTING RNA SECONDARY STRUCTURE USING PROBABILISTIC METHODS

David H. Mathews

RNA is increasingly understood to play many important cellular roles. This talk focuses on computational methods for modeling and understanding RNA structure. At the secondary structure resolution, i.e. the resolution of canonical base pairing, a free energy nearest neighbor model can be used to quantify the stability of a structure at folding equilibrium. Using this model and a dynamic programming algorithm, partition functions can be calculated, which can be used to predict base pairing probabilities. This talk will focus on the use of base pairing probabilities to improve secondary structure prediction. Base pairing probabilities suggest predicted pairs that are more likely to be accurate. Structures can be assembled from highly probable pairs, and these structures can include pseudoknots, which are known to be difficult to predict. Finally, for a set of homologous sequences, base pairing probabilities and alignment probabilities can be used to model a conserved structure.

NONCODING RNA DATABASE AND FUNCTIONAL RESEARCH

Yi Zhao, Tengfei Xiao, Jiao Yuan, Runsheng Chen

Non-coding RNAs (ncRNAs) constitute a significant fraction of the transcriptome . Widespread application of high-throughput RNA sequencing (RNA-seq), with the aid of computational methods, has revealed increasing number of ncRNAs identified from various organisms. Owing to their functional significance, we built two databases, NONCODE and NPInter. NONCODE (http://www.bioinfo.org/NONCODE/) is an integrated knowledge database dedicated to non-coding RNAs (excluding tRNAs and rRNAs). NPInter (http://www.bioinfo.org/NPInter) is a database that integrates experimentally verified functional interactions between noncoding RNAs (excluding tRNAs and rRNAs) and other biomolecules (proteins, RNAs and genomic DNAs). Although previous studies have found several lncRNAs specifically regulated during adipogenesis no lncRNA is yet reported to be involved in a genetic pathway regulating adipogenic differentiation. we found a 2.4-kilobase lncRNA, named ADINR (adipogenic differentiation induced noncoding RNA), activates the C/EBPα gene by impacting on the H3K4me3 and H3K27me3 histone modification of C/EBPα locus during adipogenic differentiation. In addition, forced expression of C/EBPα rescued the ADINR depletion phenotype, suggesting that they are involved in the same genetic pathway.

A

Achrol, Achal S., 17
Aguiar, Derek, 10
Ahn, Kwangmi, 58
Albu, Mihai, 74
Aldrich, Joshua T., 48
Altman, Russ B., 28, 66, 99
An, Steven S, 58
Anderson, Peter, 53
Aristarova, Zanna, 70
Arooj, Mahreen, 52

B

Badea, Liviu, 44
Baggerly, Keith A., 65
Baldassi, Carlo, 12
Barbarino, Julia, 66
Barnholtz-Sloan, Jill, 51
Bartha, Gabor, 32
Beasley, Ellen, 78
Bebek, Gurkan, 51
Birdwell, Kelly A., 30
Blucher, Aurora S., 50
Blyth, Bruce, 89
Bonora, Giancarlo, 59
Boston, Jonathan, 60, 64, 73, 88, 96
Bourne, Philip E., 19
Bowton, Erica A., 30
Boyer, Laurie, 80
Boyle, Sean M, 28
Braunstein, Alfredo, 12
Brenner, Steven E., 87
Brilliant, Murray, 25
Bronstad, Matthew, 75
Brown, Myles, 101
Brubaker, Douglas, 51
Bush, William S., 30, 60, 64, 73, 96
Bustamante, Carlos D, 28, 29

C

Cancer Genome Atlas Research Network, 46
Carter, Gregory W., 26
Chaibub Neto, Elias, 11, 14, 61, 79
Chance, Mark, 51
Chandratillake, Gemma, 32
Chayes, Jennifer, 12
Chen, Richard, 32
Chen, Runsheng, 106
Chen, Yanwen, 51
Chen, Yiwen, 101
Chervitz, Stephen, 32
Cheshier, Samuel H., 17
Cheung, Philip, 63
Chiba, Kenichi, 91
Ching, Jinny, 75
Cho, Kwang-Hwi, 83
Chowdhury, Salim, 16, 72
Christian Borgs, 12
Chronis, Constantinos, 59
Chu, Rosalie K., 48
Chu, Tim, 89
Chute, Christopher, 22
Clark, Michael, 32
Clark, Wyatt T., 35
Clarke, Bertrand, 62
Clarke, Jennifer, 62
Clauss, Therese R., 48
Cleves, Ann E., 21
Cohen, K. Bretonnel, 36
Coleman, Robert L., 65
Corey, David R., 102
Cowan, Jay, 60
Crane, Erin K., 65
Crawford, Dana C., 25, 30, 40, 64, 73, 88, 96
Czarnecki, Jan, 95

D

Daneshjou, Roxana, 28
Darabos, Christian, 24
Datta, Gargi, 92
Davidson, Rebecca, 92
de Boer, Thon, 78
de Laat, Wouter, 59
Denholtz, Matthew, 59
Denny, Joshua C., 30, 60
Difeo, Analisa, 51
Dilks, Holli H., 30
Dobra, Adrian, 62
Drescher, Charles W., 65
Du, Pan, 15
Du, Zhou, 101
Dudek, Scott M., 25
Dumitrescu, Logan, 88

