INNOVATIVE APPROACHES TO COMBINING GENOTYPE, PHENOTYPE, EPIGENETIC, AND EXPOSURE DATA FOR PRECISION DIAGNOSTICS

MELISSA A. HAENDEL
Library and Department of Medical Informatics and Epidemiology, Oregon Health and Science University
Portland, OR 97239, USA
Email: haendel@ohsu.edu

MARICEL G. KANN
Department of Biological Sciences, University of Maryland, Baltimore County
Baltimore, MD 21250, USA
Email: mkann@umbc.edu

NICOLE L. WASHINGTON
Genomics Division, Lawrence Berkeley National Laboratory
Berkeley, CA 94530, USA
Email: nlwashington@lbl.gov

1. Introduction

Despite the increasing prevalence of Whole Genome Sequencing (WGS) and Whole Exome Sequencing (WES) in clinical settings, it is still very difficult to determine the causal variant for any given disease and most such explorations fall into research contexts rather than routine clinical diagnostics. There are literally thousands of studies and talks at human genetics conferences regarding the determination of causality, and a plethora of techniques available for statistical association of variants to disease phenotypes (e.g. GWAS). However, for rare diseases, the small number of individuals prevents statistical correlation techniques used for more common diseases. For complex diseases without Mendelian inheritance patterns, the challenge is even greater. Because we know the phenotypic consequences of mutation in approximately less than 40% of human coding genes, it is necessary to utilize a diversity of other data sources and algorithms to help determine causality. What is clinically actionable is an even more difficult assessment. Further, the methods and provenance of the data by which determinations of causality and actionability are often lacking and/or produce conflicting results. Finally, to realize truly precision medicine, we must embrace the idea that all diseases are rare in that each person has their own diversity of genotypic, phenotypic, and environmental variation.

Recent work has highlighted some of the exciting new possibilities to inform rare disease diagnostics. For example, use of model organism phenotyping data, interactome data, orthology, phylogenetic inference, and epigenomics can help fill some of the gaps. Further, methods that utilize semantic inference and probabilistic modeling have also been shown to aid diagnostics. Such methods combine standard WES or WGS prioritization techniques with an increasing diversity of phenotyping data and approaches. However, all of these combined approaches depend upon highly curated data and a diversity of software tools and algorithms, all of which provide the provenance for making any sort of causal or actionability judgment, the conclusions of which may

change over time as the data and algorithms change. In addition, the quality of the phenotyping data varies widely and is not always accessible via clinical notes. Finally, few such combined approaches attempt to include life history data such as exposures and chronological representation of disease progression.

2. Session Summary

This session includes an invited talk, five reviewed oral presentations, and two additional accepted papers. The studies presented in this session explore problems in combining genotype and phenotype data to support rare disease and/or precision diagnostics and treatment, and spanning multiple types of data. In particular, we selected contributions from those whose methodologies leveraged multiple data modalities in their analysis of genetic variation, such as clinical measures, imaging, natural language processing, semantics, homology, mined electronic health records (EHR) and manually curated data.

2.1. *Invited Talk*

The invited talk is given by Elissa Chessler, Ph.D. an Associate Professor of Bioinformatics and Computational Biology at the Jackson Laboratory, whose work spans a diversity of biological, genomic, and behavioral data toward identifying the biological basis for the relationships among behavioral traits, particularly in mouse models of disease. The resemblance of objectively measured phenotypic characteristics across species is limited by the extent to which the phenotypic inferences supported by these assays are relevant to the disease under investigation and reflect similar characteristics across species. 'Construct validity' is a more important criterion for the matching of phenotypes across species, and to the matching of phenotypes to disease. Construct-valid assays are expected to be associated with similar molecular and other biological characteristics across species, even when the external manifestation of the disease related phenotypes is quite different in humans and model organisms. There is a wealth of relevant data consisting of gene-phenotype associations obtained through high throughput, whole genome experimentation, including genetic mapping, expression correlation, differential expression, systems genetics, mutant screens, proteomic assays and curated functional genomics experiments. A variety of statistical and combinatorial approaches may then be applied to match data from various experiments and known gene-disease or gene phenotype associations. This approach to data driven inference of the relationships among the biological characteristics of animal models, assays and disease features has been implemented in the GeneWeaver.org system, a web service consisting of a database and analytic tools for collaborative integration of functional genomic experiments.

2.2. *Papers*

In *Discovering Patient Phenotypes Using Generalized Low Rank Models,* **Schuler et al**. develop a methodology for capturing phenotypic information within EHRs. The authors show that inherited challenges on the analysis of EHRs for phenotype discovery, such as missing data, sparsity, and

data heterogeneity, can be overcome by using the generalized low ranking model framework for such analysis.

In Diagnosis-guided method for identifying multi-modality neuroimaging biomarkers associated with genetic risk factors in Alzheimer's Disease, **Hao et al** present a novel, diagnosis-oriented, framework for selecting multi-modality quantitative traits associated with SNPs in the context of Alzheimer's Disease. This method has the potential to improve classification of patients with respect to their likelihood of developing Alzheimer's, by leveraging new data types and variables in their analysis algorithms.

In *Metabolomics Differential Correlation Network Analysis of Osteoarthritis*, **Hu et al.** describe a differential network approach to analyzing the metabolomics of an osteoarthritis (OA) cohort. The authors identified key metabolites that differ in OA and subsequently the cellular processes in which they are involved, with the goal of eventually leveraging these markers for the development of targeted therapies.

In *Integrating Clinical Laboratory Measures and ICD-9 Code Diagnoses in Phenome-wide Association Studies*, **Verma et al** describe a workflow that associates SNPs with clinical lab measures extracted from EHRs as well as ICD-9 codes. The suggested workflow would enable the use of clinical measures and their association with disease toward bringing clinical diagnoses and treatment to the level of individuals in the clinic for precision medicine.

In *Investigating the importance of anatomical homology for cross-species phenotype comparisons using semantic similarity*, **Manda et al** studies the influence of anatomical homology information on gene semantic similarity measures for phenotypic comparisons across species. Their findings are relevant to merging functional and anatomy-based gene homologue analyses.

In *Personalized Drug Targets via Network Propagation,* **Shnaps et al** present a computational strategy to simulate drug treatment in a personalized setting. The method is based on integrating patient mutation and differential expression data with a protein-protein interaction network.

In *Testing population-specific quantitative trait associations for clinical outcome relevance in a biorepository linked to electronic health records: LPA and myocardial infarction in African Americans* **Dumitrescu et al** combine genomic variant assessment (variants in LPA) and EHR phenotyping to determine risk in an unevaluated population, African Americans. This is important from the perspective of understanding how quantitative trait studies differ in different populations and highlights the challenges for complex clinical outcomes such as myocardial infarction.

2.3. *Acknowledgements*

We would like to thank all of the reviewers who provided valuable feedback for the authors of this session.