PRECISION MEDICINE: DATA AND DISCOVERY FOR IMPROVED HEALTH AND THERAPY

ALEXANDER A. MORGAN,

Stanford University School of Medicine, Stanford, CA 94305 USA, alexmo@stanford.edu

DANA C. CRAWFORD,

Epidemiology and Biostatistics, Institute for Computational Biology, Case Western Reserve University, Cleveland, OH, 44106 USA, dana.crawford@case.edu

JOSH C. DENNY,

Vanderbilt University Medical Center, Nashville, TN 37203 USA, josh.denny@vanderbilt.edu

SEAN D. MOONEY,

University of Washington Seattle, WA 98105 USA, sdmooney@uw.edu

BRUCE J. ARONOW, and

Center for Computational Medicine, Cincinnati Children's Hospital Medical Center and the University of Cincinnati, Cincinnati, OH 45229 USA, bruce.aronow@cchmc.org

STEVEN E. BRENNER

University of California, Berkeley, CA 94720-3012 USA, brenner@compbio.berkeley.edu

Abstract

The major goal of precision medicine is to improve human health. A feature that unites much research in the field is the use of large datasets such as genomic data and electronic health records. Research in this field includes examination of variation in the core bases of DNA and their methylation status, through variations in metabolic and signaling molecules, all the way up to broader systems level changes in physiology and disease presentation. Intermediate goals include understanding the individual drivers of disease that differentiate the cause of disease in each individual. To match this development of approaches to physical and activity-based measurements, computational approaches to using these new streams of data to better understand improve human health are being rapidly developed by the thriving biomedical informatics research community. This session of the 2017 Pacific Symposium of Biocomputing presents some of the latest advances in the capture, analysis and use of diverse biomedical data in precision medicine.

1. Introduction

The major goal of precision medicine is to improve human health. The researchers presenting work in the 2017 PSB conference session on precision medicine represent a wide range of approaches this challenge. The work ranges from examination of variation in the core bases of DNA and their methylation status, through variations in metabolic and

signaling molecules, all the way up to broader systems level changes in physiology and disease presentation. Recent advances in areas as diverse as microfluidics, solid phase chemistry, optics, wireless communication, battery technology, and social networking are supporting the collection and analysis of a whole host of highly multiplex biomedical measurements in increasingly fine temporal resolution of sampling. Whether it is understanding the individual drivers of disease that differentiate the cause of disease in each individual, to the creation of customized drug dosing algorithms, the researchers in this session are advancing data-driven medicine from applying to populations down to individuals.

One common thread that unites much of this work is the value of large datasets combining a wide range of features that encompass causal factors, state measures, and differential outcomes. Whether using a large patient registry focused on specific phenotypes and pathologies (such as autism or cancer) or broad spectrum electronic medical record systems, the linking of data collected as part of healthcare delivery combined with molecular and genomic features has provided an invaluable resource to help create data-precision models for disease understanding and improving care.[1-3] Without these data resources, most of the work in this session would essentially be impossible. Although those who make maximal use of these large datasets have been criticized for taking undue advantage of the labor of others,[4] it is clear that making these large datasets available to biomedical informatics researchers is enabling new methodological developments and new insights to advance clinical care. One recently reported study on the genetics of hypertension used samples from over 300,000 people;[5] at this scale, data should be considered a resource of global importance to health and wellbeing, not part of the academic fiefdom of a single researcher. Newborn screening extends this to its largest scale, addressing every member of a population (e.g., nearly 500,000 per year in California) without bias.[6]

The extensive work reported in this session reflects the diversity of activity in precision medicine and the enthusiasm in the field. However, this enthusiasm must be tempered with healthy caution and skepticism. Last year, the PSB session on precision medicine[7] was accompanied by another session focused on aspects and challenges in reproducibility[8] in research, and this continues to be a challenge in our efforts to develop an individualized understanding of physiology and disease as each person is in effect a sample size of one. This is a challenge across science, and much of the research across the psychological sciences has recently been criticized for its poor level of reproducibility.[9] In parallel to the methodological challenges in reproducibility, there continues be healthy skepticism and cautious evaluation of the continuously evolving techniques and approaches to collecting samples, measuring their properties, and evaluating their biomedical significance in isolation or in combination with other data and properties. A recent evaluation of a direct-to-consumer lab testing technology[10] revealed that any claim of technological advance without appropriate controls, comparisons, and supporting evidence must be examined in open formats by external parties before launching its widespread use in clinical care.

The burden of very careful experimental design and reproducibility does not mean that the field of precision medicine is advancing slowly. For example, recent work has shown that it is possible to develop predictive, customized models of blood glucose level in response to

different forms of dietary intake, a huge advance in precision, personalized nutrition.[11] Further research will demonstrate whether these models are stable over time.

