Abstracts for Presentation and Poster

Plenary Sessions (Days 1 and 2)

Day 1 - Plenary Session #1

Complexity of Decision-Making in Surgical Care Provided in Patient Portals

Authors: Jamie R Robinson[1], Alissa Valentine[2], Cathy Carney[1], Daniel Fabbri[1], Gretchen P. Jackson[1], 1. Vanderbilt University, 2. Vanderbilt University Medical Center

Abstract: Messages exchanged between patients and providers through patient portals contain diverse types of communications. The types of communications and complexity of medical decision-making in portal messages sent to surgeons has not been studied. We obtained all message threads initiated by patients and exchanged with surgical providers through the Vanderbilt University Medical Center patient portal from June 1 to December 31, 2014. Five hundred randomly selected messages were manually analyzed. In total, 9,408 message threads were sent to 401 surgical providers during the study period. In the 500 threads selected for detailed analysis, 453 message threads (90.6%) involved medical needs or communications. Further, 339 (67.8%) of message threads contained medical decision-making. The overall complexity of medical decision-making was straightforward in 210 threads (62%), low in 102 threads (30%), and moderate in 27 threads (8%). No highly complex decisions were made over portal messaging. Through patient portal messages, surgeons deliver substantial medical care with varied levels of medical complexity. Models for compensation of online care must be developed as consumer and surgeon adoption of these technologies increases.

Longitudinal Changes in Psychological States in Online Health Community Members: Understanding the Long-Term Effects of Participating in Online Depression Health Community

Authors: Albert Park, Mike Conway, University of Utah

Abstract: Online health communities have shown the potential to reduce the symptoms of depression, however, emotional contagion theory suggests that negative emotion can spread within a community and prolonged interactions with other depressed individuals has potential to worsen the symptoms of depression. We investigate longitudinal changes in psychological states that are manifested through linguistic changes in depression community members who are interacting with other depressed individuals. We examine emotion-related language usages using the Linguistic Inquiry and Word Count (LIWC) program for each member of a depression community from Reddit. To measure the changes, we applied the linear least-squares regression to the LIWC scores against the interaction sequence for each member. We measured the differences in linguistic changes against three online health communities focusing on positive emotion, diabetes, and irritable bowel syndrome from Reddit as control for comparison purpose only. On average, members of an online depression community showed improvement in 9 of 10 pre-specified linguistic dimensions. Moreover, these members improved either significantly or at least as much as members of other online health communities. These results are an important step towards providing insights into the long-term psychosocial well-being of members.

Visual Analytics Approach for Identifying Diabetes Self-Management Trends Using Patient-Generated Data

Authors: Daniel Feller, Lena Mamykina, PhD, Columbia University

Abstract: Mobile health technologies have generated vast amounts of patient-generated data (PGD) raising significant concerns regarding information overload among clinicians. We developed a visual analytics tool to reveal patterns of glycemic response in diabetes self-monitoring data.

Using participatory design, we developed Glucotype, an interactive tool utilizing hierarchical clustering and heatmap visualization to allow dietitians to visually inspect patterns of associations between nutrition in meals and blood glucose levels. Ten registered dietitians evaluated the interactive tool in comparison to a static tabular representation of 1 month of diabetes self-monitoring data. The 'think-aloud' protocol was employed to compare approaches to data analysis across conditions.

Observations generated using Glucotype were more detailed and numerous compared to those generated by a static HTML-based representation. Participants using the HTML-based representation reported being overwhelmed by the volume of PGD and often based their conclusions on a small sample of meals. 9/10 participants considered hierarchical clustering helpful in revealing patterns of glycemic response.

Our findings suggest that visual analytics and computational learning can support the meaningful use of patient-generated data in diabetes. Interfaces that allow clinicians to filter and manipulate large amounts of patient-generated data may enhance user engagement and limit the cognitive biases resulting from information overload.

Using Modern Optical See-Through Augmented Reality (OSTAR) to Improve Medical Practices

Author: Ryan C James, University of Washington

Abstract: Medical images are the cornerstone of pre-procedural planning and surgical guidance. However, these images are viewed in two-dimensions (2D), which limits comprehension of the true 3-dimensional (3D) structure for the majority of viewers. Research has demonstrated that stereoscopic displays can improve comprehension, leading to clearer delineation of near and far structures, improved identification of tumours, more accurate preparation for complex operations, and improved navigation of surgical tools. However, prior research is just the tip of the iceberg, and not enough to safely integrate optical see-through augmented reality (OSTAR), a relatively new and less-studied type of stereoscopic display, into medical workflows.

We hypothesize that OSTAR will improve the interpretation of medical imaging data. To test this hypothesis, we will 1) develop an application that displays medical imaging data as truly 3D holograms seen through an OSTAR head mounted display (HMD), 2) evaluate OSTAR's effect on decision making by conducting a comparative study with 6-8 interventional cardiologists (ICs) who plan with and without the developed application, and 3) understand OSTAR's potential influence on surgical guidance by developing an OSTAR surgical guidance system, then formatively evaluating it while 6-8 ICs use it to perform a transseptal puncture (TS).

Day 1 Plenary Session #2

Integrative Cancer Patient Stratification via Subspace Merging

Authors: Michael Sharpnack, The Ohio State University

Abstract: Technologies that generate high-throughput 'omics data are flourishing, creating enormous, publicly available repositories of multi-omics data. As many data repositories continue to grow, there is an urgent need for computational methods that can leverage these data to create comprehensive clusters of patients with a given disease. Our proposed approach creates a patient-to-patient similarity graph for each data type as an intermediate representation of each omics data type and merges the graphs through subspace analysis on a Grassmann manifold. We hypothesize that this approach generates more informative clusters by preserving the complementary information from each level of 'omics data. We applied our approach to a TCGA breast cancer data set and show that by integrating gene expression, microRNA, and DNA methylation data, our proposed method can produce clinically useful subtypes of breast cancer. We then investigate the molecular characteristics underlying these subtypes. We discover a highly expressed cluster of genes on chromosome 19p13 that strongly correlates with survival in TCGA breast cancer patients and validate these results in three additional breast cancer datasets. We also compare our approach with previous integrative clustering approaches and obtain comparable or superior results.

Novel Pedigree Analysis Implicates DNA Repair and Chromatin Remodeling in MM Risk

Authors: Rosalie G Waller, Todd M. Darlington, Myke Madsen, Karen Curtin, Nicola J. Camp, University of Utah

Abstract: In an age of precision medicine, understanding the genetic factors of common, complex disease is paramount. Missing heritability exists for the majority of complex disease; rare, risk-variants have been suggested as a missing source. Identifying these variants remains a challenge in complex disease due to genetic heterogeneity and complex inheritance models. To address these issues, we developed a new high-risk pedigree (HRP) strategy that identifies segregating segments. These segments capture inherited coding variants, and non-coding variants important for mapping regulatory risk. We applied this strategy in 11 large, multiple myeloma (MM) HRPs, with subsequent exome sequencing in the 11 HRPs and 57 smaller pedigrees from a collaborative resource. One genome-wide significant, 1.8 Mb shared segment was found at 6q16. Exome sequencing in this region revealed predicted deleterious variants in *USP45*, a gene known to influence DNA repair through endonuclease regulation. Additionally, a 1.2 Mb segment at 1p36.11 was found to segregate in two Utah HRPs, with coding variants identified in *ARID1A*, a key gene in the SWI/SNF chromatin remodeling complex. Our results provide compelling, segregating risk genes and variants for MM and demonstrate a novel strategy to use large HRPs for risk-variant discovery more generally in complex disease.

A Spatiotemporal Model to Simulate Chemotherapy Regimens for Heterogeneous Bladder Cancer Metastases to the Lung

Authors: Kimberly R Kanigel Winner, James C Costello, University of Colorado Anschutz Medical Campus

Abstract: In tumors, somatic genetic aberrations are one form of heterogeneity that allows clonal cells to adapt to chemotherapeutic stress, providing a path for resistance to arise. *In silico* tumor models provide a platform for rapid, quantitative experiments to inexpensively study how compositional heterogeneity contributes to drug resistance. We have designed a spatiotemporal model of a lung metastasis originating from a primary bladder tumor, incorporating data for *in vivo* drug concentrations of first-line chemotherapy, vascular density of lung metastases, and resistance gains in bladder cancer cell lines that survive chemotherapy. In metastatic bladder cancer, a first-line drug regimen includes six 21-day cycles of gemcitabine plus cisplatin (GC) on day one and gemcitabine on day eight. Simulations under this regimen or regimen variations produce tumor cell populations that are mixtures of originally resistant cells and new clones that have gained resistance to cisplatin, gemcitabine, or both. This emergence of tumors with increased resistance is qualitatively consistent with the five-year survival of 6.8% for patients with metastatic transitional cell carcinoma of the urinary bladder treated with a GC regimen. The model is being expanded to explore the parameter space for clinically relevant variables, including the timing of drug delivery to optimize cell death and patient-specific data, and can be adapted to other cancers.

Correlates of Cognitive Phenotype Severity in Autism Spectrum Disorders

Authors: Andrew H Chiang, Jonathan Chang, Dennis Vitkup, Columbia University

Abstract: An important goal of genetic research is to understand the phenotypic consequences of diverse genetic mutations. However, progress towards this goal is difficult for common disorders with complex phenotypes, such as cognitive phenotypes in autism spectrum disorders (ASDs).

We analyzed *de novo* likely gene-disrupting (LGD) mutations identified in ASD probands from the Simons Simplex Collection. Specifically, we explored correlations between molecular and genetic properties of these mutations and cognitive phenotypes in affected probands, measured by intellectual quotients (IQ).

To explore the generality of our findings, we analyzed genomic and transcriptomic data across multiple human tissues, obtained from the GTEx consortium, and asked whether relationships could be found between properties of LGD variants and their effects on gene expression.

