PERSONALIZED MEDICINE: FROM GENOTYPES AND MOLECULAR PHENOTYPES TOWARDS COMPUTED THERAPY

OLIVER STEGLE

Max Planck Institutes Tübingen, 72076 Tübingen, Germany
Email: oliver.stegle@tuebingen.mpg.de

FREDERICK P. ROTH

University of Toronto, Donnelly Centre, 160 College Street, Toronto, ON M5S 3E1, Canada
Email: fritz.roth@utoronto.ca

QUAID MORRIS

University of Toronto, Donnelly Centre,160 College Street, Toronto, ON M5S 3E1, Canada
Email: quaid.morris@utoronto.ca

JENNIFER LISTGARTEN

Microsoft Research, 110 Glendon Avenue, Suite PH1, Los Angeles, CA
Email: jennl@microsoft.com

Joint genotyping and large-scale phenotyping of molecular traits are currently available for a number of important patient study cohorts and will soon become feasible in routine medical practice. These data are one component of several that are setting the stage for the development of personalized medicine, promising to yield better disease classification, enabling more specific treatment, and also allowing for improved preventive medical screening. This conference session explores statistical challenges and new opportunities that arise from application of genome-scale experimentation for personalized genomics and medicine.

1. Overview of major statistical and computational challenges

Realizing the promises of personalized medicine requires robust analysis approaches, handling the breadth of data types and addressing key statistical challenges. Examples of these challenges include hidden structure within the data that can confound analysis results and lead to loss of power; missing or incomplete information; data heterogeneity; and the burden of multiple testing.

While these statistical challenges are not new, *per se*, the scale of genomic datasets comes along with additional difficulties but also opportunities for methodological innovation. For example, genome-wide association studies (GWAS) generate millions of hypotheses; it requires special consideration to reduce the burden of multiple testing so that the rate of false discoveries can be controlled[1] while retaining sufficient statistical power to detect true genetic associations, for example with single nucleotide polymorphisms (SNPs). One can begin to tackle these issues by incorporation of prior information (e.g. Lee et al.[2] and Sun et al.[3]), or using multivariate modeling[4]. Tied in with these techniques are also methods that combine groups of candidate features (*e.g.*, SNPs) in such a way as to obtain higher power, thereby attributing larger effect sizes, and uncovering a more complete picture of the underlying sources of heritability (e.g. Yang et al.[5] and Tatonetti et al.[4]).

Very large-scale datasets also support analysis strategies not available on smaller datasets. These include the ability to deduce and model hidden confounders from high-dimensional measurements, by way of Principal Components Analysis (e.g. Eigenstrat[6]), Factor Analysis[7,8], and Linear Mixed Models[9,10,11], for example. All of these approaches leverage on high data dimensionality, assuming that confounders act similarly on a large fraction of SNPs or phenotypes, which allows these factors to be reconstructed solely from the observed data.

Another challenge is the development of high quality software for the community, so that resources and expertise can be appropriately leveraged and shared. Use of such software exposes weaknesses that can then be addressed by further developments. For example, Linear Mixed Models software for GWAS has been increasingly made faster and faster, as the speed and memory constraints of each new software release becomes the bottleneck for larger and larger data sets[9,10,11].

2. Open statistical challenges

Statistical genomics is further complicated by the fact that, in real world settings, multiple confounders with intertwined impacts affect data, and as such, heterogeneous data need to be analyzed together rather than independently. Thus, tools are needed that tackle these statistical challenges in a joint fashion.

For example, when relating genotype to phenotype in a GWAS, population structure and family relatedness can reduce power to detect true associations and cause spurious associations[6]. Most molecular phenotypes, such as gene expression, are additionally contaminated with experimental artifacts or environmental influences. Such confounding factors, sometimes termed *expression heterogeneity*, have been shown to severely corrupt results of naïve analyses[7,8,12]. When seeking the genetic underpinnings of gene expression, such as in an expression quantitative trait loci analysis, problems of population structure, family relatedness and expression heterogeneity can be jointly present, and therefore models that address all of them simultaneously are required[12]. Additionally, individual readings of high-dimensional cellular phenotypes cannot be considered as independent, and thus hypothesizing and learning hidden regulatory causes of co-expression, such as cell type or transcription factor activity, has been shown to shed light on otherwise incomprehensible expression patterns[13]. However, further work in this area is still needed.

