PERSONALIZED MEDICINE: FROM GENOTYPES, MOLECULAR PHENOTYPES AND THE QUANTIFIED SELF, TOWARDS IMPROVED MEDICINE

JOEL T DUDLEY
Icahn School of Medicine at Mount Sinai, 1425 Madison Ave., New York, NY
Email: joel.dudley@mssm.edu

JENNIFER LISTGARTEN
Microsoft Research, One Memorial Drive, Cambridge, MA, 02142
Email: jennl@microsoft.com

OLIVER STEGLE
European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom
Email: oliver.stegle@ebi.ac.uk

STEVEN E BRENNER
Department of Plant & Microbial Biology, 111 Koshland Hall, University of California, Berkeley 94720
Email: brenner@compbio.berkeley.edu

LEOPOLD PARTS
University of Toronto, Donnelly Centre for Cellular and Biomolecular Research, 160 College Street, Toronto, ON M5S 3E1, Canada
Email: leopold.parts@utoronto.ca

Advances in molecular profiling and sensor technologies are expanding the scope of personalized medicine beyond genotypes, providing new opportunities for developing richer and more dynamic multi-scale models of individual health[1,2]. Recent studies demonstrate the value of scoring high-dimensional microbiome[3], immune[4], and metabolic[5] traits from individuals to inform personalized medicine. Efforts to integrate multiple dimensions of clinical and molecular data towards predictive multi-scale models of individual health and wellness are already underway[6-8]. Improved methods for mining and discovery of clinical phenotypes from electronic medical records[9] and technological developments in wearable sensor technologies present new opportunities for mapping and exploring the critical yet poorly characterized "phenome" and "envirome" dimensions of personalized medicine[10,11]. There are ambitious new projects underway to collect multi-scale molecular, sensor, clinical, behavioral, and environmental data streams from large population cohorts longitudinally to enable more comprehensive and dynamic models of individual biology and personalized health[12]. Personalized medicine stands to benefit from inclusion of rich new sources and dimensions of data. However, realizing these improvements in care relies upon novel informatics methodologies, tools, and systems to make full use of these data to advance both the science and translational applications of personalized medicine.

Genotyping and large-scale molecular phenotyping are already available for large patient cohorts and may soon become available for many patients. Exome or complete genome sequences are increasingly being collected, and in some cases are now covered by insurance. Prenatal diagnosis has been improved by genotyping fetal DNA circulating in mother's blood tissue. Robust statistical and computational methods for analyzing these data will be critical to realizing the promise of personalized medicine. The challenges span from accurate low-level analyses of high throughput datasets to high-level synthesis of mechanisms of action, and identification of causal links between different abstract layers of molecular information, before, finally, incorporating them into health-care such as diagnostics. Important analysis problems include accurate phenotypic characterization, identifying and correcting for latent structure, dealing with missing data, deciding at what level to test (e.g., within genomes, whether to use single base pair values, sets of polymorphisms, exonic regions, etc.), data heterogeneity, the problem of multiple testing, integrating various modalities, deducing functional consequences *in silico*, addressing data quality, and making sense of new data types as they become available.

The path from genotype to disease state goes through intermediate phenotypes. To modulate the disease risk or trait, one of the molecular intermediates must be changed in a controlled way using small molecules or changes in environment. Finding the right intermediate molecule to target for these interventions remains a key challenge. A first level of understanding should come from genetic mapping studies – that is, to determine to which extent do the loci responsible for heritable disease risk affect intermediate traits. Much progress has been made on this front over the last years, especially for genetic control of RNA levels[13,14]—so-called "eQTL" analysis, but also for protein, metabolite and epigenetic modification abundances[15-18], with much remaining to be done. The next task is distinguishing the actual drivers of ailment from traits that do respond to genotype, but do not cause disease. Causal models, such as those based on Mendelian randomization and mediation analysis, will play a crucial role in separating out the molecular causes of disease from the high-dimensional state of the organism[19,20].

Medicine is gradually moving away from the traditional model of reactive sick-care towards wellness and all-time learning healthcare systems that aim to prevent individuals from perturbing their individual biology towards states of disease[1,2]. Personalized medicine aims to soon allow dynamic, quantitative representation of an individual patient's heath "GPS coordinates" estimated from multiple modalities of personal health data[1]. Still, much work is required in all areas, from basic discovery of molecular mechanisms of disease pathology, to statistical methods of causality and publicly available computational infrastructure to deliver on the promise of genetic and other personalized information in the clinic and beyond.

Session contributions

Dr. Nathan Price gives the invited talk**.** Dr. Price, along with colleagues at the Institute for Systems Biology, is spearheading the innovative Wellness 1K program that aims to take personalized medicine from "sick care" to maintenance of wellness by way of democratized healthcare[2].

