ROCKY 2018

Bioinformatics Conference

SIXTEENTH ROCKY MOUNTAIN BIOINFORMATICS CONFERENCE

DECEMBER 6 TO 8, 2018

SNOWMASS/ASPEN
COLORADO

ISCB
INTERNATIONAL SOCIETY FOR COMPUTATIONAL BIOLOGY
OFFICIAL CONFERENCE

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

WELCOME

Dear Rocky 2018 participant

Welcome to the 16th Rocky Mountain Bioinformatics Conference.

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

We want to thank and acknowledge the ongoing support of IBM who has provided significant sponsorship funds to this conference for the past sixteen years; to SomaLogic for their financial and administrative support; and PatientsLikeMe for their continued support of this meeting. We are so fortunate to have such support and we hope to have them support this conference for many years to come. It is only with the help of these sponsors that we can make this meeting as affordable as it is.

Please seek out attendees from the sponsoring organizations, and let them know that their participation is important to you! Finally, the meeting would simply not be possible without organizational help from Stephanie Hagstrom, Kathy Campbell, Elizabeth Wethington, and Caitlin Moloney.

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

Welcome!

Larry Hunter
Conference Chair

AGENDA AT-A-GLANCE

All sessions will take place at Viceroy Hotel

THURSDAY – December 6, 2018

TIME	SESSION TYPE
8:00 AM – 6:00 PM	**REGISTRATION**
8:00 AM – 9:00 AM	**BREAKFAST**
9:00 AM – 9:45 AM	**KEYNOTE 1: ZHIYONG LU**, *PhD* *Deputy Director for Literature Search* *National Center for Biotechnology Information (NCBI)* *Senior Investigator, NCBI/NLM/NIH* **Machine Learning in Biomedicine: from PubMed Search to Autonomous Disease Diagnosis**
9:45 AM – 10:25 AM	**OP01 – OP04**
10:25 AM – 10:45 AM	**BREAK**
10:45 AM – 11:15 AM	**KEYNOTE 2: AARON VON HOOSER**, *PhD* *Principal Scientist Computational Biology PatientsLikeMe, Inc.* *Massachusetts, USA* **Building a Learning System that Helps Individuals to Thrive by Connecting Their Experiences and Goals with Molecular Measures of Health**
11:15 AM – 11:55 AM	**OP05 — OP08**
12:00 PM – 04:00 PM	**SKI BREAK**
4:00 PM – 4:30 PM	**KEYNOTE 3: JOSLYNN S. LEE**, *PhD* *Science Education Fellow, Howard Hughes Medical Institute* **Training and Engaging URM Undergraduate Students in Genomics Research Through a Place-based Microbiome Research Project**
4:30 PM – 5:10 PM	**OP09 – OP12**
5:10 PM – 5:30 PM	**BREAK**
5:30 PM – 6:00 PM	**OP13 – OP15**
6:30 PM – 9:30 PM	**DINNER — IL POGGIO**

AGENDA-AT-A-GLANCE

All sessions will take place at Viceroy Hotel

FRIDAY – December 7, 2018

TIME	SESSION TYPE
8:00 AM – 6:00 PM	**REGISTRATION**
8:00 AM – 9:00 AM	**BREAKFAST**
9:00 AM – 9:45 AM	**KEYNOTE 4: DEBORAH L. MCGUINNESS**, *PhD* *Tetherless World Senior Constellation Chair, Professor of Computer Science and Cognitive Science, Rensselaer Polytechnic Institute*
9:45 AM – 10:25 AM	**OP16 – OP19**
10:25 AM – 10:45 AM	**BREAK**
10:45 AM – 11:15 AM	**KEYNOTE 5: DAVID ASTLING, PHD** *Scientist, Bioinformatics, SomaLogic Inc.* **Beyond Genomics: Deriving Actionable Health Insights from the Human Proteome**
11:15 AM – 11:55 AM	**OP20 - OP23**
12:00 PM – 4:00 PM	**SKI BREAK**
4:00 PM – 4:30 PM	**KEYNOTE 6: NICOLE A. VASILEVSKY**, *PhD* *Research Assistant Professor, Department of Medical Informatics and Clinical Epidemiology (DMICE), Oregon Health & Science University* **LOINC2HPO: Improving Translational Informatics by Standardizing EHR Phenotypic Data Using the Human Phenotype Ontology**
4:30 PM – 5:10 PM	**OP24 – OP27**
5:10 PM – 5:30 PM	**BREAK**
5:30 PM – 6:00 PM	**OP28 – OP30**
6:30 PM – 8:30 PM	**POSTER SESSION**

AGENDA-AT-A-GLANCE

All sessions will take place at Viceroy Hotel

SATURDAY – December 8, 2018

TIME	SESSION TYPE
8:00 AM – 11:00 AM	**REGISTRATION**
8:00 AM – 9:00 AM	**BREAKFAST**
9:00 AM – 9:30 AM	**KEYNOTE 7: BENJAMIN M. GOOD**, *PhD* *Consultant, Lawerence Berkeley National Labs* **Integrating Pathway Databases with Gene Ontology Causal Activity Models**
9:30 AM – 10:20 AM	**OP31 – OP35**
10:20 AM – 10:40 AM	**BREAK**
10:40 AM – 11:30 AM	**OP36 – OP40**
11:30 AM – 12:00 PM	**KEYNOTE 8: KIRK E. JORDAN**, *PhD* *IBM Distinguished Engineer, Data Centric Solutions, IBM T.J. Watson Research & Chief Science Officer, IBM Research UK* **Algorithm Exploitation & Evolving AI/Cognitive Examples on IBM's Data Centric Systems**
12:00 PM – 12:15 PM	**CLOSING RAFFLE AND AWARDS**

DETAILED AGENDA

All sessions will take place at Viceroy Hotel

THURSDAY – December 6, 2018

TIME	SESSION TYPE
8:00 AM – 6:00 PM	**REGISTRATION**
8:00 AM – 9:00 AM	**BREAKFAST**
9:00 AM – 9:45 AM	**KEYNOTE 1: ZHIYONG LU**, *PhD* *Deputy Director for Literature Search* *National Center for Biotechnology Information (NCBI)* *Senior Investigator, NCBI/NLM/NIH* **Machine Learning in Biomedicine: from PubMed Search to Autonomous Disease Diagnosis**
9:45 AM – 9:55 AM	**OP01: Proteomics of natural bacterial isolates powered by deep learning-based de novo identification** *Presenting Author: Samuel Payne, Brigham Young University*
9:55 AM – 10:05 AM	**OP02: A platform for community-scale transcriptome-wide association studies** *Presenting Author: YoSon Park, Perelman School of Medicine University of Pennsylvania*
10:05 AM – 10:15 AM	**OP03: Harmonizing and Analyzing Clinical Trials Data in the AHA Precision Medicine Platform** *Presenting Author: Carsten Goerg, University of Colorado*
10:15 AM – 10:25 AM	**OP04: anexVis: visual analytics framework for analysis of RNA expression** *Presenting Author: Diem-Trang Tran, University of Utah*
10:25 AM – 10:45 AM	**BREAK**
10:45 AM – 11:15 AM	**KEYNOTE 2: AARON VON HOOSER**, *PhD* *Principal Scientist Computational Biology PatientsLikeMe, Inc. Massachusetts, United States* **Building a Learning System that Helps Individuals to Thrive by Connecting Their Experiences and Goals with Molecular Measures of Health**
11:15 AM – 11:25 AM	**OP05: The characterization of different cell types using the Benford law** *Presenting Author: Sne Morag, Ariel University*
11:25 AM – 11:35 AM	**OP06: Use of metadata and Bag-of-words to map measurements across observational study data** *Presenting Author: Laura Stevens, University of Colorado Anschutz Medical Campus*
11:35 AM – 11:45 AM	**OP07: ExtRamp: A novel algorithm for extracting the ramp sequence based on the tRNA adaptation index or relative codon adaptiveness** *Presenting Author: Justin Miller, Brigham Young University*

THURSDAY – December 6, 2018

TIME	SESSION TYPE
11:45 AM – 11:55 AM	**OP08: Using machine learning algorithms for classification of medulloblastoma subgroups based on gene expression data** *Presenting Author: Sivan Gershanov, Ariel University*
12:00 PM – 4:00 PM	**SKI BREAK**
4:00 PM – 4:30 PM	**KEYNOTE 3: JOSLYNN S. LEE**, *PhD* *Science Education Fellow, Howard Hughes Medical Institute* **Training and Engaging URM Undergraduate Students in Genomics Research Through a Place-based Microbiome Research Project**
4:30 PM – 4:40 PM	**OP09: A human disease network from gene-publication relationships on PubMed** *Presenting Author: Edward Lau, Stanford University*
4:40 PM – 4:50 PM	**OP10: Transcriptome analysis of cancer adjacent normal tissues reveal genes co-expressed with LINE elements** *Presenting Author: Mira Han, University of Nevada Las Vegas*
4:50 PM – 5:00 PM	**OP11: Highly accurate computational characterization of protein kinase family-specific phosphorylation sites** *Presenting Author: Chen Li, ETH Zürich*
5:00 PM – 5:10 PM	**OP12: ORCHID: a method for detecting short-range chromatin interactions in high-resolution 5C and Hi-C datasets** *Presenting Author: Fei Ji, Massachusetts General Hospital*
5:10 PM – 5:30 PM	**BREAK**
5:30 PM – 5:40 PM	**OP13: Using Adversarial Deep Neural Networks to Remove Nonlinear Batch Effects from Expression Data** *Presenting Author: Jonathan Dayton, Brigham Young University*
5:40 PM – 5:50 PM	**OP14: Med2Mech: Neural-Symbolic Representation of Molecular Mechanisms Underlying Pediatric Disease** *Presenting Author: Tiffany Callahan, University of Colorado Denver Anschutz Medical Campus*
5:50 PM – 6:00 PM	**OP15: A Machine Learning Classifier for Assigning Individual Patients with Systemic Sclerosis to Intrinsic Molecular Subsets** *Presenting Author: Jennifer Franks, Geisel School of Medicine at Dartmouth*
6:30 PM – 9:30 PM	**DINNER - IL POGGIO**

FRIDAY – December 7, 2018

TIME	SESSION TYPE
8:00 AM – 6:00 PM	**REGISTRATION**
8:00 AM – 9:00 AM	**BREAKFAST**
9:00 AM – 9:45 AM	**KEYNOTE 4: DEBORAH L. MCGUINNESS**, *PhD* *Tetherless World Senior Constellation Chair, Professor of Computer Science and Cognitive Science, Rensselaer Polytechnic Institute*
9:45 AM – 9:55 AM	**OP16: A Data Quality Testing Tool for Cross-institutional OMOP Electronic Health Record Data Repositories** *Presenting Author: Timothy Bergquist, University of Washington*
9:55 AM – 10:05 AM	**OP17: Comparative Analysis of Germline Microsatellites in the 1,000 Genomes Project** *Presenting Author: Nicholas Kinney, Virginia College of Osteopathic Medicine*
10:05 AM – 10:15 AM	**OP18: Unbiased Pathway Detection Expands Cancer Pathways** *Presenting Author: Chih-Hsu Lin, Baylor College of Medicine*
10:15 AM – 10:25 AM	**OP19: A systems biology approach to define essential kinases in small cell lung cancer** *Presenting Author: Jihye Kim, University of Colorado Denver Anschutz Medical Campus*
10:25 AM – 10:45 AM	**BREAK**
10:45 AM – 11:15 AM	**KEYNOTE 5: DAVID ASTLING**, *PhD* *Scientist, Bioinformatics, SomaLogic Inc.* **Beyond Genomics: Deriving Actionable Health Insights from the Human Proteome**
11:15 AM – 11:25 AM	**OP20: Clustering of Protein Conformations using Parallelized Dimensionality Reduction** *Presenting Author: Arpita Joshi, University of Massachusetts, Boston*
11:25 AM – 11:35 AM	**OP21: PredHPI: an integrated web-server platform for the prediction and visualization of host-pathogen interactions** *Presenting Author: Rakesh Kaundal, Utah State University*
11:35 AM – 11:45 AM	**OP22: Searching for translatable alternative splice isoforms in the human proteome** *Presenting Author: Maggie Pui Yu Lam, University of Colorado Anschutz Medical Campus*
11:45 AM – 11:55 AM	**OP23: Text Mining Novel Disease- and Drug-Specific Pathways** *Presenting Author: Minh Pham, Baylor College of Medicine*
12:00 PM – 4:00 PM	**SKI BREAK**

FRIDAY – December 7, 2018

TIME	SESSION TYPE
4:00 PM – 4:30 PM	**KEYNOTE 6: NICOLE A. VASILEVSKY**, *PhD* *Research Assistant Professor, Department of Medical Informatics and Clinical Epidemiology (DMICE), Oregon Health & Science University* **LOINC2HPO: Improving Translational Informatics by Standardizing EHR Phenotypic Data Using the Human Phenotype Ontology**
4:30 PM – 4:40 PM	**OP24: Optimizing nontuberculous mycobacteria (NTM) de novo genome assemblies for application in clinical case studies** *Presenting Author: Sara Kammlade, National Jewish Health*
4:40 PM – 4:50 PM	**OP25: REAL-neo, a comprehensive neoantigen prediction and prioritization pipeline using tumor sequencing data** *Presenting Author: Yesesri Cherukuri, Mayo Clinic*
4:50 PM – 5:00 PM	**OP26: Measuring chromosome conformation** *Presenting Author: Brian Ross, University of Colorado Anschutz Medical Campus*
5:00 PM – 5:10 PM	**OP27: Modeling the Structure of BioGRID PPI Networks** *Presenting Author: Sridevi Maharaj, University of California-Irvine*
5:10 PM – 5:30 PM	**BREAK**
5:30 PM – 5:40 PM	**OP28: Addressing the compositional data problem in sequencing with a novel, robust normalization method** *Presenting Author: James St. Pierre, University of Toronto*
5:40 PM – 5:50 PM	**OP29: A Case Study on the Effects of Noisy, Long-read Correction Approaches on Assembly Contiguity** *Presenting Author: Brandon Pickett, Brigham Young University*
5:50 PM – 6:00 PM	**OP30: Integrative analysis of transcriptomics and proteomics to detect novel protein isoforms from alternatively spliced transcripts induced by SF3B1 spliceosomal mutations** *Presenting Author: Kelsey Nassar, University of Colorado Anschutz Medical Campus*
6:30 PM – 8:30 PM	**POSTER SESSION**

SATURDAY – December 8, 2018

	SESSION TYPE
8:00 AM – 11:00 AM	**REGISTRATION**
8:00 AM – 9:00 AM	**BREAKFAST**
9:00 AM – 9:30 AM	**Keynote 7: BENJAMIN M. GOOD**, *PhD* *Consultant, Lawerence Berkeley National Labs* **Integrating Pathway Databases with Gene Ontology Causal Activity Models**
9:30 AM – 9:40 AM	**OP31: Exploring the Fabric of Breast Cancer Using Gene Sets** *Presenting Author: Judith Blake, The Jackson Laboratory*
9:40 AM – 9:50 AM	**OP32: Measuring and Protecting the Sensitive Linking Information Leakage Across Epigenetic and Transcriptomics Datasets Through Genotype and Assay Prediction** *Presenting Author: Arif Harmanci, University of Texas Health Science Center*
9:50 AM – 10:00 AM	**OP33: Education, Networking and Building Next Generation Prototypes -- Hackathons and Analyzeathons for Bioinformaticians, Biomedical Informaticians and Computational Biologists!** *Presenting Author: Ben Busby, NCBI*
10:00 AM – 10:10 AM	**OP34: Alternative Splicing of Single Cells in Squamous Cell Lung Cancer Premalignancy** *Presenting Author: Hyunmin Kim, University of Colorado Anschutz Medical Campus*
10:10 AM – 10:20 AM	**OP35: Inferring trade-offs in protein folding networks** *Presenting Author: Sebastian Pechmann, Université de Montréal*
10:20 AM – 10:40 AM	**BREAK**
10:40 AM – 10:50 AM	**OP36: Fully Bayesian model for non-random missing data in qPCR** *Presenting Author: Valeriia Sherina, University of Rochester Medical Center*
10:50 AM – 11:00 AM	**OP37: Homologous Inter-Domain Segments in Protein Families** *Presenting Author: Dylan Barth, University of Nevada Las Vegas*
11:00 AM – 11:10 AM	**OP38: Integrating extracted relations into existing knowledge bases** *Presenting Author: Harrison Pielke-Lombardo, University of Colorado Anschutz Medical Campus*

SATURDAY – December 8, 2018

	SESSION TYPE
11:10 AM – 11:20 AM	**OP39: The Use of Scientific Ignorance to Drive Literature-Based Discovery in Prenatal Nutrition Across Disciplinary Boundaries** *Presenting Author: Mayla Boguslav, University of Colorado Anschutz Medical Campus*
11:20 AM – 11:30 AM	**OP40: Highly accurate computational characterization of protein kinase family-specific phosphorylation sites** *Presenting Author: Kari A. Stephens, University of Washington*
11:30 AM – 12:00 PM	**KEYNOTE 8: KIRK E. JORDAN**, *PhD* *IBM Distinguished Engineer, Data Centric Solutions, IBM T.J. Watson Research & Chief Science Officer, IBM Research UK* **Algorithm Exploitation & Evolving AI/Cognitive Examples on IBM's Data Centric Systems**
12:00 PM – 12:15 PM	**CLOSING RAFFLE AND AWARDS**

KEYNOTE SPEAKERS

DAVID ASTLING, PhD

Scientist, Bioinformatics, SomaLogic Inc., Colorado, USA

Beyond Genomics: Deriving Actionable Health Insights from the Human Proteome

The circulating human proteome offers a unique and dynamic perspective into a person's physiological and health status and presents a great opportunity for rapid and accurate health diagnostics. Genomics by contrast fails in applications where diagnostic fingerprints of environmental impact, disease progression, or infection are needed. SomaLogic's proteomic assay utilizes a library of over 5,000 SOMAmers for the simultaneous measurement of thousands of protein-analytes in a single blood sample. SomaLogic has shown that analysis of the proteome can provide indicators of patient risk for occurrence of a secondary cardiovascular event. To further this work, SomaLogic has embarked on a collaboration with major academic institutions to discover indicators of primary cardiovascular events, type 2 diabetes, kidney function, and lifestyle characteristics of pre-diabetic patients, as targets for incorporation into actionable insights that are of medical significance. Machine learning and statistical modeling techniques are used to develop insights that can provide rapid feedback to patients to inform strategies of managing aspects of cardio-metabolic syndrome. Additional collaborations are underway to discover insights for other disease states, physiological indicators of health and wellness, and non-blood related sample types. This presentation will examine co-regulatory networks to further our understanding of the existing models, to explore and understand the biomarkers underlying each disease model.

BENJAMIN M. GOOD, PhD

Consultant, Lawerence Berkeley National Labs, California, USA

Integrating Pathway Databases with Gene Ontology Causal Activity Models

The Gene Ontology (GO) Consortium (GOC) is developing a new knowledge representation approach called 'causal activity models' (GO-CAM). A GO-CAM describes how one or several gene products contribute to the execution of a biological process. In these models (implemented as OWL instance graphs anchored in Open Biological Ontology (OBO)

classes and relations), gene products are linked to molecular activities via semantic relationships like 'enables', molecular activities are linked to each other via causal relationships such as 'positively regulates', and sets of molecular activities are defined as 'parts' of larger biological processes. This approach provides the GOC with a more complete and extensible structure for capturing knowledge of gene function. It also allows for the representation of knowledge typically seen in pathway databases.

Here, we present details and results of a rule-based transformation of pathways represented using the BioPAX exchange format into GO-CAMs. We have automatically converted all Reactome pathways into GO-CAMs and are currently working on the conversion of additional resources available through Pathway Commons. By converting pathways into GO-CAMs, we can leverage OWL description logic reasoning over OBO ontologies to infer new biological relationships and detect logical inconsistencies. Further, the conversion helps to increase standardization for the representation of biological entities and processes. The products of this work can be used to improve source databases, for example by inferring new GO annotations for pathways and reactions and can help with the formation of meta-knowledge bases that integrate content from multiple sources.

AARON VON HOOSER, PhD

Principal Scientist, Computational Biology PatientsLikeMe, Inc.
Massachusetts, USA

Building a Learning System that Helps Individuals to Thrive by Connecting Their Experiences and Goals with Molecular Measures of Health

Through the PatientsLikeMe (PLM) network, patients connect with others who have the same disease or condition and track and share their experiences. In the process, they generate data about the real-world nature of disease that help researchers, pharmaceutical companies, regulators, providers, and non-profits develop more effective products, services and care. Studies have shown that members of PLM report tangible benefits from the connectedness and sharing that is part of the PLM community experience. With more than 500,000 members, PLM is a trusted source for real-world disease information and a clinically robust resource that has published more than 60 peer-reviewed research studies.

The Biocomputing team at PLM is leveraging the digitization of person-generated experiential data with deep molecular analyses and machine learning to help patients understand and evaluate their own molecular biology and how they may be able to change their daily lives to optimally thrive. To this end, participants in PLM's DigitalMeTM

program have donated 1000s of biospecimens, building a massive health data set that spans dozens of disease conditions, including SLE, Fibromyalgia, MS, ALS, PD, and RA; captured on an ever-increasing list of big data platforms, including DNAseq, RNAseq, metabolomics, proteomics, and antibody immunosignatures. Here, we report results from several pilot "n of 1" studies, providing deep molecular biological characterization of longitudinal timepoints from the same individuals, tracking normal physiological systems perturbed by health interventions, as well as indications that a spectrum of processes tightly associated with specific disease activities may be perturbed in "healthy" individuals under various circumstances.

KIRK E. JORDAN, Ph.D.

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

Algorithm Exploitation & Evolving AI/ Cognitive Examples on IBM's Data Centric Systems

The volume, variety, velocity and veracity of data is pushing how we think about computer systems. IBM Research's Data Centric Solutions organization has been developing systems that handle large data sets shortening time to solution. This group has created a data centric architecture initially delivered to the DoE labs at the end of 2017 and being completed in 2018. As various features to improve data handling now exist in these systems, we need to begin to rethink the algorithms and their implementations to exploit these features. This data centric view is also relevant for Artificial Intelligence (AI) and Machine Learning (ML). In this talk, I will briefly describe the architecture and point out some of hardware and software features ready for exploitation. I will show how we are using these data centric AI/cognitive computing systems to address some challenges in the life sciences in new ways as case studies.

JOSLYNN S. LEE, Ph.D.

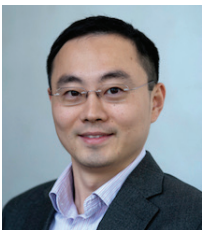
Science Education Fellow, Howard Hughes Medical Institute, Maryland, USA

Training and Engaging URM Undergraduate Students in Genomics Research Through a Place-based Microbiome Research Project

The participation of American Indian/Alaskan Native (AIAN) people and other underrepresented minority (URM) populations in STEM fields remains shockingly low. In the computational field, it is even lower. AIAN face various barriers that impede them from pursuing or continuing careers in genomics. Alongside, there is a demand for Integrating bioinformatics and data science into the life sciences curriculum. I am presenting a workshop training that allows students to gain hands-on laboratory and computational experience to understand the diversity of local environmental microbiomes in Colorado and New Mexico. This workshop targets early-career undergraduate students from Southwest regional PUIs, two-year and tribal colleges. Core competencies incorporated in the workshop are computational concepts (algorithms and file formats), statistics, accessing genomic data and running bioinformatics tools to analyze data. I will discuss some of the successes and pitfalls that I have encountered and the adaption for a one-semester course.

ZHIYONG LU, Ph.D.

Deputy Director for Literature Search, National Center for Biotechnology Information (NCBI), Senior Investigator, NCBI/NLM/NIH, Maryland, USA



Machine Learning in Biomedicine: from PubMed Search to Autonomous Disease Diagnosis

The explosion of biomedical big data and information in the past decade or so has created new opportunities for discoveries to improve the treatment and prevention of human diseases. But the large body of knowledge—mostly exists as free text in journal articles for humans to read—presents a grand new challenge: individual scientists around the world are increasingly finding themselves overwhelmed by the sheer volume of research literature and are struggling to keep up to date and to make sense of this wealth of textual information. Our research aims to break down this barrier and to empower scientists towards accelerated knowledge discovery. In this talk, I will present our work on developing open-source NLP and image analysis tools based on machine learning. Moreover, I will demonstrate their uses in some

real-world applications such as improving PubMed searches, scaling up human curation for precision medicine, and enabling image-based autonomous disease diagnosis.

DEBORAH L. MCGUINNESS, Ph.D.

Tetherless World Senior Constellation Chair, Professor of Computer Science and Cognitive Science, Rensselaer Polytechnic Institute, New York, USA

Semantic Data Resources Enabling Science: Building, Using, and Maintaining Ontology-Enabled Biology Data Resources

Ontologies are seeing a resurgence of interest and usage as big data proliferates, machine learning advances, and integration of data becomes more paramount. The previous models of sometimes labor-intensive, centralized ontology construction and maintenance do not mesh well in today's interdisciplinary world that is in the midst of a big data, information extraction, and machine learning explosion. Today many high quality ontologies exist that can and should be utilized. We will describe our approach to building maintainable and reusable semantics-enabled health and life science data ecosystems. We will introduce our method in the context of our National Institutes of Environmental Health Science-funded Child Health Exposure Analysis Resource and we will describe how how our community built and maintains a broad interdisciplinary ontology that spans exposure science and health and integrates with numerous long standing, well used ontologies. We will also describe how this ontology powers an integrated data resource. We will also give examples of how the same methodology is being used in an IBM-funded Health Empowerment using Analysis, Learning and Semantics project as well as a semantics-aware drug repurposing effort. We will conclude by discussing today's requirements for choosing, reusing, and interlinking existing, evolving resources and the resulting requirements for new methodologies and their resulting systems that can be used and maintained by large diverse communities to accelerate science discovery.

NICOLE A. VASILEVSKY, Ph.D.

