PERSONALIZED MEDICINE: FROM GENOTYPES AND MOLECULAR PHENOTYPES TOWARDS THERAPY

JENNIFER LISTGARTEN
Microsoft Research, 110 Glendon Avenue, Suite PH1, Los Angeles, CA
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

QUAID MORRIS
University of Toronto, Donnelly Centre for Cellular and Biomolecular Research,160 College Street, Toronto, ON M5S 3E1, Canada
Email: quaid.morris@utoronto.ca

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

Genotyping and large-scale molecular phenotyping are already available for large patient cohorts and will soon become routinely available for all patients. Exome or complete genome sequences are being increasingly collected and are explored as newborn screening technologies. These data are setting the stage for rapid advances in personalized medicine, enabling better disease classification, more precise treatment, and improved disease prevention. Robust statistical and computational methods for analyzing these data are critical to realizing the promise of genome-based medicine. The challenges span from accurate low level analyses of high throughput datasets to identification of causal links between different layers of molecular information, and incorporating them into 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. 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.

For example, in genome-wide association studies, population structure and family relatedness can reduce power and cause spurious associations. In gene expression and epigenetic studies, experimental artifacts and environmental influences have been shown to corrupt results of naive analyses. All of these problems can be tackled by various classes of latent variables, such as those related to Principal Components Analysis and probabilistic variations thereof, linear and non-linear mixed models. These models learn latent factors from the large scale of the data---that is, patterns which permeate many of the features, and therefore speak to wide-spread “contamination”. By removing these broad patterns, we hope to be left with the true associations; however, being certain of this is difficult [1-14].

Using patient genotype to inform treatment in the clinic is limited by our ability to accurately predict the impact of genetic variation, and the lack of models for its mechanistic effect. While whole genome sequencing has been successfully used to identify causal mutations for severe developmental disorders and other Mendelian diseases, use of genotype information has not yet permeated clinical practice, save for a handful of single locus tests [15]. Personalized approaches, however, are becoming increasingly common in applications to cancer treatment, albeit these are at present mostly limited to a research setting. Questions that remain are whether to treat the sequence data as clinical test, and only report known causal locus results for any phenotype under heavy regulation, or whether to broadly disclose any incidental findings. Many found variants are of unknown effect, and precise statistical models, as well as convenient software are needed to help practitioners make decisions [16]. To this end, efforts such as Critical Assessment of Genome Interpretation [17,18] are performing controlled experiments to probe the limits of our ability to predict phenotype from genotype.

The path from genotype to disease state goes through intermediate phenotypes [19]. 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, but one current limitation is finding out the right targets for these interventions. A first level of understanding should come from genetic mapping studies - to which extent do the loci responsible for heritable disease risk affect intermediate traits? Some progress has been made on this front over the last years, especially for RNA levels [20,21], but also protein and metabolite abundances [22,23], 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 Mendelian randomization methods, will play a crucial role in separating out the molecular causes of disease from the high-dimensional state of the organism [24,25].

Clinical grade confidence in data and methods to use genome information for providing better treatment has been difficult to establish, with heterogeneous and imperfect medical records also remaining a real bottleneck. However, each year brings more rigor and agreement in applications of genome-based personalized medicine in the field. 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 information in the clinic. The payoffs will be large.

Session contributions

The session keynote is given by **Robert Gentleman**, who has spearheaded the use of computational methods in biology and medicine [26], and is currently employing them to design cancer therapeutics.

The availability of inexpensive partial genotype data, and increasingly cheaper full genome sequencing to complement traditional diagnostic markers has fuelled the promise of personalised genomic medicine. However, genetic tests inform the diagnosis and treatment for only a minority of heritable disease cases in clinic today. This is partly due to low explanatory power of common small effect variants that underlie the common disease risk, but also due to larger effect alleles not being well captured by standard genotyping arrays as they have low frequency. In our session, **Martin et al.** analyse the performance of different genotyping platforms for imputing rare coding variation. Perhaps somewhat surprisingly, they find that genotyping arrays dedicated to measuring rare exome variants can be less useful in imputing unobserved rare variants than dense common variant arrays. This occurs because the latter are actually able to tag unmeasured variants (including rare ones) better than the specialized rare variant chips.

Interpreting incidental findings in whole-genome sequencing is difficult, and can take up considerable time of clinicians and genetic counselors. **Daneshjou et al.** will present PATH-SCAN, a publicly available tool that automatically annotates the variants that have been designated as pathogenic by ClinVar. The tool is expected to accelerate the analysis of genes that have been recommended by the American College of Medical Genetics and Genomics to be followed up and reported to the patient.

Also in our session, **Zhe et al.** tackle the problem of employing genotype and endophenotypes (intermediate phenotype) in disease diagnosis. Focussing on dissecting the genetic basis of Alzheimer's disease, a neurodegenerative disorder, they apply a latent variable model to the genotypes, magnetic resonance imaging, and diagnosis label where all the three types of observed features are sparse manifestations of a single continuous underlying disease state. After learning the model parameters, they then use them to predict disease state in a patient cohort, achieving better performance compared to current alternatives, and also uncovering several potentially causal links between genotype and the measured endophenotypes.