Surprisingly, we found that LGD mutations in the same gene often resulted in different phenotypes. However, we also identified properties of mutations that appear to correlate with phenotypic severity. Furthermore, these patterns were reproducible in gene expression studies across multiple human tissues, suggesting that the findings may be general.

Our results suggest that large cohort sequencing could achieve clinically useful predictions of cognitive phenotypes in ASDs. Also, similar patterns may be observed in other diseases with substantial contributions from *de novo* mutations.

The Cancer Targetome: A Critical Step Towards Evidence-Based Precision Oncology

Authors: Aurora S Blucher, Gabrielle Choonoo, Molly Kulesz-Martin, Guanming Wu, Shannon K McWeeney, Oregon Health & Science University

Abstract: A core tenet of precision oncology is the rational selection of pharmaceutical therapies to interact with patient-specific biological targets of interest, but it is currently difficult for researchers to obtain consistent and well-supported target information for pharmaceutical drugs. To address this gap we have aggregated drug-target interaction and bioactivity information for FDA-approved antineoplastic drugs across four publicly available resources to create the Cancer Targetome. Our work offers a novel contribution due to both the inclusion of putative target interactions encompassing multiple targets for each antineoplastic drug and the introduction of a framework for categorizing the supporting evidence behind each drug-target interaction. We provide use cases for the drugs imatinib and vandetanib to demonstrate the utility of this resource and map the full antineoplastic target space to Reactome pathways for an estimate of "light" or potentially targetable pathways. This resource provides researchers access to clearly-evidenced drug-target interaction data in a manner that facilitates informed decision-making within the highly contextual nature of drug and target prioritization in precision oncology. We also highlight the use of this resource as a foundation for our modeling efforts with respect to predicting drug response.

Day 2 Plenary Session #3

MetaSRA: Normalized Sample-Specific Metadata for the Sequence Read Archive

Authors: Matthew N Bernstein[1], AnHai Doan[1], Colin N. Dewey[1,2]
[1]Computer Sciences Department; [2]Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison

Abstract: ***Motivation:*** The NCBI's Sequence Read Archive (SRA) promises great biological insight if one could analyze the data in the aggregate; however, the data remain largely underutilized, in part, due to the poor structure of the metadata associated with each sample. The rules governing submissions to the SRA do not dictate a standardized set of terms that should be used to describe the biological samples from which the sequencing data are derived. As a result, the metadata include many synonyms, spelling variants, and references to outside sources of information. Furthermore, manual annotation of the data remains intractable due to the large number of samples in the archive. For these reasons, it has been difficult to perform large-scale analyses that study the relationships between biomolecular processes and phenotype across diverse diseases, tissues, and cell types present in the SRA.

Results: We present MetaSRA, a database of normalized SRA sample-specific metadata following a schema inspired by the metadata organization of the ENCODE project. This schema involves mapping samples to terms in biomedical ontologies, labeling each sample with a sample-type category, and extracting real-valued properties. We automated these tasks via a novel computational pipeline.

The Precarious Wisdom of Communal Science

Authors: Arjun K Manrai[1,2], Chirag J Patel[1], John PA Ioannidis[3,4,5,6], Isaac S Kohane[1,2]

Affiliations:
[1]Department of Biomedical Informatics, Harvard Medical School, Boston, MA.
[2]Division of Health Sciences and Technology, Harvard-MIT, Cambridge, MA.
[3]Stanford Prevention Research Center, Department of Medicine, Stanford University.
[4]Department of Health Research and Policy, Stanford University School of Medicine.[5]Department of Statistics, Stanford University School of Humanities and Sciences.[6]Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, CA.

Abstract: Data sharing has been proposed as a means to address the reproducibility crisis in science. To study the impact of current data sharing practices on reproducibility, we introduce a quantity called the "dataset positive predictive value" (*dPPV*), the proportion of true claims amongst all claims made during many reuses of a shared dataset. We simulate repeated uses of a shared dataset by separate research teams and different data sharing policies. We derive quantitative relationships to connect *dPPV* to the actions of key scientific stakeholders—from individual researchers to funders—and demonstrate several surprising facts about the reproducibility of communal science, while allowing for varying levels of dependence between investigations. For example, small pilot studies may produce more confusion than guidance, and reproducibility often declines as more teams study the same topic, yet these structural factors may neither be controlled by nor visible to individual investigators. We also find that restrictive data access policies may obscure evaluations of reproducibility more than open policies. Finally, we outline how *dPPV* may be estimated and controlled prospectively and provide a web application for investigators to explore reproducibility in communal use of their own datasets. These findings offer suggestions on how to optimize the reproducibility of communal science.

Identifying and Characterizing Near-Duplicates in Big Clinical Note Datasets

Authors: Rodney A Gabriel, MD; Sanjeev Shenoy, MS; Tsung-Ting Kuo, PhD; Chun-Nan Hsu, PhD, University of California, San Diego

Abstract: The widespread presence of near-duplication in clinical data is problematic when training predictive algorithms that model the language or attributes in notes. For example, duplicates may cause a predictive model to erroneously identify correlations between symptoms if these are repeated many times. We developed scalable algorithms to characterize sources of near-duplication in clinical notes as: exact copies, common output notes (e.g., device output), or general templates.

We developed a method to identify clusters of highly similar notes – it uses an approximation algorithm to minimize pairwise comparisons and consists of three phases: 1) Minhashing (first used in AltaVista search engine to detect duplicate webpages from the entire World Wide Web) with Locality Sensitive Hashing; 2) a clustering method using tree-structured disjoint sets; and 3) classification of near-duplicates via pairwise comparison of notes in each cluster. The algorithm can be used to analyze large clinical note corpora with finite available memory space. Results demonstrated that from 10,902,795 notes from UCSD, there were a total of 14,539 clusters, in which 6,945 notes were exact copies, 11,509 were common output notes, and 44,218 were general templates. There were no false positive clusters when clusters were created for notes with Jaccard similarity of 100%.

Beyond a One-Size-Fits-All Evaluation of Causal Inference Methods

Authors: Alejandro Schuler, Ken Jung, Nigam Shah, Stanford University

Abstract: Most medical decisions are made without the support of rigorous evidence, in large part due to the cost and complexity of performing randomized trials for most clinical situations. Although clinical informaticists have successfully used EHR data to classify diseases and assess personalized risks, a problem with using this data to answer causal questions is that patients who receive different treatments often differ systematically in other ways that affect outcomes, creating bias in estimates of treatment effects. Causal inference methods (most of which match similar observations together for comparison) have been developed in the statistics community to address these issues and estimate causal treatment effects. To compare the performance of these methods, it is necessary to know the data-generating process that created the dataset they are evaluated on so that the true treatment effect is known. Traditional methodological studies from the statistics community have relied on simple parametric models that generate small datasets that do not reflect the complexity of real clinical data. Because of this gap, it is not known if the results of analyses that use these datasets to benchmark causal inference methods are applicable to studies that use EHR data. To bridge the divide, I have developed a formal procedure for generating EHR-realistic simulated data in which the true treatment effects are known. My algorithm creates simulated data based on an arbitrarily complex seed dataset by fitting a machine learning model that best fits the outcome while simultaneously achieving the user-specified treatment effect. That model is then used to generate simulated outcomes, which in conjunction with the real covariate and treatment data comprise the simulated dataset. Comparative analysis of causal inference methods benchmarked on realistic simulated datasets will identify the methods that are best-performing on the real dataset at hand and characterize their performance, eschewing the need for the meticulous design of one-size-fits-all evaluations. This will allow for the confident and defensible use of causal inference methods in observational studies, boosting clinicians' confidence in their results and enabling the learning healthcare system.

Antihypertensives, Statins, and Risk for Dementias in Elders

Authors: Fabrício S P Kury, Seo Hyon Baik, Clement J McDonald, National Library of Medicine

Abstract: Existing studies provide mixed results about whether antihypertensives and statins influence the risk of age-related dementias, and it is difficult to reconcile the evidence. In this study we analyzed the association of all such drugs with dementia in 798,610 Medicare beneficiaries. We took advantage of our large cohort to examine the effect of individual drug classes broken down by whether each drug acts on the central nervous system. We employed two measures of drug exposure – duration and quantity – and filtered 98,634 outliers using the drugs' Defined Daily Doses published by the World Health Organization. Models were adjusted for patient gender, race, socioeconomic status, and history of Medicare's 27 chronic health conditions including hypertension and hyperlipidemia. In our results, use of any antihypertensive was associated with a 2% smaller risk of dementia, in contrast to the 19-36% reduction previously reported. Statins had protective association (3.3%) in only one statistical model. Five of the 12 drug subtypes studied had nonsignificant effect in spite of the large cohort sizes. This study addresses shortcomings of observational studies and secondary analyses of data from randomized trials, unifies disparate investigations into one comprehensive assessment, and demonstrates the power and limitations of the Medicare data for pharmacoepidemological studies.

Day 2 Plenary Session #4

Drug Safety Monitoring with Composite Representations and Machine Learning

Authors: Justin Mower[1,2], Devika Subramanian[3], Trevor Cohen[1,2]

[1]Baylor College of Medicine; [2]UT Health; [3]Rice University

Abstract: Adverse drug events (ADEs) are a leading cause of preventable patient morbidity and mortality. Post-marketing drug surveillance consists of identifying and validating potential ADEs utilizing data from reporting systems, Electronic Health Records, and the biomedical literature. These data are noisy, and identified drug/ADE associations must be manually reviewed by domain experts

which scales poorly with large numbers of possibly dangerous associations. There is a pressing unmet need for informatics tools to assist in the review process. We focus on developing new representations for knowledge extracted from the rapidly growing biomedical literature by the SemRep natural language processing system. We compose high-dimensional distributed representations of drug/ADE pairs, which serve as inputs to machine learning classifiers. By doing so, we leverage large amounts of unlabeled data for unsupervised pre-training to enhance drug/ADE recognition. Evaluation against manually curated reference standards shows that applying a classifier to such representations significantly improves performance over previous approaches. Information available to the classifier includes implicit relational information and similarity between drug vector representations. These trained systems can reproduce outcomes of the extensive manual literature review process used to create the reference standards, paving the way for assisted, automated review as an integral component of the pharmacovigilance process.