The electronic medical record (EMR) captures clinical phenotype information and is being used increasingly as an important source of research data for precision medicine discovery. In our session, **Glicksberg *et al.*** discuss a novel integrative method combing genetic and EMR data for data-driven

discovery of disease relationships. The authors integrate disease-associated variants reported in the literature with EMR data from a large metropolitan hospital. The method evaluates statistical overlaps between patients sharing disease diagnoses in the EMR and disease phenotypes sharing overlapping associated loci. The results identify 19 putatively novel disease pairs supported by both EMR and genetic data that suggest possible shared etiological factors or novel risk factors.

Equipping clinical investigators with the ability to perform large-scale analysis of integrated clinical and molecular data is a key challenge in precision medicine. Clinical scientists sit at the interface of patient care and medical research and thereby serve as critical translators of clinical needs into specific research questions. Integrative methods combining genomic and clinical data offer powerful high-dimensional approaches for clinical hypothesis testing and patient cohort exploration. However, clinical investigators often lack the technical skills required to build, manage, and query integrated genomic and clinical data. In our session, **Hinterberg *et al.*** present the Phenotype-Expression Association eXplorer (PEAX) software enabling interactive data exploration of relationships between multivariate phenotype models and gene expression. The PEAX software interface enables visual, interactive definition of sub-phenotyping using clinical parameters and the system performs background statistical analysis to identify and plot gene expression correlates of sub-phenotype definitions. The PEAX software implementation uses open-source frameworks and source code is made available for download.

Also in our session, ***Diggans et al.*** describe a translational bioinformatics study identify and validate pre-operative mRNA based diagnostic test for V600E DNA mutations in thyroid nodules. A machine learning approach was applied in the discovery phase to identify a predictive 128-gene linear support vector machine from a feature space 3,000 transcripts measured from 716 thyroid fine needle aspirate biopsies (FNABs). The authors evaluate the 128-gene predictor against qPCR data in an independent test set and observe high positive and negative percent agreement with the qPCR test set. The results provide support for further clinical validation of the predictor and the potential for a first-of-a-kind diagnostic test for an unmet clinical need in thyroid cancer.

Efficient methods for inferring causal relationships across multiple scales of molecular traits are critical for modeling the complexity of biological systems. In our session, **Chang *et al.*** describe a novel method using Bayesian belief propagation for inferring the responses of perturbation events on molecular traits given a hypothesized graph structure. The method is not constrained by the conditional dependency arguments that limit the ability of statistical causal inference methods to resolve causal relationships within sets of graphical models that are Markov equivalent. The authors infer causal relationships from synthetic microarray and RNA sequencing data, and also apply their method to infer causality in real metabolic network with v-structure and feedback loop. Their approach is found to recapitulate the causal structure and recover the feedback loop given only steady-state data.

Accurate detection and modeling of tumor heterogeneity is a central challenge in understanding tumorigeneis and individual patient tumor characteristics. In our session, **Sengupta *et al.*** present a novel approach for modeling tumor heterogeneity (TH) using next-generation sequencing (NGS) data. The authors take a Bayesian approach that extends the Indian buffet process (IBP) to define a class of nonparametric models. Instead of partitioning somatic mutations into non-overlapping clusters with similar cellular prevalences, the authors do not assume somatic mutations with similar cellular prevalence

must be from the same subclone and allow overlapping mutations shared across subclones. The authors argue that this representation is closer to the underlying theory of phylogenetic clonal expansion, where somatic mutations occurred in parent subclones should be shared across the parent and child subclones. Their method yields posterior probabilities of the number, genotypes, and proportions of subclones in a tumor sample, thereby providing point estimates as well as variabilities of the estimates for each subclone. The method is implemented in a software package called BayClone that is made available for download.

Technological advances and increased public availability of data offer new opportunities to gain insights into the complexity of the eukaryotic transcriptome. Alternative cleavage of 3' UTRs has numerous functional consequences pertaining to the stability, transport, and translocation of transcripts. 3' UTR cleavages site analysis is also important clinically, particularly in cancer, where proto-oncogene can be activated by mRNA isoforms having shorter cleaved 3' UTRs. Thus both biological investigations and clinical applications benefit from more accurate methods for cleavage site analysis from transcriptional profiling data. In our session, **Birol *et al.*** describe KLEAT, a novel analysis tool that uses de novo assembly of RNA-sequencing data to search for and prioritize cleavage sites in poly(A) tails. The authors apply KLEAT to RNA-sequence data from ENCODE cell lines for which RNA-PET libraries are also available to compare predicted and actual 3' poly(A) signatures. The authors find that KLEAT exhibits > 90% positive predictive value when there are at least three RNA-sequencing reads supporting a poly(A) using the validation criteria of a minimum of three RNA-PET reads mapping within 100 nucleotides. The KLEAT software may accelerate biological and clinical applications of 3' UTR cleavage site analysis by enabling accurate analysis from more standard RNA-sequencing data and obviating the need for specialized wet lab techniques or sequencing libraries.

