Precision Medicine: Improving health through high-resolution analysis of personal data

Steven E. Brenner[†]
University of California, Berkeley

Martha Bulyk
Brigham & Women's Hospital and Harvard Medical School

Dana C. Crawford
Case Western Reserve University

Jill P. Mesirov
University of California, San Diego

Alexander A. Morgan
Khosla Ventures

Predrag Radivojac
Indiana University

For the 2019 Pacific Symposium on Biocomputing's session on precision medicine, we present new research on computational techniques in range of areas including data curation, whole genome analysis, transcriptomics, microbiome profiling, EHR data-mining, and histological image processing.

Keywords: genomics, trancriptomics, personalized medicine, precision medicine

1. Introduction

For this session we sought submissions that drive research forward in the development of techniques in high resolution data science to advance personalization in clinical care based on quantified models. The roots of using data to improve healthcare and to personalize medicine are ancient and run deep in medicine. Hippocrates recommended that physicians learn to read so they could keep records and learn how to treat new patients by studying the case histories their colleagues compiled. James Lind, a naval surgeon, performed the first controlled clinical trial of a therapeutic intervention in 1747, with a multi-arm study of six different possible interventions for scurvy. William Osler (1849-1919), originator of the modern system of training physicians, said "The good physician treats the disease, the great physician treats the patient who has the disease." However, it has only been within the last few decades that we have had the tools to change the approach to understanding a patient from a somewhat subjective art to a deeply quantified science. We have advanced rapidly in molecular profiling from expensive single genomes to increasingly low cost genomic, transcriptomic, and proteomic profiling of single human cells. The massive switch to electronic health records, including the rise of large volumes of electronic imaging data in such forms as CT and MRI, has created huge volumes of computationally tractable data within the healthcare system. With an ever increasingly connected world, biosensors and mobile health tracking devices are providing new streams of phenotypic data. Inspired by the very earliest efforts in pushing medicine toward being a system of constant improvement and innovation based on data and experimentation (planned and naturally occurring), data is being collected in ever larger

[†] Supported by U41 HG007346 and U19 HD077627

volumes. However, we increasingly need innovation in bioinformatic techniques that help organize this data, discern the multi-omic characterization of disease, elucidate pathophysiology at the level of cells and tissues, and create actionable insights for points of intervention. The papers in this session span this gamut, and we hope will help drive the field from being not only precise, but also accurate in promoting the health and wellbeing improvements that can have widespread impact.

2. Session Papers

2.1. *Data Curation Tools and Techniques*

Peyton Greenside and colleagues [1] have developed a tool, CrowdVariant, based on Google's crowdsourcing platform to allow non-experts to annotate genomic data. They demonstrate with data from the Genome In A Bottle Consortium that the general public can be quickly trained to annotate deletions as a proof of concept. As the authors note, the images derived from genomic data, such as NGS read alignment create visual patterns that non-experts can be quickly trained to identify and interpret, opening up plenty of opportunity for future efforts to leverage the "wisdom of the crowd" in the expensive task of genome annotation, and potentially other forms of biomedical data.

Moving from the human crowd to the internal crowd of microbial flora, Wontack Han and Yuzhen Ye [2] have developed a repository of microbial marker genes and a set of tools to link microbial markers with human host phenotype, with an initial focus on diabetes, liver cirrhosis, and cancer. Their computational pipeline, Mi2P (Microbiome to Phenotype) is a publicly available project in Sourceforge.[a]

Another project helping to manage data related to precision medicine is the work of Zhiyue Tom Hu and colleagues [3], where they describe a framework for addressing inconsistency in large pharmacogenomic data sets, where individual potential therapeutics are screened against cancer cell lines. The method, Alternating Imputation and Correction Method (AICM), uses shared overlap of a handful of tested medications to bring divergent datasets into alignment for comparison across the full span of data. They show the validity of this approach with three large pharmacogenomic datasets.

2.2. *Techniques in Probing Complex Genome-Phenome Interactions*

Autism is a complex phenotype, with a strongly heritable component little explained by known genetic variants. Maya Varma and colleagues [4] have made creative use of a creative control group (progressive supranuclear palsy) to probe the genomic dark matter of non-coding regions to identify a set of genetic markers associated with autism. Despite significant work to remove

[a] https://sourceforge.net/projects/mi2p/

potential batch effects, they are able to achieve very strong classification accuracy (0.96 AUC) based on genetic features for identifying autism cases, suggesting the features they have identified in non-coding regions may be causal in ways that we have yet to identify.