The Outcome of Opioid/Antibiotic Medication Standardization

Authors: Erin Hickman, Vishnu Mohan, Oregon Health & Science University

Abstract: Pediatric patients are at increased safety risk secondary to medication dosing errors. Two of the most common prescribed medications in the pediatric population are opioid derivatives and antibiotics. These medications pose significant risks to children if prescribed incorrectly or if overly prescribed. The purpose of this pre- and post-intervention research study is to standardize a subset of opioids and antibiotics at Randall Children's Hospital (RCH) to determine if it influences current prescribing and dispensing, practice culture, and physician and pharmacy staff satisfaction. Our hypothesis is that standardization of medications will lead to improved staff satisfaction, to physician acceptance of the proposed dose and increased consistency between hospital and home doses.

Data will be collected from the medical records of all children with antibiotic and opioid orders between October 2016 and March 2017, by use of a Pharmacy Ordering System and Web

Intelligence. Surveys will be distributed through RedCap. Data will be analyzed using descriptive statistics, Fisher's Exact Test and one-sided proportion tests.

Deep Learning Derived Novel Representation of Omics Data Enhances Predictive Modeling of Cancer Drug Sensitivity

Authors: Michael Q Ding, MS, and Xinghua Lu, MD, PhD, University of Pittsburgh

Abstract: **P**recision oncology requires the prescription of effective therapies based on tumor specific characteristics. The current practice of using the genomic status of a drug target as a therapeutic indicator has significant limitations. Most chemotherapy drugs do not have biomarkers to guide their application, and for the ones that do, biomarker prediction is often inaccurate. In this study, we used pharmacogenomics data from the Genomics of Drug Sensitivity in Cancer Project to identify informative genomic features via feature selection and deep learning techniques. We used these features to train machine learning classifiers to predict the effectiveness of drugs against various cell lines. Our classifiers significantly outperformed single-gene indicators in this task. Furthermore, we were able to accurately predict the efficacy of drugs without clearly defined molecular targets. These results demonstrate the capability of advanced machine learning methods to gain insights of cellular systems using genome-scale omics data for predicting the drug sensitivity of cancer cells.

Genomic and Clonal Alterations of Drug Resistance in the NeoALTTO Trial

Authors: Ryan Powles, Weiwei Shi, Christos Hatzis, Lajos Pusztai, Hongyu Zhao, Yale University

Abstract: The anti-HER2 monoclonal antibody trastuzumab and the small molecule tyrosine kinase inhibitor lapatinib have non-overlapping but complementary mechanisms of action in treating HER2-positive breast cancer. Early results in the NeoAdjuvant Lapatinib and/or Trastuzumab Treatment Optimisation (NeoALTTO) trial have shown that dual inhibition of HER2 through combined neoadjuvant treatment increases the rates of pathologically complete response (pCR) in patients. Additionally, next-generation sequencing of pre-treatment tumor biopsies identified pathways that affect patient response to either lapatinib or trastuzumab when mutated. More recently, post-treatment tumor biopsies collected in patients with residual disease after treatment offer a unique opportunity to examine the molecular profile of those early stage HER2-positive breast cancer samples that developed resistance to HER2-blockage treatments. Here, we perform whole-exome sequencing of 76 post-treatment residual disease samples (64 with paired pre-treatment biopsies from the same patient) from patients enrolled in the NeoALLTO study. By analyzing changes between pre- and post-treatment samples in overall mutational load, gene- and pathway-level genomic alterations, and clonal heterogeneity, we characterize the molecular profile of tumors that exhibit resistance to trastuzumab, lapatinib, and paclitaxel neoadjuvant treatment in HER2-positive breast cancers.

Focus Sessions (Days 1 and 2)

Day 1 Parallel Paper Focus Session A

Focus Session A1

Detecting Missing Hierarchical Relations in SNOMED CT using Lexical Features

Authors: Satyajeet Raje, Olivier Bodenreider, National Library of Medicine

Abstract: SNOMED CT is the world’s largest clinical ontology. Recent investigations have revealed issues in its hierarchical structure and the possible harmful effects in clinical applications.

We identify potentially missing hierarchical relations in SNOMED CT from logical definitions based on the lexical features of concept names. First, we create logical definitions for each concept using its preferred name and synonyms (“descriptions” in SNOMED CT). We also leverage the semantic tag (part of the “fully specified name”) of the concept for this logical definition. We represent these logical definitions using description logics (in OWL-EL). We infer hierarchical (subClassOf) relations among these concepts using ELK reasoner. Finally, we compare the hierarchy obtained from lexical features to the original SNOMED CT.

We applied this method to each top level hierarchy in SNOMED CT. We compared the results of using preferred names only versus all the synonyms of the concepts. We also used NLM’s SemRep tools to enrich the lexical features. We are currently conducting manual review of (a subset) of the relations found. Initial reviews indicate that applying SemRep to all synonyms to create the logical definitions produces best results. The lexical approach to finding hierarchical relations is easy to implement, efficient and scalable.

The Newman Test: Small Sample Size Statistic to Personalize Transcriptomic Data Analysis

Authors: Zachary B Abrams, Kevin Coombes, The Ohio State University

Abstract: A problem in conducting transcriptomic analyses is small sample size. To address this problem we have developed a statistic called the Newman test. The Newman test was developed from Newman's 1939 standardized range statistic, which is an outlier detection test. This statistic focuses on the difference in the tails of distributions rather than on the difference in mean. Modifying the statistic to pool power from the large number of measurements per sample in RNA-Seq data, we have developed a statistical test that does not require replicate samples; i.e., the Newman test allows for statistical analyses on a sample size of one sample. This power leveraging is important in the context of omics data since there are more measurements per sample then number of samples in omics datasets. Although this is normally a weakness in omics experiments, we have leveraged this aspect to become a statistical advantage. While useful for small sample sizes, its true power lies in generating a personalized transcriptomic profile for individual patients. Consequently, when a sequenced tumor sample is analyzed employing the Newman test, it generates a truly personal RNA sequencing profile since there is no additional noise from other individuals from within the tested experimental cohort.

Focus Session A2

Breast Cancer Gene Expression Differences between Pre- and Post-Menopausal Women

Authors: Yuzhe Liu, Vanathi Gopalakrishnan, PhD, Pittsburgh University

Abstract: Modern diagnosis of breast cancer includes molecular subtyping based on genetic markers like estrogen receptor (ER), progesterone receptor (PR), and HER2. The onset of menopause brings significant changes to systemic levels of hormones that bind to these receptors. The effect of this change on the development of breast cancer has largely been studied in the limited context of these known hormones and receptors. We sought to identify overall genetic differences in tumors from pre- versus post-menopausal women using Bayesian network learning on RNA-seq gene expression data from breast cancer tumors in The Cancer Genome Atlas (TCGA) Genomics Data Commons (GDC). We learned that GREB1, a key regulator of ESR1, was likely to be high when ESR1 was high and low when ESR1 was low in post-menopausal tumors, but remained likely to be high regardless of ESR1 expression in pre-menopausal tumors. These results suggest that menopause status is an additional factor to consider when using GREB1 as a predictor for good clinical outcomes, as has previously been proposed.

Identifying the Role of Non-Coding Regulatory Variants in Retinal Disease

Authors: Evan M Jones[1], Timothy A Cherry[2], Aleksandar Milosavljevic[1], Rui Chen[1]

[1]Baylor College of Medicine

[2]Harvard Medical School

Abstract: Increasing evidence supports the hypothesis that non-coding variants may play a functional role in inherited retinal disease (IRD). Through a comprehensive analysis of tissue-specific epigenomic

profiling data we have identified 763 putative regulatory regions across 240 currently known retinal disease genes. We designed a capture panel for these regions and have sequenced 185 individuals with single pathogenic coding mutations. From this sequencing data we have identified both functional deep-intronic splicing variants as well as recurrent variants in predicted enhancer regions. Validation of our deep-intronic splicing variants has identified their functional pathogenic effects. We are currently testing the role of enhancer variants on gene expression through cell-line luciferase reporter assays. We further utilize a neural network machine learning (DeepSEA) approach applied to public epigenomic profiling data in order to score and prioritize variants found within regulatory regions.

Our analysis includes comprehensive identification and annotation of regulatory elements across over 240 retinal disease and in combination with targeted next-generation sequencing of these regions we have identified functional non-coding variants that alter either splicing or gene expression. Through this work we show the role of non-coding variants in the functional dysregulation of retinal genes in IRD patients.

Focus Session A3

Identification of Repeatable Research using a Machine Learning Classifier

Authors: Negacy D Hailu, Lawrence E. Hunter; University of Colorado-Denver

Abstract: Recently, there has been a deep interest in repeatable and reproducible research in various research fields and research institutions including NIH. Collberg et al. (2016) conducted a repeatability test on 601 computer science papers. Even though their work has introduced repeatability and reproducibility awareness to the research community, human did most of it manually. Manual repeatability test is extremely time consuming, slow and expensive. On this research, we make an effort to automate the test. When fully developed, authors and research organizations can use the tool to conduct repeatability test on their research articles. We hypothesize that the problem of knowing whether a scientific paper is repeatable or not can be automated and it can be formulated as a binary machine learning classification problem. Two models are developed---the first model uses Collberg et al. (2016) data as features and the second model uses automatically derived features from scientific literature. We trained an SVM classifier with RBF kernel that achieved an F1 score of 0.97 and 0.70 on an average 5-fold cross validation for the first and second model respectively. We are using the first model as a guidance to perform error analysis to improve performance of the second model.

