PRECISION MEDICINE: DATA AND DISCOVERY FOR IMPROVED HEALTH AND THERAPY

ALEXANDER A. MORGAN
Stanford University School of Medicine & Khosla Ventures
Stanford, CA, USA
Email: alexmo@stanford.edu

SEAN D. MOONEY
Department of Biomedical Informatics and Medical Education
Seattle, WA 98105, USA
Email: sdmooney@uw.edu

BRUCE J. ARONOW
Biomedical Informatics, Developmental Biology and Computer Science,
University of Cincinnati and Cincinnati Children's Hospital Medical Center
Cincinnati, OH 45229, USA
Email: bruce.aronow@cchmc.org

STEVEN E. BRENNER
Department of Plant and Microbial Biology, University of California, Berkeley
Berkeley, CA, USA
Email: brenner@compbio.berkeley.edu

Rapid advances in personal, cohort, and population-scale data acquisition, such as via sequencing, proteomics, mass spectroscopy, biosensors, mobile health devices and social network activity and other apps are opening up new vistas for personalized health biomedical data collection, analysis and insight. To achieve the vaunted goals of precision medicine and go from measurement to clinical translation, substantial gains still need to be made in methods of data and knowledge integration, analysis, discovery and interpretation. In this session of the 2016 Pacific Symposium on Biocomputing, we present sixteen papers to help accomplish this for precision medicine.

1. Introduction

Ultimately, precision medicine represents the significant enhancement of evidence-based medicine, where clinical guidelines gleaned from population-level studies are able to be precisely modified based on the attributes of the individual patient to both learn about new significant biological determinants of individual subtypes, and to then optimally treat that

individual. The age of precision medicine is already upon us. The evolution of medicine from an art and craft to science was facilitated through the development of methods of careful data collection and statistics for clinical trials, leading to medicine guided by population level evidence. In an analogous way that industrialization in manufacturing increased production volumes with standardization and systematic improvements in quality metrics, medicine has been moving from a heuristic based craft to mechanistically-tethered measures and guidelines. However, as we learn more about heterogeneity among the strongest factors determining disease risk, progression, and response to therapies, we can now identify highly significant factors that can forge new standards and individual-level customization. In an analogous way that evidence based medicine now guides standard of care practice at the population level, newer techniques will use data to guide practice at the level of individuals. Informatics methods in this space need to take advantage of highly multiplex heterogeneous mixes of categorical and numerical data, leverage related studies taking advantage of approaches in meta-analysis and transfer learning, be robust to missing data elements and sparsity, scale with superlinear interaction complexity, and be able to deal with a feature space much greater than the number of patients/samples by using approaches such as regularization and efficient use of priors.

Major efforts to create precision medicine datasets include the new national cohort as part of the US precision medicine initiative,[1] the 100k genomes being sequenced by each of the Geisinger-Regeneron collaboration[a] and the UK 100k genomes[2] projects and linked to clinical data, the US Veteran's Administration's Million Veterans initiative,[3] the many new ongoing trials using Apple's Research Kit,[b,4] Google's ambitious Baseline Study,[c] Vanderbilt's BioVU repository,[5] Craig Venter's Human Longevity Inc.,[d] and the massive cancer molecular profiling initiatives including The Cancer Genome Atlas.[6,7] Some of these data are already publicly available, but some of these projects are clearly not intended to be made public. In its most ambitious, precision medicine will require integration of data created by clinicians, biomedical labs, and commercial devices. How the academic research community, healthcare industry, commercial device industry, diagnostic test industry, and patient advocacy groups will negotiate the challenges of collaboration and privacy in the face of sometimes conflicting interests will be a challenge. However, recent efforts by groups such as Sage Bionetworks have highlighted the value of network effects between researchers and how new collaborative frameworks can accelerate and improve the discovery and

[a] https://www.genomeweb.com/sequencing/regeneron-launches-100k-patient-genomics-study-geisinger-forms-new-genetics-cent

[b] http://www.apple.com/pr/library/2015/03/09Apple-Introduces-ResearchKit-Giving-Medical-Researchers-the-Tools-to-Revolutionize-Medical-Studies.html

[c] http://www.wsj.com/articles/google-to-collect-data-to-define-healthy-human-1406246214

[d] http://www.humanlongevity.com/human-longevity-inc-hli-launched-to-promote-healthy-aging-using-advances-in-genomics-and-stem-cell-therapies/

innovation process.[8,9] Importantly, there is a moral imperative to accelerate the process of healthcare innovation and improvement, as the successes are measured in lives.