Research Assistant Professor, Department of Medical Informatics and Clinical Epidemiology (DMICE), Oregon Health & Science University, Oregon, USA

LOINC2HPO: Improving Translational Informatics by Standardizing EHR Phenotypic Data Using the Human Phenotype Ontology

Electronic Health Record (EHR) data are often encoded using Logical Observation Identifier Names and Codes (LOINC), which is a universal standard for coding clinical laboratory tests. LOINC codes encode clinical tests and not the phenotypic outcomes, and multiple codes can be used to describe laboratory findings that may correspond to one phenotype. However, LOINC encoded data is an untapped resource in the context of deep phenotyping with the Human Phenotype Ontology (HPO). The HPO describes phenotypic abnormalities encountered in human diseases, and is primarily used for research and diagnostic purposes. As part of the Center for Data to Health (CD2H)'s effort to make EHR data more translationally interoperable, our group developed a curation tool that is used to convert EHR observations into HPO terms for use in clinical research. To date, over 1,000 LOINC codes have been mapped to HPO terms. To demonstrate the utility of these mapped codes, we performed a pilot study with de-identified data from asthma patients. We were able to convert 70% of real-world laboratory tests into HPO-encoded phenotypes. Analysis of the LOINC2HPO-encoded data showed that the HPO term eosinophilia was enriched in patients with severe asthma and prednisone use. This preliminary evidence suggests that LOINC data converted to HPO can be used for machine learning approaches to support genomic phenotype-driven diagnostics for rare disease patients, and to perform EHR based mechanistic research.

ORAL PRESENTATION LIST

OP01: Proteomics of natural bacterial isolates powered by deep learning-based de novo identification
Presenting Author: Samuel Payne, Brigham Young University

OP02: A platform for community-scale transcriptome-wide association studies
Presenting Author: YoSon Park, Perelman School of Medicine University of Pennsylvania

OP03: Harmonizing and Analyzing Clinical Trials Data in the AHA Precision Medicine Platform
Presenting Author: Carsten Goerg, University of Colorado

OP04: anexVis: visual analytics framework for analysis of RNA expression
Presenting Author: Diem-Trang Tran, University of Utah

OP05: The characterization of different cell types using the Benford law
Presenting Author: Sne Morag, Ariel University

OP06: Use of metadata and Bag-of-words to map measurements across observational study data
Presenting Author: Laura Stevens, University of Colorado Anschutz Medical Campus

OP07: ExtRamp: A novel algorithm for extracting the ramp sequence based on the tRNA adaptation index or relative codon adaptiveness
Presenting Author: Justin Miller, Brigham Young University

OP08: Using machine learning algorithms for classification of medulloblastoma subgroups based on gene expression data
Presenting Author: Sivan Gershanov, Ariel University

OP09: A human disease network from gene-publication relationships on PubMed
Presenting Author: Edward Lau, Stanford University

OP10: Transcriptome analysis of cancer adjacent normal tissues reveal genes co-expressed with LINE elements
Presenting Author: Mira Han, University of Nevada Las Vegas

OP11: Highly accurate computational characterization of protein kinase family-specific phosphorylation sites
Presenting Author: Chen Li, ETH Zürich

OP12: ORCHID: a method for detecting short-range chromatin interactions in high-resolution 5C and Hi-C datasets
Presenting Author: Fei Ji, Massachusetts general hospital

OP13: Using Adversarial Deep Neural Networks to Remove Nonlinear Batch Effects from Expression Data
Presenting Author: Jonathan Dayton, Brigham Young University

OP14: Med2Mech: Neural-Symbolic Representation of Molecular Mechanisms Underlying Pediatric Disease
Presenting Author: Tiffany Callahan, University of Colorado Denver Anschutz Medical Campus

OP15: A Machine Learning Classifier for Assigning Individual Patients with Systemic Sclerosis to Intrinsic Molecular Subsets
Presenting Author: Jennifer Franks, Geisel School of Medicine at Dartmouth

OP16: A Data Quality Testing Tool for Cross-institutional OMOP Electronic Health Record Data Repositories
Presenting Author: Timothy Bergquist, University of Washington

OP17: Comparative Analysis of Germline Microsatellites in the 1,000 Genomes Project
Presenting Author: Nicholas Kinney, Virginia College of Osteopathic Medicine

OP18: Unbiased Pathway Detection Expands Cancer Pathways
Presenting Author: Chih-Hsu Lin, Baylor College of Medicine

OP19: A systems biology approach to define essential kinases in small cell lung cancer
Presenting Author: Jihye Kim, University of Colorado Denver Anschutz Medical Campus

OP20: Clustering of Protein Conformations using Parallelized Dimensionality Reduction
Presenting Author: Arpita Joshi, University of Massachusetts, Boston

OP21: PredHPI: an integrated web-server platform for the prediction and visualization of host-pathogen interactions
Presenting Author: Rakesh Kaundal, Utah State University

OP22: Searching for translatable alternative splice isoforms in the human proteome
Presenting Author: Maggie Pui Yu Lam, University of Colorado Anschutz Medical Campus

OP23: Text Mining Novel Disease- and Drug-Specific Pathways
Presenting Author: Minh Pham, Baylor College of Medicine

OP24: Optimizing nontuberculous mycobacteria (NTM) de novo genome assemblies for application in clinical case studies
Presenting Author: Sara Kammlade, National Jewish Health

OP25: REAL-neo, a comprehensive neoantigen prediction and prioritization pipeline using tumor sequencing data
Presenting Author: Yesesri Cherukuri, Mayo Clinic

OP26: Measuring chromosome conformation
Presenting Author: Brian Ross, University of Colorado Anschutz Medical Campus

OP27: Modeling the Structure of BioGRID PPI Networks
Presenting Author: Sridevi Maharaj, University of California-Irvine

OP28: Addressing the compositional data problem in sequencing with a novel, robust normalization method
Presenting Author: James St. Pierre, University of Toronto

OP29: A Case Study on the Effects of Noisy, Long-read Correction Approaches on Assembly Contiguity
Presenting Author: Brandon Pickett, Brigham Young University

OP30: Integrative analysis of transcriptomics and proteomics to detect novel protein isoforms from alternatively spliced transcripts induced by SF3B1 spliceosomal mutations
Presenting Author: Kelsey Nassar, University of Colorado Anschutz Medical Campus

OP31: Exploring the Fabric of Breast Cancer Using Gene Sets
Presenting Author: Judith Blake, The Jackson Laboratory

OP32: Measuring and Protecting the Sensitive Linking Information Leakage Across Epigenetic and Transcriptomics Datasets Through Genotype and Assay Prediction
Presenting Author: Arif Harmanci, University of Texas Health Science Center

OP33: Education, Networking and Building Next Generation Prototypes -- Hackathons and Analyzeathons for Bioinformaticians, Biomedical Informaticians and Computational Biologists!
Presenting Author: Ben Busby, NCBI

OP34: Alternative Splicing of Single Cells in Squamous Cell Lung Cancer Premalignancy
Presenting Author: Hyunmin Kim, University of Colorado Anschutz Medical Campus

OP35: Inferring trade-offs in protein folding networks
Presenting Author: Sebastian Pechmann, Université de Montréal

OP36: Fully Bayesian model for non-random missing data in qPCR
Presenting Author: Valeriia Sherina, University of Rochester Medical Center

OP37: Homologous Inter-Domain Segments in Protein Families
Presenting Author: Dylan Barth, University of Nevada Las Vegas

OP38: Integrating extracted relations into existing knowledge bases
Presenting Author: Harrison Pielke-Lombardo, University of Colorado

Op39: The Use of Scientific Ignorance to Drive Literature-Based Discovery in Prenatal Nutrition Across Disciplinary Boundaries
Presenting Author: Mayla Boguslav, University of Colorado Anschutz Medical Campus

OP40: Governance Innovations for Promoting Cross-institutional Electronic Health Data Sharing
Presenting Author: Kari Stephens, University of Washington

ORAL PRESENTATION ABSTRACTS

OP01: Proteomics of natural bacterial isolates powered by deep learning-based de novo identification

Presenting Author: Samuel Payne, Brigham Young University

Co-Author(s):

Joon-Yong Lee, Pacific Northwest National Laboratory; Hugh Mitchell, Pacific Northwest National Laboratory; Meagan Burnet, Pacific Northwest National Laboratory; Sarah Jenson, Pacific Northwest National Laboratory; Eric Merkley, Pacific Northwest National Laboratory Anil Shukla, Pacific Northwest National Laboratory; Ernesto Nakayasu, Pacific Northwest National Laboratory

ABSTRACT: The fundamental task in proteomic mass spectrometry is identifying peptides from their observed spectra. Where protein sequences are known, standard algorithms utilize these to narrow the list of peptide candidates. If protein sequences are unknown, a distinct class of algorithms must interpret spectra de novo. Despite decades of effort on algorithmic constructs and machine learning methods, de novo software tools remain inaccurate when used on environmentally diverse samples. Here we train a deep neural network on 5 million spectra from 55 phylogenetically diverse bacteria. This new model outperforms current methods by 25-100%. The diversity of organisms used for training also improves the generality of the model, and ensures reliable performance regardless of where the sample comes from. Significantly, it also achieves a high accuracy in long peptides which assist in identifying taxa from samples of unknown origin. With the new tool, called Kaiko, we analyze proteomics data from six natural soil isolates for which a proteome database did not exist. Without any sequence information, we correctly identify the taxonomy of these soil microbes as well as annotate thousands of peptide spectra

OP02: A platform for community-scale transcriptome-wide association studies

Presenting Author: YoSon Park, Perelman School of Medicine University of Pennsylvania

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

ABSTRACT: Transcriptome-wide association studies (TWAS) infer causal relationships between genes, phenotypes and tissues using strategies such as 2-sample Mendelian randomization (MR). Such methods largely eliminate the need to access individual-level data and allow openly sharing data and results. Nonetheless, to our knowledge, there are no public platforms automating quality assurance and continuous integration of TWAS results. Consequently, finding, replicating, and validating causal relationships among millions of similar non-causal

relationships remain enormously challenging and are often time- and resource-consuming with many duplicated efforts.

To address this shortcoming, we develop a platform that uses version control software and continuous integration to construct a data resource for the components of TWAS. Community members can contribute additional association studies or methods. We use automated testing to catch formatting mistakes and use pull request functionality to review contributions. We provide a set of tools, available in a Docker container, that perform common downstream analyses using these resources.

Researchers who contribute summary-level datasets substantially increase the impact of their work by making it easy to integrate with complementary datasets. Those who contribute analytical tools will benefit by providing users with numerous off-the-shelf use cases. For this proof-of-concept, we integrate a set of eQTLs provided by the Genotype-Tissue Expression (GTEx) project and a set of curated GWAS summary statistics using 2-sample MR. Our long-term goal for this project is a public community-driven repository where users contribute new summary-level data, download complementary data, and add new analytical methods that enables the field to rapidly translate new studies into actionable findings.

OP03: Harmonizing and Analyzing Clinical Trials Data in the AHA Precision Medicine Platform

Presenting Author: Carsten Goerg, University of Colorado

Co-Author(s):
Christophe Roeder, University of Colorado
Bethany Doran, University of Colorado
Ann Marie Navar, Duke University
Michael Hinterberg, SomaLogic
John Graybeal, Stanford University
Mark Musen, Stanford University
Jennifer Hall, American Heart Association
David Kao, University of Colorado

ABSTRACT: Clinical trials have produced many highly valuable datasets, but their potential to support discovery through meta-analysis has not been fully realized. Answering biomedical questions often requires integrating and harmonizing data from multiple trials to increase statistical power. Due to the lack of supporting computational approaches, this challenging and time-consuming integration process is currently performed manually, which leads to scalability and reproducibility issues. We present a framework and prototype implementation within the cloud-based American Heart Association Precision Medicine Platform as a first step towards addressing this problem. Our framework provides (1) a metadata-driven mapping process from study-specific variables to the OMOP common data

model, (2) a metadata-driven extraction process for creating analysis matrices of harmonized variables, and (3) an interactive visual interface to define and explore cohorts in harmonized studies. To demonstrate our approach, we present a prototype use case that investigates the relationship between blood pressure and mortality in patients treated for hypertension. Using our framework, we harmonized five publicly available NIH-funded studies (ALLHAT, ACCORD, BARI-2D, AIM-HIGH, and TOPCAT), assessed distributions of blood pressure by study, and using harmonized data performed individual patient-data meta analyses to show the statistical relationship between all-cause mortality and systolic blood pressure, for individual studies as well as the aggregated data. We discuss how the cloud-based implementation supports reproducibility as well as transparent co-development between collaborators over time and space. Future work will entail development of a generalized workflow for acquisition and semantic annotation of new datasets based on the CEDAR metadata management system.

OP04: anexVis: visual analytics framework for analysis of RNA expression

Presenting Author: Diem-Trang Tran, University of Utah

Co-Author(s):
Tian Zhang, University of Utah
Ryan Stutsman, University of Utah
Matthew Might, University of Alabama at Birmingham
Umesh Desai, Virginia Commonwealth University
Balagurunathan Kuberan, University of Utah

ABSTRACT: Although RNA expression data are accumulating at a remarkable speed, gaining insights from them still requires laborious analyses, which hinder many biological and biomedical researchers. We introduce a visual analytics framework that applies several well-known visualization techniques to leverage understanding of an RNA expression dataset. Our analyses on glycosaminoglycan-related genes have demonstrated the broad application of this tool, anexVis (analysis of RNA expression), to advance the understanding of tissue-specific glycosaminoglycan regulation and functions, and potentially other biological pathways.

The application is publicly accessible at https://anexvis.chpc.utah.edu/, source codes deposited on GitHub.

OP05: The characterization of different cell types using the Benford law

Presenting Author: Sne Morag, Ariel University

Co-Author(s):
Mali Salmon-Divon, Ariel University

ABSTRACT: Abstract publication declined

OP06: Use of metadata and Bag-of-words to map measurements across observational study data

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

Co-Author(s):
Tiffany Callahan, University of Colorado Anschutz Medical Campus
Sonia Leach, University of Colorado Anschutz Medical Campus
David Kao, University of Colorado Anschutz Medical Campus

ABSTRACT: Data integration is an important strategy for validating research results or increasing sample size in biomedical research. Integration is made challenging by metadata and data differences between studies, and is often done manually by a clinical expert for a highly select set of measurements. Unfortunately, this process is rarely documented, and when it is, the details are not accessible, interoperable, or reusable. We explored the utility of using bag-of-words, an information retrieval model, to map medical conditions, characteristics, and lifestyle measurements among multiple studies such as diabetes, age, blood pressure, or alcohol intake. We hypothesized applying cosine similarity to features extracted as a bag-of-words model from observational study measurement annotations would yield accurate recommendations for mapping measurements within and between studies and increase scalability compared to manual mapping. Each measurement's metadata, including descriptions, units, and value-coding, were extracted and then combined for all 105,611 measurements in four cardiovascular-health observational studies. The measurement's combined metadata was input to the bag-of-words model. Cosine similarity of word vectors was used to score similarity between measurement pairs. The highest scoring matches for each measurement were compared to 612 unique expert-vetted, manual mappings. Among the vetted measurement pairings, 99.8% had the correct mapping in the top-10 scored matches, 92.5% had the correct mapping in the top-5, and 55.7% had the correct mapping as the top score. This approach provides a scalable method for recommending measurement mappings in observational study data. Next steps include incorporating additional metadata such as measurement type or a synonyms dictionary for concept recognition.

OP07: ExtRamp: A novel algorithm for extracting the ramp sequence based on the tRNA adaptation index or relative codon adaptiveness

Presenting Author: Justin Miller, Brigham Young University

Co-Author(s):
Logan Brase, Brigham Young University
Perry Ridge, Brigham Young University

ABSTRACT: Different species, genes, and locations within genes use different codons to fine-tune gene expression. Within genes, the ramp sequence assists in ribosome spacing and decreases downstream

collisions by incorporating slowly-translated codons at the beginning of a gene. Although previously reported as occurring in some species, no previous attempt at extracting the ramp sequence from specific genes has been published. We present ExtRamp, a software package that quickly extracts ramp sequences from any species using the tRNA adaptation index or relative codon adaptiveness. Different filters facilitate the analysis of codon efficiency and enable researchers to identify genes with a ramp sequence. We validate the existence of a ramp sequence in most species by running ExtRamp on 229,742,339 genes across 23,428 species. We evaluate differences in reported ramp sequences when we use different parameters. Using the strictest ramp sequence cut-off, we show that across most taxonomic groups, ramp sequences are approximately 20-40 codons long and occur in about 10% of gene sequences. We also show that as gene expression increases, more ramp sequences are identified in Drosophila melanogaster. We provide a framework for performing this analysis on other species and present our algorithm at https://github.com/ridgelab/ExtRamp.

OP08: Using machine learning algorithms for classification of medulloblastoma subgroups based on gene expression data

Presenting Author: Sivan Gershanov, Ariel University

Co-Author(s):
Igor Vainer, Ariel University
Helen Toledano, Schneider Children's Medical Center of Israel
Albert Pinhasov, Ariel University
Nitza Goldenberg-Cohen, Bnai Zion Medical Center
Mali Salmon-Divon, Ariel University

ABSTRACT: Medulloblastoma (MB), the commonest malignant pediatric brain tumor, is divided into four molecular subgroups: WNT, SHH, Group 3 and Group 4. Clinical practice and treatment design are becoming subgroup-specific. Nowadays clinicians use a 22-gene signature set to diagnose the subgroups. While WNT and SHH subgroups are well-defined differentiating Group 3 from Group 4 is less obvious.

The aim of this study is to improve the diagnosis process in the clinic by identifying the most efficient list of biomarkers for accurate, fast and cost-effective MB subgroup classification.

We tested five machine learning based algorithms, four are well known and one is a novel method we developed. We applied them on a public microarray expression data set and compared their performance to that of the known 22-gene set.

Both decision tree and decision rules resulted in a reduced set with similar accuracy to the 22-gene set. Random forest and SVM-SMO methods showed improved performance, without applying feature-

selection. When implementing our novel SARC (SVM Attributes Ranking and Combinations) classifier, allowing feature-selection, the resulted accuracy level was the highest and better than using the 22-gene set as input. The number of attributes in the best-performing combinations range from 13 to 32, including known MB related genes such as WIF1, NPR3 and GRM8, along with LOC440173 a long non-coding RNA.

To summarize we identified sets of attributes that have the potential to improve MB subgroup diagnosis. Broad clinical use of this classification may accelerate the design of patient's specific targeted therapies and optimize clinical decision.

OP09: A human disease network from gene-publication relationships on PubMed

Presenting Author: Edward Lau, Stanford University

Co-Author(s):
Cody Thomas, University of Colorado AMC
Maggie Pui Yu Lam, University of Colorado AMC

ABSTRACT: Human diseases can be represented as a network connecting similar disorders based on their shared phenotypic and molecular characterizations. Network analysis of disease-disease relationships can yield insights into important biological processes and pathogenic pathways. We recently described a method to determine the semantic similarity between a gene or protein and the literature publications related to a disease, by combining PubMed web queries and curated/text-mined annotations of gene-PMID links from NCBI. We devised a weighted co-publication distance metric to score gene-disease co-occurrences in PubMed, where genes with many non-specific publications are down-ranked whereas recent and high-impact publications are given more weight. We show that this method outperforms existing bibliometric analysis in predicting benchmark gene lists of disease terms. Using this method, we have now compiled significant protein lists from over 20,000 human disease or disease phenotype terms from three standardized vocabularies, namely Disease Ontology (DO), Human Phenotype Ontology (HPO), and Pathway Ontology (PWO). We find that disease terms are associated with specific popular protein lists that inform on protein-disease relationships. The PubMed-based disease network recapitulates several known properties from previous "diseasomes" constructed from OMIM or phenotypic similarity data (e.g., Barabási 2007), including the centrality of metabolic diseases and clustering of related diseases around high-level hub terms. We discuss applications for the disease network, including (i) finding commonly associated diseases from a list of differentially expressed genes in a RNA-seq experiment, and (ii) using gene-disease relationship to predict hidden disease genes in a particular disease

OP10: Transcriptome analysis of cancer adjacent normal tissues reveal genes co-expressed with LINE elements

Presenting Author: Mira Han, University of Nevada Las Vegas

Co-Author(s):
Nicky Chung, University of Nevada, Las Vegas
G.M. Jonaid, University of Nevada, Las Vegas
Sophia Quinton, University of Nevada, Las Vegas
Austin Ross, University of Nevada, Las Vegas

ABSTRACT: Despite the long-held assumption that transposons are normally only expressed in the germ-line, recent evidence shows that transcripts of LINE sequences are frequently found in the somatic cells. However, the extent of variation in LINE transcript levels across different tissues and different individuals, and the genes and pathways that are co-expressed with LINEs are unknown. Here we report the variation in LINE transcript levels across tissues and between individuals observed in the normal tissues collected for The Cancer Genome Atlas. Mitochondrial genes and ribosomal protein genes were enriched among the genes that showed negative correlation with L1HS in transcript level. We hypothesize that oxidative stress is the factor that leads to both repressed mitochondrial transcription and LINE over-expression. KRAB zinc finger proteins (KZFPs) were enriched among the transcripts positively correlated with older LINE families. The correlation between transcripts of individual LINE loci and individual KZFPs showed highly tissue-specific patterns. There was also a significant enrichment of the corresponding KZFP's binding motif in the sequences of the correlated LINE loci, among KZFP-LINE locus pairs that showed co-expression. These results support the KZFP-LINE interactions previously identified through ChIP-seq, and provide information on the in vivo tissue context of the interaction.

OP11: Highly accurate computational characterization of protein kinase family-specific phosphorylation sites

Presenting Author: Chen Li, ETH Zürich

Co-Author(s):
Fuyi Li, Monash University
Tatiana Marquez-Lago, University of Alabama at Birmingham
Andre Leier, University of Alabama at Birmingham
Tatsuya Akutsu, Kyoto University
Anthony Purcell, Monash University
A. Ian Smith, Monash University
Trevor Lithgow, Monash University
Roger Daly, Monash University
Jiangning Song, Monash University
Kuo-Chen Chou, Gordon Life Science Institute

ABSTRACT: Kinase-regulated phosphorylation is a ubiquitous type of post-translational modification (PTM) in both eukaryotic and prokaryotic cells. Numerous experimental studies have demonstrated that phosphorylation is involved in regulation of a variety of fundamental cellular processes, such as protein-protein interaction,

protein degradation, signal transduction and signaling pathways. It also has been revealed that signaling defects caused by aberrant phosphorylation are highly associated with a variety of human diseases, especially cancers. In light of this, a number of computational methods aiming to accurately predict protein kinase family-specific or kinase-specific phosphorylation sites have been established, thereby facilitating phosphoproteomic data analysis. In this work, we present Quokka, a novel bioinformatics tool that allows users to rapidly and accurately identify human kinase family-regulated phosphorylation sites. Quokka was developed by using a variety of sequence scoring functions combined with an optimized logistic regression algorithm. We evaluated Quokka based on well-prepared up-to-date benchmark and independent test datasets, curated from a variety of databases. The independent test demonstrates that Quokka improves the prediction performance compared with state-of-the-art computational tools for phosphorylation prediction. In summary, our tool provides users with high-quality predicted human phosphorylation sites for hypothesis generation and biological validation.

OP12: ORCHID: a method for detecting short-range chromatin interactions in high-resolution 5C and Hi-C datasets

Presenting Author: Fei Ji, Massachusetts general hospital

Co-Author(s):
Sharmistha Kundu, Massachusetts general hospital
Robert Kingston, Massachusetts general hospital
Ruslan Sadreyev, Massachusetts general hospital

ABSTRACT: The chromatin interaction assays 5C and Hi-C are robust techniques to investigate spatial organization of the genome by capturing interaction frequencies between genomic loci. Although 5C and Hi-C resolution is theoretically restricted only by the length of digested DNA fragments (1Kb-4Kb), intrinsic stochastic noise and high frequencies of background interactions at the distances below 100 Kbp present a significant challenge to understanding short-distance chromatin organization. Here we present the shOrt Range Chromosomal Interaction Detection method (ORCHID) for a comprehensive high-resolution analysis of chromatin interactions in 5C and Hi-C experiments. This method includes background correction of raw interaction frequencies for individual primers or genomic bins, empirical correction for distance dependency of background noise, and detection of areas of significant interactions. When applied to publicly available datasets, ORCHID improves the identification of small (20-200Kb) interaction domains. Unlike larger classic TADs, these chromatin domains are often specific to cell type and functional state of the genomic region. In addition to the expected associations (e.g. with CTCF, cohesin, and mediator complexes), these domains show significant associations with other DNA-binding proteins.

An important subtype of these small domains is fully covered and controlled by Polycomb Repressive Complex 1 (PRC1), which mediates transcriptional repression of many key developmental genes. As a separate unexpected example of a potential new mode of regulating chromatin interactions, the binding of RING1B, an essential subunit of the PRC1 complex, is also enriched near domain boundaries at the focused loci that do not necessarily correspond to repressed promoters.

OP13: Using Adversarial Deep Neural Networks to Remove Nonlinear Batch Effects from Expression Data

Presenting Author: Jonathan Dayton, Brigham Young University
Co-Author(s): Stephen Piccolo, Brigham Young University

ABSTRACT: Batch effects and other confounding effects can skew research results when working with quantitative molecular data (e.g. RNA-Seq). Most existing batch adjustment methods only take into account linear effects, but modern analysis tools such as machine learning can still identify and be influenced by nonlinear batch effects, even after linear effects have been removed. We introduce Confounded, a method that uses adversarial deep neural networks to identify and remove linear and nonlinear batch effects. Confounded is composed of 1) a discriminator designed to detect confounding effects and 2) an autoencoder designed to replicate the input data while identifying and removing confounding effects in order to fool the discriminator. Once the data have been faithfully reproduced and the confounders have been removed, the adjusted data are output for use in analysis. We have tested Confounded on image vectors with artificial nonlinear batch effects. We show that Confounded removes these batch effects more effectively than ComBat, the most commonly used batch-effect adjustment method, while still retaining most of the true signal as measured by several classification algorithms. We are also validating Confounded with molecular datasets, both with artificial and real batch effects, and publishing our software to enable other scientists to use Confounded in their bioinformatics pipelines. In addition to batch correction, Confounded may also be used for data integration between multiple databases or between different technologies (e.g. microarray and RNA-Seq) or for removing general confounding effects from data.