Recognizing Question Entailment for Consumer Health Question Answering

Authors: Asma Ben Abacha, Dina Demner-Fushman, National Library of Medicine

Abstract: Similar questions are becoming more and more frequent online, making them a rich and relevant resource for Question Answering. However, similar and related questions are often expressed differently even if they have equivalent answers. Efficient automatic approaches are therefore required to detect questions that share the same answers. We propose a new approach for the detection of similar questions based on Recognizing Question Entailment (RQE). In particular, we consider Frequently Asked Questions (FAQs) as a valuable and widespread source of information. Our final goal is to automatically provide an existing answer if a FAQ similar to a Consumer Health Question (CHQ) exists.

We give a formal definition of RQE and describe our feature-based classifier for this task. We propose an automatic method to construct learning data using the question focus and question type to identify an inference relation between questions. We apply our RQE approach on CHQs received by the National

Library of Medicine to retrieve relevant (entailed) FAQs and therefore their associated answers. The obtained results show that our method outperforms open-domain models with 80% Accuracy. Our results are promising and suggest the feasibility of our approach as a valuable complement to classic question answering approaches.

Day 2 Parallel Paper Focus Session B

Focus Session B1

Annual Hospitalization-Related Burden and Costs of Inflammatory Bowel Diseases: Quest for High-Value Care

Authors: Siddharth Singh, Nghia Nguyen, William J. Sandborn, Lucila Ohno-Machado, MD, PhD; University of California San Diego

Abstracts: Assessing the cumulative burden of hospitalization in patients with inflammatory bowel diseases (IBD), and identifying 'high consumers' is the first step in devising population health management strategies to improve quality of care and reduce healthcare costs. Using the Nationwide Readmissions Database (NRD) 2013, we identified 42,435 patients with IBD (ICD9 555.x or 556.x) who were hospitalized at least once between Jan-June 2013, and calculated monthly burden and costs of hospitalizations. Over a 10m follow-up, IBD patients spend median 0.58 days/month (IQR, 0.30-1.22) in the hospital, at a rate of $4584/month (IQR, 2208-10036). Patients in the top quintile based on total days spent in the hospital (n=9,235; 'high consumers') spent median 2.9 days/month in hospital (IQR, 2.3-4.3), with monthly, per-patient hospitalization-related costs of $21,466 (IQR, 13,811-36,841). In contrast, patients in the bottom quintile (n=6,082) spent median 0.10 days/month in hospital (IQR, 0.09-0.11), with monthly, per-patient hospitalization-related costs of $1,387 (IQR, 802–3,237). Hence, there are considerable differences in burden and costs of hospitalization within IBD patients. In next steps, we will identify reasons for hospitalizations, and develop and validate prediction models to identify 'high-consumers' of hospitalization to facilitate risk stratification for population health management to provide high-value care.

Identifying Needs for a Health Policy Dashboard in South Africa

Authors: Lauren E Snyder, Jeffrey Lane, Aaron Katz, Anne Turner, University of Washington

Abstract: Health policy is not traditionally considered an informatics issue. Yet, the number and complexity of actors, systems, metrics, and the large amount textual data lends itself to the application of data science with an ultimate goal of improved health outcomes. The standardization and surveillance of policy-related terms and processes would greatly benefit from a data-driven quality improvement exercise, enabling an understanding of the current policy landscape and providing tools to allow for strengthening throughout the policy lifecycle. Further, no single repository or standardized process for tracking currently exists in the National District of Health (NDoH) for South Africa.

In the absence of such a data system, the necessity to construct one is high. Colleagues at the University of Washington recently extensively reviewed the health policy literature and found very infrequent applicable of informatics activities such as data management, analysis, and visualization. Key topic areas have been identified, including: policy development, costing analysis, stakeholder analysis, situational analysis, implementation plan, dissemination strategy, and monitoring and evaluation metrics. Our process for identifying key policy activities and tasks for the development of a potential policy dashboard will be discussed include plans for a business process analysis to be completed in August 2017.

Focus Session B1

Performance of End Stage Renal Disease (ESRD) Anemia Identification Algorithm in Veterans

Authors: Celena B Peters, Jared L Hansen, Ahmad Halwani, Monique E Cho and Brian C Sauer, Department of Veterans Affairs/VA Salt Lake City Health Care System and University of Utah

Abstract:

Background: Recent studies associate FDA regulatory and CMS reimbursement changes affecting use of erythropoietin-stimulating agents (ESAs) with reduction in ESA doses, lower hemoglobin levels and increases in intravenous iron and red blood cell (RBC) transfusions. Research is needed to further understand unanticipated consequences of these policy and reimbursement changes.

Objective: To validate a database algorithm designed to identify the use of RBC transfusions for inpatient anemia management in ESRD patients.

Methods: The database algorithm was applied to Veterans with ESRD hospitalized between 1/1/2008 and 12/31/2013, and excluded patients hospitalized with alternative reasons for anemia (e.g., bleeding). The algorithm required a hemoglobin less than 9.0 mg/dL 24-hours prior to and up to 24 hours after hospital admission AND a RBC transfusion within 24-hours of admission and after evidence of anemia. The algorithm was evaluated against human chart-review guided by an electronic abstraction form.

Results: Two reviewers abstracted 533 hospitalizations that were used as the reference standard to determine algorithm performance. Overall the algorithm agreed with human reviews 85% of the time with 93% sensitivity and 83% specificity.

Conclusions: The database algorithm along with information about measurement error will be used to study anemia management practices in the VA.

Focus Session B2

Digital Phenotyping in Schizophrenia: Preliminary Analysis of data from the SMART Study

Authors: John Torous, Ian Barnett, Patrick Staples, Jukka-Pekka Onnela, Matcheri Keshavan, Harvard University

Abstract: It is now feasible to gather real time symptom surveys from smartphones as well as passive data including: GPS, accelerometer, call logs, text logs, screen on/off time, wifi and Bluetooth connections, and voice. Utilizing the Beiwe platform, we investigated whether survey and behavioral data derived from smartphone passive data use in patients with schizophrenia could augment symptom reporting and relapse prediction. Seventeen patients diagnosed with schizophrenia are currently partaking in this ongoing three-month study. Each subject downloaded the app onto their personal smartphone and used the app in their daily life for three months. Subjects took biweekly symptom surveys on the phone and returned for monthly face-to-face clinical assessments. Smartphone passive data was collected at all times. Generalized estimation equations methods were used to assess for correlations between clinical, phone self-report, and passive phone sensor data. Significant correlations were found between psychotic symptoms recorded in clinical interviews and receiving fewer incoming phone calls ($p = .016$), not returning missed phone calls ($p <.01$), and changes in locations visited from baseline recorded by GPS ($p <.01$). **Smartphones offer a feasible tool to monitor symptoms in schizophrenia and provide new streams of real time data on social and spatial behavior that may be of clinical and research value.**

Patient-Centered Technology use for Cancer Among the Underserved: A Review

Authors: Will L Tarver & David A Haggstrom, Department of Veterans Affairs/Richard L Roudebush, VA Medical Center, Indianapolis

Abstract: Introduction: A wide range of Internet-based [eHealth] technologies and mobile [mHealth] applications are available to patients that are designed to improve their access to care and empower them to participate actively in their care providing a means to reduce cancer disparities; yet, little is known of their use among underserved populations.

Objective: To systematically review the current evidence on the use of cancer-specific patient-centered technologies among the underserved.

Methods: Computer-based searches were conducted in four academic databases to identify peer reviewed articles that were published in the English language and conducted in the US. We used a 3-step inclusion process to identify 47 articles in which we examined study titles, abstracts, and full-text articles for assessment of inclusion criteria.

Results: Underserved populations are receptive to patient-centered technologies; however, a lack of knowledge of how to use technologies is a major challenge. More complex technologies (e.g., a tablet-based risk assessment tool) were also found favorable to use, but also resulted in more significant barriers.

Conclusion: Despite the potential of patient-centered technologies and their receptivity among disparate populations, challenges still exist which can be mitigated by education and training, as well as tailoring the technologies to the populations of interest.

Focus Session B3

Toxin Diversity Evolves via Genomic Reorganization and Transcriptional Rewiring

Authors: Abigail Lind, Jennifer Wisecaver, Antonis Rokas, Vanderbilt University

Abstract: Filamentous fungi produce a diverse array of secondary metabolites (SMs) that play ecological roles in defense, virulence, and inter- and intra-species communication. Fungal SMs have both deleterious and beneficial impacts on human health; some are carcinogenic toxins found in contaminated food supplies, while others, such as lovastatin and penicillin, have been repurposed as successful therapeutics. SMs are narrowly taxonomically distributed and highly diverse between species, and the biosynthetic pathways that produce them are among the most fast-evolving genes in filamentous fungal genomes. SM production is triggered by both biotic and abiotic factors and is controlled by widely conserved transcriptional regulators. To understand how the transcriptional regulators of SM regulate such divergent pathways under different conditions, we examined the genome-wide regulatory role of several master SM regulators in different fungal species and in different environmental conditions. To further gain insight into the evolution of SM pathways, we leveraged population genomics in the human pathogen Aspergillus fumigatus to determine the genetic drivers of SM diversity. Our findings indicate that master SM regulators undergo rapid transcriptional rewiring and interact with multiple abiotic signals to control SM production, and that novel SMs evolve through extensive genomic reorganization and through incorporation of foreign DNA.

Towards Identifying Host Genes that Control Gut Microbial Functions and M Metabolism

Authors: Lindsay L Traeger, Karl Broman, Alan Attie, Federico E Rey, University of Wisconsin Madison

Abstract: The population of microbes that inhabit mammalian gut ecosystems (gut microbiota) have profound effects on host physiology. Alterations in the gut microbiota contribute to metabolic disease, including obesity and diabetes. Major factors influencing the composition of gut microbiota include host genetics and diet; however, a comprehensive investigation of the interrelationship among gut microbial functions, metabolism, and host genetics is lacking. To address this, we leveraged a powerful genetic model, the Diversity Outbred mice (DO), which exhibits tremendous genetic diversity, yet is derived from eight inbred strains, and each mouse is genotyped at high-resolution. We used “shotgun” metagenomics and metabolomics approaches to assess the functional capacity of the distal-gut microbiomes from 300 DO mice maintained in a Western-type diet. Our analysis revealed large differences in gut microbial gene content and metabolite profiles across all mice. Quantitative trait loci analysis identified host genes that associate with both gut microbial functions and metabolites. Some of these genes are known to play important roles in host disease, or encode host extracellular proteins that could facilitate interaction with microbes. Together, our analysis will yield novel insights into the forces that shape the functional capacity of the gut microbiota and how it modulates disease.