Xinyuan Zhang and colleagues [5] tackle a different kind of complexity, as they look for pleiotropy in cardiovascular and neurological diseases in a dataset of 530,000 SNPs coupled with phenotypes extracted from EHR data for 43,870 individuals from the eMERGE network. Genes certainly play different roles in different contexts, such as different tissue types, different environmental stimuli, and different life histories; however, pleiotropy has been hard to detect in prior studies, due to a mix of factors including small datasets barely powered to find even simple single variant-phenotype interactions and poor phenotypic characterization. Here, leveraging a large dataset and the rich clinical annotations, they present a framework mixing a range of approaches to detect pleiotropy.

2.3. *Molecular Biology of the Tissues*

The natural extension for precision medicine discovery from the genome is moving into functional data and specifically gene expression. However, gene expression is very context specific, as noted in the work in this session. Derek Reiman and colleagues [6] look at the relationship between histopathology and gene expression in cancer, with a special focus on immune infiltration in the tumor micro-environment, of potential relevance to immune therapies in oncology. Applying a neural net based approach, they show that integrating features derived from digital surgical pathology imaging and RNA-Seq can automatically predict infiltration of the tumor by NK cells, macrophages, and CD8+ T-cells.

Binglan Li and colleagues [7] also focus on gene expression, and did tissue specific transcriptome wide association studies on clinical phenotypes in set of 4,360 individuals in an AIDS clinical trial, leveraging data on the context specificity of gene expression and eQTL's from the GTEx (Genotype Tissue Expression Project). This work has a poster at the conference and a paper in the proceedings.

2.4. *Creating Actionable Insights*

Precision medicine is about moving beyond just discovery to changing clinical practice with precise, personalized data. This session includes two pieces of work in this direction. The first is similar in direction with the previously mentioned work in that it focuses on eQTLs and gene expression regulatory relationships, but its focus is on therapeutic discovery and drug repositioning. Francesca Vitali and colleagues [8] use a network biology and semantic similarity approach to look for putative shared functional relationships between diseases to propose opportunities for drug repurposing.

Rounding out our session is work positioned to directly make recommendations around care decisions, particularly around the problem of when to order lab tests for critically ill patients. Patients in the ICU can have rapidly worsening clinical status, and blood-based diagnostic testing can help detect early signs of dangerous conditions such as sepsis or kidney failure. However, testing is not free, both in actual expense, but also patients do not have an infinite blood volume. Although patients in the ICU can have continuous venous access, in the general case, a blood draw is a form of invasive procedure, with discomfort and some risk involved. Li-Fang Cheng and colleagues [9] have developed a reinforcement learning framework to train a system for an optimal testing policy. This type of approach can both reduce unnecessary lab testing, but also suggests testing earlier than is currently done, in advance of critical events, ideally enabling early intervention to prevent poor outcomes.

References

1. Greenside, Peyton, et al. "CrowdVariant: a crowdsourcing approach to classify copy number variants," *Pac Symp Biocomput, 2019.*
2. Ye, Yuzhen and Han, Wontack, "A repository of microbial marker genes related to human health and diseases for host phenotype prediction using microbiome data," *Pac Symp Biocomput, 2019.*
3. Hu, Zhiyue Tom, et al. "AICM: A Genuine Framework for Correcting Inconsistency Between Large Pharmacogenomics Datasets," *Pac Symp Biocomput, 2019.*
4. Varma, Maya, et al. "Machine Learning Approach Identifies Single Nucleotide Variants in Noncoding DNA Associated with Autism Spectrum Disorder," *Pac Symp Biocomput, 2019.*
5. Zhang, Xinyuan, et al. "Detecting pleiotropy across cardiovascular and neurological diseases using univariate, bivariate and multivariate methods on 43,870 individuals from the eMERGE network," *Pac Symp Biocomput, 2019.*
6. Reiman, Derek, et al. "Integrating RNA expression and visual features for immune infiltrate prediction," *Pac Symp Biocomput, 2019.*
7. Li, Binglan, et al. "Influence of tissue context on gene prioritization for predicted transcriptome-wide association studies," *Pac Symp Biocomput, 2019.*
8. Vitali, Francesco, et al. "Precision drug repurposing via convergent eQTL-based molecules and pathway targeting independent disease-associated polymorphisms," *Pac Symp Biocomput, 2019.*
9. Cheng, Li-Fang, et al. "Learning an Optimal Policy for Ordering Patient Laboratory Tests in Intensive Care Units," *Pac Symp Biocomput, 2019.*