OP14: Med2Mech: Neural-Symbolic Representation of Molecular Mechanisms Underlying Pediatric Disease

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

Co-Author(s):
Adrianne Stefanski, University of Colorado Denver Anschutz Medical Campus
Michael Kahn, University of Colorado Denver Anschutz Medical Campus
Lawrence Hunter, University of Colorado Denver Anschutz Medical Campus

ABSTRACT: Subphenotyping aims to cluster patients with a particular disease into clinically distinct groups. Genomic and related molecular signatures, such as mRNA expression, have shown great promise for subphenotyping, but such molecular data is not and will not be available for most patients. Here, we present Med2Mech, a method for linking knowledge from generalized molecular data to specific patients' electronic patient records, and demonstrate its utility for subphenotyping. We hypothesized that integrating knowledge of molecular mechanisms with patient data would improve subphenotype classification. Med2Mech employs neural-symbolic representation learning to generate patient-level embeddings of molecular mechanisms using publicly available biomedical knowledge. Using clinical terminologies and biomedical ontologies, the mechanisms can then be mapped to patient data at scale. Med2Mech was developed and tested using clinical data from a subset of rare disease and other similarly medically complex patients from the Children's Hospital Colorado. A one-vs-the-rest multiclass classification strategy was used to evaluate the discriminatory ability of embeddings generated using Med2Mech versus only clinical data. Clinical embeddings were built for 2,464 rare disease and 10,000 similarly complex patients using 6,382 conditions, 2,334 medications, and 272 labs. Molecular mechanism embeddings were generated from a knowledge graph (116,158 nodes and 3,593,567 edges) built with 23,776 genes, 3,744 diseases, 49,185 gene ontology concepts, 13,159 phenotypes, 11,124 pathways, and 15,019 drugs. For classification, the molecular mechanism embeddings (precision=0.95, recall=0.94) out-performed all parameterizations of clinical embeddings (precision=0.83, recall=0.82). The Med2Mech representation of patient data improves subphenotype classification relative to standard subphenotyping approaches by incorporating knowledge of molecular mechanisms.

OP15: A Machine Learning Classifier for Assigning Individual Patients with Systemic Sclerosis to Intrinsic Molecular Subsets

Presenting Author: Jennifer Franks, Geisel School of Medicine at Dartmouth

Co-Author(s):
Viktor Martyanov, Geisel School of Medicine at Dartmouth
Guoshuai Cai, Arnold School of Public Health at University of South Carolina
Yue Wang, Geisel School of Medicine at Dartmouth
Tammara Wood, Geisel School of Medicine at Dartmouth
Michael Whitfield, Geisel School of Medicine at Dartmouth

ABSTRACT: High-throughput gene expression profiling of skin biopsies from patients with systemic sclerosis (SSc) has identified four "intrinsic" gene expression subsets (inflammatory, fibroproliferative, normal-like, limited) conserved across multiple cohorts and tissues. In order to characterize patients in clinical trials or for diagnostic purposes, supervised methods that can classify single samples are required.

Three gene expression cohorts were curated and merged for the training dataset. Supervised machine learning algorithms were trained using repeated three-fold cross-validation. We performed external validation using three additional datasets, including one generated by an independent laboratory on a different microarray platform. WGCNA and g:Profiler were used to identify and functionally characterize gene modules associated with the intrinsic subsets.

The final model, a multinomial elastic net, performed with average classification accuracy of 88.1%. All intrinsic subsets were classified with high sensitivity and specificity, particularly inflammatory (83.3%, 95.8%) and fibroproliferative (89.7%, 94.1%). In external validation, the classifier achieved an average accuracy of 85.4%. In a re-analysis of GSE58095, we identified subgroups of patients that represented the canonical inflammatory, fibroproliferative, and normal-like subsets. Inflammatory gene modules showed upregulated biological processes including inflammatory response, lymphocyte activation, and stress response. Similarly, fibroproliferative gene modules were enriched in cell cycle processes.

We developed an accurate, reliable classifier for SSc intrinsic subsets, trained and tested on 427 skin biopsies from 213 individuals. Our method provides a robust approach for assigning samples to intrinsic gene expression subsets and can be used to aid clinical decision-making and interpretation for SSc patients and in clinical trials.

OP16: A Data Quality Testing Tool for Cross-institutional OMOP Electronic Health Record Data Repositories

Presenting Author: Timothy Bergquist, University of Washington

Co-Author(s):
Hossein Estiri, Harvard University
Justin Prosser, University of Washinton
Adam Wilcox, University of Washington
Kari Stephens, University of Washington

ABSTRACT: Data quality testing is critical to cross-institutional data sharing, a key component of health innovations produced through translational research. Harmonizing electronic health record (EHR) data is a resource intensive strategy used in many data sharing efforts, involving extraction, translation, and loading activities that can perpetuate and add to pre-existing data quality issues. Yet, we lack standards and tools for testing the quality of datasets produced through these complex harmonization processes. Given its large scale adoption, the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) standard is primed as a front running CDM to target establishing a standard set of executable data quality tests to support cross institutional data sharing. We adapted a prototype tool, DQe-c, to OMOP CDM V5 with scalability across database platforms. Namely it examines completeness in all data tables and columns, calculates the percentage of patients who have key clinical variables present (e.g., blood pressure, height), detects the presence of orphan keys (i.e., foreign keys that are not present in their reference table), reports on the size of the databases, and assesses conformance to the standard. All test results are produced as data visualizations in a single HTML dashboard. This prototype is being explored for use in multiple data sharing pilot projects supported by the Clinical Translational Science Award (CTSA) Program Data to Health (CD2H) Coordinating Center, with an aim towards configuring a robust set of completeness, conformance, and plausibility tests that confirm OMOP CDM V5 datasets are fit for cross-institutional data sharing.

OP17: Comparative Analysis of Germline Microsatellites in the 1,000 Genomes Project

Presenting Author: Nicholas Kinney, Virginia College of Osteopathic Medicine

Co-Author(s):
Kyle Titus-Glover, Virginia Tech
Jonathan Wren, Oklahoma Medical Research Foundation
Robin Varghese, Edward Via College of Osteopathic Medicine
Pawel Michalak, Edward Via College of Osteopathic Medicine
Han Liao, Virginia Tech
Ramu Anandakrishnan, Edward Via College of Osteopathic Medicine
Arichanah Pulenthiran, Edward Via College of Osteopathic Medicine
Lin Kang, Edward Via College of Osteopathic Medicine
Harold Garner, Edward Via College of Osteopathic Medicine

ABSTRACT: Microsatellites are regions of DNA characterized by short – one to six base pair – motifs repeated in tandem to form an array.

Over 600,000 unique microsatellites exist in the human genome embedded in gene introns, gene exons, and regulatory regions. Indeed they are well established as an important source of genetic variation. A number of databases provide searchable interfaces to microsatellites within the human reference genome; however, none provide data on actual polymorphism rates among and within human populations. We introduce the Comparative Analysis of Germline Microsatellites (CAGm) Database. The database is designed to assist with future studies of germline microsatellites and enhance our understanding of human genetic variation. Samples can be easily grouped by population, ethnicity, and gender. Microsatellites can be searched by gene, functional element, and location. Users can query genotypes, view multiple sequence alignments, and easily download data for further analysis. The database has a wide range of additional capabilities. Database content is fully described with examples and future directions are discussed. The database is freely available at http://www.cagmdb.org/.

OP18: Unbiased Pathway Detection Expands Cancer Pathways

Presenting Author: Chih-Hsu Lin, Baylor College of Medicine

Co-Author(s):
Stephen Wilson, Baylor College of Medicine
Teng-Kuei Hsu, Baylor College of Medicine
Minh Pham, Baylor College of Medicine
Olivier Lichtarge, Baylor College of Medicine

ABSTRACT: Pathways are a type of functional gene group and they help to understand biological systems by representing how signals are transmitted/received and which genes/proteins interact. Conventionally, domain experts manually annotate pathways based on the literature. Thus, the unbiased detection of functional gene groups solely based on the gene-gene interaction network structure may provide novel insights. Here, we hypothesized that gene members in a functional gene group interact within the group more than outside the group. We developed Recursive Louvain algorithm to detect communities (i.e., clustered gene groups) on a human protein-protein interaction network. 85.2 % of the communities overlapped with known functional pathways and disease pathways significantly compared to a random gene group control, whereas 452 communities didn't and may be potentially novel functional gene groups. In addition, variants of genes overlapping with communities are more likely to be pathogenic in ClinVar and have high evolutionary impact quantified by Evolutionary Action (chi-squared test $p << 0.0001$). As a case study in head and neck cancer, we found the RNA-seq profiles of 10 communities could separate survival by K-means clustering significantly (log-rank $q <= 0.1$). Also, those 10 communities are linked to cancer hallmarks. More importantly, one community related to cell adhesion could stratify patient survival independent

of clinical data and immune response (Cox multivariate analysis $q = 0.022$). In conclusion, the communities recover known functional and disease pathways, and could be used as cancer survival predictors. This study will help understanding of cancer pathways and provide biomarkers for cancer patients.

OP19: A systems biology approach to define essential kinases in small cell lung cancer

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

Co-Author(s):
Daniel Foster, National Jewish Health
Rangnath Mishra, National Jewish Health
James Finigan, National Jewish Health
Jeffrey Kern, National Jewish Health
Aik Choon Tan, University of Colorado Denver

ABSTRACT: Small cell lung cancer (SCLC) is a deadly cancer where its five-year survival rate is $< 7\%$ and kills approximately 30,000 lives this year. Treatment of SCLC using the chemotherapy combination of cisplatin and etoposide with radiation therapy has not changed in almost 30 years. Therefore, novel therapies are needed for this disease. Building on the role of kinases and their regulation of cell growth and survival, we hypothesized that kinases regulate cell survival pathways in SCLC (essential kinases) and they may be effective targets as novel monotherapy, or act synergistically with standard chemotherapy, and improve therapeutic outcome. To test this hypothesis, we employed a systems biology approach to identify essential kinases in SCLC. We performed in vivo kinome-wide screening using an shRNA library targeting human kinases on seven chemonaïve SCLC patient derived xenografts (PDX). We developed a suite of bioinformatics tools to deconvolute the kinome screening data, and identified 23 essential kinases found in two or more PDX models. The top essential kinases were RET, MTOR and ATM. We connected these kinases to our drug database to identify specific inhibitors as potential therapy and performed in vitro and in vivo validation of their efficacy. Notably, monotherapy with a small molecule inhibitor targeting mTOR significantly reduced SCLC tumor growth in vivo proving mTOR's essential kinase function. In addition, mTOR inhibition synergized with standard chemotherapy to significantly augment tumor responses in SCLC PDX models. These results warrant the further investigation of MTOR inhibitors combined with chemotherapy as novel treatment for SCLC.

OP20: Clustering of Protein Conformations using Parallelized Dimensionality Reduction

Presenting Author: Arpita Joshi, University of Massachusetts, Boston
Co-Author(s): Nurit Haspel, Umass Boston

ABSTRACT: Analyzing the conformational pathways that a macromolecule undergoes is imperative to understanding its function and dynamics. We present a combination of techniques to sample the conformational landscape of proteins better and faster. Datasets representing these landscapes of protein folding and binding are complex and high dimensional. Therefore, there is a need for dimensionality reduction methods that best preserve the variance in the data, and facilitate the analysis of the data. The crux of this work lies in the way this is done. We start with a non-linear dimensionality reduction technique, Isomap, which has been shown to produce better results than linear dimensionality reduction in approximating the complex niceties of protein folding. However, the algorithm is computationally intensive for large proteins or a large number of samples (samples here refer to the various conformations that are used to ascertain the pathway between two distinctively different structures of a protein). We present a parallel algorithm written in C, using OpenMP, with a speed-up of approximately twice. The results obtained are coherent with the ones obtained using sequential Isomap. Our method uses a distance function to calculate the distance between the points that in turn measures the similarity between the conformations that each of these points represent. The output is a lower-dimensional projection that can be used later for purposes of visualization and analysis. A proof of quantitative validation comes with the least RMSD computation for the two embeddings. The algorithm also makes efficient use of the available memory.

OP21: PredHPI: an integrated web-server platform for the prediction and visualization of host-pathogen interactions

Presenting Author: Rakesh Kaundal, Utah State University
Co-Author(s): Cristian Loaiza, Utah State University

ABSTRACT: Understanding the mechanisms underlying infectious diseases is fundamental to develop prevention strategies. Host-pathogen interactions, which includes from the initial invasion of host cells by the pathogen through the proliferation of the pathogen in their host, have been studied to find potential genomic targets for the development of novel drugs, vaccines, and other therapeutics. Few in silico prediction methods have been developed to infer novel host-pathogen interactions, however, there is no single framework which combines those approaches to produce and visualize a comprehensive analysis of host-pathogen interactions. We present a web server

platform named PredHPI available at http://bioinfo.usu.edu/PredHPI/. PredHPI is composed of independent sequence-based tools for the prediction of host-pathogen interactions. The Interolog module, including some of the IMEX databases (HPIDB, MINT, DIP, BioGRID and IntAct), provides three comparison flavors using the BLAST homology results (best-match, ranked-based and generalized). The Domain module, which performs the predictions of the domains, using Pfam and HMMer, and the interactions using the 3DID and IDDI databases. And the GO Similarity module which uses some of the Bioconductor species databases to calculate similarities using GOSemSim R package of the GO terms detected using InterProScan. PredHPI incorporates the functionally to visualize the resulting interaction networks plus the integration of several databases with enriched information about the proteins involved in it. To our knowledge, PredHPI is the first system to build and visualize interaction networks from sequence-based methods as well as curated databases. We hope that our prediction tool will be useful for researchers studying infectious diseases.

OP22: Searching for translatable alternative splice isoforms in the human proteome

Presenting Author: Maggie Pui Yu Lam, University of Colorado Anschutz Medical Campus
Co-Author(s): Edward Lau, Stanford University

ABSTRACT: The human genome contains over 100,000 alternative splice isoform transcripts, but the biological functions of most isoform transcripts remain unknown and many are not translated into mature proteins. A full appreciation of the biological significance of alternative splicing therefore requires knowledge of isoforms at the protein level, such as using mass spectrometry-based proteomics. One described is to perform in-silico translation of alternative transcripts, and then to use the resulting custom FASTA protein sequence databases with a database search engine for protein identification in shotgun proteomics. However, challenges remain as custom protein databases often contain many sequences that are in fact not translated as proteins inside the cell, thus contributing to a high false discovery rate in proteomics experiments.

We describe here a computational workflow and software to generate custom protein databases of alternative isoform sequences using RNA-seq data as input. The workflow is designed with the explicit goal to minimize untranslated sequences to rein in false positives. To evaluate its performance, we processed public RNA sequencing data from ENCODE to build custom FASTA databases for 10 human tissues (adrenal gland, colon, esophagus, heart, lung, liver, ovary, pancreas, prostate, testis). We applied the databases to identify unique splice junction peptides from public mass spectrometry data of the same human tissues on ProteomeXchange. We identified 1,984 protein isoforms including 345 unique splice-specific peptides not currently

documented in common proteomics databases. We suggest that the described proteotranscriptomics approach may help reveal previously unidentified alternative isoforms, and aid in the study of alternative splicing.

OP23: Text Mining Novel Disease- and Drug-Specific Pathways

Presenting Author: Minh Pham, Baylor College of Medicine

Co-Author(s):
Stephen Wilson, Baylor College of Medicine
Chih-Hsu Lin, Baylor College of Medicine
Olivier Lichtarge, Baylor College of Medicine

ABSTRACT: In response to the exponential growth of scientific publications, text mining is increasingly used to extract biological pathways and processes. Though multiple tools explore individual connections between genes, diseases, and drugs, not many extensively examine contextual biological pathways for specific drugs and diseases. In this study, we extracted more than 3,000 functional gene groups for specific diseases and drugs by applying a community detection algorithm to a literature network. The network aggregated co-occurrences of Medical Subject Headings (MeSH) terms for genes, diseases, and drugs in publications. The detected literature communities were groups of highly associated genes, diseases, and drugs. The communities significantly captured genetic knowledge of canonical pathways and recovered future pathways in time-stamped experiments. Furthermore, the disease- and drug-specific communities recapitulated known pathways for those given diseases and drugs. In addition, diseases in same communities had high comorbidity with each other and drugs in same communities shared great numbers of side effects, suggesting that they shared mechanisms. Indeed, the communities robustly recovered mutual targets for drugs (AUROC = 0.75) and shared pathogenic genes for diseases (AUROC = 0.82). These data show that the literature communities not only represented known biological processes but also suggested novel disease- and drug-specific mechanisms, facilitating disease gene discovery and drug repurposing.

OP24: Optimizing nontuberculous mycobacteria (NTM) de novo genome assemblies for application in clinical case studies

Presenting Author: Sara Kammlade, National Jewish Health

Co-Author(s):
Nabeeh Hasan, National Jewish Health
L. Elaine Epperson, National Jewish Health
Michael Strong, National Jewish Health
Rebecca Davidson, National Jewish Health

ABSTRACT: To enable studies related to bacterial acquisition and clinical infections of nontuberculous mycobacteria (NTM), we developed a

standardized bioinformatic analysis pipeline to process sequenced bacterial isolates from paired-end Illumina reads to fully annotated genomes and a companion PostgreSQL genomic database. Our NTM Genomes Database includes 1200+ isolates from 20 different NTM species which have been processed through our automated and optimized steps for read-trimming, de novo genome assembly, species identification using the average nucleotide identity (ANI) method, contig-ordering against a reference genome, and comprehensive annotation of genomic features. To optimize genome assembly methods and explore the theoretical potential of assembling complete genomes in the context of NTM, we performed experiments testing different parameter combinations in Skewer, SPAdes, and Unicycler on sequences from Illumina MiSeq (2x300bp) and HiSeq (2x250bp) platforms as well as on synthetic reads of varying read lengths and sequencing depths derived from published complete genomes. Assemblies from Illumina data revealed a negative effect of high GC content on assembly quality as measured by NG50. SPAdes and Unicycler yielded similar quality assemblies with Unicycler yielding fewer small (<1Kbp) contigs. From the synthetic reads we found diminished returns on NG50 improvement beyond 25x coverage at 250bp, and failed to assemble a single contig genome using 50Kbp reads at 60x coverage. Using our high quality genomes we are able to identify core and accessory genes and investigate clinically relevant genotype-phenotype relationships. As an example, we will share findings from a case study of bacterial genomic evolution during a long-term pulmonary infection.

OP25: REAL-neo, a comprehensive neoantigen prediction and prioritization pipeline using tumor sequencing data

Presenting Author: Yesesri Cherukuri, Mayo Clinic

Co-Author(s):
Yingxue Ren, Mayo Clinic
Vivekananda Sarangi, Mayo Clinic
Yi Lin, Mayo Clinic
Keith Knutson, Mayo Clinic
Yan Asmann, Mayo Clinic

ABSTRACT: Neoantigens are immunogenic peptides from tumor-specific somatic mutations. The expressed neoantigens can be presented to class-I or class-II MHC molecules and induce robust and enduring anti-tumor T-cell responses. Recent studies have demonstrated the great potential of personalized neoantigen vaccines as a new type of immunotherapy.

In general, identification of neoantigens from tumor sequencing data includes the following steps: (1) call somatic mutations from tumor genomic sequencing data; (2) derive neo-peptide sequences containing somatic mutations; (3) predict binding affinities between neo-peptides and MHC molecules. However, the current bioinformatics practices

ignore transcript splicing isoforms, expressed fusion gene products, and often times only focus on non-synonymous single nucleotide mutations but not frame-shifting INDELs. In addition, the MHC binding affinity prediction mainly focuses on class-I but not class-II MHC molecules. Furthermore, studies have shown that substantial numbers of neo-peptides predicted to have low MHC affinities are actually immunogenic, suggesting the necessity of alternative approaches for neoantigen discovery. Finally, nominated neoantigens need to be further filtered to ensure tumor specificity.

We have improved and optimized each step of the bioinformatics workflow for neoantigen identification from tumor sequencing data to address the complexity and current limitations of the process.

OP26: Measuring chromosome conformation

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

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

ABSTRACT: The in-vivo conformation of chromosomes is an outstanding unsolved problem in structural biology. Most structural information is currently inferred indirectly from Hi-C data, as direct measurements of chromosomal positioning have not been possible for more than a handful of genetic loci. We have previously demonstrated a computational method for scaling direct positioning measurements up to the whole-chromosome scale. Here we present our latest results from simulations and experiments.

OP27: Modeling the Structure of BioGRID PPI Networks

Presenting Author: Sridevi Maharaj, University of California-Irvine

Co-Author(s):
Pedro Silva, University of California-Irvine
Zarin Ohiba, University of California-Irvine
Wayne Hayes, University of California-Irvine

ABSTRACT: Protein-protein interaction (PPI) networks are being continuously updated but are still incomplete, sparse, and have false positives and negatives. Amongst the heuristics employed to describe network topology, graphlets have emerged successful in quantifying local structure of biological networks. Some studies analyzing the graphlet degree distributions and relative graphlet frequency, found Geometric (GEO) networks to be a reasonable basis for modeling PPI networks. However, all extensive studies to model PPI networks as a whole utilized older PPI network data. While there are numerous techniques through which PPI data can be curated, in this study, we re-evaluate these models on the newest PPI data available from BioGRID for the following nine species: AThaliana, CElegans, DMelanogaster, EColi, HSapiens, MMusculus, RNorvegicus, SCerevisiae, and SPombe.

To the best of our knowledge, this has not yet been performed, as the data is relatively new. We compare the graphlet distributions of several models to distributions of the updated networks and analyze their fit using several measures that have been shown to be suitable for measuring network distances (or similarities): RGFD, GDDA, Graphlet Kernel, and GCD. Despite minor behavioral differences amongst the comparison measures, we find that other than the Sticky model, the Scale-Free Gene Duplication and Divergence (SFGD) and Scale-Free (SF) models unanimously outperform other traditional models (including GEO and GEOGD) in matching the structure of these 9 BioGRID PPI networks. We further corroborate these results using machine learning classifiers to categorize each species as a network model and visualize these results using t-SNE plots. *

OP28: Addressing the compositional data problem in sequencing with a novel, robust normalization method

Presenting Author: James St. Pierre, University of Toronto
Co-Author(s): John Parkinson, Hospital for Sick Children, Toronto

ABSTRACT: A problem that faces high-throughput sequencing datasets is that raw sequencing data is semi-quantitative due to the random sampling procedure of the sequencing process itself. The raw counts produced only give relative abundances of various genes and must be appropriately normalized to give an approximation of the absolute abundances of genes in the samples. This 'compositional data problem' in sequencing is especially apparent in the microbiome field. Normalization methods developed for RNA-seq data have been shown to fail when used on 16S microbiome sequencing data, leading to inflated false discovery rates when performing differential abundance analysis. Moreover, the effectiveness of these normalization techniques when used on metagenomics and metatranscriptomics data has yet to be systematically evaluated. We present a novel normalization method that shows improved performance over previous methods (DESeq2, edgeR, and metagenomeSeq) when applied to simulated sequencing data. All current normalization methods have the statistical assumption that most genes (or taxa) are not differentially abundant between experimental groups. The new technique does not have this assumption and is the only method that successfully controls false positive rates during differential abundance testing on a simulated 16S dataset where 50% of taxa were set to be differentially abundant. Even ANCOM and ALDEx2, two compositional data analysis tools previously shown to be more robust than other methods, are shown here to have inflated false positive rates. This new normalization method will be an asset to microbiome researchers, leading to more robust discoveries.

OP29: A Case Study on the Effects of Noisy, Long-read Correction Approaches on Assembly Contiguity

Presenting Author: Brandon Pickett, Brigham Young University

Co-Author(s):
Justin Miller, Brigham Young University
Perry Ridge, Brigham Young University

ABSTRACT: Third-generation sequencing technologies are advancing our ability to sequence increasingly long DNA sequences in a high-throughput manner. Pacific Biosciences (PacBio) Single-molecule, Real-time (SMRT) sequencing and Oxford Nanopore Technologies (ONT) nanopore sequencing routinely produce raw sequencing reads averaging 20-30kbp in length. Maximum read lengths have, in some cases, exceeded 100kbp. Unfortunately, these long reads are expensive to generate and have a high error rate (10-15%) when compared with Illumina short reads (1%). The limitation on assembly from high error rates can be mitigated by (a) co-assembling high-error, long reads with low-error, short reads (e.g., MaSuRCA) or (b) correcting the errors prior to assembly. Pre-assembly error correction typically happens by either (a) self-correction or (b) hybrid correction. Self-correction requires increased sequencing depth (and thus expense) and can be done with stand-alone software (e.g., Racon) or via a module in an assembler (e.g., Canu). Hybrid correction involves alignment of low-error, short reads to the raw long reads to generate the consensus (e.g., CoLoRMap). Note that low-error, short reads can also be used to polish the assembled contigs, i.e., correct misassemblies and errors. To investigate how self-correction, hybrid correction, or both correction methods affect assembly contiguity, we tried each approach in a case study. Bonefish (Albula glossodonta) DNA was extracted and sequenced on PacBio Sequel to theoretical 70x coverage and on Illumina HiSeq 2500 to theoretical 100x coverage with paired-end (PE) 2x250 in Rapid run mode. Our assembly results demonstrate that a combination of both approaches generates the most contiguous bonefish assembly.

OP30: Integrative analysis of transcriptomics and proteomics to detect novel protein isoforms from alternatively spliced transcripts induced by SF3B1 spliceosomal mutations

Presenting Author: Kelsey Nassar, University of Colorado Anschutz Medical Campus

Co-Author(s):
Hyunmin Kim, University of Colorado Anschutz Medical Campus
Jihye Kim, University of Colorado Anschutz Medical Campus
Maggie Lam, University of Colorado Anschutz Medical Campus
Aik Choon Tan, University of Colorado Anschutz Medical Campus

ABSTRACT: Alternative splicing (AS) contributes to transcriptional complexity and is hypothesized to alter the proteome. AS events have been found to be increased in various cancers, however, the functional

consequences of AS events on tumorigenesis remains unclear. Recently, mutations in core spliceosomal proteins, such as SF3B1, have been identified at a high frequency in multiple cancers. Next generation RNA-sequencing (RNA-seq) has identified that SF3B1 mutations result in global transcriptomic alterations in AS, primarily an increase in alternative 3′ splice site recognition. We hypothesized that mutations in SF3B1 increases proteome diversity through alternative splice variants that contribute to tumorigenesis. To test this hypothesis, we performed deep RNA-Sequencing on an SF3B1-Mutant and SF3B1-WildType uveal melanoma cell lines. We developed SALSA (Systemic ALternative Splice Analysis), for RNA-seq analysis to identify novel AS events as a result of SF3B1 mutations. In addition, we conducted proteome-wide mass spectrometry (MS) to identify novel protein isoforms detected from RNA-seq. We curated a novel peptide database from our custom AS events identified by SALSA to detect novel protein isoforms. From this integrative analysis, we identified 76 novel peptides enriched in SF3B1-Mutant cells detected at both RNA-seq and MS levels. From the MS peptide list, we validated SETD5, an 3′ alternatively spliced transcript. To our knowledge, this is the first description of a novel alternatively spliced transcript that results in a novel protein in SF3B1-mutant cells. This preliminary analysis lays the ground work for further identification of novel protein isoforms resulting from SF3B1 mutations that ultimately may contribute to tumorigenesis.