Using the Microbial Composition within Sputum to Stratify Patients by Asthma Severity

Authors: Daniel J Spakowicz, Qing Liu, Yale University; George M Weinstock, The Jackson Laboratory for Genomic Medicine; Mark Gerstein, and Geoff L Chupp, Yale University

Abstract: Sputum is a complex mixture of cells that can report on the pathophysiologic heterogeneity observed in individuals with asthma. Previous work has used sputum transcriptomics to stratify asthmatics by disease severity. Here we analyze the non-human composition of sputum RNAseq data to determine to what extent exogenous sequences can contribute to patient stratification and inference on disease. We analyzed sputum transcriptomes of 157 patients with varying asthma severities. We applied a custom version of the exceRpt pipeline to processes reads for quality, align to the human genome and then align unmapped reads to non-human databases. Microbial composition was validated using barcode sequencing of the V1-V3 16S rRNA in bacteria and ITS in fungi.

A median 70% of reads aligned to the human reference genome and the exogenous signal ranged from 0.01 - 29% of the reads. Microbial abundances by barcode sequencing significantly correlated with RNAseq results (Spearman, p-value < 0.001). Preliminary models indicate that incorporating exogenous sequences as predictors of asthma severity have higher accuracy than those with some, but not all, clinical data alone. This work demonstrates the utility of analyzing sputum transcriptome for both human and non-human sequences and may contribute to the understanding of asthma heterogeneity.

Posters (Day 1 and 2)

Topic 1 – Clinical Research Translational Informatics

Poster #101

Integration of Genomic Results in EHR Requires Multidisciplinary Workflow Harmonization

Authors: Jung Hoon Son, Chunhua Weng, Columbia University

Abstract: To report barriers to integrating genomic reports into an EHR system.
We conducted individual interviews with multidisciplinary stakeholders of genomic clinical decision support (CDS), including clinicians, clinical geneticists, pathologists, hospital IT or EHR developers, clinical trials investigators, and research coordinators.
During the inception phase of returning clinically relevant genomic results to clinicians and patients, the subsequent EHR integration of these results was regarded primarily as a translational informatics task. The prevailing misconception among stakeholders was that the testing laboratory's CLIA-certification obviates other regulatory requirements for integrating these genomic results. As a consequence, pre-established clinical workflows designed to satisfy the stringent local (hospital/organizational) and state-level (NYSDOH) regulatory requirements had been unintentionally disregarded. Only when the project evolved to enhance multidisciplinary stakeholder representation via cross-sharing of departmental workflows did these misconceptions resolve. This refocused our goals and discussions to center around how to provide genomic decision support within the appropriate scope and the nuanced boundaries imposed by the pre-existing systems and workflows.
A novel workflow development of a genomic CDS integration necessitate a prior investigation and harmonization of pre-existing organizational workflows in research, clinical practice, and pathology laboratories. Genomic decision support requires the engagement of these multidisciplinary stakeholders.

Poster #102

Using Recurrent Neural Networks for EHR Phenotyping via Clinical Free-Text and Word Embedding

Authors: Joseph D Romano, Phyllis Thangaraj, Mitchell Elkind, and Nicholas P Tatonetti, Columbia University

Abstract: Acute stroke is a clinically important neurological disease. Difficulty in identification of acute stroke patients–largely due to coding inconsistencies–limits retrospective studies. New algorithms and data sources present opportunities for improving stroke phenotyping. EHR phenotyping is the process of identifying patient cohorts via their electronic health record data. Here, we use a recurrent neural network to phenotype stroke patients based on their clinical notes and compare its performance to logistic regression, random forest, and support vector machine classifiers trained on the same set of notes. For a curated set of 4,584 stroke patients, we extracted 285,896 notes from the Columbia University Medical Center EHR. We identified 8,828 controls (454,245 notes) as other patients admitted to neurology without stroke. We processed these notes using cTAKES (for the structured learning algorithms) and word2vec (for deep learning) and evaluated performance using hold-out validation. Deep learning performed best: accuracy of 0.94 and AUROC of 0.96, followed by logistic regression, accuracy of 0.86 and AUROC of 0.94. This algorithm can form the basis of cohort discovery for future retrospective studies of stroke outcomes, and the model may be expanded to other phenotyping tasks given the presence of an adequate reference standard.

Poster #103

Toward Shareable Individualized Drug Interaction Alerts

Authors: Samuel C Rosko, Richard D Boyce, University of Pittsburgh
Philip D Hansten, John R Horn, University of Washington
Daniel C Malone, University of Arizona

Abstract: Override rates for potential drug-drug interaction (PDDI) alerts remain high, largely due to excessive and non-informative alerting. One contributing problem is that alert acceptance depends on the specific clinical situation. To address this, computable evidence-based clinical algorithms were developed that consider a patient’s electronic health record information to provide clinicians with actionable information tailored to the specific context.

Data was collected at the University of Arizona Medical Center to identify PDDIs with high override rates. Drug interaction experts (JRH, PDH, and DM) performed a primary literature search to identify drug- and patient-related factors that could enhance or mitigate the risk of harm from exposure to these PDDIs. The results were used to develop clinical algorithms that defined when alerts should be fired, the seriousness of the potential interaction, management options, and supporting evidence.

The resulting clinical algorithms were translated into sharable decision support rules implemented as JBoss Drools using concept sets defined with standardized terminologies. The rules were tested on a simulated Medicare population represented in an open source common data model created by the Observational Health Data Science and Informatics collaborative. Future work will test if the rules are more predictive of adverse outcomes than the current, non-individualized, approach.

Poster #104

Analysis of Type 2 Diabetes Etiology through PheWAS and Functional Studies

Authors: Jamie C Fox, Brain A Hoch, Elizabeth McPherson, & Scott J Hebbring, Marshfield Clinic, Marshfield WI, University of Wisconsin-Madison

Abstract: Type 2 diabetes is an escalating global health dilemma. Genome wide association studies (GWASs) have identified hundreds of single nucleotide polymorphisms (SNPs) associated with T2D, indicating that this disease may be influenced by contributions of many low penetrant variants. Utilizing this genetic foundation, we analyzed the association of GWAS significant SNPs with phenotypes recorded in electronic health records through Phenome Wide Association studies (PheWASs).

Understanding the functional implications of genetics variants and disease associations identified through PheWAS is a major component of our study. We identified several associations between diabetes SNPs and various common disorders, including Alzheimer's disease, hyperlipidemia, obesity, diabetes, and visual impairment. Ongoing analysis suggests that these diseases may exhibit common etiologies helping us pinpoint relevant biological pathways influenced by genetic variants in genes such as *APOC1*, *TCF7L2, HMGA2.* We have also identified 2 novel and likely clinically relevant variants within the gene *SLC5A2*, a gene highly targeted by type 2 diabetes therapeutics. As precision medicine takes precedence in health care it is hoped that studies like these, which begin with a genetic informatics approach, will have a clinical impact on patient care.

Topic 2 – Healthcare Informatics

Poster #105

The Efficacy and Unintended Consequences of Hard-Stop Electronic Alerts in Electronic Health Record Systems: A Systematic Review

Authors: Emily Powers, Andrew Hickner, Richard N Shiffman, Mona Sharifi, Yale University

Abstract: Clinical decision support (CDS) within an Electronic Medical Record (EMR) in the form of alerts has a spectrum of rigidity ranging from passive knowledge links to rigid "hard stops" in which the clinician is prevented from executing an action. While effective, the hard-stop modality may have unintended consequences with the potential to impact patient safety and to date no review has been published outlining risks, benefits, and appropriate use cases. This systematic review seeks to summarize evidence on hard-stop alerts to determine their effectiveness, unintended consequences, appropriate use cases. We will pull from published and grey literature in both the medical and computer science domains. Meta-analysis of efficacy of hard-stop alerts versus soft-stop alerts will be performed if appropriate. Results anticipated by May, 2017.

Poster #106

Assessment of Laboratory System-Assigned LOINC Codes for Common Tests

Authors: Sharidan K Parr, Shuanghui Luo, Sina Madani, Jacqueline Kirby, S Trent Rosenbloom.
Dept. of Veterans Affairs Tennessee Valley Healthcare System, Vanderbilt University Medical Center, University of Texas at Houston

Abstract: Background: Representing clinical data accurately and consistently across sites requires mapping institution-specific information to a standardized terminology such as The Logical Observation Identifiers Names and Codes (LOINC). However, mapping can be hindered by idiosyncratic, ambiguous local identifiers.

Methods: Using 52 common laboratory tests collected within the Vanderbilt University Medical Center, we assessed LOINC code accuracy by examining the property, time, system, and scale attributes of the assigned code.

Results: We examined 14,139,111 laboratory results. Every laboratory test had an associated LOINC code recorded in the database. Forty-three (83%) of the 52 laboratory tests were associated with a single, correct LOINC code, and one correct specimen type. One test (2%) was associated with multiple LOINC codes and multiple specimen types. Two tests (4%) each mapped to two LOINC codes but contained one specimen type, and six tests (11%) mapped to a single LOINC code with multiple specimen types.

Conclusion: Optimizing LOINC code accuracy at our institution will require harmonizing the laboratory mapping schema with the locally-defined laboratory tests. This process will require reducing redundancy in the laboratory mapping schema, further examining the role of the technician and the laboratory device in generating mapping features, and improving concept coverage.