The recent CAGI (Critical Assessment of Genome Interpretation)* evaluations have shown the power of the Common Task Framework to allow researchers to compare techniques that make predictions of phenotype from genotype, a key element of precision medicine. However, one of the trade-offs is that improved prediction accuracy often comes at the cost of human interpretability. For example, in the most recent CAGI of 2016, an approach using deep neural networks to predict psychiatric disease status from exome data performed better than other approaches that used far more interpretable models and those that integrated far more human knowledge. Unfortunately, the maturity of performance of these techniques of machine learning currently exceeds the maturity of the tools the to help interpret their predictions, limiting our ability to correct the apparent biases in our human understanding. Consequently, the significance and application of these findings are unclear. Much work has been done in fields like natural language processing and image processing to help visualize and unpack complex predictive AI models;[12] however, successful approaches and visualizations to fully support this increased understanding of many of the currently "black box" models of genomics and precision medicine are continuing to be developed.[13, 14] The many pieces of work presented in this session use a range of visualizations and evaluation metrics, but this continues to be an active area of endeavor in need of new advances.

Concomitant with advances in predictive and analytic approaches, informatics, and machine learning techniques are learning how to perform goal directed tasks, often at better than human levels of performance.[15] It is hoped that we can go beyond simple tasks like playing complex games to guidance of the steps and actions in the delivery of healthcare; however, as noted this will require healthy and active skepticism along with insight into the models developed.

2. Podium presentations

When Hippocrates espoused the idea that physicians should be literate and keep records of patient care and outcomes[16], it was so that these records might be used to improve the understanding of disease and help future patients. It is therefore not a new idea that medical records might be a powerful source of data for advancing biomedicine; the widespread use of electronic medical records systems has allowed several researchers in this session to use these data to deepen our understandings of disease and possible new methods of precision treatment. In particular, **CR Bauer and colleagues** investigate the relationship between genetic variation and 29 common laboratory values. Importantly, they go beyond simply viewing each of the laboratory values as simple quantitative traits, but look at the relationships between those quantitative traits and start to examine compositional quantitative traits derived from those measurements. Although it is common to think about the multiplicity of possible hypotheses derived from examining many genetic variants, little effort is typically spent examining the multiple hypotheses that can be derived from how we partition and divide phenotypes. In the closely related work of **SS Verma and colleagues**,

*https://genomeinterpretation.org

the focus is on the genetic drivers of the variability of common laboratory measurements. Going beyond the conventional central tendency of the laboratory values, they examine genetic associations with heteroscedasticity. This shift in focus from average value or pure prediction accuracy, toward models of higher moments and a focus on understanding what drives dispersion is another theme that runs through several of the papers in this session.

Laboratory values are part of assigning diagnoses, and **MK Beck and colleagues** mine records from 6,923,707 Danish patients to examine issues around the temporal ordering of diagnoses. They focus on the conditions of diabetes and sleep apnea, which often co-occur, but their presence can be hidden from the sufferer for years, and identification of one can lead to ascertainment bias of the other. When mining clinical records, researchers have access to when a disease was diagnosed but little data as to why, which may be impacted by a range of externalities, including differing access to care, but Beck and colleagues investigate patterns of age trajectory and of subsequent disease diagnoses, and data driven methods to stratify patients into subgroups. Moving up from laboratory measurements and diagnoses to directly guiding clinical decision making, but still using data derived from records of clinical care, **LK Wiley and colleagues** evaluate models that determine dosing of a medication with a narrow therapeutic window (warfarin) based on genetic variations in admixed populations, particularly those with African ancestry.

In addition to large datasets of mixed-type patient records, disease registries around specific diseases are a powerful data resource for precision medicine informatics. Three pieces of work in this session focus on techniques for identifying how variants in groups of genes may work together to contribute to phenotype, and much of this work relies on disease specific registry data. **GR Venkataraman and colleagues** use data from an autism patient registry to examine the way *de novo* mutations diffusely spread across sets of genes of shared function and how they may contribute to disease risk. The challenge of polygenic phenotypes is also the focus of the work of **D He and L Parida,** who presently work on disentangling epistasis underlying quantitative traits. **J Gallion and colleagues** have also been working on examining genetic variations in families of genes, in this case families of kinases in cancer, highlighting the shared disease association of variations in homologous locations across genes in a particular family.

Digging in more deeply into cancer, particularly using the data provided by The Cancer Genome Atlas,[17] **JA Thompson and CJ Marsit** present their work combining methylation with gene expression data to predict cancer survival; mixing heterogeneous data with colinearities being a hallmark of much of the cutting edge of precision medicine work. **G Speyer and colleagues** turn focus to drug response in cancer cells. Their work investigates the way expression dependency graphs vary between responsive and non-responsive cells; continuing the theme of mining differences in dispersion, here spread around network connectivity and likelihood, between subgroups. They also make the results of their work available online as a searchable resource.[†]

[†]http://biocomputing.tgen.org/software/EDDY/CTRP/home.html

3. Posters with published papers

The work presented in our poster session with published papers represents a broad range of interests by research groups, with some fairly technical work delving in deeply to new methods of analysis of biomedical data. **A Beck and colleagues** present an approach for using genome uncertainty to modify thresh-holding for tests of Hardy-Weinberg equilibrium; highly relevant to some of the most basic analysis done in population genomics and often serving as a filter for all the analysis downstream of the variant calling.

