Recent progress in high-throughput genomic technologies has revolutionized the field of human genetics and promises to lead to important scientific advances. With new improvements in massively parallel biotechnologies, it is becoming increasingly more efficient to generate vast amounts of information at the genomics, transcriptomics, proteomics, metabolomics etc. levels, opening up as yet unexplored opportunities in the search for the genetic causes of complex traits. Despite this tremendous progress in data generation, it remains very challenging to analyze, integrate and interpret these data. The resulting data are high-dimensional and very sparse, and efficient statistical methods are critical in order to extract the rich information contained in these data. The major focus of the mini-workshop, entitled “Recent Developments in Statistical Methods with Applications to Genetics and Genomics”, has been on integrative methods. Relevant research questions included the optimal study design for integrative genomic analyses; appropriate handling and pre-processing of different types of omics data; statistical methods for integration of multiple types of omics data; adjustment for confounding due to latent factors such as cell or tissue heterogeneity; the optimal use of omics data to enhance or make sense of results identified through genetic studies; and statistical and computational strategies for analysis of multiple types of high-dimensional data.

Introduction by the Organisers

The mini-workshop “Recent Developments in Statistical Methods with Applications to Genetics and Genomics”, organized by Iuliana Ionita-Laza (New York), Michael Krawczak (Kiel), Xihong Lin (Harvard), Michael Nothnagel (Köln), was attended by 16 participants with broad geographic representation from North America and Europe. This workshop was interdisciplinary, and had a nice blend of junior and senior researchers with diverse backgrounds in theoretical/applied statistics, and genomics. The small scale and focused meeting has allowed for plenty of time for discussions and brainstorming new ideas, and has started several new collaborative projects. During the week, 15 lectures have been given by the participants. The lectures were accompanied by lively and interesting discussions. This report contains extended abstracts of all the talks.

The major focus of the mini-workshop has been on the efficient integration of different sources of data to gain a better understanding of the genetic mechanisms that lead to complex diseases. Multiple omics data (genome, epigenome, transcriptome, proteome, metabolome, phenome) can now be easily collected simultaneously on a genome-wide scale, yet remarkably little is known about how to integrate these different data types in a knowledge-based way. Integrative analysis of multiple omics data types can help the search for the underlying biological mechanisms in disease by discovering genomic features that tend to be dysregulated by multiple mechanisms. Because many of these technologies only recently became feasible on a genomic scale, the data are only now becoming available on a large scale, making the timing of this workshop ideal for sharing the emergent results, and for beginning to address the many challenges associated with these complex sources of data. New statistical methods have been discussed to make full use of the multi-source data for clustering, classification and prediction. Most integrative methods do not take into account known biological relations between different data sources. For example, there are well known regulatory relations between genomic data sets; e.g. gene expression levels can be regulated by both genetic aberrations and epigenetic factors. Integrating multiple data sets without accounting for their intrinsic relationships may unnecessarily increase the degrees of freedom in data and fail to contribute new information to existing variables. Therefore, new methods are needed to address these issues and others. Additional relevant research questions that have been addressed during the workshop include causal inference methods to identify causal mechanisms in disease, adjustment for confounding due to latent factors, the optimal use of omics data to enhance interpretation of results of genome-wide association studies (GWAS) and the integration of multiple GWAS datasets on different, correlated phenotypes.

Overall, the topics of the workshop are very important, and the talks and discussions have attempted to provide a survey of the current state of the field, and to explore new ideas and directions for future data integration approaches. The organizers would like to thank the Institute staff for providing such a great environment for our meeting.
Acknowledgement: The MFO and the workshop organizers would like to thank the National Science Foundation for supporting the participation of junior researchers in the workshop by the grant DMS-1049268, “US Junior Oberwolfach Fellows”.
Mini-Workshop: Recent Developments in Statistical Methods with Applications to Genetics and Genomics

Table of Contents

Stefan Böhringer (joint with Brunilda Balliu, Eleni Karasami)
The role of joint cumulants in genetic analysis .......................... 2975

Heather J. Cordell
Moving beyond genome-wide association studies through the modelling of more complex mechanisms ................................. 2977

Florence Demenais
Integration of biological knowledge, SNP and omics data for gene discovery in multifactorial diseases. .............................. 2978

Michael Epstein (joint with K. Alaine Broadaway)
Assessing Cross-Phenotype Effects of Rare Variants .................. 2982

Jeanine J. Houwing-Duistermaat (joint with Said el Bouhaddani, Hae-Won Uh)
Integrative analysis of two omics datasets from several heterogeneous studies using probabilistic O2-PLS .............................. 2984

Iuliana Ionita-Laza (joint with Kenneth McCallum, Bin Xu, Joseph Buxbaum)
A Spectral Approach Integrating Functional Genomic Annotations for Coding and Noncoding Variants ............................ 2985

Suzanne M. Leal (joint with Gao Wang, Di Zhang, Hang Dai, Zongxiao He, Biao Li)
Pitfalls of Rare Variant Data Association Analysis and Method Development ................................................................. 2987

Hongzhe Li (joint with S. Dave Zhao, Tony Cai)
Simultaneous Sparse Signal Detection with Applications in Genomics 2988

Xihong Lin (joint with Zhonghua Liu)
Multiple Phenotype Association Tests using GWAS Summary Statistics 2991

Sach Mukherjee
Learning molecular networks: interventions, joint estimation and causal interpretation ......................................................... 2992

Dan Nicolae (joint with Carole Ober, Oren Livne, Sahar Mozaffari and Matthew Reimherr)
Evolving designs in disease genetics ........................................ 2994
Michael Nothnagel (joint with S. Siegert, A. Wolf, D.N. Cooper and M. Krawczak)
  Confounding in omics data analysis: an example .......................... 2995

Catalina A. Vallejos (joint with John C. Marioni, Sylvia Richardson)
  Disentangling transcriptional heterogeneity among single-cells: a
  Bayesian approach ............................................................... 2998

Noah Zaitlen (joint with Hugues Aschard, Peter Kraft)
  Playing musical chairs in multi-phenotype studies improves power and
  identifies novel associations .................................................... 3000

Andreas Ziegler (joint with Marvin N. Wright, Inke R. König)
  An Orientational walk in the random forest: About first steps, solid
  grounds and interactions in a random forest ............................. 3002