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

STEVEN E. BRENNER
Department of Plant & Microbial Biology, 111 Koshland Hall, University of California, Berkeley 94720-3102
Email: brenner@compbio.berkeley.edu

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

Introduction

Sequencing, genotyping, and large-scale phenotyping are currently available for a number of important patient cohorts and will soon be available as a result of routine medical practice. These molecular data, in conjunction with electronic medical records and rich, on-line resources, are setting the stage for truly personalized medicine. Personalized medicine promises to yield better disease classification, enable patient-specific treatment, and also allow for improved preventive medical screening. This session explores technical challenges and new opportunities that arise from the application of genome-scale experimentation for personalized genomics and medicine.

Realizing the promises of personalized medicine requires robust analysis approaches that handle a breadth of data, addressing key statistical challenges, and understanding how to leverage the wealth of information that is available. Examples of some 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 limitations; and the burden of multiple testing.

While these challenges are not new, per se, the scale of genomic datasets comes with added difficulties, but also offers new opportunities for methodological innovation. For example, genome-wide association studies (GWAS) generate millions of hypotheses, requiring

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]). These challenges are magnified as personal genomics moves to using genome sequence data.

Statistical genomics is further complicated by the fact that, in real world settings, multiple confounders are intertwined, affecting the data in ways which require complex models and the need for heterogenous data to be analyzed together rather than independently. 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 when naïve analyses are performed [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]. The trend we see in the problems and solutions just described is that large-scale data sets, while potentially problematic, also support analysis strategies not available on smaller datasets. In particular, they allow for us to deduce and then 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 [9-11], for example. All of these approaches leverage 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.

Ultimately, personalized medicine needs to make its way into the clinic--results of statistical inference need to be communicated to both clinicians and patients. In such a setting, how knowledgeable do end-users need to be about statistics, molecular genetics, and machine learning in order to interpret results in a way that is useful to that user? 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 require assistance in acting on these inferences to guide medical and lifestyle decisions that maximize expected benefit to the patient.

Session contributions

Our session explores these challenges within the context of personalized medicine.

The keynote lecture will be from **Atul Butte**, who has extensively demonstrated how comprehensive information about impacts of genetic variation have an important role in the interpretation of individual genomes, with strong implications for the clinic.

In **Province et al.,** a statistical method is developed to allow for robust combination of analyzed data sets for meta-analysis. In particular, the authors develop a framework for combining the results from different genome-wide scans when hidden dependency structures (may) couple together the various data sets. For example, when the same individuals appear in multiple data sets, these data sets are not completely independent and should not be treated so. Similarly, if siblings appear across data sets, these data sets are not completely independent. The authors use the reported p-values from each data set to estimate the full pairwise correlation matrix between all data sets that are to be combined, and then use this correlation matrix to correct for the dependency structure. With increasing data set sizes, relatedness of individuals will become an even more pervasive problem than it currently is; the methodology introduced in this paper will enable more general meta-analyses of such data sets.

Identifying clinical risk factors related to difficult-to-diagnose diseases remains a daunting problem. Such risk factors are important for early diagnosis, prognostics and preventative care. Using a case-study for one such disease, Alzheimer's, **Li et al.**, present a strategy to identify novel clinical markers using a manually curated database containing patient phenotype data and genome-wide associations. The author's driving hypothesis is that
traits that share genetic underpinning with Alzheimer's, as inferred by shared GWAS results, could serve as clinical risk factors. They find six clinical traits significantly associated with Alzheimers, of which one was not previously known as a clinical risk factor. This newly discovered association was then validated using electronic medical records, suggesting that it could be used as a new and effective prognostic marker.

Although genome-wide association scans are now routinely turning up important and reproducible associations, finding the actual causal variants responsible for disease generally requires further genotyping. **Crawford et al.** describe the properties of a custom content BeadChip designed for fine-mapping metabolic diseases and traits. Through application of this chip to 360 HapMap samples of European, African, Asian and Mexican descent, they explore the allele frequency distribution of these SNPs in these populations, and overall population differentiation. Also, they were able to identify, by way of pathway enrichment, a single SNP which indicates a difference in the functional properties of glutathione and drug metabolism through cytochrome P450 between the European and Mexican populations.

In addition to direct genetic factors, the state of microbiomes has also been shown to be predictive of phenotype and can help to understand patient well-being. To this end, it is necessary to extract useful information from metagenomic data, for example originating from the human gut. **Biswas**

et al. develop a hierarchical dictionary-based model to discover metagenomic units from pooled DNA-sequencing reads. The authors consider various likelihood models, including negative-binomial models, which are well suited for overdispersed count data. The resulting model is able to outperform several state-of-the-art assembly methods, both on simulated data and human gut metagenome datasets.

Several genomic analyses on health-related data require clustering of molecular data such as gene expression profiles. A key challenge in this context is to make an appropriate choice of the number of clusters. **Huang et al.** propose an efficient clustering approach that is suitable for heterogeneous molecular datasets as from disease studies. The developed approach is substantially faster than previous methods and does not require setting the number of clusters *a priori.* As a result, the approach yields clusterings that are better enriched for interpretable GO categories when applied to cancer genome data sets.

Finally, once molecular patterns indicative of disease have been identified, the next step is to understand the mechanisms that lead to disease. **Flores et al.** consider mutations in telomerase complexes, which can disrupt either nucleic acid binding or catalysis, thereby causing numerous human diseases. The authors tackle the underlying process by building a partial model of the human telomerase complex. Several predictions can be made from the model, elucidating disease-associated mutations.

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).