OP31: Exploring the Fabric of Breast Cancer Using Gene Sets

Presenting Author: Judith Blake, The Jackson Laboratory

Co-Author(s):
Carol Bult, The Jackson Laboratory
Leigh Carmody, The Jackson Laboratory
Mary E Dolan, The Jackson Laboratory
Harold J Drabkin, The Jackson Laboratory
Akenna Harper, The Jackson Laboratory
Joan Malcolm, The Jackson Laboratory
Monica S McAndrews, The Jackson Laboratory
Peter Robinson, The Jackson Laboratory
Sara Patterson, The Jackson Laboratory
Susan Mockus, The Jackson Laboratory
Erich Baker, Baylor University
David P Hill, Baylor University

ABSTRACT: A key requirement in understanding the complexities of biological systems is being able to move from a single gene approach to understanding how genes interact to give rise to complex phenomena. One way of analyzing how multiple genes interact is to study gene sets that represent a given aspect of biology. By comparing and contrasting sets, we determine whether sets describing different aspects are related, share common members or can be refined based on their metadata. The GeneWeaver suite of analysis tools allows comparison and manipulation of gene sets to identify and understand

commonalities and differences and to understand the underlying biology that genes in the sets share. We took a targeted approach to gene set analysis by curating genes sets related to breast cancer. We used our curated sets to identify and quantify mutations that are common in breast cancer, and to explore how differentially expressed genes may influence chemotherapy response and how they can be used to identify underlying biology that might contribute to breast cancer progression.

This work supported by NIH grants: NHGRI grant R25 HG007053 Diversity action Plan for Mouse Genome Database (C. Bult, PI); NCI grant P30 CA034196 Jackson Laboratory Cancer Center (E. Liu, PI); NHGRI U41 HG000330 Mouse Genome Informatics (C. Bult and J. Blake PIs); and NIAA/NIDA AA018776 Data Structures, algorithms and tools for ontological discovery (E. Chesler, PI). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health

OP32: Measuring and Protecting the Sensitive Linking Information Leakage Across Epigenetic and Transcriptomics Datasets Through Genotype and Assay Prediction

Presenting Author: Arif Harmanci, University of Texas Health Science Center

ABSTRACT: The next generation sequencing (NGS) is used to measure cellular phenotypes across many different levels such as epigenetic and transcriptomic datasets. Currently there are hundreds of functional genomics assays based on NGS technologies and more assays are being proposed for measuring diverse set of epigenetic and transcriptomic states of cells, even at some individual levels. While the main purpose of these data are to probe and reveal important biological knowledge, such as cancer epigenetics and transcriptomics, the privacy aspect of the data is not well studied. Much of the data is being distributed We have previously shown that gene expression matrices can be used to predict the genotypes of expression quantitative loci (eQTL) and these can be used in linking attacks to link expression (such as GTEx Project) and genotype (such as The 1000 Genomes Project) datasets. In this study we extend the possible linkages that can be performed and study whether the correlations in epigenetic, transcriptomic, and genetic datasets can be exploited to perform accurate linking attacks. Specifically, we evaluate linkage of epigenetic-transcriptomic and epigenetic-transcriptomic-genetic datasets. We first propose robust rank-based measurement statistics to measure leakages originating from epigenetic datasets. We then present a practical linking attack and evaluate linking accuracies under different scenarios. We finally present a method for sanitizing the leakage of genetic and transcriptomic information from epigenetic data. We demonstrate the effectiveness of sanitization method on histone modification ChIP-Seq signal profiles and matrices.

OP33: Education, Networking and Building Next Generation Prototypes — Hackathons and Analyzeathons for Bioinformaticians, Biomedical Informaticians and Computational Biologists!

Presenting Author: Ben Busby, NCBI

ABSTRACT: Over the past three years, NCBI has run or been involved in 31 data science hackathons. In these hackathons, participants assemble into teams of five or six to work collaboratively for three days on pre-scoped projects of general interest to the bioinformatics community. On average, about 80% of teams produce an alpha or beta working prototype, and approximately ten percent ultimately publish a manuscript describing their work. Thus, NCBI hackathons have generated over 150 products, and about 50% of them are stable, and/or continue to be developed. Some of these can be found at http://biohackathons.github.io. In addition to the production aspect, hackathons provide an immersive learning environment and promote networking opportunities. This presentation will discuss the educational aspects of these hackathons, options for setting up hackathons at your own institution, and tricks to make them successful. NCBI and other parts of NLM and NIH are also involved in other programs pertaining to project-based data science education. These include the NIH data science mentorship program, the visiting bioinformatician program, and the microbial metagenomics discovery challenge. In spring 2019, we are scaling up to run analyzeathons aimed at indexing massive data collections and making them amenable to modern analysis techniques, including machine learning.

OP34: Alternative Splicing of Single Cells in Squamous Cell Lung Cancer Premalignancy

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

Co-Author(s):
Moumita Ghosh, National Jewish Health
Jihye Kim, University of Colorado Anschutz Medical Campus
Aik-Choon Tan, University of Colorado Anschutz Medical Campus

ABSTRACT: Single-cell RNA-seq sequencing (scRNA-seq) is a rapidly evolving technology for studying transcriptomic landscape at the resolution of individual cells. Unique Molecular Identifier (UMI)-based approach such as 10x scRNA-seq system allows researchers to reconstitute gene expression to cell (G2C) association by extracting the UMI and cell barcode from the reads assigned to the genes. One of the common data analytics for G2C matrix is to present the reduced dimensionality making a visualization plot. However, the underlying biological mechanisms such as cell viability and transcription efficiency were not addressed by the current approaches. We hypothesized that alternative splicing (AS) could provide a complementary approach to address these limitations. To test this hypothesis, we performed

deep scRNA-seq to deconvolute the premalignancy of Squamous Cell Lung Cancer (SCC). SCC often develops from a premalignant field that includes dysplasia. Identification of genomic change in the dysplasia epithelium could aid both early detection and improved prevention strategies. The dysplastic lung is enriched with heterogeneous cell types complicating identification of cell types in the premalignancy of SCC. We performed scRNA-seq on two endo-bronchial biopsies from a high-risk patient. We performed standard scRNA-seq analysis to deconvolute the premalignancy landscape of SCC. We also develop a novel pipeline that considered AS in the analysis pipeline. We will report comparison results by showing cell subgroups obtained by the standard approach and the AS pipeline. We believe that the AS pipeline for scRNA-seq is providing additional information to decompose the mixture of cells in a heterogeneity population.

OP35: Inferring trade-offs in protein folding networks

Presenting Author: Sebastian Pechmann, Université de Montréal

ABSTRACT: How proteins fold inside the cell remains a fundamental open question. The cell employs a complex regulatory and quality control system, the protein homeostasis network, that keeps proteins in their correct shape. Failure of protein homeostasis is directly linked to so-called protein misfolding diseases such as Alzheimer's and Parkinson's. However, how proteins interact with their quality control mechanisms remains poorly understood. Here, we demonstrate how the integration of genomic data into the systematic analysis of protein families can decouple contributions to sequence-structure relationships within proteins that define their interactions with cellular quality control mechanisms. Our work highlights overlapping constraints and trade-offs between protein synthesis, folding, and quality control. Joint inference of these trade-offs outlines quantitative principles underlying protein folding in the cell. We conclude by discussing how our results help to understand how mutations may perturb protein homeostasis and ultimately lead to ageing and neurodegenerative diseases.

OP36: Fully Bayesian model for non-random missing data in qPCR

Presenting Author: Valeriia Sherina, University of Rochester Medical Center

Co-Author(s):
Matthew McCall, University of Rochester Medical Center
Tanzy Love, University of Rochester Medical Center

ABSTRACT: We propose a new statistical approach to obtain differential gene expression of non-detects in quantitative real-time PCR (qPCR) experiments through Bayesian hierarchical modeling. We propose to treat non-detects as non-random missing data, model the missing data mechanism, and use this model to impute Ct values or obtain direct

estimates of relevant model parameters. A typical laboratory does not have unlimited resources to perform experiments with a large number of replicates; therefore, we propose an approach that does not rely on large sample theory. We aim to demonstrate possibilities that exist for analyzing qPCR data in the presence of non-random missingness through the use of Bayesian estimation. Bayesian analysis typically allows for smaller data sets to be analyzed without losing power while retaining precision. The heart of Bayesian estimation is that everything that is known about a parameter before observing the data (the prior) is combined with the information from the data itself (the likelihood), resulting in updated knowledge about the parameter (the posterior). In this work we introduce and describe our hierarchical model and chosen prior distributions, assess the model sensitivity to the choice of prior, perform convergence diagnostics of the Markov Chain Monte Carlo, and present the results of a real data application.

OP37: Homologous Inter-Domain Segments in Protein Families

Presenting Author: Dylan Barth, University of Nevada Las Vegas

ABSTRACT: We are interested in sequences between conserved domains of multi-domain proteins. These sequences have historically been ignored in evolutionary analysis because they are not conserved between species and therefore cannot be aligned effectively. To study the evolution of the lengths of these segments, we first need to define homologous inter-domain segments across species. We gathered gene trees from the Ensembl database to provide information on homologous gene families and the evolutionary relationships of the genes. Gene trees were divided into subtrees that are less than 400 million years old. Domain data for each human protein within each gene family have been gathered from both the Superfamily and Pfam databases. Using the boundaries of human domains, we inferred the homologous domain positions across the alignment of the gene family, and defined the homologous inter-domain segments. We have found that these inter-domain segments approximately follow an exponential distribution with a mean and median length of 46 and 23 bp respectively. Based on these data, we plan to study how the lengths of these segments have evolved through insertions and deletions.

OP38: Integrating extracted relations into existing knowledge bases

Presenting Author: Harrison Pielke-Lombardo, University of Colorado

ABSTRACT: The KaBOB knowledge base (Livingston et al., "KaBOB.") was built using structured data sources. However, these data sources must be manually curated by consulting existing literature sources,

each of which contains only a few fragments of knowledge. Literature sources are unstructured and distributed information.

Here, we present a method for extracting such unstructured information and integrating it into an existing knowledge base. First, concept embeddings are generated using ConceptMapper (Tanenblatt et al., "The ConceptMapper Approach to Named Entity Recognition.") covering ten OBO ontologies, followed by relation extraction using a modification to the Snowball algorithm (Agichtein and Gravano, "Snowball.") which considers the syntax and dependency within a sentence. Relations are matched to the Relation Ontology such that the domains and ranges of the relations are satisfied. Next, relations are matched by the coreference chains of their subjects and objects to form a graph of relations which can then be integrated into the knowledge base.

Introducing statements from the literature can lead to logical inconsistencies. Reasoners like ELK and HermiT can be used to evaluate whether new statements violate the current state of knowledge. Additionally, a confidence for each relation is assigned based on the number of literature sources that make each claim.

To evaluate these methods, we attempt to demonstrate that we can recreate the existing model of cholesterol clearance in the liver from Reactome by using literature sources. The concept, coreference, and relation annotations are evaluated using the CRAFT corpus (Bada et al., "Concept Annotation in the CRAFT Corpus.").

OP39: The Use of Scientific Ignorance to Drive Literature-Based Discovery in Prenatal Nutrition Across Disciplinary Boundaries

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

Co-Author(s):
Lawrence Hunter, University of Colorado Anschutz Medical Campus
Sonia Leach, National Jewish Health

ABSTRACT: Researchers are interested in the gaps in knowledge (unknowns, speculations, hypotheses). We aim to reframe the literature in terms of such gaps: we hypothesize that knowledge gaps exist in the scientific literature, we can automatically identify and classify them, and we can use them to drive further scientific research. Knowledge gaps exist in the literature because current natural language processing (NLP) explicitly attempts to discard them and instead focuses on the literature as a knowledge source. For example, the statement "little is known about whether calcium interacts with iron" would be discarded. However, researchers include such statements of open questions, research goals, and current controversies (ignorance statements) in scientific publications because scientists and clinicians are interested. Following the example above, the scientific goal remains to determine if calcium interacts with iron or not. Utilizing such statements, we propose that formal computational representations of them, analogous

to knowledge representations, will allow new computational tools to support research, identify and integrate relevant information from articles in other disciplines, and enhance clinical research. We hypothesize that this new form of literature-based discovery, finding new relevant information from other articles, will find novel connections across disciplinary boundaries in the biomedical domain. To evaluate this approach, we apply it the field of prenatal nutrition because of its public health relevance and clearly stated knowledge gaps. The creation of such NLP tools will allow researchers, funders, medical professionals, and the public to explore the dynamic landscape of research and discover novel insights into current knowledge gaps.

OP40: Governance Innovations for Promoting Cross-institutional Electronic Health Data Sharing

Presenting Author: Kari Stephens, University of Washington

Co-Author(s):
Adam Wilcox, University of Washington
Philip Payne, Washington University
Jason Morrison, University of Washington
Jennifer Sprecher, University of Washington
Rania Mussa, University of Washington
Randi Foraker, Washington University
Sarah Biber, Oregon Health Sciences University
Sean Mooney, University of Washington

ABSTRACT: Cross institutional electronic health data sharing is an essential requirement for health innovation research. Healthcare organizations across the country are governed separately by state and local laws and policies that complicate research related data sharing. Electronic health record (EHR) data are not only highly protected via federal laws (i.e., HIPAA) and regional Internal Review Boards (IRBs), but are also often protected as assets by individual organizations. No clear pathway exists for organizations to execute governance for rapid EHR data sharing, stifling research efforts ranging from simple observational studies to complex multi-institutional trials. Universal governance solutions are essential to provide pathways for data sharing to address the rapid pace of research. The Clinical Translational Science Award (CTSA) Program Data to Health (CD2H) Coordinating Center has launched a cloud data sharing pilot project to begin addressing this complex issue. In order to configure a web-based data sharing software tool, Leaf, that can cross-query comprehensive harmonized EHR data generated by multiple healthcare organizations, we are exploring a singular governance solution (i.e., embodied in a data use agreement (DUA) and Internal Review Board (IRB) solution) to accommodate both a general and research specific use. While DUAs and IRBs are not streamlined governance solutions, this is an essential first step in creating broader sustainable national governance solutions (i.e., master consortium agreements, access governance policies).

POSTERS

LOCATION

Viceroy Hotel Ballroom

POSTER SESSION HOURS

Posters will only be available for viewing Friday. The Poster Session with authors present will be on Friday evening. Poster Presenters must be available for presentation during the scheduled poster session.

POSTER NUMBER ASSIGNMENTS

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

(Maximum size 4 feet high x 4 feet wide)

SCHEDULE

FRIDAY, DECEMBER 7	**12:00 PM – 6:00 PM**	**SET UP POSTERS**
	6:30 PM – 7:30 PM	**POSTER SESSION WITH AUTHORS** *Even Number Posters*
	7:30 PM – 8:30 PM	**POSTER SESSION WITH AUTHORS** *Odd Number Posters*

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

POSTER PRESENTATION LIST

P01: Simplifying biological data access for data science
Presenting Author: David Adams, Brigham Young University

P02: Dynamical comparison between Hemoglobin and Myoglobin reveals the affects of quaternary structure of Hemoglobin on the intrinsic dynamics of its subunits
Presenting Author: Rotem Aharoni, Ariel University

P03: Convolutional Neural Networks In Classifying Cancer Through DNA Methylation
Presenting Author: Satya Avva, Saama

P04: Homologous Inter-Domain Segments in Protein Families
Presenting Author: Dylan Barth, University of Nevada Las Vegas

P05: GPCR-PEnDB: A database of protein sequences and derived features to facilitate prediction and classification of G protein-coupled receptors
Presenting Author: Khodeza Begum, University of Texas at El Paso

P06: A Data Quality Testing Tool for Cross-institutional OMOP Electronic Health Record Data Repositories
Presenting Author: Timothy Bergquist, University of Washington

P07: Med2Mech: Neural-Symbolic Representation of Molecular Mechanisms Underlying Pediatric Disease
Presenting Author: Tiffany Callahan, University of Colorado Denver Anschutz Medical Campus

P08: Cell4D: a Spatial Stochastic Simulator for Biological Modeling
Presenting Author: Donny Chan, University of Toronto

P09: REAL-neo, a comprehensive neoantigen prediction and prioritization pipeline using tumor sequencing data
Presenting Author: Yesesri Cherukuri, Mayo Clinic

P10: NaVARgator: A bioinformatics program to cluster phylogenetic trees and identify representative variants
Presenting Author: David Curran, Hospital for Sick Children

P11: Identifying candidate druggable targets in canine cancer cell lines using whole exome sequencing
Presenting Author: Sunetra Das, Colorado State University

P12: Using Adversarial Deep Neural Networks to Remove Nonlinear Batch Effects from Expression Data
Presenting Author: Jonathan Dayton, Brigham Young University

P13: SumSec: Accurate Prediction of Sumoylation Sites using Predicted Secondary Structure
Presenting Author: Abdollah Dehzangi, Morgan State University

P14: Leaf: A self service cohort discovery and extraction browser for mining clinical enterprise data warehouses for research and quality improvement
Presenting Author: Nicholas Dobbins, University of Washington

P15: The GA4GH/DREAM Workflow Execution Challenge
Presenting Author: James Eddy, Sage Bionetworks

P16: A Machine Learning Classifier for Assigning Individual Patients with Systemic Sclerosis to Intrinsic Molecular Subsets
Presenting Author: Jennifer Franks, Geisel School of Medicine at Dartmouth

P17: ShapeShifter: Making it Easy to Transform Genomic and Transcriptomic Data from One File Format to Another
Presenting Author: Brandon Fry, Brigham Young University

P18: Using machine learning algorithms for classification of medulloblastoma subgroups based on gene expression data
Presenting Author: Sivan Gershanov, Ariel University

P19: Computational Identification and Analysis of Bacterial Virulence Factors Embedded Into Bacteriophage Genomes
Presenting Author: Cody Glickman, University of Colorado Anschutz Medical Campus

P20: Harmonizing and Analyzing Clinical Trials Data in the AHA Precision Medicine Platform
Presenting Author: Carsten Goerg, University of Colorado

P21: Integrating pathway databases with Gene Ontology Causal Activity Models
Presenting Author: Benjamin Good, Berkeley Labs

P22: Human Skin Biopsy Culture Model Maintains Psoriasis Disease Function and Demonstrate Pathway Engagement by Dexamethasone
Presenting Author: Shaun Grosskurth, AbbVie

P23: How open is open? The (Re)usable Data Project assesses data licensing
Presenting Author: Melissa Haendel, Oregon Health & Science University

P24: Custom database for identifying coral symbionts
Presenting Author: Graham Hamilton, University of Glasgow

P25: Transcriptome analysis of cancer adjacent normal tissues reveal genes co-expressed with LINE elements
Presenting Author: Mira Han, University of Nevada Las Vegas

P26: Refining orthology determination in RNA-seq phylogenetics
Presenting Author: Madison Hansen, American Museum of Natural History

P27: Optimized hybrid assembly of mycobacterial genomes using MinION and Illumina next generation sequence reads
Presenting Author: Jo Hendrix, University of Colorado Anschutz Medical Campus

P28: Predicting cancer outcomes based on gene-expression profiles more accurately with deep neural networks
Presenting Author: Kimball Hill, Brigham Young University

P29: Computational and cultural aspects of improved attribution
Presenting Author: Kristi Holmes, Northwestern University

P30: A new computational pipeline for PAR-CLIP characterizes a key immune regulatory mechanism
Presenting Author: Rachel Hovde, Chimera Bioengineering

P31: Food preservatives induce Proteobacteria dysbiosis of the human gut microbiota
Presenting Author: Tomas Hrncir, Czech Academy of Sciences

P32: ORCHID: a method for detecting short-range chromatin interactions in high-resolution 5C and Hi-C datasets
Presenting Author: Fei Ji, Massachusetts general hospital

P33: Clustering of Protein Conformations using Parallelized Dimensionality Reduction
Presenting Author: Arpita Joshi, University of Massachusetts, Boston

P34: Optimizing nontuberculous mycobacteria (NTM) de novo genome assemblies for application in clinical case studies
Presenting Author: Sara Kammlade, National Jewish Health

P35: PredHPI: an integrated web-server platform for the prediction and visualization of host-pathogen interactions
Presenting Author: Rakesh Kaundal, Utah State University

P36: An Automated Case Notes System for Psychiatrists Using Text Mining
Presenting Author: Nazmul Kazi, Montana State University

P37: A systems biology approach to define essential kinases in small cell lung cancer
Presenting Author: Jihye Kim, University of Colorado Denver Anschutz Medical Campus

P38: Comparative Analysis of Germline Microsatellites in the 1,000 Genomes Project
Presenting Author: Nicholas Kinney, Virginia College of Osteopathic Medicine

P39: Omic profiling in healthy volunteers taking celecoxib reveals novel biomarkers regulated by cyclooxygenase-2
Presenting Author: Nicholas Kirkby, Imperial College London

P40: An In Silico Approach For Finding Vaccines
Presenting Author: Siddharth Krishnakumar, Thomas Jefferson high school for science and technology

P41: Searching for translatable alternative splice isoforms in the human proteome
Presenting Author: Maggie Pui Yu Lam, University of Colorado Anschutz Medical Campus

P42: A human disease network from gene-publication relationships on PubMed
Presenting Author: Edward Lau, Stanford University

P43: Unbiased Pathway Detection Expands Cancer Pathways
Presenting Author: Chih-Hsu Lin, Baylor College of Medicine

P44: Identifying HCV-Host Interactions from Amino Acids Sequences Using SVM Identifying HCV-Host Interactions from Amino Acids Sequences Using SVM
Presenting Author: Xin Liu, XuZhou Medical University

P45: Modeling the Structure of BioGRID PPI Networks
Presenting Author: Sridevi Maharaj, University of California-Irvine

P46: Select: A SQL-based, high-resolution selection scanning tool to Identify genomic selection signals using next-generation sequencing data
Presenting Author: Hannah Maltba, Brigham Young University

P47: Codon Pairs are Phylogenetically Conserved: Codon pairing as a novel phylogenetic character state for parsimony and alignment-free methods
Presenting Author: Lauren McKinnon, Brigham Young University

P48: ExtRamp: A novel algorithm for extracting the ramp sequence based on the tRNA adaptation index or relative codon adaptiveness
Presenting Author: Justin Miller, Brigham Young University

P49: Metabolic profiling using a UHPLC-MS/MS-based platform to quantify amines, amino acids and methylarginines in plasma from cyclooxygenase-2 knockout mice
Presenting Author: Jane Mitchell, Imperial College London

P50: The characterization of different cell types using the Benford law
Presenting Author: Sne Morag, Ariel University

P51: Integrative analysis of transcriptomics and proteomics to detect novel protein isoforms from alternatively spliced transcripts induced by SF3B1 spliceosomal mutations
Presenting Author: Kelsey Nassar, University of Colorado Anschutz Medical Campus

P52: Mining Heterogenous Relationships from Pubmed Abstracts Using Distant Supervision
Presenting Author: David Nicholson, University of Pennsylvania

P53: Mutational impact on protein structure and function in endometrial cancer
Presenting Author: Amanda Oliphant, Brigham Young University

P54: A platform for community-scale transcriptome-wide association studies
Presenting Author: YoSon Park, Perelman School of Medicine University of Pennsylvania

P55: Good Nomen: An Interactive Web Application for Cleaning Clinical Data Using Standardized Terminologies
Presenting Author: Alyssa Parker, Brigham Young University

P56: Proteomics of natural bacterial isolates powered by deep learning-based de novo identification
Presenting Author: Samuel Payne, Brigham Young University

P57: Text Mining Novel Disease- and Drug-Specific Pathways
Presenting Author: Minh Pham, Baylor College of Medicine

P58: A Case Study on the Effects of Noisy, Long-read Correction Approaches on Assembly Contiguity
Presenting Author: Brandon Pickett, Brigham Young University

P59: Measuring chromosome conformation
Presenting Author: Brian Ross, University of Colorado Anschutz Medical Campus

P60: Challenges Using Electronic Medical Record for Pharmacokinetic Analysis
Presenting Author: Matthew Shotwell, Vanderbilt University Medical Center

P61: Measuring Transcription Factor Activity with Nascent RNA Sequencing
Presenting Author: Rutendo Sigauke, University of Colorado Anschutz Medical Campus

P62: Addressing the compositional data problem in sequencing with a novel, robust normalization method
Presenting Author: James St. Pierre, University of Toronto

P63: Governance Innovations for Promoting Cross-institutional Electronic Health Data Sharing
Presenting Author: Kari Stephens, University of Washington

P64: Use of metadata and Bag-of-words to map measurements across observational study data
Presenting Author: Laura Stevens, University of Colorado Anschutz Medical Campus

P65: Visualization Tool for interactive deciphering complex genetic regulation from multi-omic data
Presenting Author: LIN TING-WEI, Linkou Chang Gung Memorial Hospital

P66: anexVis: visual analytics framework for analysis of RNA expression
Presenting Author: Diem-Trang Tran, University of Utah

P67: Toxicant-protein relation extraction
Presenting Author: Ignacio Tripodi, University of Colorado, Boulder

P68: LOINC2HPO: Improving translational informatics by standardizing EHR phenotypic data using the Human Phenotype Ontology
Presenting Author: Nicole Vasilevsky, Oregon Health & Science University

P69: Exploratory Analysis of Diseased Male and Female Gene Expression Levels
Presenting Author: Clarissa White, Brigham Young University

P70: BioThings API: Building a FAIR API Ecosystem for Biomedical Knowledge
Presenting Author: Chunlei Wu, The Scripps Research Institute

POSTER PRESENTATION ABSTRACTS

P01: Simplifying biological data access for data science

Subject: Data management methods and systems

Presenting Author: David Adams, Brigham Young University, United States

Co-Author(s):
Sean Beecroft, Brigham Young University, United States
Amanda Oliphant, Brigham Young University, United States
Emily Hoskins, Brigham Young University, United States

ABSTRACT: Data collection methods are improving at an increasingly rapid rate, leading to larger and larger datasets. For molecular omics data, many national and international consortia have begun to collect population-scale datasets for the characterization of DNA, RNA, protein, and metabolites of numerous diseases. An explicit goal of these consortia is the dissemination and re-use of their data. Given the recent explosion of data science enthusiasts, a natural target for data re-use is non-biologist computational scientists. Unfortunately, practical use of these data by those outside of the narrow sub-domain is hindered by the steep learning curve necessary to understand data formats, experimental assumptions, raw data processing techniques, etc. Here we present a data dissemination mechanism designed to interface more fluidly with tools and expectations of the data science community. We have packaged the proteo-genomic data from a large uterine cancer cohort in a Python package, accessible natively as dataframes. In addition to the ready-for-use dataframes, the package's API provides a variety of utilities for multi-omics comparison including the metaclinical information. With an understanding of dataframes, any data scientist can follow our Jupyter-based tutorials to explore the deep molecular profiling of cancer data and participate in scientific discovery.