Though thousands of genes are implicated as underlying factors of disease it remains challenging to identify highest-value targets for novel drug development. In our session, **Gao *et al.*** address the question whether genes affected by strong genetic or environmental effects present better proxy therapeutic drug targets. To address this question, the authors propose a modeling approach that recovers both regulatory networks and estimates of environmental and genetic effects on gene expression. They apply their method to a gene expression data measured from blood samples from monozygotic and dizygotic twins and use the Connectivity Map database to assess whether genetic or environmental effects are more informative of gene's competency as a proxy target. The study findings suggest that a gene with strong genetic effects is more likely to act as a proxy target than a gene with strong environmental effects. This raises the intriguing hypothesis that diversity of a gene's expression across a genetically diverse population that makes it a suitable proxy rather than its sensitivity to environmental effects.

Finally, **Fan-Minogue *et al.*** evaluate the effectiveness of the differential expression (DE), disease-associated single nucleotide polymorphisms (SNPs) and combination of the two in recovering known therapeutic targets across 56 human diseases. They find that the performance of each feature varies across diseases and generally the features have more recovery power than predictive power. The systematic study results offer compelling evidence that the combination of the two features has more predictive power than each feature alone.

References

1 Topol, E. J. Individualized medicine from prewomb to tomb. *Cell* **157**, 241-253, doi:10.1016/j.cell.2014.02.012 (2014).

2 Hood, L. & Price, N. D. Demystifying disease, democratizing health care. *Science translational medicine* **6**, 225ed225, doi:10.1126/scitranslmed.3008665 (2014).

3 Schloissnig, S. *et al.* Genomic variation landscape of the human gut microbiome. *Nature* **493**, 45-50, doi:10.1038/nature11711 (2013).

4 Gaudilliere, B. *et al.* Clinical recovery from surgery correlates with single-cell immune signatures. *Science translational medicine* **6**, 255ra131, doi:10.1126/scitranslmed.3009701 (2014).

5 Kuehnbaum, N. L., Gillen, J. B., Gibala, M. J. & Britz-McKibbin, P. Personalized metabolomics for predicting glucose tolerance changes in sedentary women after high-intensity interval training. *Scientific reports* **4**, 6166, doi:10.1038/srep06166 (2014).

6 Chen, R. *et al.* Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell* **148**, 1293-1307, doi:10.1016/j.cell.2012.02.009 (2012).

7 Stanberry, L. *et al.* Integrative analysis of longitudinal metabolomics data from a personal multi-omics profile. *Metabolites* **3**, 741-760, doi:10.3390/metabo3030741 (2013).

8 Li-Pook-Than, J. & Snyder, M. iPOP goes the world: integrated personalized Omics profiling and the road toward improved health care. *Chemistry & biology* **20**, 660-666, doi:10.1016/j.chembiol.2013.05.001 (2013).

9 Newton, K. M. *et al.* Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. *Journal of the American Medical Informatics Association : JAMIA* **20**, e147-154, doi:10.1136/amiajnl-2012-000896 (2013).

10 Ozdemir, A. T. & Barshan, B. Detecting falls with wearable sensors using machine learning techniques. *Sensors* **14**, 10691-10708, doi:10.3390/s140610691 (2014).

11 Viventi, J. *et al.* A conformal, bio-interfaced class of silicon electronics for mapping cardiac electrophysiology. *Science translational medicine* **2**, 24ra22, doi:10.1126/scitranslmed.3000738 (2010).

12 Barr, A. in *Wall Street Journal* (New York, N.Y., 2014).

13 Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506-511, doi:10.1038/nature12531 (2013).

14 Parts, L. *et al.* Extent, causes, and consequences of small RNA expression variation in human adipose tissue. *PLoS genetics* **8**, e1002704, doi:10.1371/journal.pgen.1002704 (2012).

15 Johansson, A. *et al.* Identification of genetic variants influencing the human plasma proteome. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 4673-4678, doi:10.1073/pnas.1217238110 (2013).

16 Kettunen, J. *et al.* Genome-wide association study identifies multiple loci influencing human serum metabolite levels. *Nature genetics* **44**, 269-276, doi:10.1038/ng.1073 (2012).

17 Quon, G., Lippert, C., Heckerman, D. & Listgarten, J. Patterns of methylation heritability in a genome-wide analysis of four brain regions. *Nucleic acids research* **41**, 2095-2104, doi:10.1093/nar/gks1449 (2013).

18 McRae, A. F. *et al.* Contribution of genetic variation to transgenerational inheritance of DNA methylation. *Genome biology* **15**, R73, doi:10.1186/gb-2014-15-5-r73 (2014).

19 Gagneur, J. *et al.* Genotype-environment interactions reveal causal pathways that mediate genetic effects on phenotype. *PLoS genetics* **9**, e1003803, doi:10.1371/journal.pgen.1003803 (2013).

20 Fall, T. *et al.* The role of adiposity in cardiometabolic traits: a Mendelian randomization analysis. *PLoS medicine* **10**, e1001474, doi:10.1371/journal.pmed.1001474 (2013).