Emerging Statistical Challenges in Genome and Translational Research

Jennifer Bryan (University of British Columbia),
Sandrine Dudoit (University of California, Berkeley),
Jane Fridlyand (Genentech Inc.),
Darlene Goldstein (Ecole Polytechnique Fédérale de Lausanne),
Sunduz Keles (University of Wisconsin, Madison),
Katherine S. Pollard (University of California, Davis)

June 1–June 6, 2008

1 Overview of Genome-scale Data Analysis

Modern high-throughput technologies are changing the face of biomedical and life science research. Biological research is moving from a hypothesis-driven focus on single genes and proteins to a high-throughput, discovery-driven strategy. Integrating the vast amounts of ever-changing types of data collected to study complicated entities, such as protein complexes and regulatory networks, requires an interdisciplinary approach.

The field of high-dimensional biology comprises several areas that are fueled by technological advances and require rigorous statistical and computational analysis. In each, there are high-dimensional multivariate data that are similar in nature. Background for these topics is given here.

Statistical Genomics and Regulation of Gene Expression

Quantitative traits are biological variables that are measured on a continuous (typically positive) scale. Examples include physical properties (e.g. height, weight, time to an event), molecular biomarkers (e.g. levels of mRNA, microRNA, protein or glycan), and chemical profiles (e.g. drug or metabolite concentrations). These traits tend to vary among individuals in a population. The observed trait variation results from genetic variation between individuals in the population. A region of the genome associated with variation in a quantitative trait is called a quantitative trait locus (QTL).

Single nucleotide polymorphism (SNPs) are a simple and prevalent source of genetic polymorphism in the human genome. SNP genotyping and haplotyping technologies are producing massive amounts of SNP data. One challenge faced by researchers is how to relate such multimillion dimensional genotypic profiles to biological and clinical phenotypes, such as disease and drug reaction. Analysis of the emerging complex data requires comprehensive statistical methodologies capable of dealing with challenging issues such as power, censoring, and causality.

Cancer Genomics

Translational aims are of paramount importance in current biomedical research. Recently, the National Cancer Institute has awarded a number of grants to generate an atlas of genomic and genetics features in cancers. While the ultimate aim is to improve cancer patient treatment, two major statistical complications arise. The first involves using high-dimensional patient data for predictions (e.g. response to standard or experimental therapies, time to recurrence). This problem falls into the class of *prediction statistical approaches*. The second involves identification of druggable markers of response to treatment, recurrence, progression or early detection. This can be viewed as a *variable selection problem*. We note that these two issues are tightly linked.

The statistical challenges include study design, building predictors based on heterogeneous cohorts, dealing with the small ratio of sample size (hundreds) to the number of variables (hundreds of thousands), multiple testing issues, computationally efficient classification, exploration of the interaction space of the variables, and handling the diverse data types in a unified rather than ad hoc manner.

Genome-scale data are also at the forefront of research into targeted therapeutics and individualized medicine. Pharmacogenomics deals with the influence of genetic variation on drug response in patients, and is overturning the "one size fits all" paradigm of drug development and treatment. Better understanding of an individual's genetic makeup may be a key element of the prescribed therapeutic regime. This multidisciplinary field combines traditional pharmaceutical sciences with large scale data and meta-data on genes, proteins, and SNPs.

The realization of the promise of personalized molecular medicine will require the efficient development and implementation of novel targeted therapeutics. The goal will be to deliver the right drug to the right patient at the right time at the right dose. This effort will require a integration of information from the DNA, RNA and protein level into predictors of which patients are likely to respond to particular therapies. The overall likelihood of response to particular drugs represents the interaction between predictors of sensitivity with predictors of resistance. Efficient clinical trials testing these precepts will require the development and implementation of novel trial designs. It is likely that the size of Phase I and II trials will need to be increased to allow the identification and validation of molecular markers at the same time as the initial evaluation the toxicity and efficacy of targeted therapeutics. This will come with the advantage of being able to deliver targeted therapeutics to enroll a much smaller population of patients selected for the likelihood to respond in phase III trials accelerating the approval of effective targeted therapeutics.

However, data analysis can be difficult due to limitations in the present state of knowledge regarding the relevant signaling pathways, as well as to high noise levels inherent in such data. New statistical developments here have the potential to play an important role in further progress toward individualized medicine.

High-Throughput Biotechnologies

Recent technological advances enable collection of many different types of data at a genome-wide scale, including: DNA sequences, gene and protein expression measurements, splice variants, methylation information, protein-protein interactions, protein structural information, and protein-DNA binding data. These data have the potential to elucidate cellular organization and function. There is now a trend for quantitative genome-wide phenotyping, with several large-scale studies currently being carried out with these new technologies: e.g. large collections of deletion mutants, cells or organisms undergoing RNA interference (RNAi). Studies of disease processes in humans often include patient clinical data and covariates as well.

Revolutionary breakthroughs in genomic technologies are enabling both the measurement of trait variation (especially molecular phenotypes) and the assaying of millions of genetic markers for large QTL studies. Microarrays, high-throughput ("next generation") sequencing, and mass spectrometry have revolutionized the field of quantitative genetics. In particular, single nucleotide polymorphism (SNP) chips have enabled genome-wide studies of genetic variation in panels of thousands of individuals. Next generation sequencing technologies (e.g. Illumina, Roche 454, SOLiD) have increased the quantity of DNA sequence data that can be produced in a laboratory by several orders of magnitude. The savings in time and cost mean that in the not very distant future it will be feasible to collect entire genome sequences of individual humans and model organisms. Advances in mass spectrometry are enabling researchers to accurately measure the concentrations of proteins (proteomics) and small molecules (metabolomics) in samples, expanding the collection of

molecular phenotypes that researchers can use to understand a biological process, such as a particular disease.