E

Ernst, Jason, 59

F

Farber-Eger, Eric, 40, 64, 73, 88, 96
Farnum, Michael, 39, 98
Fei, Teng, 101
Fenger, Douglas D., 63
Ferdig, Michael T., 54
Forster, Michael, 45
Fraenkel, Ernest, 12
Franke, Andre, 45
Frey, Brendan J., 74
Friend, Stephen H., 11, 14, 61
Funahashi, Akira, 67
Funk, Christopher S., 36

G

Gaiteri, Chris, 79
Garber, Manuel, 103
Garcia, Ben, 92
Garcia, John, 89
Garcia, Sarah, 32
Gentleman, Robert C., 15
Gharpure, Kshipra, 65
Gibb, Bill, 78
Gitter, Anthony, 12
Gitter, Tony, 76
Gnad, Florian, 15
Goodloe, Robert, 25, 64, 73, 88, 96
Greenblatt, Jack, 74
Greene, Casey S., 84
Gritsenko, Marina A., 48
Guinney, Justin, 14

H

Haines, Jonathan L., 60
Hall, Jacob, 64, 73
Hall, Molly A., 25
Han, David K, 70
Han, Henry, 37
Harmon, Samantha H., 24
Hart, Reece K., 89
Hasan, Nabeeh, 92
Hauptman, Ruth, 19
Haverty, Peter M., 15
Hayashi, Yasuhide, 91
Herbrich, Shelley M., 65
Hewett, Darla, 66
Hiraiwa, Takumi, 67
Hiroi, Noriko, 67
Hirst, Martin, 90
Hoff, Bruce, 79
Hogenesch, John, 104
Homma, Yukio, 91
Hu, Hao, 13
Huff, Chad D., 13
Hughes, Timothy R., 74
Hunter, Lawrence E., 36
Huntley, Melanie A., 15
Hwang, Sun-Il, 70

I

Istrail, Sorin, 10
Ivan, Christina, 65
Iwasaki, Tsuyoshi, 67

J

Jackson, Mary, 92
Jain, Ajay N., 21
Jang, In Sock, 11, 14, 61
Jeon, Jouhyun, 68
Jeong, Jennie, 78
Jiang, Zhaoshi, 15
Jin, Daeyong, 69
Jojic, Nebojsa, 86
Jones, Steven J.M., 90
Ju, Hyunjung, 83

K

Kahles, André, 46
Kasaian, Katayoon, 90
Kellis, Manolis, 80
Kenny, Eimear E., 29
Khatri, Purvesh J., 17
Kim, Jaesang, 83
Kim, Philip M., 68
Kim, Sang Cheol, 83
Kim, Sangok, 83
Kim, Seok Jin, 83
Kim, Shinhyuk, 69
Kim, Wankyu, 83
Kim, Won Seog, 83
Kimura, Tadamasa, 67
Kirby, Jacqueline, 60
Klein, Robert, 46
Klein, Teri E., 28, 66
Ko, Young Hyeh, 83
Kolonel, Laurence N., 81
Komandur Elayavilli, Ravikumar, 97
Ku, Jeremy, 78
Kukurba, Kim, 28
Kume, Haruki, 91
Kuo, David, 46
Kwei, Kevin, 78

L

Lam, Kathy N., 74
Latifi, Ardian, 70
Le Marchand, Loic, 81
Lee, Byungwook, 83
Lee, Haeseung, 83
Lee, Hyunju, 69
Lee, Jong-Eun, 83
Lee, Keun Woo, 52
Lee, Sanghyuk, 83
Lee, Seung Ho, 83
Lee, William, 46
Lehmann, Kjong-Van, 46
Leng, Nan, 32
Levy, Sam, 78
Li, Matthew D., 17
Li, Ruiqiang, 15
Li, Yingrui, 15
Liu, Hongfang, 38, 97
Liu, Jinfeng, 15
Liu, Mei-Lan, 78
Liu, Tao, 48
Liu, Wen, 47
Liu, X Shirley, 101
Liu, Yashu, 39, 98

Lopez-Berestein, Gabriel, 65
Lovato, Pietro, 86
Lundgren, Deborah H, 70

M

Ma, Qi, 47
Mahoney, J. Matthew, 71
Malinowski, Jennifer, 40
Margolin, Adam A., 11, 14, 61
Marra, Marco A., 90
Martin, Alicia R., 29
Mathews, David H., 105
Mayba, Oleg, 15
McCarty, Catherine A., 25
McDermott, Jason E., 48
McWeeney, Shannon K., 50
Min, Martin Renqiang, 16, 72
Minna, John D., 15
Mitchell, Sabrina, 64, 73
Miyano, Satoru, 91
Mnaimneh, Sanie, 74
Monroe, Matthew E., 48
Montgomery, Stephen B, 28
Moore, Jason H., 24
Moore, Richard A., 90
Moore, Ronald J., 48
Morgan, Alexander A., 17, 32
Morris, Quaid, 41
Moult, John, 87
Mungall, Karen L., 90