P02: Dynamical comparison between Hemoglobin and Myoglobin reveals the affects of quaternary structure of Hemoglobin on the intrinsic dynamics of its subunits

Subject: other

Presenting Author: Rotem Aharoni, Ariel University, Israel

Co-Author(s): Dror Tobi, Ariel University, Israel

ABSTRACT: Myoglobin and hemoglobin are globular hemeproteins, when the former is a monomer and the latter a heterotetramer. Despite the structural similarity of myoglobin to α and β subunits of hemoglobin, there is a functional difference between the two proteins, owing to the quaternary structure of hemoglobin. The effect of the quaternary structure of hemoglobin on the intrinsic dynamics of its subunits is explored by dynamical comparison of the two proteins. Anisotropic Network Model modes of motion were calculated for hemoglobin and myoglobin. Dynamical comparison between the

proteins was performed using global and local Anisotropic Network Model mode alignment algorithms based on the algorithms of Smith-Waterman and Needleman–Wunsch for sequence comparison. The results indicate that the quaternary structure of Hemoglobin substantially alters the intrinsic dynamics of its subunits, an effect that may contribute to the functional difference between the two proteins. Local dynamics similarity between the proteins is still observed at the major exit route of the ligand.

P03: Convolutional Neural Networks In Classifying Cancer Through DNA Methylation

Subject: Machine learning

Presenting Author: Satya Avva, Saama, United States

Co-Author(s):
Soham Chatterjee, Saama, India
Archana Iyer, Saama, India
Abhai Kollara, Saama, India
Malaikannan Sankarasubbu, Saama, United States

ABSTRACT: DNA Methylation has been the most extensively studied epigenetic mark. Usually a change in the genotype, DNA sequence, leads to a change in the phenotype, observable characteristics of the individual. But DNA methylation, which happens in the context of CpG (cytosine and guanine bases linked by phosphate backbone) dinucleotides, does not lead to a change in the original DNA sequence but has the potential to change the phenotype. DNA methylation is implicated in various biological processes and diseases including cancer. Hence there is a strong interest in understanding the DNA methylation patterns across various epigenetic related ailments in order to distinguish and diagnose the type of disease in its early stages. In this work, the relationship between methylated versus unmethylated CpG regions and cancer types is explored using Convolutional Neural Networks (CNNs). A CNN based Deep Learning model that can classify the cancer of a new DNA methylation profile based on the learning from publicly available DNA methylation datasets is then proposed.

P04: Homologous Inter-Domain Segments in Protein Families

Subject: inference and pattern discovery

Presenting Author: Dylan Barth, University of Nevada Las Vegas, United States

ABSTRACT: We are interested in sequences between conserved domains of multi-domain proteins. These sequences have historically been ignored in evolutionary analysis because they are not conserved between species and therefore cannot be aligned effectively. To study the evolution of the lengths of these segments, we first need to define homologous inter-domain segments across species. We gathered gene trees from the Ensembl database to provide information on

homologous gene families and the evolutionary relationships of the genes. Gene trees were divided into subtrees that are less than 400 million years old. Domain data for each human protein within each gene family have been gathered from both the Superfamily and Pfam databases. Using the boundaries of human domains, we inferred the homologous domain positions across the alignment of the gene family, and defined the homologous inter-domain segments. We have found that these inter-domain segments approximately follow an exponential distribution with a mean and median length of 46 and 23 bp respectively. Based on these data, we plan to study how the lengths of these segments have evolved through insertions and deletions.

P05: GPCR-PEnDB: A database of protein sequences and derived features to facilitate prediction and classification of G protein-coupled receptors

Subject: Data management methods and systems

Presenting Author: Khodeza Begum, University of Texas at El Paso, United States

Co-Author(s):
Jonathon Mohl, University of Texas at El Paso, United States
Ming-Ying Leung, University of Texas at El Paso, United States

ABSTRACT: G protein-coupled receptors (GPCRs) constitute the largest group of membrane receptor proteins in eukaryotes. Due to their significant roles in physiological processes such as vision, smell, and inflammation, GPCRs are targets of many prescription drugs. However, the functional and sequence diversity of GPCRs has kept their prediction and classification based on amino acid sequence data as a challenging bioinformatics problem. There are existing computational approaches, mainly using machine learning and statistical methods, to predict and classify GPCRs based on amino acid sequence and sequence derived features. In this project, we have constructed a searchable MySQL database and web application, named GPCR-PEnDB, of confirmed GPCRs and non-GPCRs for users to compile and download reliable training and testing datasets for different combinations of computational tools. This database contains over 2800 GPCRs and 3500 non-GPCR sequences (including transmembrane proteins) collected from the UniProtKB/Swiss-Prot protein database, covering more than 1100 species. Each protein is assigned a unique identification number and linked to information about its source organism, sequence length, and other features including amino acid and dipeptide compositions. For the GPCRs, family classifications according to the GRAFS and IUPHAR systems and the lengths of characteristic structural regions are also included. The web-based user interface allows researchers to compile and customize datasets with adjustable sequence diversity using the clustering tool CD-HIT, and output them as FASTA files. The current database provides a framework for future expansion to include predicted but unconfirmed GPCRs that

would help the development and assessment of GPCR prediction and classification tools.

P06: A Data Quality Testing Tool for Cross-institutional OMOP Electronic Health Record Data Repositories

Subject: Data management methods and systems

Presenting Author: Timothy Bergquist, University of Washington, United States

Co-Author(s):
Hossein Estiri, Harvard University, United States
Justin Prosser, University of Washinton, United States
Adam Wilcox, University of Washington, United States
Kari Stephens, University of Washington, United States

ABSTRACT: Data quality testing is critical to cross-institutional data sharing, a key component of health innovations produced through translational research. Harmonizing electronic health record (EHR) data is a resource intensive strategy used in many data sharing efforts, involving extraction, translation, and loading activities that can perpetuate and add to pre-existing data quality issues. Yet, we lack standards and tools for testing the quality of datasets produced through these complex harmonization processes. Given its large scale adoption, the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) standard is primed as a front running CDM to target establishing a standard set of executable data quality tests to support cross institutional data sharing. We adapted a prototype tool, DQe-c, to OMOP CDM V5 with scalability across database platforms. Namely it examines completeness in all data tables and columns, calculates the percentage of patients who have key clinical variables present (e.g., blood pressure, height), detects the presence of orphan keys (i.e., foreign keys that are not present in their reference table), reports on the size of the databases, and assesses conformance to the standard. All test results are produced as data visualizations in a single HTML dashboard. This prototype is being explored for use in multiple data sharing pilot projects supported by the Clinical Translational Science Award (CTSA) Program Data to Health (CD2H) Coordinating Center, with an aim towards configuring a robust set of completeness, conformance, and plausibility tests that confirm OMOP CDM V5 datasets are fit for cross-institutional data sharing.

P07: Med2Mech: Neural-Symbolic Representation of Molecular Mechanisms Underlying Pediatric Disease

Subject: Machine learning

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

Co-Author(s):
Adrianne Stefanski, University of Colorado Denver Anschutz Medical Campus
Michael Kahn, University of Colorado Denver Anschutz Medical Campus
Lawrence Hunter, University of Colorado Denver Anschutz Medical Campus

ABSTRACT: Subphenotyping aims to cluster patients with a particular disease into clinically distinct groups. Genomic and related molecular signatures, such as mRNA expression, have shown great promise for subphenotyping, but such molecular data is not and will not be available for most patients. Here, we present Med2Mech, a method for linking knowledge from generalized molecular data to specific patients' electronic patient records, and demonstrate its utility for subphenotyping. We hypothesized that integrating knowledge of molecular mechanisms with patient data would improve subphenotype classification. Med2Mech employs neural-symbolic representation learning to generate patient-level embeddings of molecular mechanisms using publicly available biomedical knowledge. Using clinical terminologies and biomedical ontologies, the mechanisms can then be mapped to patient data at scale. Med2Mech was developed and tested using clinical data from a subset of rare disease and other similarly medically complex patients from the Children's Hospital Colorado. A one-vs-the-rest multiclass classification strategy was used to evaluate the discriminatory ability of embeddings generated using Med2Mech versus only clinical data. Clinical embeddings were built for 2,464 rare disease and 10,000 similarly complex patients using 6,382 conditions, 2,334 medications, and 272 labs. Molecular mechanism embeddings were generated from a knowledge graph (116,158 nodes and 3,593,567 edges) built with 23,776 genes, 3,744 diseases, 49,185 gene ontology concepts, 13,159 phenotypes, 11,124 pathways, and 15,019 drugs. For classification, the molecular mechanism embeddings (precision=0.95, recall=0.94) out-performed all parameterizations of clinical embeddings (precision=0.83, recall=0.82). The Med2Mech representation of patient data improves subphenotype classification relative to standard subphenotyping approaches by incorporating knowledge of molecular mechanisms.

P08: Cell4D: a Spatial Stochastic Simulator for Biological Modeling

Subject: Qualitative modeling and simulation

Presenting Author: Donny Chan, University of Toronto, Canada

Co-Author(s):
Graham Cromar, The Hospital for Sick Children, Canada
Billy Taj, The Hospital for Sick Children, Canada
John Parkinson, The Hospital for Sick Children, Canada

ABSTRACT: With high-throughput datasets revealing complexity in many biological pathways, computational models can allow biologists to study a large variety of biochemical processes within complicated systems; a process that would otherwise be time consuming if investigated in vitro or in vivo. In silico simulations of metabolic or signaling pathways can be utilized as an approach to identify novel interactions that can then be experimentally validated. We are developing a graphical cell simulator, Cell4D, that can capture how spatial effects change the behavior of biological systems. The program can be used to simulate complex biological systems of interest, such as the role of CEACAM-related proteins in immune surveillance and evasion. Cell4D is an improved version of an older simulator from the lab by Sanford et al., with expanded infrastructure changes. The new features include support for rule-based modeling, robust compartment rules, and more efficient neighbor-searching algorithms. These improvements will also allow us to model CEACAM-mediated signaling, a family of membrane receptors that mediate intercellular adhesion and have roles in cellular growth, differentiation, and inflammation. I will demonstrate Cell4D functionality by modeling a simple biological pathway, along with some preliminary results from a simulated CEACAM1 signaling pathway. The goal of this project is to develop a robust and extendable cell simulator that is biologically accurate and easily applicable to the modeling of diverse types of cell-based processes and biological systems.

P09: REAL-neo, a comprehensive neoantigen prediction and prioritization pipeline using tumor sequencing data

Subject: other

Presenting Author: Yesesri Cherukuri, Mayo Clinic, United States

Co-Author(s):
Yingxue Ren, Mayo Clinic, United States
Vivekananda Sarangi, Mayo Clinic, United States
Yi Lin, Mayo Clinic, United States
Keith Knutson, Mayo Clinic, United States
Yan Asmann, Mayo Clinic, United States

ABSTRACT: Neoantigens are immunogenic peptides from tumor-specific somatic mutations. The expressed neoantigens can be presented to class-I or class-II MHC molecules and induce robust and enduring anti-tumor T-cell responses. Recent studies have demonstrated the great potential of personalized neoantigen vaccines as a new type of immunotherapy.

In general, identification of neoantigens from tumor sequencing data includes the following steps: (1) call somatic mutations from tumor genomic sequencing data; (2) derive neo-peptide sequences containing somatic mutations; (3) predict binding affinities between neo-peptides and MHC molecules. However, the current bioinformatics practices ignore transcript splicing isoforms, expressed fusion gene products,

and often times only focus on non-synonymous single nucleotide mutations but not frame-shifting INDELs. In addition, the MHC binding affinity prediction mainly focuses on class-I but not class-II MHC molecules. Furthermore, studies have shown that substantial numbers of neo-peptides predicted to have low MHC affinities are actually immunogenic, suggesting the necessity of alternative approaches for neoantigen discovery. Finally, nominated neoantigens need to be further filtered to ensure tumor specificity.

We have improved and optimized each step of the bioinformatics workflow for neoantigen identification from tumor sequencing data to address the complexity and current limitations of the process.

P10: NaVARgator: A bioinformatics program to cluster phylogenetic trees and identify representative variants

Subject: Optimization and search

Presenting Author: David Curran, Hospital for Sick Children, Canada

Co-Author(s):
Jamie Fegan, University of Toronto, Canada
John Parkinson, Hospital for Sick Children, Canada

ABSTRACT: Phylogenetic trees are representations of the relatedness of a group of variants. No matter what the variants represent – genes, proteins, genomes, species, etc – the task of identifying clusters of similar variants arises in many different fields. NaVARgator performs clustering by identifying k variants as cluster centers such that the total phylogenetic distance from all variants to their nearest cluster center is minimized. The software can be run on any phylogenetic tree and allows the clustering procedure to be customized by classifying variants. If the tree contains an outgroup, or other variants to be removed from the clustering procedure, they can be assigned as "ignored". If there are variants that should be selected as cluster centers – perhaps because they are biologically important or already well studied – they can be assigned as "chosen". The variants that the remaining cluster centers will be chosen from should be assigned as "available". Unassigned variants will still impact the clustering calculations but cannot be selected as cluster centers.

NaVARgator provides a rich graphical user interface designed to aid the user in evaluating a cluster configuration, as well as comparing different configurations or numbers of clusters. Clustering data can be exported in a number of ways: as a customizable image of the tree, a list of variant names in the clusters or other subsets, a list of the distance between each variant and its cluster center, or a histogram of those distances. The software is available for installation or as a web tool.

P11: Identifying candidate druggable targets in canine cancer cell lines using whole exome sequencing

Subject: inference and pattern discovery

Presenting Author: Sunetra Das, Colorado State University, United States

Co-Author(s):
Rupa Idate, Colorado State University, United States
Dawn Duval, Colorado State University, United States

ABSTRACT: The FACC canine cell line panel is a valuable resource to study genome variations that drive cancer in dogs and assess pharmacogenomic correlations through in vitro testing of new targeted therapies. The goal of this study is to create a database of somatic mutations in canine cancer cell lines using whole exome sequencing (WES) technology. WES data of 33 cell lines from ten different cancer types were mapped against the canine genome using BWA tool. Variant calling and annotation was conducted with Freebayes and SnpEff resources, respectively. Following removal of germline variants and known polymorphisms a total of 66,344 somatic variants were identified. Mutational load throughout the FACC panel ranged from 15.79 to 129.37 per MB, and 13.2% of all variants were located in protein coding region of 5,085 genes. Using the Cancer Gene Census (COSMIC), 232 curated genes that play a role in cancer, were identified in this dataset. Upon cross-checking with human driving mutations, 62 variants were collated as candidate cancer drivers across 30 cell lines. To identify other protein coding variants that may play a role in cancer progression, following functional annotation of genes, a two prong approach was used to select functional terms: A. associated with activating and maintaining cancer (GO, PFAM, KEGG); B. with at-least one cancer-causing gene. This yielded 502 genes that are not currently in COSMIC database, with an enrichment of MAPK, and PI3K-Akt pathways. This functionally annotated database will be useful in conducting hypothesis-driven research based on the cell line mutational landscape.

P12: Using Adversarial Deep Neural Networks to Remove Nonlinear Batch Effects from Expression Data

Subject: Machine learning

Presenting Author: Jonathan Dayton, Brigham Young University, United States
Co-Author(s): Stephen Piccolo, Brigham Young University, United States

ABSTRACT: Batch effects and other confounding effects can skew research results when working with quantitative molecular data (e.g. RNA-Seq). Most existing batch adjustment methods only take into account linear effects, but modern analysis tools such as machine learning can still identify and be influenced by nonlinear batch effects, even after linear effects have been removed. We introduce Confounded, a method that uses adversarial deep neural networks to identify and remove linear and nonlinear batch effects. Confounded is composed

of 1) a discriminator designed to detect confounding effects and 2) an autoencoder designed to replicate the input data while identifying and removing confounding effects in order to fool the discriminator. Once the data have been faithfully reproduced and the confounders have been removed, the adjusted data are output for use in analysis. We have tested Confounded on image vectors with artificial nonlinear batch effects. We show that Confounded removes these batch effects more effectively than ComBat, the most commonly used batch-effect adjustment method, while still retaining most of the true signal as measured by several classification algorithms. We are also validating Confounded with molecular datasets, both with artificial and real batch effects, and publishing our software to enable other scientists to use Confounded in their bioinformatics pipelines. In addition to batch correction, Confounded may also be used for data integration between multiple databases or between different technologies (e.g. microarray and RNA-Seq) or for removing general confounding effects from data.

P13: SumSec: Accurate Prediction of Sumoylation Sites using Predicted Secondary Structure

Subject: Machine learning

Presenting Author: Abdollah Dehzangi, Morgan State University, United States

Co-Author(s):
Yosvany Lopez, Genesis Healthcare Co.,, Japan
Ghazaleh Taherzadeh, Griffith University, Australia
Tatsuhiko Tsunoda, RIKEN Center for Integrative Medical Sciences, Japan
Alok Sharma, RIKEN Center for Integrative Medical Sciences, Japan

ABSTRACT: Post Translational Modification (PTM) is defined as the interaction of amino acids along the protein sequences with different macromolecules after the translation process. These interactions significantly impact on the functioning of proteins and can range from strongly deleterious to strongly advantageous. Therefore, understanding the underlying mechanism of PTMs can play critical role in understanding the functioning of proteins. Among a wide range of PTMs, Sumoylation is one the most important ones due to its important functioning which includes, transcriptional regulation, protein stability and protein subcellular localization. Despite its importance, determining sumoylation sites using experimental methods is time consuming and costly. Therefore, there is a crucial demand for the development of fast computational methods able to accurately determine the sumoylation sites in proteins. In this study, we develop a new machine learning based method to predict sumoylation sites called SumSec. To do this, we employ the predicted secondary structure of amino acids to extract two types of structural features from neighboring amino acids along the protein sequence which has never been used for this task. We also employ the concept of profile-bigram to extract local information about the interaction of the amino acids based on structural information. As a result, our proposed method is able to enhance the sumoylation site

prediction task better than previously proposed methods found in the literature. SumSec demonstrates high sensitivity (0.91), accuracy (0.94) and MCC (0.88). The prediction accuracy achieved in this study is 21% better than previous studies found in the literature.

P14: Leaf: A self service cohort discovery and extraction browser for mining clinical enterprise data warehouses for research and quality improvement

Subject: Data management methods and systems

Presenting Author: Nicholas Dobbins, University of Washington, United States

Co-Author(s):
Anthony Black, University of Washington, United States
Cliffard Spital, University of Washington, United States
Robert Harrington, University of Washington, United States
Bas de Veer, University of Washington, United States
Xlyao Yang, University of Washington, United States
Robert Meizlik, University of Washington, United States
Beth Britt, University of Washington, United States
Jason Morrison, University of Washington, United States
Kari Stephens, University of Washington, United States
Adam Wilcox, University of Washington, United States
Peter Tarczy-Hornoch, University of Washington, United States

ABSTRACT: Academic medical centers and health systems are increasingly challenged with supporting appropriate secondary uses of data from a multitude of sources. To that end, the UW Medicine Enterprise Data Warehouse (EDW) has emerged as a central port for all data that can include clinical, research, administrative, financial and other datatypes. Although EDW's have been popular and successful in providing a single stop for data, they are often non-self service and require an informatician or clinical informatics expert to access. To address this challenge, we have developed an easy to use, self service web-based tool for querying, browsing and extracting clinical cohorts from the UW Medicine EDW, called Leaf. Leaf enables querying by data dictionaries or ontologies and allows both de-identified and identified access to patient data and grants access to these datasets in a compliant manner. Leaf is an interface that is being built upon multiple data models and is independent of a specific data model. While Leaf provides basic visualizations, it contains robust tools for exporting directly to REDCap projects. Leaf is different from existing query tools (e.g. i2b2, SlicerDicer) because it does not specify a specific data model and is intended to only be a reusable lightweight modern web interface. The users of Leaf include both quality improvement and research investigators and has been developed using an Agile development process with a soft production rollout to identify and address software, support and data quality concerns.

P15: The GA4GH/DREAM Workflow Execution Challenge

Subject: System integration

Presenting Author: James Eddy, Sage Bionetworks, United States

Co-Author(s):
Brian O'Connor, University of California, Santa Cruz, United States
Denis Yuen, Ontario Institute for Cancer Research, Canada
Justin Guinney, Sage Bionetworks, United States

ABSTRACT: Software and platforms for workflow sharing and execution are increasingly utilized in massive data generation efforts. In turn, groups are developing standards, APIs, and best practices for running portable and reproducible pipelines. In order to ensure that the promises of reproducibility are being met by these standards and projects, we must critically assess workflows and workflow management systems.

With the GA4GH/DREAM Infrastructure Challenges, we aim to bring groups together to test and demonstrate tool portability while continuing to develop common standards. We also aim to bring together workflow authors and execution platform engineers to increase communication and accelerate the resolution of compatibility issues. In the GA4GH/DREAM Workflow Execution Challenge (synapse.org/WorkflowChallenge), participants downloaded a Dockerized, CWL/WDL-described workflow—along with any required input, reference, or parameter files—from Synapse. Participants ran the workflow in their environment and uploaded results to Synapse along with a description of their methods.

Through this challenge, we began to formalize methods for evaluating workflow portability. We not only piloted the use of centralized and systematic validation through Synapse, but defined standard procedures for authoring, registering, and onboarding workflows. The processes and frameworks used in this challenge resulted in a collection of stress-tested workflows, a rich body of examples and documentation, and a well annotated record of workflow/platform compatibility. This work has also informed efforts of the GA4GH Cloud Work Stream (ga4gh.cloud) to establish a "testbed" framework for centralized benchmarking of workflows and platforms.

P16: A Machine Learning Classifier for Assigning Individual Patients with Systemic Sclerosis to Intrinsic Molecular Subsets

Subject: Machine learning

Presenting Author: Jennifer Franks, Geisel School of Medicine at Dartmouth, United States

Co-Author(s):
Viktor Martyanov, Geisel School of Medicine at Dartmouth, United States
Guoshuai Cai, Arnold School of Public Health at University of South Carolina, United States
Yue Wang, Geisel School of Medicine at Dartmouth, United States
Tammara Wood, Geisel School of Medicine at Dartmouth, United States
Michael Whitfield, Geisel School of Medicine at Dartmouth, United States

ABSTRACT: High-throughput gene expression profiling of skin biopsies from patients with systemic sclerosis (SSc) has identified four "intrinsic" gene expression subsets (inflammatory, fibroproliferative, normal-like, limited) conserved across multiple cohorts and tissues. In order to characterize patients in clinical trials or for diagnostic purposes, supervised methods that can classify single samples are required.

Three gene expression cohorts were curated and merged for the training dataset. Supervised machine learning algorithms were trained using repeated three-fold cross-validation. We performed external validation using three additional datasets, including one generated by an independent laboratory on a different microarray platform. WGCNA and g:Profiler were used to identify and functionally characterize gene modules associated with the intrinsic subsets.

The final model, a multinomial elastic net, performed with average classification accuracy of 88.1%. All intrinsic subsets were classified with high sensitivity and specificity, particularly inflammatory (83.3%, 95.8%) and fibroproliferative (89.7%, 94.1%). In external validation, the classifier achieved an average accuracy of 85.4%. In a re-analysis of GSE58095, we identified subgroups of patients that represented the canonical inflammatory, fibroproliferative, and normal-like subsets. Inflammatory gene modules showed upregulated biological processes including inflammatory response, lymphocyte activation, and stress response. Similarly, fibroproliferative gene modules were enriched in cell cycle processes.

We developed an accurate, reliable classifier for SSc intrinsic subsets, trained and tested on 427 skin biopsies from 213 individuals. Our method provides a robust approach for assigning samples to intrinsic gene expression subsets and can be used to aid clinical decision-making and interpretation for SSc patients and in clinical trials.

P17: ShapeShifter: Making it Easy to Transform Genomic and Transcriptomic Data from One File Format to Another

Subject: Data management methods and systems
Presenting Author: Brandon Fry, Brigham Young University, United States
Co-Author(s): Stephen Piccolo, Brigham Young University, United States

ABSTRACT: Despite bioinformatics' emphasis on handling and interpreting large data files, there is a distinct lack of uniformity in file formats for sharing this data among researchers. Because biomolecular data is stored in many different and frequently incompatible formats, researchers spend a frustrating amount of time transforming data from one format to another, which impedes solving the more interesting biological problems that researchers wish to address. To ease and simplify this process, we have developed ShapeShifter, a Python module and accompanying command-line tool that enable researchers

to quickly transform preprocessed, tabular data from one format to another. Additionally, researchers can perform queries on the data, select specific columns, merge multiple files into one, and gzip the resulting data using simple commands from a terminal. ShapeShifter currently supports transforming 15 different file types, ranging from formats used across many disciplines—like CSV, Excel, and SQLite—to those used specifically in bioinformatics applications for processing genomic and transcriptomic data, including Kallisto, Salmon, and GenePattern. Support for additional file formats is ongoing, and we encourage requests for such to be made to our open-source repository at https://github.com/srp33/ShapeShifter. To demonstrate ShapeShifter's utility, we performed a benchmark evaluation, comparing the time taken to import and export small, medium, and large data files when various filters are applied. Our evaluation shows that ShapeShifter excels at transforming and filtering small and medium sized files; currently, we are developing a solution to efficiently process files that are too large to store in memory.

P18: Using machine learning algorithms for classification of medulloblastoma subgroups based on gene expression data

Subject: Machine learning

Presenting Author: Sivan Gershanov, Ariel University, Israel

Co-Author(s):
Igor Vainer, Ariel University, Israel
Helen Toledano, Schneider Children's Medical Center of Israel, Israel
Albert Pinhasov, Ariel University, Israel
Nitza Goldenberg-Cohen, Bnai Zion Medical Center, Israel
Mali Salmon-Divon, Ariel University, Israel

ABSTRACT: Medulloblastoma (MB), the commonest malignant pediatric brain tumor, is divided into four molecular subgroups: WNT, SHH, Group 3 and Group 4. Clinical practice and treatment design are becoming subgroup-specific. Nowadays clinicians use a 22-gene signature set to diagnose the subgroups. While WNT and SHH subgroups are well-defined differentiating Group 3 from Group 4 is less obvious.

The aim of this study is to improve the diagnosis process in the clinic by identifying the most efficient list of biomarkers for accurate, fast and cost-effective MB subgroup classification.

We tested five machine learning based algorithms, four are well known and one is a novel method we developed. We applied them on a public microarray expression data set and compared their performance to that of the known 22-gene set.

Both decision tree and decision rules resulted in a reduced set with similar accuracy to the 22-gene set. Random forest and SVM-SMO methods showed improved performance, without applying feature-selection. When implementing our novel SARC (SVM Attributes Ranking

and Combinations) classifier, allowing feature-selection, the resulted accuracy level was the highest and better than using the 22-gene set as input. The number of attributes in the best-performing combinations range from 13 to 32, including known MB related genes such as WIF1, NPR3 and GRM8, along with LOC440173 a long non-coding RNA.