The diversity of papers in this session reflect some of the exciting range of topics in precision medicine. Informatics techniques for interpreting rare variation in complex genomes are presented alongside approaches that leverage links between clinical data stores and those genomic features. Methods for quantifying and analyzing complex phenotypes in patients in their daily lives are presented along with techniques for creating patient subgroups for targeting therapies. These papers reflect a sampling of the advances in informatics that are needed as we move into the age of precision medicine. Forums such as the Pacific Symposium for Biocomputing enable researchers to share ideas and help accelerate the process of discovery.

"The future is already here — it's just not very evenly distributed."
- William Gibson, National Public Radio interview

2. Session Contributions

2.1. *Methods for managing data complexity and limited sample size*

The explosion of rich and complex data in the age of precision medicine demands fundamentally new methods of analysis. The number of discrete data points collected from a patient can easily exceed the number of patients it would be possible to enroll in a single study, and can even exceed the population of the planet[10]. **Victor Bellón and colleagues** describe using a regularized transfer learning approach using task descriptors to address this problem. In addition to the sheer volume of data, the multivariate inter-relationships and connections mean that the complexity can scale at a rate much greater than simple linearity. **Nattapon Thanitorn and colleagues** present an approach using RDF Sketch Maps to reduce representational complexity.

2.2. *Probing rare genomic variation*

Much of what drives individual differences requiring precision, personalized treatments originates in the genome. However, rare or unique variants present an exceedingly difficult challenge in genome analysis and interpretation. **Anna Okula and colleagues** present the BioBin tool which builds on previous work[11] to support variant aggregation and statistical analyses. It is being used in the Marshfield Personalized Medicine Research Project. Expanding out from SNP's and indels, **Dokyoon Kim and colleagues** develop an annotation pipeline for copy number variants to support analysis of rare CNV's, also part of the Marshfield Personalized Medicine Research project. Going beyond the genome, **Yong Fuga Li and colleagues** describe diseaseExPatho, a tool that integrates transcription data with

genomic variants to develop regulatory modules to aid in the interpretation of genetic variation in rare diseases.

2.3. *Leveraging demographic and clinical data, challenges in precision medicine*

Performing studies and analyses in varying patient populations is challenging for many reasons, including biases in levels of representation and the phenotypic data collected. Three papers in this session describe how databases containing demographic and clinical data can highlight some of these challenges and provide new opportunities for research. **Sarah Laper and colleagues** discuss their inability to replicate previously well-established genetic links with cardiovascular phenotypes represented in hospital clinical records; suggesting that either the enthusiasm for the potential of clinical datasets linked to biorepositories needs to be tempered by the significant challenge of using this data for basic science research, or that this highlights the big gap between prior precision medicine results and their translation into clinical significance, or some mixture of these two. **Nophar Geifman and Atul Butte** focus their attention on a mismatch between the demographics of patients sampled in clinical outcomes studies and high molecular resolution for studies like The Cancer Genome Atlas and the general demography of cancer. **Jessica N. Cooke Bailey and colleagues** present their work looking at genetic variants associated with kidney disease in diverse ethnic subgroups, highlighting the particular challenges in investigating the genetic basis of disease pathologies that disproportionately affect particular ethnic subgroups.

2.4. *High-throughput holistic functional phenotypic profiling*

One of the most exciting aspects of precision medicine is the ability to go beyond the high-throughput molecular assays for genomics, transcriptomics, proteomics, and metabolomics, and to move into measuring phenotypes at the level of the whole individual. Rather than probing the function of a protein, we can probe the functioning of a whole human individual and how they interact with their world. Two papers in this session present efforts at phenotype profiling using mobile devices. **Elias Chaibub Neto and colleagues** describe their work profiling patients with Parkinson's disease using smartphone sensor data. **Maulik R. Kamdar and Michelle Wu** present their tool PRISM for monitoring mental wellness using a smart, sensor laden commercial wrist-based wearable. In both cases, new vistas for profiling patients are being opened up by these new data-streams and the informatics techniques to analyze them.