N

Nagata, Yasunobu, 91
Najafabadi, Hamed S., 74
Narayan, Vaibhav A, 39, 98
Ng, Clara, 19
Ng, Ho Leung, 75
Nick, Alpa M., 65
Nie, Zhi, 39, 98
Nim, Satra, 68
Nobeli, Irene, 95

O

Oetjens, Matthew, 30
Ogawa, Seishi, 91
Okuno, Yusuke, 91
Ormond, Kelly E, 28
Ostroff, Rachel, 16, 72
Ozpolat, Bulent, 65

P

Pagnani, Andrea, 12
Parikh, Ankur P., 31
Park, Seongjin, 83
Patwardhan, Anil, 32
Payne, Samuel H., 48
Pearl, Taylor, 51
Peissig, Peggy, 25
Pelham, Robert, 78
Pellegrini, Matteo, 59
Pendergrass, Sarah, 25, 64, 73
Percha, Bethany, 99
Perina, Alessandro, 86
Petyuk, Vladislav A., 48
Peyton, Michael, 15
Philip, Vivek M., 26
Plath, Kathrin, 59
Poon, Hoifung, 76
Pratt, Mark, 32

Q

Qendro, Veneta, 70
Qi, Yanjun, 16, 72
Qi, Yuan, 33
Qu, Kunbin, 78
Quigley, Ann, 77
Quirk, Chris, 76

R

Radivojac, Predrag, 35
Radovani, Ernest, 74
Rapoport, Judith L., 58
Rätsch, Gunnar, 46
Ravikumar K.E., 38
Rezaul, Karim, 70
Rho, Kyoohyoung, 83
Ritchie, Marylyn D., 25, 60
Roden, Dan M., 30
Rodland, Karin D., 48
Rosenfeld, Michael. G., 47

S

Sakarya, Onur, 78
Sanada, Masashi, 91
Sato, Yusuke, 91
Sauerwine, Ben, 79
Schein, Jacquie, 90
Schepmoes, Athena A., 48
Schmitges, Frank W., 74
Schoeniger, Joseph, 53
Schultz, Nikolaus, 46
Shames, David S., 15
Shaw, Matthew, 63
Shen, Feichen, 22
Shepherd, Adrian J, 95
Shiba, Norio, 91
Shiraishi, Yuichi, 91
Shukla, Anil K., 48
Sinha, Rileen, 46
Siwo, Geoffrey Henry, 54
Smith, Adrian M, 95
Smith, Richard D., 48
Smith, Roger, 54

Snyder, Michael, 28
Sood, Anil K., 65
Splinter, Erik, 59
Stegle, Oliver, 46
Stewart, Alex, 16, 72
Strong, Michael, 92
Su, Zhen, 101
Sung, Min Kyung, 83

T

Takiguchi, Ryo, 67
Tan, Asako, 54
Tao, Cui, 22
Tasaki, Shinya, 79
Terribilini, Michael, 77
Teyra, Joan, 68
Toyoshiba, Hiroyoshi, 79
Tse, Gerard, 29
Tucker, Susan L., 65
Tully, Tim, 63
Tyler, Anna L., 26, 71

U

Unruh, Anna, 65

V

Valdes, Camillo, 62
Vembu, Shankar, 41
Verhaak, Roel G W, 101

W

Wagholikar, Kavishwar B., 38, 97
Walker, Blair A., 90
Wallace, John, 60
Walter, Kimberly, 15
Wang, Hansong, 81
Wang, Xinchen, 80
Weirauch, Matthew T., 74
Whaley, Ryan, 66
Whirl-Carrillo, Michelle, 66
Wiittenberg, Gayle, 39
Wilke, Russell A., 30
Wiseman, Sam M., 90
Wittenberg, Gayle, 98
Wong, Wendy S.W., 10
Woon, Mark, 66
Wu, Sherry, 65
Wu, Wei, 31

X

Xiao, Tengfei, 106
Xie, Lei, 19
Xing, Eric P., 31
Xu, Jinbo, 20
Xu, Zenglin, 33

Y

Yadav, Gitanjali, 56
Yamazaki, Kazuto, 67
Yang, Ally, 74
Yang, Fan, 20
Yang, Feng, 48
Ye, Jieping, 39, 98
Yera, Emmanuel R., 21
Yoo, Hae Yong, 83
Yoshida, Kenichi, 91
Yu, Peng, 33
Yuan, Jiao, 106

Z

Zappala, Zachary, 28
Zecchina, Riccardo, 12
Zelaya, Rene A., 84
Zeng, Jianyang, 20
Zhai, Kaide, 51
Zhang, Fan, 15
Zhang, Yinliang, 19
Zhang, Yong, 101
Zhang, Zemin, 15
Zhao, Yi, 106
Zhao, Yongjun, 90
Zhe, Shandian, 33
Zhou, Jiayu, 39, 98
Zhu, Qian, 22
Zitnik, Marinka, 94
Zupan, Blaz, 94