3. Open and usable software tools

Ultimately, personalized medicine needs to make its way into the clinic and the results of statistical inference need to be communicated to both clinicians and patients. How much statistics, molecular genetics and machine learning do users need to know to be able to interpret the results? Should software come with user-friendly tutorials on overfitting, multiple testing issues, p-values, false discovery rates and the 'winner's curse'? Although physicians and patients may be interested in inferences about health and disease, what they most require is assistance in acting on these inferences, i.e., making medical and lifestyle decisions that maximize expected benefit to the patient. So, is there another way to communicate an intuition about what it was about the primary data that led to the inference so that users can place their results in the context of current knowledge and evolving expertise? On top of addressing these difficult requirements, software must also safeguard patient privacy.

4. Session contribution

Our session explores these statistical and software development challenges within the context of personalized medicine.

In this session, **Karczewski et al.** propose new software, which allows the lay user to view and draw understanding from their genome sequence. It allows users to browse individual genome information in an interface that pays heed to security and privacy concerns. Of particular interest to computational biologists and statisticians engaged in tools development for personal genomics is a plugin interface that promises to ease the transition from ideas to implementation to wide use by physicians, patients and curious consumers.

Other directions of development addressed in the session are uncovering relationships between genotype and phenotype. Established modeling approaches are predominantly based on a linear model to predict the phenotypic readout from genotype, whereas heterogeneous disorders such as diabetes group distinct genetic diseases together. **Warde-Farley et al.** investigate using a mixture model of phenotypes and genotypes to map the age of diagnosis of type 2 diabetes. Using this simple, non-linear mapping, allows them to fit a set of simple genotype models that collectively predict phenotype. Perhaps many complex disorders with related phenotypes can be decomposed into multiple simple genetic disorders. Complementary to this work, **Curtis et al.** propose a new method to identify important transcriptome and phenome relations in the GWAS data by incorporating relational information on the gene and trait level. This new framework is based on a two-step procedure, where the first step identifies the important transcriptome relations (from SNPs to genes) and the second step identifies gene-to-trait relationships, thereby simplifying the problem of mapping from SNPs to traits.

Finally, our session explores computational methods to predict the sensitivity of tumor disease systems to external drug therapy. **Pal et al.** propose a novel approach that derives an abstract representation of cancer pathways from data, yielding Boolean drug kinase inhibition maps, which can then be used to predict sensitivity of systems with respect to a previously unseen drug treatment.

References

1. JD. Storey, R. Tibshirani, *Proc Nat Acad Sci* **16**, 9440 (2003).
2. S.-I. Lee, A.M. Dudley, D. Drubin, P.A. Silver, N.J. Krogan, D. Pe'erand D. Koller, *PLoS Genet* **5**, e1000358 (2009).
3. L. Sun, RV. Craiu, AD. Paterson, SB. Bull, *Genet Epidemiol* **6,** 519-30 (2006).
4. NP. Tatonetti, JT. Dudley, H. Sagreiya, AJ. Butte, RB. Altman, *BMC Bioinf* **11**, S9 (2010).
5. J. Yang, B. Benyamin, BP. McEvoy, S. Gordon *et al.*, *Nat Genet* **42**, 565–569 (2010).
6. AL. Price, NA. Zaitlen, D. Reich, N. Patterson, *Nat Rev Genet* **11**, 459–463 (2010).
7. JT. Leek, JD. Storey JD., PLoS Genet **3**, e161 (2007).
8. O. Stegle, L. Parts, R. Durbin, J. Winn, PLoS Comp Biology **6**, e1000770 (2010).
9. HM. Kang, J.H. Sul *et al.*, *Nat Genet* **42**, 348–354 (2010).
10. Z. Zhang, Z. Ersoz, CQ. Lai, *et al., Nat Genet* **42**, 355–360 (2010).
11. C. Lippert, J. Listgarten, Y. Liu, C. Kadie, R. Davidson, D. Heckerman, *Nat Methods* **8**, 833-835 (2011).
12. J. Listgarten, C. Kadie, EE. Schadt, D. Heckerman D., *Proc Nat Acad Sci* **107**, 16465 (2010).
13. L. Parts, O. Stegle, J. Winn, R. Durbin, *PLoS Genet* **7**, e1001276 (2011).