Poster #107

Current Practices in Integrating Social and Behavioral Data into Patient Care at a VAMC

Authors: Cherie Luckhurst & Michael Weiner, Department of Veterans Affairs/Richard L Roudebush VA Medical Center, Indianapolis

Abstract: Research conducted over the last several decades has firmly established that social and behavioral (S&B) factors influence physical health. Recently, the Institute of Medicine recommended that such data be collected from every patient, and it proposed that S&B data will aid in recognizing, diagnosing, and treating medical problems. However, the VA care providers that we have spoken with do not see the usefulness of such data in their clinical practices.

While the medical community agrees that S&B factors impact health, a mechanism has not been developed to integrate such data into clinical workflow and decision-making. Before we develop such a mechanism, however, the principles of informatics direct us to examine current practices.

This study uses mixed methods to understand any current use of S&B data in the diagnosis and treatment of medical issues at the Roudebush VAMC. We are interviewing and surveying 24 clinicians who serve in a variety of roles in primary care and specialty clinics. This small exploratory study will provide a cross-sectional snapshot of current practices. The findings will inform the development of a mechanism that integrates S&B data into clinical diagnosis and treatment, thereby addressing the gap between evidence and practice.

Poster #108

Treatment and management of smoking at Health Clinics, ICUs and Quitlines

Authors: Andrey Soares[1,2], Joan M Davis[2], Sonia Leach[1,3],
[1]University of Colorado Anschutz Medical Campus [2]Southern Illinois University, [3]National Jewish Health

Abstract: This research investigates innovative approaches to address the complexity of behavioral, social, and biochemical aspects of nicotine dependence, and to support healthcare providers during brief tobacco cessation interventions at the point of care, such as at health clinics, intensive care units (ICUs), and smoking Quitline programs. Smoking is known to cause numerous tobacco-related diseases such as cancer, heart disease, respiratory disease, as well as death. With 70% of smokers interacting with a healthcare provider each year, there is great potential for change and direct impact helping smokers to quit. In fact, smokers who receive a brief intervention from healthcare providers have a six times higher rate of success quitting tobacco use than smokers who try to quit on their own. With the goal of translating discoveries into practice, this research aims to analyze electronic health records to identify smoking status and habits, analyze patient health conditions and risks, identify quitting patterns and predictors of success for tobacco

cessation, and provide evidence-based treatment recommendations tailored to patients. In particular, with support from rule-based systems, knowledge representation, machine learning, and natural language processing, we are developing clinical decision support systems that focuses on helping healthcare providers to offer quality personalized services to patients.

Poster #109

Learning a Comorbidity-Driven Taxonomy of Pediatric Pulmonary Hypertension

Authors: Mei-Sing Ong, Mary Mullen, Eric Austin, Peter Szolovits, Kenneth Mandl, Harvard Medical School

Abstract: Pediatric pulmonary hypertension (PH) is a heterogeneous condition with varying natural course and therapeutic response. Precise classification of PH subtypes is therefore crucial for individualizing care. However, gaps remain in our understanding of the spectrum of PH in children. Here, we seek to assess the feasibility of applying a network-based approach to discern disease subtypes from comorbidity data recorded in longitudinal datasets.

We conducted a retrospective cohort study comprising 6,943,263 children (<18 years of age) enrolled in a commercial health insurance plan in the US between 2010 and 2013. A total of 1,583 (0.02%) children met the criteria for PH. We identified comorbidities significantly associated with PH compared with the general population of children without PH. A Bayesian comorbidity network was constructed to model the interdependencies of these comorbidities, and network clustering analysis was applied to identify disease subtypes comprising subgraphs of highly-connected comorbidities. The derived network captured most of the major PH subtypes with known etiological basis, thus validating the approach. The analysis further identified a number of subtypes documented in only a few case studies, including several well-described genetic syndromes, providing impetus for deeper investigation of the disease subtypes that will enrich current disease classifications.

Poster #110

MetaMapLite in Excel: Named-Entity Recognition for Non-technical Users

Authors: Ravi Teja Bhupatiraju, Kin Wah Fung, Olivier Bodenreider, National Library of Medicine

Abstract: With the current tools, Natural Language Processing can be daunting for biomedical researchers, untrained in computer science. We developed an easy-to-use tool for non-technical biomedical researchers to conveniently conduct Named-Entity Recognition (NER) on biomedical text, in a familiar spreadsheet system. With a simple standalone installer, the system deploys a client-server system that presents a web service that is consumed from an Excel spread sheet with an embedded macro. The system is highly responsive for interactive use and can be further scripted from both spreadsheet macros and any external scripts.

Poster #111

Exploration of Operational Data Streams as a Source of Actionable Insight

Authors: Dana M Womack, Michelle Hribar, Paul Gorman, Oregon Health & Science University

Abstract: Elimination of preventable harm in healthcare settings is an urgent priority that requires an understanding of workflow activities and stress points. Modern hospital care involves the use of operational software systems for medication and supply dispensing, communications, and other bedside care activities. These systems produce transactional records and log files that are not included in the electronic health record, but may provide important insights about care processes and serve as a new source of data in healthcare improvement research.

Research goals include development of methods for analysis of temporal care activity data and translation of activity insight into proactive patient safety capabilities. Preliminary data exploration has confirmed that operational data reflects daily rhythms. Current research seeks to find hidden patterns that are early indicators of safety risk or adverse events. Through exploration of relationships between temporal activity patterns and patient safety, this study will lay a foundation for proactive monitoring of latent safety conditions in the hospital setting.

Poster #112

Assessing Ghana's eHealth workforce: Implications for Planning and Training Programs

Authors: Henry A Ogoe, Harry Hochheiser, Gerald Douglas, University of Pittsburgh

Abstract: In their quest to provide affordable, and quality health services, health systems in both developing and developed countries are adopting strategies for eHealth— "the cost-effective and secure use of information and communication technology in support of health care." One of the key barriers for a successful adoption of eHealth is a competent and well-trained workforce required to design, implement, maintain, and evaluate eHealth systems.

In settings, like the low and middle income countries (LMIC), which are saddled with limited resources and rising population, there is a critical need for health service organizations to identify the requisite eHealth cadres and their training needs that could ensure an effective and efficient deployment and management of eHealth systems. The goal of this on-going work is to apply the workload indicator of staffing needs (WISN) to characterize, quantify, and assess the eHealth workforce requirement in a LMIC like Ghana.

When completed, this work will positively affect the management and education of eHealth services in several ways. First, its findings will help us understand gaps in eHealth staffing needs. Second, it will serve as scientific basis for effective policy planning. Last, it will inform educators on the relevant competencies to consider for informatics training programs.

Poster #113

A Comparative Social Network Analysis of Breast Cancer Provider Collaboration

Authors: Bryan Steitz, Mia Levy, Vanderbilt University

Abstract: Cancer treatment often consists of multiple therapeutic modalities delivered by specialists. A previous study by Smith and colleagues found that cancer patients see an average of 32 patients over the course of their treatment. As reimbursement paradigms shift to a value-based model focusing on quality outcomes and bundled payments, extensive care coordination between healthcare providers is imperative. Analyzing provider relationships as a social network is one method to evaluate coordination and collaboration. Our study aims to assess differences in provider networks between teams of oncology providers and teams of all clinicians caring for stage I through stage III breast cancer patients.

We gathered a cohort of 3924 patients with stage I through stage III breast cancer diagnosed between 2000 and 2014 from the Vanderbilt University Medical Center tumor registry and obtained billing data from the electronic health record for each patient. We computed connections between providers associated with the patient's medical care. Using social network analysis, we calculated statistics by patient stage. Initial results have identified statistical differences in care team connectedness and specialty importance by patient stage.

Poster #114

Examining the Video Quality of a Hospice Telehealth Intervention

Authors: Claire Jungyoun Han, Kenneth Pike, George Demiris, University of Washington

Abstract: Hospice care aims to deliver personalized palliative care for patients at the end of life and their families. As part of a randomized clinical trial testing the value of a telehealth based delivery of a problem solving therapy intervention for hospice family caregivers called PISCES, we examined the technical quality of video-based interactions. The goal was to understand how technical quality affects the content of communication. We used a validated technical quality observation form that was completed by the interventionist. Out of 518 caregivers, 171 were randomly assigned to the video group; the other two arms were an attention control and a group receiving the intervention in person. A total of 480 video sessions were evaluated. The mean session duration was 34 minutes. The technical quality was given an average rating of 92% (range 45-100%). There were no technical problems in 72% of all sessions whereas in the remaining cases there were difficulties in establishing a connection or audio/ video interruptions. While the technical problems overall were minor and infrequent, they could be disruptive when they occurred unexpectedly during an emotional conversation. Our findings inform recommendations for the effective use of videoconferencing in telehealth interventions.

Poster #115

A Pipeline for Classifying Clinical Phenotypes in the National Health and Nutrition Examination Survey

Authors: Diane Walker and Julio Facelli, University of Utah

Abstract: The National Health and Nutrition Examination Survey (NHANES) is a nationally representative, continuous, cross-sectional survey of the United States (US) noninstitutionalized civilian resident population. NHANES provides US health surveillance and guidance on national health policy and regulation. Conducted in two-year cycles with ~5,000 participants, NHANES collects individual- and household-level data on nearly 1200 health, dietary and environmental variables. NHANES datasets are publicly available for secondary analysis.

NHANES data variables can be used alone or in combination to classify individuals by clinical phenotype. These phenotypes become predictor and outcome variables in NHANES studies. Characterizing phenotypes, which are physiological abnormalities underlying conditions and diseases, enhances our understanding of disease pathophysiology and risk. The range of clinical phenotypes supported by the NHANES is currently unknown. We aim to identify clinical phenotypes in published NHANES studies using a text-mining approach. We will use MetaMap, a tool developed by the National Library of Medicine, to extract Unified Medical Language System (UMLS) concepts in PubMed abstracts. NHANES variables will be found in abstracts using exact keyword matching. Clinical phenotypes will be defined as UMLS concepts matching terms within the Human Phenotype Ontology. The results of this analysis will be used to develop a graphic representation of clinical phenotypes in NHANES literature.