To summarize we identified sets of attributes that have the potential to improve MB subgroup diagnosis. Broad clinical use of this classification may accelerate the design of patient's specific targeted therapies and optimize clinical decision.

P19: Computational Identification and Analysis of Bacterial Virulence Factors Embedded Into Bacteriophage Genomes

Subject: Metogenomics

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

Co-Author(s):
Michael Strong, National Jewish Health, United States
Josephina Hendrix, University of Colorado Anschutz Medical Campus

ABSTRACT: Pathogenic bacteria utilize gene products called virulence factors to circumvent host immunity and promote colonization. Bacteria are adept at acquiring genetic information capable of producing virulence factors from their environment through horizontal gene transfer. One understudied mechanism of horizontal gene transfer is the integration of viral elements called bacteriophages into a host bacterial genome. These integrated bacteriophages engage in a passive lifestyle called lysogeny, replicating in parallel with the host bacteria. Propagative success of these viral elements are dependent upon the ability of the host bacteria to thrive in a niche. Thus there is a selective advantage for bacteriophages that carry genetic elements apt at increasing fitness and propagative success of the host.

In this study, we utilized sequence similarity techniques to establish a baseline distribution of virulence factor genes embedded within viral genomes. The bacterial taxonomy listed in the name of viral genomes are used to discretize the bacteriophages into genus level categories. Comparisons between the categories suggest that viral elements known to infect pathogenic bacteria contain different percentages of virulence factor genes. In addition, we use network based methods to explore the functional potential of virulence factors between the categories. Finally, we compare our baseline distribution against the percentage of virulence factor genes embedded in bacteriophages isolated from clinical non-tuberculosis mycobacterial samples.

This study expands our understanding of horizontal gene transfer by bacteriophages and provides a resource for further research. In addition, the study provides information into the abundance of virulence factors within lysogenic bacteriophages of clinical mycobacteria.

P20: Harmonizing and Analyzing Clinical Trials Data in the AHA Precision Medicine Platform

Subject: other

Presenting Author: Carsten Goerg, University of Colorado, United States

Co-Author(s):
Christophe Roeder, University of Colorado, United States
Bethany Doran, University of Colorado, United States
Ann Marie Navar, Duke University, United States
Michael Hinterberg, SomaLogic, United States
John Graybeal, Stanford University, United States
Mark Musen, Stanford University, United States
Jennifer Hall, American Heart Association, United States
David Kao, University of Colorado, United States

ABSTRACT: Clinical trials have produced many highly valuable datasets, but their potential to support discovery through meta-analysis has not been fully realized. Answering biomedical questions often requires integrating and harmonizing data from multiple trials to increase statistical power. Due to the lack of supporting computational approaches, this challenging and time-consuming integration process is currently performed manually, which leads to scalability and reproducibility issues. We present a framework and prototype implementation within the cloud-based American Heart Association Precision Medicine Platform as a first step towards addressing this problem. Our framework provides (1) a metadata-driven mapping process from study-specific variables to the OMOP common data model, (2) a metadata-driven extraction process for creating analysis matrices of harmonized variables, and (3) an interactive visual interface to define and explore cohorts in harmonized studies. To demonstrate our approach, we present a prototype use case that investigates the relationship between blood pressure and mortality in patients treated for hypertension. Using our framework, we harmonized five publicly available NIH-funded studies (ALLHAT, ACCORD, BARI-2D, AIM-HIGH, and TOPCAT), assessed distributions of blood pressure by study, and using harmonized data performed individual patient-data meta analyses to show the statistical relationship between all-cause mortality and systolic blood pressure, for individual studies as well as the aggregated data. We discuss how the cloud-based implementation supports reproducibility as well as transparent co-development between collaborators over time and space. Future work will entail development of a generalized workflow for acquisition and semantic annotation of new datasets based on the CEDAR metadata management system.

P21: Integrating pathway databases with Gene Ontology Causal Activity Models

Subject: inference and pattern discovery

Presenting Author: Benjamin Good, Berkeley Labs, United States

Co-Author(s):
Paul Thomas, USC, United States
David Hill, The Jackson Laboratory, United States
Huaiyu Mi, USC, United States
Kimberly van Auken, Caltech, United States
Seth Carbon, Berkeley Labs, United States
Laurent-Philippe Albou, Berkeley Labs, United States
Nomi Harris, Berkeley Labs, United States
Suzanna Lewis, Berkeley Labs, United States
Chris Mungall, Berkeley Labs, United States
James Balhoff, RENCI, United States
Peter Deustachio , NYU, United States

ABSTRACT: The Gene Ontology (GO) Consortium (GOC) is developing a new knowledge representation approach called 'causal activity models' (GO-CAM). A GO-CAM describes how one or several gene products contribute to the execution of a biological process. In these models (implemented as OWL instance graphs anchored in Open Biological Ontology (OBO) classes and relations), gene products are linked to molecular activities via semantic relationships like 'enables', molecular activities are linked to each other via causal relationships such as 'positively regulates', and sets of molecular activities are defined as 'parts' of larger biological processes. This approach provides the GOC with a more complete and extensible structure for capturing knowledge of gene function. It also allows for the representation of knowledge typically seen in pathway databases.

Here, we present details and results of a rule-based transformation of pathways represented using the BioPAX exchange format into GO-CAMs. We have automatically converted all Reactome pathways into GO-CAMs and are currently working on the conversion of additional resources available through Pathway Commons. By converting pathways into GO-CAMs, we can leverage OWL description logic reasoning over OBO ontologies to infer new biological relationships and detect logical inconsistencies. Further, the conversion helps to increase standardization for the representation of biological entities and processes. The products of this work can be used to improve source databases, for example by inferring new GO annotations for pathways and reactions and can help with the formation of meta-knowledge bases that integrate content from multiple sources.

P22: Human Skin Biopsy Culture Model Maintains Psoriasis Disease Function and Demonstrate Pathway Engagement by Dexamethasone

Subject: other

Presenting Author: Shaun Grosskurth, AbbVie, United States

Co-Author(s): Susan Huang, AbbVie, United States
Loan Miller, AbbVie, United States
Hetal Patel, AbbVie, United States
Lauren Olson, AbbVie, United States
Joseph Wetter, AbbVie, United States
Mark Reppell, AbbVie, United States
Marc Domanus, AbbVie, United States
Christopher Miller, AbbVie, United States
Marie Honore, AbbVie, United States
Victoria Scott, AbbVie, United States

ABSTRACT: Early indication of efficacy or pathway engagement in relevant disease models is valuable during drug discovery. While animal models of skin disease are useful, not all features of human skin are recapitulated. In collaboration with AbbVie Clinical Pharmacology Research Unit, we obtained skin biopsies from psoriasis patients to develop and evaluate an ex vivo human skin biopsy culture model. Full thickness skin biopsies from psoriasis patients and healthy donors were bisected and cultured for 3, 6, or 24 hours with and without dexamethasone treatment. Conditioned media and skin biopsies were harvested to characterize cytokine levels and assess transcriptomics. Confirming elevated inflammation, higher cytokine levels were seen in the media from psoriasis lesional versus control skin samples after 24 hours of culturing. For transcriptomic analysis, skin biopsy gene expression profiling was performed on Affymetrix Human Gene ST 1.0 arrays and differential expression was performed with linear modeling. The transcriptomic psoriatic lesional phenotypic status of the cultured biopsies were confirmed with gene lists identified from a meta-analysis incorporating 9 skin biopsy public datasets from psoriasis patients. Also as expected, biopsy samples treated with dexamethasone exhibited features consistent with decreased inflammation and activation of the glucocorticoid receptor NR3C1. Here we show that the human psoriasis skin biopsy culture model maintains many clinical phenotypes of fresh psoriasis skin biopsies. More importantly, we show that the skin biopsy model is a valuable tool for interrogation of pharmacodynamic pathway engagement and potential efficacy for future candidate therapeutics.

P23: How open is open? The (Re)usable Data Project assesses data licensing

Subject: Data management methods and systems

Presenting Author: Melissa Haendel, Oregon Health & Science University, United States

Co-Author(s):
Seth Carbon, LBNL, United States
Robin Champieux, OHSU, United States
Lilly Winfree, OHSU, United States
Letisha Wyatt, OHSU, United States
Julie Mcmurry, Oregon State University , United States

ABSTRACT: Complex licensing and data reuse restrictions hinder most publicly-funded, seemingly "open," biomedical and biological data from being used, modified, and redistributed to its full potential. Such issues include missing or non-standard licenses, restrictive provisions that do not allow for resources to be redistributed after modification, and terms that limit synthesis with other resources. Further, navigating legal compliance with data licensing and use agreements is complicated, as data is often manipulated, shared, and redistributed by many types of research groups and users in various and subtle ways. The community is plagued by complex licensing and legal terms of reuse when integrating data from a broad array of publicly funded resources. This struggle spurred the creation of the (Re)usable Data Project (http://reusabledata.org), an open source and open data project in which we created a five-part rubric to evaluate data resources' licensing information. Here we present our rubric and evaluations of sources based on the findability and clarity of the terms of use, how accessible the data is, and the degree to which unnegotiated and unrestricted reuse and redistribution can occur. We have tested the (Re)usable Data Project's rubric against over 50 biological data sources. Approximately 40% of the resources rank poorly, and more than half of the sources lack a clear, easily findable license. We hope that this systematic review of the data licensing landscape will build awareness, engage the community, and ultimately improve licensing practices that impact data reuse.

P24: Custom database for identifying coral symbionts

Subject: Metogenomics

Presenting Author: Graham Hamilton, University of Glasgow, United Kingdom
Co-Author(s): Nick Kamenos, University of Glasgow, United Kingdom

ABSTRACT: Determine which species of symbiotic photosynthetic algae are present in coral samples and to assess population changes due to rising ocean temperatures.

P25: Transcriptome analysis of cancer adjacent normal tissues reveal genes co-expressed with LINE elements

Subject: inference and pattern discovery

Presenting Author: Mira Han, University of Nevada Las Vegas, United States

Co-Author(s):
Nicky Chung, University of Nevada, Las Vegas, United States
G.M. Jonaid, University of Nevada, Las Vegas, United States
Sophia Quinton, University of Nevada, Las Vegas, United States
Austin Ross, University of Nevada, Las Vegas, United States

ABSTRACT: Despite the long-held assumption that transposons are normally only expressed in the germ-line, recent evidence shows that transcripts of LINE sequences are frequently found in the somatic cells. However, the extent of variation in LINE transcript levels across different tissues and different individuals, and the genes and pathways that are co-expressed with LINEs are unknown. Here we report the variation in LINE transcript levels across tissues and between individuals observed in the normal tissues collected for The Cancer Genome Atlas. Mitochondrial genes and ribosomal protein genes were enriched among the genes that showed negative correlation with L1HS in transcript level. We hypothesize that oxidative stress is the factor that leads to both repressed mitochondrial transcription and LINE over-expression. KRAB zinc finger proteins (KZFPs) were enriched among the transcripts positively correlated with older LINE families. The correlation between transcripts of individual LINE loci and individual KZFPs showed highly tissue-specific patterns. There was also a significant enrichment of the corresponding KZFP's binding motif in the sequences of the correlated LINE loci, among KZFP-LINE locus pairs that showed co-expression. These results support the KZFP-LINE interactions previously identified through ChIP-seq, and provide information on the in vivo tissue context of the interaction.

P26: Refining orthology determination in RNA-seq phylogenetics

Subject: Optimization and search

Presenting Author: Madison Hansen, American Museum of Natural History, United States
Co-Author(s): Ward Wheeler, American Museum of Natural History, United States

ABSTRACT: Orthologous genes are genes in different species which have evolved from the same gene in the common ancestor of those species. Determining groups of orthologous genes is important to evolutionary biology research, including phylogenetics and gene function studies. While orthology determination has historically relied on alignment to reference genome sequences, now next-generation sequencing techniques can generate large quantities of genome-wide sequences from organisms that do not yet have a reference genome. When no reference genome is available, orthology determination for multiple species is computationally expensive. Assorted methods for orthology

determination have been developed; each uses particular sequence similarity measurements and clustering algorithms to organize the sequences into potential orthologous groups. However, the groups sometimes conflict with the inferred evolution of the species, indicating that the groups may not comprise true orthologs. Here, we refine orthology determinations using phylogenetic inference methods and heuristics, in order to produce more consistent and precise orthologous groups.

P27: Optimized hybrid assembly of mycobacterial genomes using MinION and Illumina next generation sequence reads

Subject: Optimization and search

Presenting Author: Jo Hendrix, University of Colorado Anschutz Medical Campus

Co-Author(s):
Elaine Epperson, National Jewish Health, United States
Nabeeh Hasan, National Jewish Health, United States
Cody Glickman, National Jewish Health, United States
Michael Strong, National Jewish Health, United States

ABSTRACT: Next generation sequencing (NGS) technology allows researchers to sequence and compare bacterial genomes at an increasingly rapid pace, but in order to assemble complete bacterial genomes, and to identify plasmids, a hybrid assembly method often proves more effective. Illumina NGS produces highly accurate reads that are less than 300 bases in length and have difficulty covering low-complexity regions of the genome such as repeats. Such regions result in gaps between assembled contigs, computationally assembled segments of the sequence. In contrast to the short reads produced by Illumina, MinION technology is capable of sequencing segments that are tens of thousands of bases in length. These reads can span entire repetitive regions; however, these long reads are more expensive per base, are lower-throughput, and have a higher error rate.

In this study, we used and tested a method of hybrid genome assembly of Illumina and MinION sequence data in order to assemble complete bacterial genomes. We used the higher-throughput and more accurate Illumina reads to make a reliable scaffold of fragmented contigs which were stitched together utilizing MinION long reads. The result is an assembly of the bacterial genome and its plasmids. We demonstrate this method on two nontuberculous mycobacterial genomes, M. kubicae and M. gordonae.

With completed reference genomes from hybrid assembly, we can better annotate the complete genomes, identifying genes involved in virulence and drug resistance. Further, these annotations can be used to infer drug susceptibility and the drug combination that may be the most effective against a bacterial infection.

P28: Predicting cancer outcomes based on gene-expression profiles more accurately with deep neural networks

Subject: Machine learning

Presenting Author: Kimball Hill, Brigham Young University, United States

Co-Author(s): Stephen Piccolo, Brigham Young University, United States

ABSTRACT: The development of next-generation sequencing technologies has led to the creation of massive "omic" datasets, the analysis of which has imposed a challenging task for bioinformatics researchers. Although "shallow" machine-learning algorithms and statistical models have contributed greatly to the successful analysis of such data, these models frequently are unable to deal with the complexity of omic data and are difficult to optimize due to the need for feature engineering. Recent developments in the field of deep learning have shown that deep neural networks (DNNs) can outperform shallow algorithms in a variety of applications. These improvements are mostly demonstrated by DNNs' performance in computer vision; however, their usefulness is being demonstrated in an increasing number of other applications such as natural language processing and diverse classification problems. DNNs are highly customizable, overcoming many limitations imposed by shallow algorithms. By utilizing the latest techniques in DNN architecture design, including the use of transfer learning, self-normalizing networks, and stacked auto-encoders, we sought to generate more accurate predictions of patient diagnoses, outcomes, and treatment responses with transcriptomic data, which could more reliably inform oncologists' patient-care decisions. In our comparison across more than 30 different classification algorithms, DNNs performed best for the majority of 40+ different classification problems. In addition, we further improved the models by transferring network layers trained on other, similar models. These results encourage future research such as the unpacking of high performing and transferable DNNs to uncover relationships between genes and to understand how they influence the models' classification decisions.

P29: Computational and cultural aspects of improved attribution

Subject: other

Presenting Author: Kristi Holmes, Northwestern University, United States

Co-Author(s):
Melissa Haendel, Oregon State University, United States
David Eichmann, University of Iowa, United States
Patty Smith, Northwestern University, United States
Nicole Vasilevsky, Oregon Health & Science University, United States
Marijane White, Oregon Health & Science University, United States
Sara Gonzales, Northwestern University, United States
Karen Gutzman, Northwestern University, United States

ABSTRACT: Open science practices, collaborative team science, and a drive to understand meaningful outcomes and impacts have transformed research at all levels. It is not sufficient to consider scholarship simply from the perspective of the number of papers written, citations garnered, and grant dollars awarded. We must enable a more nuanced characterization and contextualization of the wide array of contributions of varying types and intensities that are necessary to move science forward. Unfortunately, little infrastructure exists to identify, aggregate, present, and (ultimately) assess the impact of these contributions. Moreover, these challenges are technical as well as social and require an approach that assimilates cultural perspectives for investigators and organizations, alike.

Here we will present ongoing work through the National Center for Data to Health (CD2H) to address these challenges, with a special emphasis on the unique needs and opportunities for trainees and early stage investigators (ESI) in translational science, especially in data science and informatics. We will discuss contributor roles, research products, and scholarly workflows that can be leveraged for ESI to put their best foot forward to more effectively communicate their science, get credit for their work, and ultimately drive knowledge to impact. We will also examine this topic from an institutional perspective to identify new opportunities for institutions to integrate workflows that will enable them to recognize and credit a diverse complement of work.

P30: A new computational pipeline for PAR-CLIP characterizes a key immune regulatory mechanism

Subject: inference and pattern discovery

Presenting Author: Rachel Hovde, Chimera Bioengineering, United States

Co-Author(s):
Gus Zeiner, Chimera Bioengineering, United States
Melissa Fardy, Chimera Bioengineering, United States
Krista McNally, Chimera Bioengineering, United States
Jay Danao, Chimera Bioengineering, United States
Charlotte Davis, Chimera Bioengineering, United States
Joe Solvason, Chimera Bioengineering, United States

ABSTRACT: RNA-binding proteins are key effectors of post-transcriptional gene regulation, but for most RNA-binding proteins, the scope and specifics of the RNA-binding landscape are unknown. In 2010, Hafner et al. described the PAR-CLIP assay, an approach that uses crosslinking, immunoprecipitation and next-generation sequencing to yield a transcriptome-wide RNA:protein interaction map at single-nucleotide resolution. To facilitate data analysis, Corcoran et al. (2011) developed PARalyzer, an algorithm for PAR-CLIP data analysis that predicts which groups of short reads are derived from true protein binding sites.

Although PAR-CLIP/PARalyzer is a powerful workflow, it was designed to produce short (20-30nt) sequence reads that often

map ambiguously to the genome, and it is not optimized for deep-sequenced datasets that span multiple timepoints. We have developed a modified version of PAR-CLIP that produces longer sequence reads and GOLDMINE, a tailored informatics pipeline that efficiently processes multiple terabytes of HiSeq data. GOLDMINE uses a characteristic read distribution pattern to separate true binding sites from random noise. Following identification of binding sites, it models the secondary structure of overlapping k-mer segments of these sites to identify conserved structures predictive of the presence of the binding protein. By using this workflow to find protein binding sites in activated and unactivated T cells, we are characterizing the regulatory activity of an RNA-binding protein that is critical to immune cell function.

P31: Food preservatives induce Proteobacteria dysbiosis of the human gut microbiota

Subject: Metogenomics

Presenting Author: Tomas Hrncir, Czech Academy of Sciences, Czech Republic

Co-Author(s):
Lucia Hrncirova, Czech Academy of Sciences, Czech Republic
Vladimira Machova, Czech Academy of Sciences, Czech Republic
Eva Trckova, Czech Academy of Sciences, Czech Republic

ABSTRACT: The incidence of autoimmune diseases is increasing worldwide. Recent data suggest that gut microbiota can modulate not only local but also systemic immune responses. In this study, we focus on environmental factors, specifically food preservatives, which may modify the composition of gut microbiota and thus influence host's immune responses. To address this issue, we have administered either sterile water or water supplemented with additives to C57BL/6 mice colonized with human microbiota. The daily intake of additives was calculated to match the maximum daily intake reached in human populations in Europe. We have analyzed the effect of additives on microbial composition and diversity by amplification and high-throughput sequencing of the hypervariable regions of the 16S rDNA genes. The resulting sequences were processed using QIIME2 software package. Our results indicate that commonly used food preservatives can decrease the diversity of the human gut microbiota and also trigger Proteobacteria dysbiosis. *

P32: ORCHID: a method for detecting short-range chromatin interactions in high-resolution 5C and Hi-C datasets

Subject: Machine learning

Presenting Author: Fei Ji, Massachusetts general hospital, United States

Co-Author(s):
Sharmistha Kundu, Massachusetts general hospital, United States
Robert Kingston, Massachusetts general hospital, United States
Ruslan Sadreyev, Massachusetts general hospital, United States

ABSTRACT: The chromatin interaction assays 5C and Hi-C are robust techniques to investigate spatial organization of the genome by capturing interaction frequencies between genomic loci. Although 5C and Hi-C resolution is theoretically restricted only by the length of digested DNA fragments (1Kb-4Kb), intrinsic stochastic noise and high frequencies of background interactions at the distances below 100 Kbp present a significant challenge to understanding short-distance chromatin organization. Here we present the shOrt Range Chromosomal Interaction Detection method (ORCHID) for a comprehensive high-resolution analysis of chromatin interactions in 5C and Hi-C experiments. This method includes background correction of raw interaction frequencies for individual primers or genomic bins, empirical correction for distance dependency of background noise, and detection of areas of significant interactions. When applied to publicly available datasets, ORCHID improves the identification of small (20-200Kb) interaction domains. Unlike larger classic TADs, these chromatin domains are often specific to cell type and functional state of the genomic region. In addition to the expected associations (e.g. with CTCF, cohesin, and mediator complexes), these domains show significant associations with other DNA-binding proteins. An important subtype of these small domains is fully covered and controlled by Polycomb Repressive Complex 1 (PRC1), which mediates transcriptional repression of many key developmental genes. As a separate unexpected example of a potential new mode of regulating chromatin interactions, the binding of RING1B, an essential subunit of the PRC1 complex, is also enriched near domain boundaries at the focused loci that do not necessarily correspond to repressed promoters.

P33: Clustering of Protein Conformations using Parallelized Dimensionality Reduction

Subject: Optimization and search

Presenting Author: Arpita Joshi, University of Massachusetts, Boston, United States

Co-Author(s):
Nurit Haspel, Umass Boston, United States

ABSTRACT: Analyzing the conformational pathways that a macromolecule undergoes is imperative to understanding its function and dynamics. We present a combination of techniques to sample the conformational landscape of proteins better and faster. Datasets representing these landscapes of protein folding and binding are complex and high dimensional. Therefore, there is a need for dimensionality reduction methods that best preserve the variance in the data, and facilitate the analysis of the data. The crux of this work lies in the way this is done. We start with a non-linear dimensionality reduction technique, Isomap, which has been shown to produce better results than linear dimensionality reduction in approximating the complex niceties of protein folding. However, the algorithm is

computationally intensive for large proteins or a large number of samples (samples here refer to the various conformations that are used to ascertain the pathway between two distinctively different structures of a protein). We present a parallel algorithm written in C, using OpenMP, with a speed-up of approximately twice. The results obtained are coherent with the ones obtained using sequential Isomap. Our method uses a distance function to calculate the distance between the points that in turn measures the similarity between the conformations that each of these points represent. The output is a lower-dimensional projection that can be used later for purposes of visualization and analysis. A proof of quantitative validation comes with the least RMSD computation for the two embeddings. The algorithm also makes efficient use of the available memory.

P34: Optimizing nontuberculous mycobacteria (NTM) de novo genome assemblies for application in clinical case studies

Subject: Optimization and search

Presenting Author: Sara Kammlade, National Jewish Health, United States

Co-Author(s):
Nabeeh Hasan, National Jewish Health, United States
L. Elaine Epperson, National Jewish Health, United States
Michael Strong, National Jewish Health, United States
Rebecca Davidson, National Jewish Health, United States

ABSTRACT: To enable studies related to bacterial acquisition and clinical infections of nontuberculous mycobacteria (NTM), we developed a standardized bioinformatic analysis pipeline to process sequenced bacterial isolates from paired-end Illumina reads to fully annotated genomes and a companion PostgreSQL genomic database. Our NTM Genomes Database includes 1200+ isolates from 20 different NTM species which have been processed through our automated and optimized steps for read-trimming, de novo genome assembly, species identification using the average nucleotide identity (ANI) method, contig-ordering against a reference genome, and comprehensive annotation of genomic features. To optimize genome assembly methods and explore the theoretical potential of assembling complete genomes in the context of NTM, we performed experiments testing different parameter combinations in Skewer, SPAdes, and Unicycler on sequences from Illumina MiSeq (2x300bp) and HiSeq (2x250bp) platforms as well as on synthetic reads of varying read lengths and sequencing depths derived from published complete genomes. Assemblies from Illumina data revealed a negative effect of high GC content on assembly quality as measured by NG50. SPAdes and Unicycler yielded similar quality assemblies with Unicycler yielding fewer small (<1Kbp) contigs. From the synthetic reads we found diminished returns on NG50 improvement beyond 25x coverage at 250bp, and failed to assemble a single contig genome using 50Kbp

reads at 60x coverage. Using our high quality genomes we are able to identify core and accessory genes and investigate clinically relevant genotype-phenotype relationships. As an example, we will share findings from a case study of bacterial genomic evolution during a long-term pulmonary infection.

P35: PredHPI: an integrated web-server platform for the prediction and visualization of host-pathogen interactions

Subject: web services

Presenting Author: Rakesh Kaundal, Utah State University, United States
Co-Author(s): Cristian Loaiza, Utah State University, United States

ABSTRACT: Understanding the mechanisms underlying infectious diseases is fundamental to develop prevention strategies. Host-pathogen interactions, which includes from the initial invasion of host cells by the pathogen through the proliferation of the pathogen in their host, have been studied to find potential genomic targets for the development of novel drugs, vaccines, and other therapeutics. Few in silico prediction methods have been developed to infer novel host-pathogen interactions, however, there is no single framework which combines those approaches to produce and visualize a comprehensive analysis of host-pathogen interactions. We present a web server platform named PredHPI available at http://bioinfo.usu.edu/PredHPI/. PredHPI is composed of independent sequence-based tools for the prediction of host-pathogen interactions. The Interolog module, including some of the IMEX databases (HPIDB, MINT, DIP, BioGRID and IntAct), provides three comparison flavors using the BLAST homology results (best-match, ranked-based and generalized). The Domain module, which performs the predictions of the domains, using Pfam and HMMer, and the interactions using the 3DID and IDDI databases. And the GO Similarity module which uses some of the Bioconductor species databases to calculate similarities using GOSemSim R package of the GO terms detected using InterProScan. PredHPI incorporates the functionally to visualize the resulting interaction networks plus the integration of several databases with enriched information about the proteins involved in it. To our knowledge, PredHPI is the first system to build and visualize interaction networks from sequence-based methods as well as curated databases. We hope that our prediction tool will be useful for researchers studying infectious diseases.