The entire *in silico* metabolic modeling of the most simple of single cells is now a reality,[18] and techniques of temporal molecular metabolic flux analysis are advancing dramatically.[19] **A Schultz and colleagues** are working to identify cancer specific metabolic signatures, and we may one day have patient and cancer-specific cellular metabolic models as tools for precision medicine.

As noted, one of the themes of this session has been on the investigation of measures of dispersion as a key biological measures, and **PF Kuan and colleagues** are examining DNA methylation, with methylDMV, a tool that compares not only measures of central tendency but also heteroscedasticity, as a way to highlighting issues like sample bias vs. biological signal.

There is a substantial amount of work in this session delving into methods to better identify cancer subgroups, both as a tool to more precision in individualized prognostic models, but perhaps more importantly to find features that unite these groups that might lead to precisely targeted therapies in those subgroups, or at least provide increased clarity on which existing therapies are likely to more or less efficacious. Cancer driver mutation identification is the focus of the work by **M Ma and colleagues**. **A Durmaz and colleagues** present work on subgraph analysis with a focus on grouping via dysregulated pathways. **H Kabbat and colleagues** use a competitive endogenous RNA based method combining DNA copy number variation, mRNA expression, and microRNA levels.

The interest in pathway analysis and uniting mRNA with microRNA is united in the work of **D Diaz and colleagues**, which focuses on just that topic. Finally, **T Kamp and colleagues** present work on the value of moving to a more Boolean view of gene expression when doing gene set enrichment analysis in improving analytical output.

Acknowledgments

5. References

1. Bush WS, Oetjens MT, Crawford DC. Unravelling the human genome-phenome relationship using phenome-wide association studies. Nat Rev Genet. 2016;17(3):129–45. [PubMed: 26875678]

2. Denny JC, Bastarache L, Roden DM. Phenome-Wide Association Studies as a Tool to Advance Precision Medicine. Annual Review of Genomics and Human Genetics. 2016;17(1):353–73.

3. Roden DM, Denny JC. Integrating electronic health record genotype and phenotype datasets to transform patient care. Clinical Pharmacology & Therapeutics. 2016;99(3):298–305. [PubMed: 26667791]

4. Longo DL, Drazen JM. Data Sharing. New England Journal of Medicine. 2016;374(3):276–7. [PubMed: 26789876]

5. Ehret GB, Ferreira T, Chasman DI, Jackson AU, Schmidt EM, Johnson T, et al. The genetics of blood pressure regulation and its target organs from association studies in 342,415 individuals. Nat Genet. 2016;48(10):1171–84. [PubMed: 27618452]

6. Brenner SE, Kingsmore S, Mooney SD, Nussbaum R, Puck J. Use of genome data in newborns as a starting point for life-long precision medicine. Pac Symp Biocomput. 2016;21:568–75. [PubMed: 26776219]

7. Morgan AA, Mooney SD, Aronow BJ, Brenner SE. Precision medicine: data and discovery for improved health and therapy. Pac Symp Biocomput. 2016;21:243–8. [PubMed: 26776190]

8. Manrai AK, Wang BL, Patel CJ, Kohane IS. Reproducible and shareable quantifications of pathogenicity. Pac Symp Biocomput. 2016;21:231–42. [PubMed: 26776189]

9. Open Science Collaboration. Estimating the reproducibility of psychological science. Science. 2015;349(6251).

10. Kidd BA, Hoffman G, Zimmerman N, Li L, Morgan JW, Glowe PK, et al. Evaluation of direct-to-consumer low-volume lab tests in healthy adults. The Journal of Clinical Investigation.126(7): 2773.

11. Zeevi D, Korem T, Zmora N, Israeli D, Rothschild D, Weinberger A, et al. Personalized Nutrition by Prediction of Glycemic Responses. Cell. 2015;163(5):1079–94. [PubMed: 26590418]

12. Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knolwedge Discovery and Data Mining 2016:1135–44.

13. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in Bioinformatics. 2013;14(2): 178–92. [PubMed: 22517427]

14. Gehlenborg N, O'Donoghue SI, Baliga NS, Goesmann A, Hibbs MA, Kitano H, et al. Visualization of omics data for systems biology. Nat Methods. 2010;7(3):S56–S68. [PubMed: 20195258]

15. Silver D, Huang A, Maddison CJ, Guez A, Sifre L, van den Driessche G, et al. Mastering the game of Go with deep neural networks and tree search. Nature. 2016;529(7587):484–9. [PubMed: 26819042]

16. Kassell L Casebooks in early modern England: medicine, astrology, and written records. Bull Hist Med. 2014;88(4):595–625. [PubMed: 25557513]

17. The Cancer Genome Atlas Research N, Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013;45(10):1113–20. [PubMed: 24071849]

18. Karr Jonathan R, Sanghvi Jayodita C, Macklin Derek N, Gutschow Miriam V, Jacobs Jared M, Bolival B, Jr, et al. A Whole-Cell Computational Model Predicts Phenotype from Genotype. Cell. 2012;150(2):389–401. [PubMed: 22817898]

19. Birch EW, Udell M, Covert MW. Incorporation of flexible objectives and time-linked simulation with flux balance analysis. Journal of Theoretical Biology. 2014;345:12–21. [PubMed: 24361328]