Poster #116

Understanding Genetic Influence on Molecular Phenotype Variation in iPSC-Derived Cells

Authors: Margaret Donovan, Paola Benaglio, Agnieszka D’Antonio-Chronowska, Erin Smith, Kelly Frazer, University of California, San Diego

Abstract: Large biobanks of induced pluripotent stem cell (iPSC) lines have enabled the study of the molecular and cellular impact of genetic variation, including the characterization of loci associated with traits and diseases in human populations. While it is known that somatic and regulatory changes occur during reprogramming and differentiation, it is unclear how these changes affect the relationship between genetic variation and traits of interest. The goal of this study is to determine the contribution of reprogramming, differentiation, and genetic background to the heterogeneity of molecular phenotypes and how this impacts our ability to map traits. We derived 22 iPSC lines from 12 donors including 5 monozygotic (MZ) twin pairs and a mother-daughter pair and differentiated them into cardiomyocytes (iPSC-CM). We characterized each line by RNA-Seq, ATAC-Seq, and ChIP-Seq (H3K27ac). Using hierarchical clustering and principle component analysis of this data set, we observed that cell lines originating from the same MZ twin pairs tended to cluster together. This suggests gene expression and epigenetic heterogeneity in iPSC and iPSC-CM are largely attributable to the genetic background, rather than changes acquired during the experimental process. These early results suggest that iPSCs and iPSC-CMs are a suitable model for quantitative trait association studies.

Poster #117

Semantic Characterization of Clinical Trial Criteria and MIMIC-3 Patient Notes

Authors: Jianyin Shao, Ramkiran Gouripeddi, Julio C Facelli, Department of Biomedical Informatics, University of Utah

Abstract: Clinical trials (CT) are essential for evaluating medical interventions on patients and/or populations. Patient recruitment is a major bottleneck in performance of CT. Automatic or even semiautomatic matching of CT criteria with electronic health records could enhance the effectiveness of recruitment for CT. The goal of this study is to find a minimal set of semantic concepts that describe CT and patient notes for efficient computational matching. We used MetaMap to extract UMLS concepts (CUIs) from 195,352 CT eligibility criteria and 2,077,755 MIMIC-III clinical notes. Our calculation of frequencies of UMLS concepts in both these corpora show classical power distributions. The CUIs that appear in both corpora and are below a certain threshold in their distributions provide a representative set of concepts for matching CT and clinical notes. Several threshold values to select informative CUI sets will be used to empirically determine the best match performance. If successful, this low-cost computational screening process could allow dedicating limited CT recruitment resources to the most promising targets for recruitment.

Topic 3 – Bioinformatics/Computational Biology:

Poster #118

Inferring Clonal Heterogeneity in Chronic Lymphocytic Leukemia

Authors: Mark Zucker, The Ohio State University

Abstract: Clonal heterogeneity is a common feature in many types of cancer and previous research has suggested it is often clinically relevant. There is therefore considerable demand for computational tools for the deconvolution of heterogeneous populations into distinct subclones from high throughput data. Several such tools already exist, but require sequencing data. We, however, are developing an R module capable of resolving the clonal heterogeneity of a cell population in copy number alterations (CNAs) from just SNP array data. Our method will employ Bayesian methods to determine the most probably clonal architecture of a sample. Our method will also be applicable to DNA sequencing data (for the detection of clones defined by subclonal mutations, rather than just CNAs), but will require only SNP array or sequencing data, not both. In addition to developing this tool, we will also apply it to samples from 173 chronic lymphocytic leukemia (CLL) patients to determine how common clonal heterogeneity is in this cancer, whether clonal heterogeneity as a feature itself corresponds with any clinical outcomes, and whether specific subclonal patterns can be detected that correlate with clinical outcome. Our method will ultimately be applicable to data from patients of all types of cancer.

Poster #119

Using Sequence-based Features to Robustly Mine Metagenomic Data for Viral Sequences

Authors: Cody Glickman[1,2] and Michael Strong[1,2]

1. Computational Bioscience Program, University of Colorado-Denver, Denver, CO 80045
2. Center for Genes, Environment, and Health, National Jewish Health, Denver, CO 80206

Abstracts: State-of-the-art metagenomic binning techniques depend on the presence of reference genomes when assigning taxonomy via mapping. Reads that are unable to be mapped are either discarded, resulting in a loss of information, or they are binned together into unlabeled categories, resulting in read mixing (e.g., viral and bacterial). Here, we introduce a method, which attempts to classify the mixed unlabeled reads using supervised machine learning techniques by leveraging sequence-based features. The goal of this method is to correctly identify a read as being viral or bacterial to tease apart ambiguity of unlabeled categories in current metagenomic binning practices.

Our method uses multi-k-mer count features derived from metagenomic reads to label the read origin. Our method leverages a model derived from an ensemble classifier, which enables the extraction of important features. The classification performance of our method is compared against other taxonomic binners using a simulated metagenome with multiple genomes not yet in the extant reference databases.

Without the development of novel and innovative approaches to identify read origin from metagenomic data sources, a vast collection of unexplored metagenomic sequences will remain uncharacterized or become discarded.

Poster #120

Chromatin Accessibility Profiling in Human Cell Models for Neuropsychiatric Illness

Authors: Marion J Riggs[1,2], Jean Fan[1], Jennifer Wang[2], Roy H. Perlis[2], Peter V. Kharchenko[1]

Institution: 1. Harvard Medical School, 2. Massachusetts General Hospital

Abstract: Assay of Transposase Accessible Chromatin (ATAC) followed by sequencing is a powerful approach to interrogate transcriptional regulatory elements of the genome that are associated with factor accessible regions. In “tagmentation”, the Tn5 enzyme fragments the genome while simultaneously loading sequencing adapters for next generation sequencing. From bioinformatics processing, peak profiles are generated to identify genomic regulatory elements, such as gene promoters and enhancers. In this study, we use ATAC-seq generated data from human *in vitro* differentiated neurons to identify accessible regions associated with transcriptional regulation. In a case study, we look at differences in cell lines derived from patients with neuropsychiatric illness versus healthy controls. We observed that the majority of differential chromatin accessibility occurs in enhancers and does not directly correspond to transcriptional differences in cis-linked genes. In an additional analysis, we query publically available single-nucleotide polymorphisms from gene-wide association studies that are associated with neuropsychiatric illness against chromatin accessible profiles and further identify candidate loci of interest. This preliminary study supports using ATAC-seq for improving our understanding of epigenetic factors in neuropsychiatric illness and the merit of using human cell models for the long-term aim of clinical translation and drug discovery for treatment of mental disorders.

Poster #121

Named Entity Recognition in Functional Neuroimaging Literature

Authors: Alba G Seco de Herrera, Asma Ben Abacha, Dina Demner-Fushman, L. Rodney Long and Sameer Antani, National Library of Medicine

Abstract: Functional Magnetic Resonance Imaging (fMRI) is a powerful non-invasive technique for collecting and analyzing information about activity in the human brain. Human neuroimaging research aims to map relationships between brain activity and a large number of broad cognitive states, i.e., to decode cognitive states from brain activity. To this end, we are connecting brain activity information extracted from fMRI data with named entities and events extracted from functional neuroimaging literature.

As a first step, we focus on Named Entity Recognition (NER) and compare different methods to extract relevant entities from the functional neuroimaging literature. We selected 14 types of named entities to describe cognitive states, anatomical areas and fMRI experiments stimuli and responses. We manually construct a gold-standard corpus by two experts to evaluate our NER methods. The resulting system will be used to annotate available functional neuroimaging literature.

Poster #122

Building a Pseudogene-Gene Network Database to Mine for Phylogenetic Relationships

Authors: Travis Johnson, Kun Huang, Yan Zhang, The Ohio State University

Abstract: Pseudogenes are fossil relatives of genes, historically considered “junk DNAs”, since they do not code proteins in normal tissues. Although most of the human pseudogenes do not have noticeable functions, ~20% of them exhibit transcriptional activity. Some pseudogenes adopt functions as lncRNAs and regulate gene expression. Some pseudogenes can be ‘activated’ generating transcripts and even proteins in cancer. All the above have shown that pseudogenes could have functional contribution to the genome.

Here we constructed a comprehensive set of pseudogene-gene families, each of which contains multiple homologous genes and pseudogenes. We identified mutation signatures in the families, and constructed phylogenetic trees that represent the relationships between family members. From the study, we have found evidence supporting the evolutionary history of olfactory family. We further developed a computational tool that utilizes the pseudogene-gene network with ontology annotations to infer potential functions of ‘activated’ pseudogenes. Finally, we will provide a user interface that allows users to query their own activated pseudogene sequences, receive the family summary and inferred function that can guide experimental validation.

Poster #123

Natural Product Targetome in Cancer: Definition and Application

Authors: Steve Chamberlin, Gabrielle Choonoo, Molly Kulesz-Martin, Shannon McWeeney, Oregon Health & Science University

Abstract: Targeted therapies for cancer act on molecular targets considered to be specific drivers for the disease, theoretically causing little damage to healthy cells and promising fewer adverse effects than cytotoxic approaches. However, limitations of these therapies include acquired drug resistance, limited treatment options for some cancers and for children, and the expense and difficulty with synthesis of effective molecules for some targets. Natural products may address some of these challenges and have also been shown to have synergistic effects with some cancer drugs. Network pharmacology offers a framework to explain and guide targeted therapies that include combination therapies of natural products and FDA approved drugs by using knowledge of critical cancer pathways.

The objective of this study is to comprehensively define the natural product targets for pan-cancer critical disease pathways by leveraging earlier work assessing the coverage of the “cancer targetome” based on FDA approved drugs. **It is our hypothesis that natural products, defined as compounds from living sources (plant, animal, microbial), can substantially increase the coverage of critical aberrational cancer pathways and assist in identifying novel therapeutic strategies**. This work lays a critical foundation needed to predict synergistic combination therapies using natural products.