P36: An Automated Case Notes System for Psychiatrists Using Text Mining

Subject: Machine learning

Presenting Author: Nazmul Kazi, Montana State University, United States
Co-Author(s): Indika Kahanda, Montana State University, United States

ABSTRACT: Current health care systems require clinicians to spend a substantial amount of time to digitally document their interactions with their patients through the use of electronic health records (EHRs), limiting the time spent on face-to-face patient care. Moreover, the use of EHRs is known to be highly inefficient due to additional time it takes for completion, which also leads to clinician burnout. In this project, we explore the feasibility of developing an automated case notes system for psychiatrists using text mining techniques that will listen to doctor-patient conversations, generate digital transcripts using speech-to-text conversion, classify information from the transcripts by identifying important keywords, and automatically generate structured case notes.

In our preliminary work, we develop a human powered doctor-patient transcript annotator and obtain a gold standard dataset through National Alliance of Mental Illness (NAMI) Montana. We model the task of classifying parts of conversations in to six broad categories such as medical and family history as a supervised classification problem and apply several popular machine learning algorithms. According to our preliminary experimental results obtained through 5-fold cross validation, Support Vector Machines are able to classify an unseen transcript with an average AUROC (area under the receiver operating characteristic curve) score of 89%. Currently, we are working on developing information extraction techniques to generate structured case notes from these classified information. At the same time, we are investigating the recording environment most effective for automatically transcribing a doctor-patient conversation using existing speech-to-text tools with built-in multi-speaker detection capability.

P37: A systems biology approach to define essential kinases in small cell lung cancer

Subject: other

Presenting Author: Jihye Kim, University of Colorado Denver Anschutz Medical Campus
Co-Author(s):
Daniel Foster, National Jewish Health, United States
Rangnath Mishra, National Jewish Health, United States
James Finigan, National Jewish Health, United States
Jeffrey Kern, National Jewish Health, United States
Aik Choon Tan, University of Colorado Denver, United States

ABSTRACT: Small cell lung cancer (SCLC) is a deadly cancer where its five-year survival rate is < 7% and kills approximately 30,000 lives this year. Treatment of SCLC using the chemotherapy combination of cisplatin and etoposide with radiation therapy has not changed in

almost 30 years. Therefore, novel therapies are needed for this disease. Building on the role of kinases and their regulation of cell growth and survival, we hypothesized that kinases regulate cell survival pathways in SCLC (essential kinases) and they may be effective targets as novel monotherapy, or act synergistically with standard chemotherapy, and improve therapeutic outcome. To test this hypothesis, we employed a systems biology approach to identify essential kinases in SCLC. We performed in vivo kinome-wide screening using an shRNA library targeting human kinases on seven chemonaïve SCLC patient derived xenografts (PDX). We developed a suite of bioinformatics tools to deconvolute the kinome screening data, and identified 23 essential kinases found in two or more PDX models. The top essential kinases were RET, MTOR and ATM. We connected these kinases to our drug database to identify specific inhibitors as potential therapy and performed in vitro and in vivo validation of their efficacy. Notably, monotherapy with a small molecule inhibitor targeting mTOR significantly reduced SCLC tumor growth in vivo proving mTOR's essential kinase function. In addition, mTOR inhibition synergized with standard chemotherapy to significantly augment tumor responses in SCLC PDX models. These results warrant the further investigation of MTOR inhibitors combined with chemotherapy as novel treatment for SCLC.

P38: Comparative Analysis of Germline Microsatellites in the 1,000 Genomes Project

Subject: Metogenomics

Presenting Author: Nicholas Kinney, Virginia College of Osteopathic Medicine, United States

Co-Author(s):
Kyle Titus-Glover, Virginia Tech, United States
Robin Varghese, Edward Via College of Osteopathic Medicine, United States
Pawel Michalak, Edward Via College of Osteopathic Medicine, United States
Han Liao, Virginia Tech, United States
Ramu Anandakrishnan, Edward Via College of Osteopathic Medicine, United States
Arichanah Pulenthiran, Edward Via College of Osteopathic Medicine, United States
Lin Kang, Edward Via College of Osteopathic Medicine, United States
Harold Garner, Edward Via College of Osteopathic Medicine, United States

ABSTRACT: Microsatellites are regions of DNA characterized by short – one to six base pair – motifs repeated in tandem to form an array. Over 600,000 unique microsatellites exist in the human genome embedded in gene introns, gene exons, and regulatory regions. Indeed they are well established as an important source of genetic variation. A number of databases provide searchable interfaces to microsatellites within the human reference genome; however, none provide data on actual polymorphism rates among and within human populations. We introduce the Comparative Analysis of Germline Microsatellites (CAGm) Database. The database is designed to assist with future studies of germline microsatellites and enhance our understanding of human genetic variation. Samples can be easily grouped by population,

ethnicity, and gender. Microsatellites can be searched by gene, functional element, and location. Users can query genotypes, view multiple sequence alignments, and easily download data for further analysis. The database has a wide range of additional capabilities. Database content is fully described with examples and future directions are discussed. The database is freely available at http://www.cagmdb.org/.

P39: Omic profiling in healthy volunteers taking celecoxib reveals novel biomarkers regulated by cyclooxygenase-2

Subject: System integration

Presenting Author: Nicholas Kirkby, Imperial College London, United Kingdom

Co-Author(s):
Sarah Mazi, Imperial College London, United Kingdom
Timothy Warner, Queen Mary University of London, United Kingdom
Jane Mitchell, Imperial College London, United Kingdom

ABSTRACT: Introduction: Nonsteroidal anti-inflammatory drugs (NSAIDs) work by blocking cyclooxygenase (COX)-2 and are amongst the most commonly taken drugs worldwide but they also cause cardiovascular toxicity. Because of their widespread use, these are a major concern but no biomarkers or detailed mechanistic pathways are known. To begin to address this we have performed a first-of-its-kind transcriptomic and proteomic analysis of blood samples from healthy volunteers taking an NSAID.

METHOD: Blood was collected from n=8 healthy male volunteers pre/post 7 days treatment with celecoxib (200mg b.i.d.). The transcriptome was measured with Illumina HumanHT-12v4 arrays and proteome using label-free UPLC-MS/MS with fragments identified using Mascot software. Data were analysed using Genespring and R/Limma by moderated t-test and interpreted using a $p<0.05$ threshold for transcriptomics and a $p<0.1$ discovery threshold for proteomics. Pathway analysis was performed using g:Profiler.

RESULTS: Transcription of 104 mapped genes were altered by celecoxib treatment. Pathway analysis revealed enrichment of genes associated with type I interferon responses, cholesterol metabolism and vasoconstriction. Levels of 26 plasma proteins (of ≈460 identifiable proteins) were also altered. In agreement with the interferon signature seen in the transcriptome, pathway analysis of the proteome data revealed altered proteins mapping to changes in acute inflammatory and acute-phase response networks.

CONCLUSION: This study is the first to apply unbiased 'omic' profiling to question of NSAID cardiovascular toxicity. This proof-of-concept study has provided viable novel targets for generation of mechanistic hypotheses as well as potential biomarkers to identify those most at risk of cardiovascular side effects.

P40: An In Silico Approach For Finding Vaccines

Subject: Machine learning

Presenting Author: Siddharth Krishnakumar, Thomas Jefferson high school for science and technology, United States

ABSTRACT: With the proliferation of antibiotic resistant microbes new novel ways to find effective antibiotics and vaccines have become the need. Immune informatics can address this need by finding vaccines by computing the epitope/antibody(produced by B-cells) interactions of the humoral immune system. B cells is one of the main constituent of the Humoral immune system in mammals. B cells fight against antigens by producing antibodies. B cells consist of B cell receptors (BCRs) on their cell membranes.Specific regions within the BCR called paratopes bind to a specific region of antigen proteins called epitopes. When the paratope exactly binds to the epitope, the immune response produces antibodies that fight the antigen. Identification of an antigen epitope for a B cell to bind and begin proliferation and differentiation is the most fundamental step in developing synthetic vaccines and therapeutic medicines. There are many experimentally available techniques to identify epitopes on an antigen, but they are expensive and difficult to apply.In silico methods offer a fast and cost-effective approach for predicting epitopes. Many insilico techniques that have been developed are based on amino acid propensity scales using a sliding window approach, which doesn't produce ideal predictions. Machine learning techniques like SVM(support vector machines)greatly improve the prediction rate. Current SVMs are based on single propensities, and their prediction rates are subpar.This project aims to find an effective method to identify the epitope sequence using an SVM machine with multiple amino acid propensities to produce a better prediction rate as opposed to other single propensity SVMs.

P41: Searching for translatable alternative splice isoforms in the human proteome

Subject: other

Presenting Author: Maggie Pui Yu Lam, University of Colorado Anschutz Medical Campus
Co-Author(s): Edward Lau, Stanford University, United States

ABSTRACT: The human genome contains over 100,000 alternative splice isoform transcripts, but the biological functions of most isoform transcripts remain unknown and many are not translated into mature proteins. A full appreciation of the biological significance of alternative splicing therefore requires knowledge of isoforms at the protein level, such as using mass spectrometry-based proteomics. One described is to perform in-silico translation of alternative transcripts, and then to use the resulting custom FASTA protein sequence databases with a database search engine for protein identification in shotgun proteomics. However, challenges remain as custom protein databases

often contain many sequences that are in fact not translated as proteins inside the cell, thus contributing to a high false discovery rate in proteomics experiments.

We describe here a computational workflow and software to generate custom protein databases of alternative isoform sequences using RNA-seq data as input. The workflow is designed with the explicit goal to minimize untranslated sequences to rein in false positives. To evaluate its performance, we processed public RNA sequencing data from ENCODE to build custom FASTA databases for 10 human tissues (adrenal gland, colon, esophagus, heart, lung, liver, ovary, pancreas, prostate, testis). We applied the databases to identify unique splice junction peptides from public mass spectrometry data of the same human tissues on ProteomeXchange. We identified 1,984 protein isoforms including 345 unique splice-specific peptides not currently documented in common proteomics databases. We suggest that the described proteotranscriptomics approach may help reveal previously unidentified alternative isoforms, and aid in the study of alternative splicing.

P42: A human disease network from gene-publication relationships on PubMed

Subject: inference and pattern discovery

Presenting Author: Edward Lau, Stanford University, United States

Co-Author(s):
Cody Thomas, University of Colorado AMC, United States
Maggie Pui Yu Lam, University of Colorado AMC, United States

ABSTRACT: Human diseases can be represented as a network connecting similar disorders based on their shared phenotypic and molecular characterizations. Network analysis of disease-disease relationships can yield insights into important biological processes and pathogenic pathways. We recently described a method to determine the semantic similarity between a gene or protein and the literature publications related to a disease, by combining PubMed web queries and curated/text-mined annotations of gene-PMID links from NCBI. We devised a weighted co-publication distance metric to score gene-disease co-occurrences in PubMed, where genes with many non-specific publications are down-ranked whereas recent and high-impact publications are given more weight. We show that this method outperforms existing bibliometric analysis in predicting benchmark gene lists of disease terms. Using this method, we have now compiled significant protein lists from over 20,000 human disease or disease phenotype terms from three standardized vocabularies, namely Disease Ontology (DO), Human Phenotype Ontology (HPO), and Pathway Ontology (PWO). We find that disease terms are associated with specific popular protein lists that inform on protein-disease relationships. The PubMed-based disease network recapitulates several

known properties from previous "diseasomes" constructed from OMIM or phenotypic similarity data (e.g., Barabási 2007), including the centrality of metabolic diseases and clustering of related diseases around high-level hub terms. We discuss applications for the disease network, including (i) finding commonly associated diseases from a list of differentially expressed genes in a RNA-seq experiment, and (ii) using gene-disease relationship to predict hidden disease genes in a particular disease.

P43: Unbiased Pathway Detection Expands Cancer Pathways

Subject: Networking

Presenting Author: Chih-Hsu Lin, Baylor College of Medicine, United States

Co-Author(s):
Stephen Wilson, Baylor College of Medicine, United States
Teng-Kuei Hsu, Baylor College of Medicine, United States
Minh Pham, Baylor College of Medicine, United States
Olivier Lichtarge, Baylor College of Medicine, United States

ABSTRACT: Pathways are a type of functional gene group and they help to understand biological systems by representing how signals are transmitted/received and which genes/proteins interact. Conventionally, domain experts manually annotate pathways based on the literature. Thus, the unbiased detection of functional gene groups solely based on the gene-gene interaction network structure may provide novel insights. Here, we hypothesized that gene members in a functional gene group interact within the group more than outside the group. We developed Recursive Louvain algorithm to detect communities (i.e., clustered gene groups) on a human protein-protein interaction network. 85.2 % of the communities overlapped with known functional pathways and disease pathways significantly compared to a random gene group control, whereas 452 communities didn't and may be potentially novel functional gene groups. In addition, variants of genes overlapping with communities are more likely to be pathogenic in ClinVar and have high evolutionary impact quantified by Evolutionary Action (chi-squared test $p << 0.0001$). As a case study in head and neck cancer, we found the RNA-seq profiles of 10 communities could separate survival by K-means clustering significantly (log-rank $q <= 0.1$). Also, those 10 communities are linked to cancer hallmarks. More importantly, one community related to cell adhesion could stratify patient survival independent of clinical data and immune response (Cox multivariate analysis $q = 0.022$). In conclusion, the communities recover known functional and disease pathways, and could be used as cancer survival predictors. This study will help understanding of cancer pathways and provide biomarkers for cancer patients.

P44: Identifying HCV-Host Interactions from Amino Acids Sequences Using SVM Identifying HCV-Host Interactions from Amino Acids Sequences Using SVM

Subject: Machine learning

Presenting Author: Xin Liu, XuZhou Medical University, China

Co-Author(s):
Wei Geng, XuZhou Medical University, China
Dan Wang, Xuzhou Medical University, China
Xue Piao, Xuzhou Medical University, China
Ting Yang, Xuzhou Medical University, China

ABSTRACT: Detecting the interactions between the hepatitis C virus (HCV) and human proteins will facilitate our understanding of the pathogenesis and is helpful in searching for new drug targets. Many researchers have focused on the computing perspective to study the protein–protein interactions (PPIs), but most of them have been designed for PPIs within the same species, which is not fit for different species. In this paper, we developed a novel computational model to predict interaction between HCV and human proteins. As the position specific scoring matrix (PSSM) not only preserves the positional information of the sequence, but also retains the chemical information of the protein, we used the local directional texture pattern (LDTP) to further extract information from the PSSM. Then, support vector machine (SVM) was used to implement the classification. When performed on the HCV dataset, the accuracy of the proposed model could achieve 86.7%, which was superior than most of the previous methods. When performed on an independent dataset, the accuracy achieved 73.9%. We also made a comparison between some state-of-the-art algorithms with our method, and the results showed that the proposed method is simple, effective, and can be used for future proteomics research.

P45: Modeling the Structure of BioGRID PPI Networks

Subject: Qualitative modeling and simulation

Presenting Author: Sridevi Maharaj, University of California-Irvine, United States

Co-Author(s):
Pedro Silva, University of California-Irvine, United States
Zarin Ohiba, University of California-Irvine, United States
Wayne Hayes, University of California-Irvine, United States

ABSTRACT: Protein-protein interaction (PPI) networks are being continuously updated but are still incomplete, sparse, and have false positives and negatives. Amongst the heuristics employed to describe network topology, graphlets have emerged successful in quantifying local structure of biological networks. Some studies analyzing the graphlet degree distributions and relative graphlet frequency, found Geometric (GEO) networks to be a reasonable basis for modeling PPI networks. However, all extensive studies to model PPI networks as

a whole utilized older PPI network data. While there are numerous techniques through which PPI data can be curated, in this study, we re-evaluate these models on the newest PPI data available from BioGRID for the following nine species: AThaliana, CElegans, DMelanogaster, EColi, HSapiens, MMusculus, RNorvegicus, SCerevisiae, and SPombe. To the best of our knowledge, this has not yet been performed, as the data is relatively new. We compare the graphlet distributions of several models to distributions of the updated networks and analyze their fit using several measures that have been shown to be suitable for measuring network distances (or similarities): RGFD, GDDA, Graphlet Kernel, and GCD. Despite minor behavioral differences amongst the comparison measures, we find that other than the Sticky model, the Scale-Free Gene Duplication and Divergence (SFGD) and Scale-Free (SF) models unanimously outperform other traditional models (including GEO and GEOGD) in matching the structure of these 9 BioGRID PPI networks. We further corroborate these results using machine learning classifiers to categorize each species as a network model and visualize these results using t-SNE plots. *

P46: Select: A SQL-based, high-resolution selection scanning tool to Identify genomic selection signals using next-generation sequencing data

Subject: inference and pattern discovery

Presenting Author: Hannah Maltba, Brigham Young University, United States

Co-Author(s):
Sean Beecroft, Brigham Young University, United States
Spencer Smith, Brigham Young University, United States

ABSTRACT: Next-generation sequencing (NGS) enables high-resolution, genomic-based evolution studies, but an exhaustive, genome-wide selection scan on NGS data is computationally intensive. We developed a comprehensive software Selection Test (SelecT) to identify specific loci under selection using NGS data at base pair resolution. SelecT calculates five statistics to identify evolutionary patterns in allele frequency (Fst, ΔDAF) and haplotype homozygosity (EHH, iHS, XPEHH) that locate regions with multiple, strong evolutionary signals. Data and results are stored in a database, making it easy to visualize results and run additional queries on the data. Multiple tests can be run on already imported populations without reloading any files and users may choose to run statistics one at a time or all at once. SelecT has an intuitive user interface, runs in only minutes, and can be used to identify selection in any organism.

P47: Codon Pairs are Phylogenetically Conserved: Codon pairing as a novel phylogenetic character state for parsimony and alignment-free methods

Subject: inference and pattern discovery

Presenting Author: Lauren McKinnon, Brigham Young University, United States

Co-Author(s): Justin Miller, Brigham Young University, United States

ABSTRACT: Identical codon pairing and co-tRNA codon pairing increase translational efficiency within genes when two codons that encode the same amino acid are located within a ribosomal window. By examining identical and co-tRNA codon pairing independently and combined across 23 423 species, we determined that both pairing techniques are phylogenetically informative using either an alignment-free or parsimony framework in all domains of life. We also determined that the minimum optimal window size for conserved codon pairs is typically smaller than the length of a ribosome. We thoroughly analyze codon pairing across various taxonomic groups. We determined which codons are more likely to pair and we analyze the frequencies of codon pairings between species. The alignment-free method does not require orthologous gene annotations and recovers species relationships that are more congruent with established phylogenies than other alignment-free techniques in all instances. Parsimony recovers trees that are more congruent with the established phylogenies than the alignment-free method in four out of six taxonomic groups. Four taxonomic groups do not have sufficient ortholog annotations and are excluded from the parsimony and/or maximum likelihood analyses. Using only codon pairing, the alignment-free or parsimony-based approaches recover the most congruent trees compared with the established phylogenies in six out of ten taxonomic groups. Since the recovered phylogenies using only codon pairing largely match established phylogenies, we propose that codon pairing biases are phylogenetically conserved and should be considered in conjunction with current techniques in future phylogenomic studies.

Availability: https://github.com/ridgelab/codon_pairing

P48: ExtRamp: A novel algorithm for extracting the ramp sequence based on the tRNA adaptation index or relative codon adaptiveness

Subject: inference and pattern discovery

Presenting Author: Justin Miller, Brigham Young University, United States

Co-Author(s):
Logan Brase, Brigham Young University, United States
Perry Ridge, Brigham Young University, United States

ABSTRACT: Different species, genes, and locations within genes use different codons to fine-tune gene expression. Within genes, the ramp sequence assists in ribosome spacing and decreases downstream

collisions by incorporating slowly-translated codons at the beginning of a gene. Although previously reported as occurring in some species, no previous attempt at extracting the ramp sequence from specific genes has been published. We present ExtRamp, a software package that quickly extracts ramp sequences from any species using the tRNA adaptation index or relative codon adaptiveness. Different filters facilitate the analysis of codon efficiency and enable researchers to identify genes with a ramp sequence. We validate the existence of a ramp sequence in most species by running ExtRamp on 229,742,339 genes across 23,428 species. We evaluate differences in reported ramp sequences when we use different parameters. Using the strictest ramp sequence cut-off, we show that across most taxonomic groups, ramp sequences are approximately 20-40 codons long and occur in about 10% of gene sequences. We also show that as gene expression increases, more ramp sequences are identified in Drosophila melanogaster. We provide a framework for performing this analysis on other species and present our algorithm at https://github.com/ridgelab/ExtRamp.

P49: Metabolic profiling using a UHPLC-MS/MS-based platform to quantify amines, amino acids and methylarginines in plasma from cyclooxygenase-2 knockout mice

Subject: other

Presenting Author: Jane Mitchell, Imperial College London, United Kingdom

Co-Author(s):
Elizabeth Want, Imperial College London, United Kingdom
Blerina Ahmetaj-Shala, Imperial College London, United Kingdom
Micahel Olanipekun, Imperial College London, United Kingdom
Abel Tesfai, Imperial College London, United Kingdom
Yu He, Imperial College London, United Kingdom
Rolf Nüsing , Imperial College London, United Kingdom
Nicholas Kirkby, Imperial College London, United Kingdom

ABSTRACT: Amine quantification is an important area in biomedical research and in patient stratification for personalised medicine. One important pathway critically reliant on amine levels is nitric oxide synthase (NOS). NOS forms NO which mediates processes in the cardiovascular, immune and nervous systems. NO release is regulated by levels of (i) the substrate, arginine, (ii) amino acids which cycle with arginine and (iii) methylarginine inhibitors of NOS. However, measurement of a wide range of amines, including methylarginines, on a common platform has been challenging. To address this, we have recently reported on an analytical method where a wide range of amines including amino acids and methylarginines can be measured in a common plasma sample. Using high-throughput ultra-high-performance liquid chromatography-tandem mass spectrometry (UHPLC-MS/MS) ≈40 amine analytes, including arginine and methylarginines were detected and quantified on a molar basis

(Ahmetaj-Shala et al., 2018; Scientific reports, 13987). Our previous work using transcriptomic analysis revealed a relationship between the enzyme cyclooxygenase-2 in the kidney, renal function and genes that regulate arginine and methylarginine levels (Ahmetaj-Shala et al., 2015, Circulation, 132, 633-642). In the current study we applied the UHPLC-MS/MS platform to the measurement of amines and methylarginines in plasma from cyclooxygenase-2 knockout mice. Principle component analysis showed separate clustering of the two groups and quantification of analytes confirmed increases in methylarginine levels in plasma from cyclooxygenase-2 knockout mice. These results illustrate the useful application of this platform and indicate that genetic deletion of cyclooxygenase-2 increases methylarginine levels and disrupts the amine metabolome.

P50: The characterization of different cell types using the Benford law

Subject: inference and pattern discovery

Presenting Author: Sne Morag, Ariel University, Israel

Co-Author(s): Mali Salmon-Divon, Ariel University, Israel

ABSTRACT: Abstract publication declined

P51: Integrative analysis of transcriptomics and proteomics to detect novel protein isoforms from alternatively spliced transcripts induced by SF3B1 spliceosomal mutations

Subject: Metogenomics

Presenting Author: Kelsey Nassar, University of Colorado Anschutz Medical Campus

Co-Author(s):
Hyunmin Kim, University of Colorado Anschutz Medical Campus
Jihye Kim, University of Colorado Anschutz Medical Campus
Maggie Lam, University of Colorado Anschutz Medical Campus
Aik Choon Tan, University of Colorado Anschutz Medical Campus

ABSTRACT: Alternative splicing (AS) contributes to transcriptional complexity and is hypothesized to alter the proteome. AS events have been found to be increased in various cancers, however, the functional consequences of AS events on tumorigenesis remains unclear. Recently, mutations in core spliceosomal proteins, such as SF3B1, have been identified at a high frequency in multiple cancers. Next generation RNA-sequencing (RNA-seq) has identified that SF3B1 mutations result in global transcriptomic alterations in AS, primarily an increase in alternative 3′ splice site recognition. We hypothesized that mutations in SF3B1 increases proteome diversity through alternative splice variants that contribute to tumorigenesis. To test this hypothesis, we performed deep RNA-Sequencing on an SF3B1-Mutant and SF3B1-WildType uveal melanoma cell lines. We developed SALSA (Systemic ALternative Splice Analysis), for RNA-seq analysis to identify novel AS events as a result

of SF3B1 mutations. In addition, we conducted proteome-wide mass spectrometry (MS) to identify novel protein isoforms detected from RNA-seq. We curated a novel peptide database from our custom AS events identified by SALSA to detect novel protein isoforms. From this integrative analysis, we identified 76 novel peptides enriched in SF3B1-Mutant cells detected at both RNA-seq and MS levels. From the MS peptide list, we validated SETD5, an 3' alternatively spliced transcript. To our knowledge, this is the first description of a novel alternatively spliced transcript that results in a novel protein in SF3B1-mutant cells. This preliminary analysis lays the ground work for further identification of novel protein isoforms resulting from SF3B1 mutations that ultimately may contribute to tumorigenesis.

P52: Mining Heterogenous Relationships from Pubmed Abstracts Using Distant Supervision

Subject: Text Mining

Presenting Author: David Nicholson, University of Pennsylvania, United States

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

ABSTRACT: Identifying mechanisms underlying disease and finding drugs that intervene or prevent such mechanisms is an important task in biomedical sciences. One approach to identify these drug targets is to combine evidence from multiple sources, including scientific publications, to model potential relationships between drugs, genes and diseases and interpolate novel relationships. Previously, a heterogeneous (hetnet) network, called hetionet pioneered such efforts by integrating various relationships from multiple data sources; however, building such network requires hours upon hours of manual curation, which is not feasible in a larger scale. We aim to remedy this bottleneck by extracting multiple relationships from Pubmed abstracts via a distant supervision approach. This approach circumvents the time-consuming task of obtaining "ground-truth" training labels via the data programming paradigm, which consists of using a simple set of programs, also called label functions, to probabilistically label large training datasets. Using these datasets, we then train machine learning classifiers to classify whether or not a sentence mentions a relationship. We evaluated this approach by assessing label function accuracy, determining if label functions can transfer between different relationship types and measuring the value of these estimated labels via downstream analyses. These analyses consisted of comparing bag of words, logistic regression and neural network-based methods. Overall performance remains modest, suggesting that label functions may need to be improved or that abstracts may not be sufficient for high-accuracy relationship extraction; however, our results also suggest that label functions can transfer across various relationships suggesting that hetnet construction through this approach may be viable.