An important role for personalized medicine is in predicting frequency of drug side effects from genotype. **Oetjens et al.** genotyped 34 genes for 127 heart transplant recipients, 35 of whom had an adverse reaction to an immune suppressor. Incorporating data from electronic medical records, known predisposition to chronic kidney disease, and broad variance components in the genotype, the authors identified a single non-synonymous variant that significantly increased the risk of renal failure. Their study serves as a nice proof-of-principle that even with limited sample size and number of genotyped loci, genotype-dependent side effects can be identified using statistical analyses of longitudinal data.

Parikh et al. consider the problem of simultaneously inferring gene expression networks from a series of evolving conditions (*e.g.*, healthy tissue *versus* cancer stages) to identify functional roles of individual genes and pinpoint the causal changes. They propose a model that finds a sparse representation of the gene co-expression patterns sharing information across the different stages in a principled manner, and one which accounts for potential differences in the network structures.

Finding individual-specific contributors to immune response can help inform therapy of viral infections. In our session, **Perina et al.** propose a bag of words model to describe the distribution of epitopes presented by cells that are targeted by immune surveillance mechanisms. Their approach is able to better explain the correlations between individual epitopes compared to alternatives. For a clinical application, they test the models on a cohort of HIV patients to find links between distribution of epitopes and the viral load.

References

1. Lippert C., et al. *The benefits of selecting phenotype-specific variants for applications of mixed models in genomics.* Sci Rep. 2013. **3**:1815.
2. Listgarten, J., et al. *FaST-LMM-Select for addressing confounding from spatial structure and rare variants.* Nat Genet, 2013. **45**(5):470-1.
3. Listgarten J., et al. *A powerful and efficient set test for genetic markers that handles confounders.* Bioinformatics, 2013. **29**(12):1526-33.
4. Listgarten J., et al. *Improved linear mixed models for genome-wide association studies.* Nature Methods, 2012, doi:10.1038/nmeth.2037.
5. Listgarten, J. *Correction for Hidden Confounders in the Genetic Analysis of Gene Expression.* PNAS, 2010.
6. Stegle, O. et al. *A Bayesian Framework to Account for Complex Non-Genetic Factors in Gene Expression Levels Greatly Increases Power in eQTL Studies.* PLoS Comp Biol, 2010.
7. Parts, L. et al. *Joint genetic analysis of gene expression data with inferred cellular phenotypes.* PLoS Genet. 201. **7**(1):e1001276.
8. Stegle O., et al. *Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses.* Nat. Protoc. 2012. **7**(3):500-7.
9. Fusi, N., et al. *Detecting regulatory gene-environment interactions with unmeasured environmental factors.* Bioinformatics. 2013. **29**(11):1382-9.
10. Fusi, N. et al. *Joint modelling of confounding factors and prominent genetic regulators provides increased accuracy in genetical genomics studies*. PLoS Comp Biol, 2012.
11. Balding, D. *A tutorial on statistical methods for population association studies.* Nat Rev Genet, 2006. **7**(): 781-91.
12. Quon, G., et al. *ISOLATE: a computational strategy for identifying the primary origin of cancers using high-throughput sequencing.* Bioinformatics, 2009. **25**(21):2882-9.
13. Quon, G., et al. *Computational purification of individual tumor gene expression profiles leads to significant improvements in prognostic prediction.* Genome Medicine, 2013. **5**:29.
14. Qiao, W., et al. *PERT: a method for expression deconvolution of human blood samples from varied microenvironmental and developmental conditions.* PLoS Computational Biology, 2012. **8**(12): e1002838.
15. McCarthy J.J., et al. *Genomic medicine: a decade of successes, challenges, and opportunities.* Sci Transl Med. 2013. **5**(189):189sr4.
16. Jacob, J.J. et al. *Genomics in clinical practice: lessons from the front lines*. Sci Transl Med., 2013. **5**(194):194cm5.

17. Radivojac, P., et al. *A large-scale evaluation of computational protein function prediction.* Nat Methods. 2013. **10**(3):221-7.

18. Callaway, E. *Mutation prediction software rewarded.* Nature, 2010. doi:10.1038/news.2010.679

19. Chakravarti, A. et al. *Distilling pathophysiology from complex disease genetics.* Cell, 2013. **55**(1):21-6.

20. Lappalainen T., et al. *Transcriptome and genome sequencing uncovers functional variation in humans.* Nature. 2013. **501**(7468):506-11.

21. Parts L., et al. *Extent, causes and consequences of small RNA expression variation in human adipose tissue*. PLoS Genet, 2012. **8**(5):e1002704.

22. Kettunen, J., et al. *Genome-wide association study identifies multiple loci influencing human serum metabolite levels.* Nat Genet. 2012.**44**(3):269-76.

23. Johansson, Å., et al. *Identification of genetic variants influencing the human plasma proteome.* PNAS, 2013. **110**(12):4673-8.

24. Gagneur J., et al. *Genotype-environment interactions reveal causal pathways that mediate genetic effects on phenotype.* PLoS Genet. 2013. **9**(9):e1003803.

25. Fall, T., et al. *The role of adiposity in cardiometabolic traits: a Mendelian randomization analysis.* PLoS Med, 2013. **10**(6):e1001474.

26. Gentleman, R.C., et al. *Bioconductor: open software development for computational biology and bioinformatics.* Genome Biol. 2004. **5**(10):R80.