2.5. *Patient stratification and sample subgrouping*

Finally, our session includes five papers on subtyping patients and patient samples. An important element of clinical research has already become differentiating subgroups based on molecular/genomic level features. A recent review identified 684 registered clinical cancer trials that required genetic profiling for enrollment.[12] Disease subtyping may be done through analysis molecular/genomic level features or through a deep analysis of differences in phenotypic presentation, but either are playing an increasingly important role in clinical trials.[13,14] Importantly, for diseases like cancer, the disease may not represent just a single subtype, but may represent a population of different subtypes all coexisting simultaneously in an afflicted patient,[15–17] subtypes which need to be treated in concert, perhaps in different ways. **Vladimir Gligorijevic and colleagues** present an approach using non-negative matrix factorization for tumor stratification. **Sahand Khakabimamaghani and Martin Ester** describe a Bayesian bi-clustering approach for patient stratification using transcriptomic data. **Alex M. Fichtenholtz and colleagues** present an approach for looking at sub-groups of glial tumors to help in analysis of variants of unknown significance in a collection of 800 tumor sequences. **Subhajit Sengupta and colleagues** describe an approach for examining tumor heterogeneity using mutation pairs. Finally, **Artem Sokolov and colleagues** describe a one-class method for identifying specific cell type signatures in mixed samples.

3. Acknowledgments

We would like to thank the PSB 2016 chairs and Tiffany Murray of Stanford University for their efforts in organizing the meeting. We would like to thank UL1TR00042309 for funding (SDM).

References

1. Collins, F. S. & Varmus, H. A New Initiative on Precision Medicine. *N. Engl. J. Med.* **372,** 793–5 (2015).
2. Marx, V. The DNA of a nation. *Nature* **524,** 503–505 (2015).
3. Roberts, J. P. Million veterans sequenced. *Nat. Biotechnol.* **31,** 470–470 (2013).
4. Friend, S. H. App-enabled trial participation: Tectonic shift or tepid rumble? *Sci. Transl. Med.* **7,** 297ed10–297ed10 (2015).
5. Denny, J. C. *et al.* PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics* **26,** 1205–10 (2010).
6. Weinstein, J. N. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45,** 1113–20 (2013).
7. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489,** 519–25 (2012).
8. Friend, S. H. & Norman, T. C. Metcalfe's law and the biology information commons. *Nat. Biotechnol.* **31,** 297–303 (2013).

9. Altshuler, J. S. *et al.* Opening up to precompetitive collaboration. *Sci. Transl. Med.* **2,** 52cm26 (2010).
10. Chen, R. *et al.* Personal Omics Profiling Reveals Dynamic Molecular and Medical Phenotypes. *Cell* **148,** 1293–1307 (2012).
11. Madsen, B. E. & Browning, S. R. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.* **5,** e1000384 (2009).
12. Roper, N., Stensland, K. D., Hendricks, R. & Galsky, M. D. The landscape of precision cancer medicine clinical trials in the United States. *Cancer Treat. Rev.* **41,** 385–90 (2015).
13. Saria, S. & Goldenberg, A. Subtyping: What It is and Its Role in Precision Medicine. *IEEE Intell. Syst.* **30,** 70–75 (2015).
14. Röcken, C. Quality assurance in clinical trials-the role of pathology. *Virchows Arch.* (2015). doi:10.1007/s00428-015-1857-x
15. Sottoriva, A. *et al.* Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 4009–14 (2013).
16. Navin, N. *et al.* Inferring tumor progression from genomic heterogeneity. *Genome Res.* **20,** 68–80 (2010).
17. Powell, A. A. *et al.* Single Cell Profiling of Circulating Tumor Cells: Transcriptional Heterogeneity and Diversity from Breast Cancer Cell Lines. *PLoS One* **7,** e33788 (2012).