P53: Mutational impact on protein structure and function in endometrial cancer

Subject: other

Presenting Author: Amanda Oliphant, Brigham Young University, United States

Co-Author(s):
Emily Hoskins, Brigham Young University, United States
Daniel Cui Zhou, Washingto University in St. Louis, United States
David Adams, Brigham Young University, United States
Sean Beecroft, Brigham Young University, United States
Li Ding, Washington University in St. Louis, United States
Samuel Payne, Brigham Young University, United States

ABSTRACT: DNA mutation is a well-known driver for cancers, including endometrial and uterine cancer. Although many mutation sites have been discovered in large population studies like TCGA and CPTAC, the functional impact of these mutations often remains unclear. The sites of mutation in endometrial cancer are often not shared between individuals in a cohort, making it difficult to interpret genomic data. Here we show how integrating protein three-dimensional structure information can provide insight into the effects of mutations on protein function. We use bioinformatics tools to locate clustered mutations in a cohort of individuals with endometrial cancer. These mutational hotspots point to functional areas of a protein that are frequently disrupted in cancerous cells. Identifying these hotspots allows us to see how mutations that are not located near each other on a genomic scale may have the same effect on protein function. For example, we found a cluster of mutations in an F-box/WD repeat-containing protein and an E3 ubiquitin ligase which may interfere with its ability to regulate target proteins such as Cyclin-E. Patients exhibiting any of the mutations in a functionally significant cluster may be treated in a similar manner. We anticipate that our findings will aid in classifying the specific type of endometrial cancer present in an individual, allowing for more personalized treatment strategies. Because generalized treatment is often ineffective, these results have the potential to help us understand hard-to-treat forms of endometrial cancer, leading to better end results.

P54: A platform for community-scale transcriptome-wide association studies

Subject: Data management methods and systems

Presenting Author: YoSon Park, Perelman School of Medicine University of Pennsylvania, United States

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

ABSTRACT: Transcriptome-wide association studies (TWAS) infer causal relationships between genes, phenotypes and tissues using strategies such as 2-sample Mendelian randomization (MR). Such methods largely eliminate the need to access individual-level data and allow openly sharing data and results. Nonetheless, to our knowledge, there

are no public platforms automating quality assurance and continuous integration of TWAS results. Consequently, finding, replicating, and validating causal relationships among millions of similar non-causal relationships remain enormously challenging and are often time- and resource-consuming with many duplicated efforts.

To address this shortcoming, we develop a platform that uses version control software and continuous integration to construct a data resource for the components of TWAS. Community members can contribute additional association studies or methods. We use automated testing to catch formatting mistakes and use pull request functionality to review contributions. We provide a set of tools, available in a Docker container, that perform common downstream analyses using these resources.

Researchers who contribute summary-level datasets substantially increase the impact of their work by making it easy to integrate with complementary datasets. Those who contribute analytical tools will benefit by providing users with numerous off-the-shelf use cases. For this proof-of-concept, we integrate a set of eQTLs provided by the Genotype-Tissue Expression (GTEx) project and a set of curated GWAS summary statistics using 2-sample MR. Our long-term goal for this project is a public community-driven repository where users contribute new summary-level data, download complementary data, and add new analytical methods that enables the field to rapidly translate new studies into actionable findings.

P55: Good Nomen: An Interactive Web Application for Cleaning Clinical Data Using Standardized Terminologies

Subject: Graphics and user interfaces
Presenting Author: Alyssa Parker, Brigham Young University, United States
Co-Author(s): Stephen Piccolo, Brigham Young University, United States

ABSTRACT: Terms used to describe medical conditions, treatments, tests, and outcomes vary widely within and across datasets. This creates difficulties when analyzing such data. For example, The Cancer Genome Atlas uses 13 different terms to describe cyclophosphamide, a drug commonly prescribed to breast-cancer patients. To address this problem, researchers have produced terminologies and thesauri that define standardized terms and synonyms for biomedical concepts. While these resources contain valuable information, restructuring clinical data to conform to these standards may be time consuming and require computational expertise to do this systematically. We have developed Good Nomen, a Web application that allows users to standardize data interactively in a high-throughput manner. Good Nomen accepts data files in a CSV, TSV, or Excel format. Then it asks the user to select a terminology to use in the standardization process. Currently, we support the National Cancer Institute Thesaurus, ICD-

10-CM, and HGNC database. The user selects a column containing categorical data to standardize. Good Nomen then examines the data and uses regular expressions to suggest standardized terms and synonyms from the terminology. If the user accepts these matches, the data values are modified to match the specified terms. The user can also manually standardize the data to ensure that misspellings and additional synonyms are not overlooked. It is our hope that harnessing the power of these terminologies into an intuitive, user-friendly Web application will enable researchers to more easily standardize their data and expedite the process of clinical data analysis.

P56: Proteomics of natural bacterial isolates powered by deep learning-based de novo identification

Subject: Machine learning

Presenting Author: Samuel Payne, Brigham Young University, United States

Co-Author(s):
Joon-Yong Lee, Pacific Northwest National Laboratory, United States
Hugh Mitchell, Pacific Northwest National Laboratory, United States
Meagan Burnet, Pacific Northwest National Laboratory, United States
Sarah Jenson, Pacific Northwest National Laboratory, United States
Eric Merkley, Pacific Northwest National Laboratory, United States
Anil Shukla, Pacific Northwest National Laboratory, United States
Ernesto Nakayasu, Pacific Northwest National Laboratory, United States

ABSTRACT: The fundamental task in proteomic mass spectrometry is identifying peptides from their observed spectra. Where protein sequences are known, standard algorithms utilize these to narrow the list of peptide candidates. If protein sequences are unknown, a distinct class of algorithms must interpret spectra de novo. Despite decades of effort on algorithmic constructs and machine learning methods, de novo software tools remain inaccurate when used on environmentally diverse samples. Here we train a deep neural network on 5 million spectra from 55 phylogenetically diverse bacteria. This new model outperforms current methods by 25-100%. The diversity of organisms used for training also improves the generality of the model, and ensures reliable performance regardless of where the sample comes from. Significantly, it also achieves a high accuracy in long peptides which assist in identifying taxa from samples of unknown origin. With the new tool, called Kaiko, we analyze proteomics data from six natural soil isolates for which a proteome database did not exist. Without any sequence information, we correctly identify the taxonomy of these soil microbes as well as annotate thousands of peptide spectra.

P57: Text Mining Novel Disease- and Drug-Specific Pathways

Subject: Text Mining

Presenting Author: Minh Pham, Baylor College of Medicine, United States

Co-Author(s):
Stephen Wilson, Baylor College of Medicine, United States
Chih-Hsu Lin, Baylor College of Medicine, United States
Olivier Lichtarge, Baylor College of Medicine, United States

ABSTRACT: In response to the exponential growth of scientific publications, text mining is increasingly used to extract biological pathways and processes. Though multiple tools explore individual connections between genes, diseases, and drugs, not many extensively examine contextual biological pathways for specific drugs and diseases. In this study, we extracted more than 3,000 functional gene groups for specific diseases and drugs by applying a community detection algorithm to a literature network. The network aggregated co-occurrences of Medical Subject Headings (MeSH) terms for genes, diseases, and drugs in publications. The detected literature communities were groups of highly associated genes, diseases, and drugs. The communities significantly captured genetic knowledge of canonical pathways and recovered future pathways in time-stamped experiments. Furthermore, the disease- and drug-specific communities recapitulated known pathways for those given diseases and drugs. In addition, diseases in same communities had high comorbidity with each other and drugs in same communities shared great numbers of side effects, suggesting that they shared mechanisms. Indeed, the communities robustly recovered mutual targets for drugs (AUROC = 0.75) and shared pathogenic genes for diseases (AUROC = 0.82). These data show that the literature communities not only represented known biological processes but also suggested novel disease- and drug-specific mechanisms, facilitating disease gene discovery and drug repurposing.

P58: A Case Study on the Effects of Noisy, Long-read Correction Approaches on Assembly Contiguity

Subject: other

Presenting Author: Brandon Pickett, Brigham Young University, United States

Co-Author(s):
Justin Miller, Brigham Young University, United States
Perry Ridge, Brigham Young University, United States

ABSTRACT: Third-generation sequencing technologies are advancing our ability to sequence increasingly long DNA sequences in a high-throughput manner. Pacific Biosciences (PacBio) Single-molecule, Real-time (SMRT) sequencing and Oxford Nanopore Technologies (ONT) nanopore sequencing routinely produce raw sequencing reads averaging 20-30kbp in length. Maximum read lengths have, in some

cases, exceeded 100kbp. Unfortunately, these long reads are expensive to generate and have a high error rate (10-15%) when compared with Illumina short reads (1%). The limitation on assembly from high error rates can be mitigated by (a) co-assembling high-error, long reads with low-error, short reads (e.g., MaSuRCA) or (b) correcting the errors prior to assembly. Pre-assembly error correction typically happens by either (a) self-correction or (b) hybrid correction. Self-correction requires increased sequencing depth (and thus expense) and can be done with stand-alone software (e.g., Racon) or via a module in an assembler (e.g., Canu). Hybrid correction involves alignment of low-error, short reads to the raw long reads to generate the consensus (e.g., CoLoRMap). Note that low-error, short reads can also be used to polish the assembled contigs, i.e., correct misassemblies and errors. To investigate how self-correction, hybrid correction, or both correction methods affect assembly contiguity, we tried each approach in a case study. Bonefish (Albula glossodonta) DNA was extracted and sequenced on PacBio Sequel to theoretical 70x coverage and on Illumina HiSeq 2500 to theoretical 100x coverage with paired-end (PE) 2x250 in Rapid run mode. Our assembly results demonstrate that a combination of both approaches generates the most contiguous bonefish assembly.

P59: Measuring chromosome conformation

Subject: Simulation and numeric computing

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

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

ABSTRACT: The in-vivo conformation of chromosomes is an outstanding unsolved problem in structural biology. Most structural information is currently inferred indirectly from Hi-C data, as direct measurements of chromosomal positioning have not been possible for more than a handful of genetic loci. We have previously demonstrated a computational method for scaling direct positioning measurements up to the whole-chromosome scale. Here we present our latest results from simulations and experiments.

P60: Challenges Using Electronic Medical Record for Pharmacokinetic Analysis

Subject: Simulation and numeric computing

Presenting Author: Matthew Shotwell, Vanderbilt University Medical Center, United States

Co-Author(s): Hannah Weeks, Vanderbilt University, United States

ABSTRACT: Hospitalized patients may benefit from individualized drug dosing that is informed by real-time blood sampling and pharmacokinetic analysis. The additional necessary dosing history information and other clinical and demographic factors can be

extracted from the electronic medical record (EMR). However, these data are prone to errors. We consider the impact of incorrect entry of dose administration times and blood sampling times. We further show that, for the intravenously infused antibiotic piperacillin, estimation of a clinically informative measure of drug exposure - the fraction of the dosing cycle in which the blood concentration of drug is above a given efficacy threshold - is robust to many types of error that occur in the EMR. In addition we demonstrate that certain drug administration techniques, including long infusion duration, ensure greater robustness.

P61: Measuring Transcription Factor Activity with Nascent RNA Sequencing

Subject: inference and pattern discovery

Presenting Author: Rutendo Sigauke, University of Colorado Anschutz Medical Campus

Co-Author(s):
Jonathan Rubin, University of Colorado Boulder, United States
Jacob Stanley, University of Colorado Boulder, United States
Robin Dowell, University of Colorado Boulder, United States

ABSTRACT: Transcription factor (TF) proteins control cellular states and functions by regulating the transcription of genes. In order to measure the activity of TFs most studies have taken advantage of Chromatin Immunoprecipitation (ChIP) Assays (Gade and Kalvakolanu, 2012). However, not all TF ChIP-bound sites result in TF activity. In this study we present Transcription Factor Enrichment Analysis (TFEA), a method to identify TF activity using nascent RNA sequencing. Nascent RNA sequencing has allowed for the sequencing of short-lived enhancer RNAs (eRNAs). Previous studies have shown that transcription of eRNAs is a direct measure of TF activity (Hah et al. 2013, Allen et al. 2014). TFEA extends the Motif Displacement Score (MDS) method which uses the colocalization of eRNAs with TF motifs as a measure of TF activity (Azofeifa et al. 2018). TFEA takes advantage of differential transcription of eRNAs, in addition to colocalization of eRNAs with TF motifs, to assess TF activity. TFEA gives a summary report of TFs predicted to be enriched in a given experiment. TFEA was able to identify TFs known to be enriched in several case studies.

P62: Addressing the compositional data problem in sequencing with a novel, robust normalization method

Subject: Simulation and numeric computing

Presenting Author: James St. Pierre, University of Toronto, Canada

Co-Author(s): John Parkinson, Hospital for Sick Children, Toronto, Canada

ABSTRACT: A problem that faces high-throughput sequencing datasets is that raw sequencing data is semi-quantitative due to the random sampling procedure of the sequencing process itself. The raw counts produced only give relative abundances of various genes and must

be appropriately normalized to give an approximation of the absolute abundances of genes in the samples. This 'compositional data problem' in sequencing is especially apparent in the microbiome field. Normalization methods developed for RNA-seq data have been shown to fail when used on 16S microbiome sequencing data, leading to inflated false discovery rates when performing differential abundance analysis. Moreover, the effectiveness of these normalization techniques when used on metagenomics and metatranscriptomics data has yet to be systematically evaluated. We present a novel normalization method that shows improved performance over previous methods (DESeq2, edgeR, and metagenomeSeq) when applied to simulated sequencing data. All current normalization methods have the statistical assumption that most genes (or taxa) are not differentially abundant between experimental groups. The new technique does not have this assumption and is the only method that successfully controls false positive rates during differential abundance testing on a simulated 16S dataset where 50% of taxa were set to be differentially abundant. Even ANCOM and ALDEx2, two compositional data analysis tools previously shown to be more robust than other methods, are shown here to have inflated false positive rates. This new normalization method will be an asset to microbiome researchers, leading to more robust discoveries.

P63: Governance Innovations for Promoting Cross-institutional Electronic Health Data Sharing

Subject: Data management methods and systems

Presenting Author: Kari Stephens, University of Washington, United States

Co-Author(s):
Adam Wilcox, University of Washington, United States
Philip Payne, Washington University, United States
Jason Morrison, University of Washington, United States
Jennifer Sprecher, University of Washington, United States
Rania Mussa, University of Washington, United States
Randi Foraker, Washington University, United States
Sarah Biber, Oregon Health Sciences University, United States
Sean Mooney, University of Washington , United States

ABSTRACT: Cross institutional electronic health data sharing is an essential requirement for health innovation research. Healthcare organizations across the country are governed separately by state and local laws and policies that complicate research related data sharing. Electronic health record (EHR) data are not only highly protected via federal laws (i.e., HIPAA) and regional Internal Review Boards (IRBs), but are also often protected as assets by individual organizations. No clear pathway exists for organizations to execute governance for rapid EHR data sharing, stifling research efforts ranging from simple observational studies to complex multi-institutional trials. Universal governance solutions are essential to provide pathways for data sharing to address the rapid pace of research. The Clinical Translational Science Award (CTSA) Program Data to Health (CD2H) Coordinating Center

has launched a cloud data sharing pilot project to begin addressing this complex issue. In order to configure a web-based data sharing software tool, Leaf, that can cross-query comprehensive harmonized EHR data generated by multiple healthcare organizations, we are exploring a singular governance solution (i.e., embodied in a data use agreement (DUA) and Internal Review Board (IRB) solution) to accommodate both a general and research specific use. While DUAs and IRBs are not streamlined governance solutions, this is an essential first step in creating broader sustainable national governance solutions (i.e., master consortium agreements, access governance policies).

P64: Use of metadata and Bag-of-words to map measurements across observational study data

Subject: inference and pattern discovery

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

Co-Author(s):
Tiffany Callahan, University of Colorado Anschutz Medical Campus
Sonia Leach, University of Colorado Anschutz Medical Campus
David Kao, University of Colorado Anschutz Medical Campus

ABSTRACT: Data integration is an important strategy for validating research results or increasing sample size in biomedical research. Integration is made challenging by metadata and data differences between studies, and is often done manually by a clinical expert for a highly select set of measurements. Unfortunately, this process is rarely documented, and when it is, the details are not accessible, interoperable, or reusable. We explored the utility of using bag-of-words, an information retrieval model, to map medical conditions, characteristics, and lifestyle measurements among multiple studies such as diabetes, age, blood pressure, or alcohol intake. We hypothesized applying cosine similarity to features extracted as a bag-of-words model from observational study measurement annotations would yield accurate recommendations for mapping measurements within and between studies and increase scalability compared to manual mapping. Each measurement's metadata, including descriptions, units, and value-coding, were extracted and then combined for all 105,611 measurements in four cardiovascular-health observational studies. The measurement's combined metadata was input to the bag-of-words model. Cosine similarity of word vectors was used to score similarity between measurement pairs. The highest scoring matches for each measurement were compared to 612 unique expert-vetted, manual mappings. Among the vetted measurement pairings, 99.8% had the correct mapping in the top-10 scored matches, 92.5% had the correct mapping in the top-5, and 55.7% had the correct mapping as the top score. This approach provides a scalable method for recommending measurement mappings in observational study data. Next steps include incorporating additional metadata such as measurement type or a synonyms dictionary for concept recognition.

P65: Visualization Tool for interactive deciphering complex genetic regulation from multi-omic data

Subject: Graphics and user interfaces

Presenting Author: LIN TING-WEI, Linkou Chang Gung Memorial Hospital, Taiwan

ABSTRACT: To decipher the complexity of the regulation relationship within certain biological intervention, an integrative analysis of the data with multiple annotation steps have to be conducted . One of the challenges behind these multi-steps analysis approach is that requirement of hypothesis generated from biologist across variety steps to modify and concentrate to certain details of the result. We provides a framework from visualizing tree relationship between pathways, which using signaling pathway impact analysis for direction. Then, a multilayer network layout to carry the master regulator analysis from prior result for representation. In the end, the biologist can use this framework and interaction with the visualization result to decipher the possible key regulation in their experimental intervention or generate new hypothesis for further experiment validation.

P66: anexVis: visual analytics framework for analysis of RNA expression

Subject: Graphics and user interfaces

Presenting Author: Diem-Trang Tran, University of Utah, United States

Co-Author(s):
Tian Zhang, University of Utah, United States
Ryan Stutsman, University of Utah, United States
Matthew Might, University of Alabama at Birmingham, United States
Umesh Desai, Virginia Commonwealth University, United States
Balagurunathan Kuberan, University of Utah, United States

ABSTRACT: Although RNA expression data are accumulating at a remarkable speed, gaining insights from them still requires laborious analyses, which hinder many biological and biomedical researchers. We introduce a visual analytics framework that applies several well-known visualization techniques to leverage understanding of an RNA expression dataset. Our analyses on glycosaminoglycan-related genes have demonstrated the broad application of this tool, anexVis (analysis of RNA expression), to advance the understanding of tissue-specific glycosaminoglycan regulation and functions, and potentially other biological pathways.

The application is publicly accessible at https://anexvis.chpc.utah.edu/, source codes deposited on GitHub.

P67: Toxicant-protein relation extraction

Subject: Text Mining
Presenting Author: Ignacio Tripodi, University of Colorado, Boulder, United States
Co-Author(s): Lawrence Hunter, University of Colorado, Denver, United States

ABSTRACT: The interaction between chemicals and proteins provides essential information regarding how exposure to certain chemicals affects cell functions. In particular, knowing how chemicals that result in toxicity are associated to the up- or down-regulation of transcription factors, can help elucidate the mechanistic details of such adverse outcomes. Some of this information can be inferred indirectly by chemical-to-gene interactions present in public databases. These resources are, however, updated at varying frequencies and generally incomplete, just as our knowledge of which of the many transcription factors regulate which genes. We propose a text-mining approach where we explore an open-access body of literature, to determine using machine learning and a set of heuristics which chemicals from a list of known toxicants are associated to an increase or decrease of specific transcription factors' activity.

P68: LOINC2HPO: Improving translational informatics by standardizing EHR phenotypic data using the Human Phenotype Ontology

Subject: Text Mining
Presenting Author: Nicole Vasilevsky, Oregon Health & Science University, United States
Co-Author(s):
Aaron Zhang, The Jackson Laboratory, United States
Jean-Philippe Gourdine, Oregon Health & Science University, United States
Amy Yates, Oregon Health & Science University, United States
Melissa Haendel, Oregon Health & Science University, United States
Peter Robinson, The Jackson Laboratory, United States

ABSTRACT: Electronic Health Record (EHR) data are often encoded using Logical Observation Identifier Names and Codes (LOINC), which is a universal standard for coding clinical laboratory tests. LOINC codes encode clinical tests and not the phenotypic outcomes, and multiple codes can be used to describe laboratory findings that may correspond to one phenotype. However, LOINC encoded data is an untapped resource in the context of deep phenotyping with the Human Phenotype Ontology (HPO). The HPO describes phenotypic abnormalities encountered in human diseases, and is primarily used for research and diagnostic purposes. As part of the Center for Data to Health (CD2H)'s effort to make EHR data more translationally interoperable, our group developed a curation tool that is used to convert EHR observations into HPO terms for use in clinical research. To date, over 1,000 LOINC codes have been mapped to HPO terms. To demonstrate the utility of these mapped codes, we performed a pilot study with de-identified data from asthma patients. We were

able to convert 70% of real-world laboratory tests into HPO-encoded phenotypes. Analysis of the LOINC2HPO-encoded data showed that the HPO term eosinophilia was enriched in patients with severe asthma and prednisone use. This preliminary evidence suggests that LOINC data converted to HPO can be used for machine learning approaches to support genomic phenotype-driven diagnostics for rare disease patients, and to perform EHR based mechanistic research.

P69: Exploratory Analysis of Diseased Male and Female Gene Expression Levels

Subject: inference and pattern discovery

Presenting Author: Clarissa White, Brigham Young University, United States

ABSTRACT: BACKGROUND: Most genetic diseases have complex genetic effects and are still not figured out. Amyloidosis is a rare and fatal disease in which an abnormal protein is produced. Transthyretin Amyloidosis (ATTR) is a type of this disease caused by inheriting a gene mutation. This study looks at the the role that gender plays in gene expression levels of subjects with and without ATTR. First, we compared gene expression level differences between males and females with the disease. Second, we used gene expression and gender to predict membership to the asymptomatic and control groups.

METHODS AND FINDINGS: We performed exploratory analyses on a publicly-available dataset of 309 patients. Each patient was either asymptomatic, symptomatic, treated, or a control. The gender and gene expression levels for over 21,000 genes were included. We performed a t-test between males and females to find differences in gene expression levels. We then used Random Forests to predict between asymptomatic patients and controls. Finally, we looked at whether the Random Forest predictions changed when including gender.

CONCLUSIONS: We found that gene expression levels were not significantly different for different genders. We also found that the Random Forest model correctly predicted disease approximately 60% of the time, evidence of slightly better predictions than a random guess. Future research should conduct a study looking for a specific molecular signature, as our results suggested there might be one, but we did not do any analysis to look at what it might be as this research was exploratory.

P70: BioThings API: Building a FAIR API Ecosystem for Biomedical Knowledge

Subject: web services

Presenting Author: Chunlei Wu, The Scripps Research Institute, United States

Co-Author(s):
Jiwen Xin, The Scripps Research Institute, United States
Cyrus Afrasiabi, The Scripps Research Institute, United States
Sebastien Lelong, The Scripps Research Institute, United States
Marco Alvarado Cano, The Scripps Research Institute, United States
Ginger Tseung, The Scripps Research Institute, United States
Trish Whetzel, EMBL-EBI, United States
Shima Dastgheib, NuMedii Inc, United States
Amrapali Zaveri, Maastricht University, Netherlands
Michel Domontier, Maastricht University, Netherlands
Andrew I. Su, The Scripps Research Institute, United States
Chunlei Wu, The Scripps Research Institute, United States

ABSTRACT: Building a web-based API (Application Programming Interface) has been rapidly adopted in the bioinformatics field as a new way of disseminating the underlying biomedical knowledge. While researchers benefit from the simplicity and the high accessibility (A) of available APIs, the findability (F), interoperability (I) and reusability (R) across APIs are largely not well-handled by the community. BioThings API project (http://biothings.io) is tasked to build a FAIR API ecosystem to better serve the underlying inter-connected biomedical knowledge. BioThings API provides three components in its API development ecosystem. First, it provides a family of high-performance APIs for accessing up-to-date annotations for genes, genetic variants, chemicals and drugs. Second, BioThings API packages its API-development best practice into a reusable SDK (Software Development Kit) to help other bioinformaticians to build the same high-quality API to distribute their own specific knowledge. Third, BioThing API provides a platform to foster the findability and interoperability across the community-developed biomedical APIs. Through the SmartAPI application (http://smart-api.info), it provides tools for authoring API metadata following the community supported OpenAPI standard and hosts standardized interactive API documentation. It also defines a set of OpenAPI extensions to provide biomedical-specific semantic annotations, such as what specific biomedical identifiers an API parameter accepts and what specific biomedical entity types an API response contains. Powered by these semantic annotations, a web application called BioThings Explorer (http://biothings.io/explorer) was developed to allow researchers to navigate the scope of the distributed biomedical API landscape and build the desired knowledge extraction workflows by identifying and combining required APIs.


NOTES

SPONSORS

PLATINUM

IBM's OpenPOWER Group is the organization that includes high performance computing (HPC) within the IBM Systems Group. This group is responsible for the strategy, marketing and identification of areas that can benefit from IBM's high end technology. The life sciences is such an area, and IBM is and will continue to bring valued solutions to life sciences.

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

GOLD

PatientsLikeMe is a free website where people with chronic health conditions get together and share their experiences living with disease. Where newly diagnosed patients can improve their outcomes by connecting with and learning from others who've gone before them. Where researchers learn more about what's working, what's not, and where the gaps are, so that they can develop new and better treatments.

SomaLogic® was founded in 2000 by Larry Gold, with the goal of improving the well-being and quality of life of every individual by transforming how diseases were detected and diagnosed. Building on the previous decade of aptamer research, SomaLogic scientists have developed a new proteomics technology that overcomes the significant challenges of current technologies, and which has multiple applications across the biological and medical sciences. Our mission is to leverage our proprietary technology to discover, develop and commercialize revolutionary new life science research tools and breakthrough clinical diagnostic products that will transform healthcare. See more at: https://somalogic.com/About-Us/

Platinum Sponsor

Gold Sponsor

patientslikeme™

SomaLogic

Rocky 2018 is supported by the Computational Bioscience Program at the University of Colorado School of Medicine