Poster #124

MeTeOR: Literature-Based Hypothesis Generation and Precision Medicine

Authors: Wilson Stephen, Wilkins AD, Palzkill T, Lichtarge O, Baylor College of Medicine

Abstract: With over twenty-four million articles and an exponential growth rate, it has become difficult to stay abreast of the PubMed literature. To address this problem, we have created a novel biological network that aggregates data from millions of PubMed articles. This network, called MeTeOR (MeSH Term Objective Reasoning network), converts manually curated MeSH terms that tag PubMed articles into a global, structured summary of biological information for data-driven discovery. When compared to the current knowledge representations in many standard curated databases regarding associations among genes, drugs, and diseases, MeTeOR contains both confirmatory as well as novel information. Furthermore, when a hypotheses-generating algorithm is applied to the MeTeOR network, it suggests new potential disease or drug associations for most genes. In the most realistic test of performance—a time-stamped analysis, hypotheses generated from a MeTeOR network based on the literature prior to 2014, were shown to have significant predictive power for discoveries that were published after 2014. These data support MeTeOR as a promising representation of the biomedical literature, that may provide ready access to high-quality information about the relationships linking genes, drugs, and diseases, and also that support novel hypotheses towards systems analysis and precision medicine.

Poster #125

A New Network-Based Method for Integrated Analysis of Biomedical Data

Authors: Andrew Laitman, Ismael Al-Ramahi, Tarik Onur, Juan Botas, Zhandong Liu, Baylor College of Medicine

Abstract: Huntington's disease (HD) is a progressive neurodegenerative disorder caused by CAG trinucleotide repeat expansion within the Huntingtin gene. Longer repeat expansion is inversely correlated with age of onset of neurological symptoms. However, repeat expansion can only explain 67% of the variance of neurological onset. This suggests that disease onset can be modified by other factors, including environmental and genetic. Thus, the discovery of genetic modifiers could lead us to pathways that can be targeted for drug treatment.

We develop a network-based method that incorporates Steiner trees to generate a network that sparsely connects significant genes. Our algorithm creates a consensus network that integrates a non-specific protein-protein interaction network with a disease specific interaction network. When applied to gene expression and screening data from HD, this method yields modules involved in cation transport and glutamate signaling, which are biological processes known to be involved in HD pathogenesis, confirming the validity of our approach. This indicates that our network-based method aids in the discovery of relevant genetic networks when applied to multi-modal genomics data. This approach can be easily refined for application to other data types such as neuroimaging, GWAS, and others.

Poster #126

Reporting of Somatic Mosaic Mutations in Brain Development Disorders

Authors: Nathaniel Delos Santos, Laurel Ball, Andreia Maer, Joseph Gleeson, University of California San Diego

Abstract: Somatic mosaic variants present in a small percentage of brain cells are enough to produce brain development defects, but issues with sequencing coverage and processing variants across a wide range of patients make it difficult to identify and prioritize high-confidence candidate variant calls for further testing in mouse models. In order to identify and prioritize these somatic variants, we have developed a method for combining results from multiple somatic variant calling tools and variant annotation databases. Through whole exome sequencing and somatic variant calling of blood and brain samples from 48 individuals (two brain sample replicates per individual) with brain development disorders including focal cortical dysplasia and hemimegalencephaly, we identified 10797 potential somatic variants across multiple brain samples. From these potential somatic variants we identified 638 variants appearing in multiple brain samples, and 359 variants appearing in COSMIC.

Variants are reported if they are called by both Mutect and Strelka in multiple samples annotated with brain developmental disorders. We alse report genes annotated with reported variants, including the samples affected and variant effect annotations. Preliminary findings also indicate the utility of this method in validating previous experiments, by identifying potential experimental errors including three mismatched blood samples.

Poster #127

Automated Mechanistic Hypothesis Generation from Qualitative Biological Networks

Authors: Michael A Kochen, Carlos F Lopez, Vanderbilt University

Abstract: Large-scale biological models with molecular-level detail are necessary to capitalize on the enormous and increasing amount of high-throughput quantitative data being generated with modern technologies. Large-scale systems are typically modeled with qualitative methodologies. Such methods lack the detail to accurately capture the precise dynamics necessary to explain molecular mechanism details and predict outcomes in fields that require high precision predictions such as drug discovery and personalized medicine. In contrast, mass-action kinetics based methods provide a suitable level of detail, but such models can quickly become unwieldy due to the number of parameters and network information necessary for model construction, simulation, calibration, and analysis. To address this gap between large-scale but coarse-grained qualitative models and highly detailed physicochemical methods we have developed HypBuilder, software to automatically convert from qualitative to physicochemical models. HypBuilder is comprised of (*i*) a molecular interaction library, a generic network representation of every molecule expected in a system and the interactions they participate in and (*ii*) algorithms to rewrite the original network into a mechanistic configuration and assemble the final mass-action kinetics based representation. We expect that HypBuilder will open the door to large mechanistic models to explain and predict the cellular outcome of complex biological processes.

Poster #128

Using Pharmacogenomic and Pharmacokinetic Data to Predict Levels of Antiretroviral Drug Resistance

Authors: Juandalyn Burke, Neil Abernethy, University of Washington

Abstract: Drug resistance is one of the primary factors contributing to virologic failure for individuals diagnosed with the human immunodeficiency virus (HIV). In recent years, healthcare agencies and organizations have observed increasing levels of HIV drug resistance (HIV-DR) within-host and between-host [1]. This has led to the generation of pharmacogenomic-pharmacokinetic data used to better understand relationship between antiretrovirals and virologic response. Genomic biomarkers generated from these analyses have led to observations of inter-individual and inter-ethnic variability among patients taking antiretroviral medication, specifically among individuals in sub-Saharan Africa [2,3]. However, there are few tools that utilize the well-documented antiretroviral pharmacogenomic data and its association with virologic failure and adverse clinical reactions to monitor HIV-DR. We aim to improve upon an existing mathematical, epidemic modeling tool using data derived from pharmacogenomic literature and databases to predict within-host and between host drug resistance levels over time. Specifically, we aim to create a treatment optimization algorithm that predicts a patient's risk for drug resistance and a personalized treatment regimen over time using the within-host modeling dynamics. Our analysis of within-host dynamics has generated antiretroviral drug resistance levels for the first-line antiretroviral treatment that complement empirical data on existing levels of HIV-DR.

Poster #129

Recent Extensions and Applications of Nested Effects Models

Authors: Yuriy Sverchkov, Audrey Gasch, Mark Craven
University of Wisconsin-Madison

Abstract: Many human diseases, including cancer and neurodegeneration, are associated with cellular processes involving pathways that are not well-characterized, such as mitochondrial dysfunction and cellular stress response. Advances in systems biology have shown that network models for capturing complex relationships in gene regulation, metabolism, and cellular signaling help better characterize such processes. A common approach to uncovering these networks involves measuring reporters of interest subject to perturbations on elements of the network, most commonly via the use of deletion mutants and other gene knockdown techniques like RNA interference or CRISPR/Cas9. We develop extensions to Nested Effects Models (NEMs), a method for inferring regulatory and signaling pathways from high-dimensional phenotypes subject to gene knockdown screens. We present the application of a novel method, context-specific NEMs, to transcriptomic data examining cellular stress-response in yeast deletion mutants. We also present an NEM-enabled analysis of the proteome, lipidome, and metabolome of yeast deletion mutants aimed at studying mitochondrial dysfunction. In the future we plan to apply these techniques to similar data obtained from knockdown experiments on human cells.

Poster #130

Identifying Abnormal Pathways from Gene Expression of Individual Samples

Authors: Michael I Klein, David F Stern, Hongyu Zhao, Yale University

Abstract: We present Gene-Ranking Analysis of Pathway Expression (GRAPE) as a novel method to identify abnormal pathways in individual samples that is robust to batch effects in gene expression profiles generated by multiple platforms. GRAPE first defines a template consisting of an ordered set of pathway genes to characterize the normative state of a pathway based on the relative rankings of gene expression levels across a set of reference samples. This template can be used to assess whether a sample conforms to or deviates from the typical behavior of the reference samples for this pathway. We demonstrate that GRAPE performs well versus existing methods in classifying tissue types within a single dataset, and that GRAPE achieves superior robustness and generalizability across different datasets. A powerful feature of GRAPE is the ability to represent individual gene expression profiles as a vector of pathways scores. We present applications of this pathway space representation to the analyses of breast

cancer subtypes and different colonic diseases. We perform survival analysis of several TCGA subtypes and find that GRAPE pathway scores perform well in comparison to other methods.

Poster #131

Computationally Unraveling the Regulatory Impacts of Neanderthal Introgression

Authors: Natalie Telis, Jonathan Pritchard, Kelley Harris, Stanford University

Abstract: There has been a long history of interest in the extent of, and selective consequences of, human interbreeding with other hominins. Existing knowledge of Neanderthal introgressed variants across the genome has revealed several deserts, as well as a general depletion in the vicinity of genes, with weak association with testes expression.

Since it has demonstrable consequences on modern clinical phenotypes and therefore fitness, it seems prudent to understand broader trends in functional consequences of introgression. Moreover, characterizing sources of strong pressures against presumably deleterious Neanderthal introgression can shed light into broader functional trends of modern selection.

We develop computational methodology to probe the evolution of these introgressed variants. By examining distributions of introgression across regulatory elements, we identify genome-wide depression of Neanderthal allele frequencies in promoters and enhancers. This genome-wide depletion increases in extremity with increasing allele frequencies. Alongside this broad depletion of Neanderthal introgression in regulatory regions, we confirm four independent in brain, stem cell, and muscle tissues. This signal is heightened in fetal regulatory regions, and suggests developmental disregulatory consequences of Neanderthal introgression.

Based our our analysis, we present a model linking modern regulatory consequence with fitness and distribution of existing Neanderthal variations.