Each technology involves computational, mathematical, and statistical issues regarding data acquisition, processing, analysis and subsequent interpretation. Statisticians have already contributed immensely to improvements in low- and high-level analyses of genomic data, e.g. generated by microarrays. Continued interdisciplinary research is crucial to achieving a high level of methodological success for analyzing these newer data types, which will only gain in importance.

Data Integration

Fundamentally sound quantitative methods for combining the very heterogeneous data types described above are required in order to give researchers power to uncover meaningful biological relationships, enabling further understanding, targeted follow-up, and efficient use of resources.

Genomic studies differ from traditional epidemiological or clinical trials in several important respects. One obvious difference is that the number of variables measured in genomic studies is usually in the thousands per sample, rather than the perhaps tens for a clinical trial. Microarray study sample sizes are also typically much smaller, putting additional impetus on effective data integration methods.

In a clinical trial, the overall goal is primarily to obtain a combined estimate treatment effect. Genomic studies more often focus on combining evidence supporting the role of a gene or to rank evidence for a large number of genes. In contrast to the estimation scenario, in this case it may be advantageous rather than harmful to draw upon multiple, heterogeneous sources. Heterogeneity should tend to increase robustness of inferences, thereby enhancing the generalizability of study conclusions. The effects of within-study bias might also be reduced, as we would expect different biases in different studies.

The possibilities for combining information across studies can be viewed as occurring along a spectrum of levels of analysis, moving roughly from combination of least to most "processed" quantities – that is, in order of decreasing information content: pooling raw or adjusted data, combining parameter estimates, combining transformed p-values, combining statistic ranks, or combining test decisions.

Findings learned *jointly* from multiple, diverse data types are likely to lead to new insights that are not as readily discovered by the analysis of just one type of data. So far computationally straightforward, mainly correlative approaches have been applied in gene expression and copy number analyses for combining study results. It seems clear, though, that traditional meta-analytic methods are not very straightforwardly applied to the problem of combining data of different types, the most obvious impediment being lack of a common parameter across a mix of letter-based (sequence), categorical (SNP), ordinal (methylation, protein expression) and continuous (expression and copy number) data types. More sophisticated approaches include hierarchical Bayesian models and variations on correlation-based approaches. However, integrating multiple data types in an automated, quantitative manner remains a major challenge. This frontier is so novel that challenges appear at even the most fundamental levels of analysis: identifying the biologically relevant questions arising from data integration; specifying applicable statistical models and corresponding parameters.

2 Recent Developments and Open Problems

The technological advances outlined above provide unprecedented opportunity for understanding the genetic basis and molecular mechanisms of disease, as well as normal biological function. At the same time, these large and complex data sets are posing serious challenges. We outline some of these, with some comments on progress in the field and open questions.

Handling Massive SNP and Phenotype Data Sets

The scale of current data sets, which far exceeds that of earlier genomic technologies such as early gene expression microarrays, generates tough computational and algorithmic problems related to data storage, normalization, and modeling. With so many variables measured in each experiment – and in light of mounting evidence that complex phenotypes (e.g. common diseases) are the result of interactions between a large number of genetic and environmental variables – methods (for multiple testing, modeling, prediction) that properly and powerfully account for correlation between genomic variables are essential.

Multiple QTLs

Most quantitative traits are controlled by the combination of multiple different QTLs, each making a contribution to observed variation in the trait. There is great interest in identifying QTL locations, quantifying their effects on the trait, and understanding their modes of interaction (e.g. additive, multiplicative, or more complex). These biological questions can be addressed with statistical models and testing procedures. The scale and complexity of quantitative genetics data sets demands rigorous statistical methods for assessing power, modeling interactions, and accounting for multiple comparisons.

Confounding

Population structure can lead to genetic associations with quantitative traits that are not causal. This confounding is particularly a concern in studies that measure many phenotypes and markers, because it is very likely that a data set will include at least a few spurious associations.

Currently and historically, quantitative genetics is facilitated by controlled breeding and/or knowledge of population history and structure. Minimizing confounding by non-genetic factors through these experimental strategies helps reduce noise in trait data and to remove false associations caused by population structure. Several speakers pointed out the importance of such study designs and reported on results from studies of gene expression (Sunduz Keles) and metabolomics (Katherine Pollard) traits measured on controlled experimental crosses.

Another approach to this problem is to undertake specific genetic manipulations and measure their effects on a trait (or traits) of interest. Studies are now being conducted on huge panels of organisms in which distinct (combinations of) genes have been knocked out, or knocked down (e.g. by RNA interference). This area of high-throughput phenotyping is another very powerful way to link genes to phenotypes and to identify the genetic interactions (i.e. epistasis) underlying multi-genic traits. Elizabeth Conibear's poster illustrated this approach and presented a large knock-out data set.

3 Presentation Highlights and Scientific Progress

Here we give highlights from the talks presented at the meeting, along with the scientific progress that they represent as it pertains to the topics and problems described above.

John Ngai discussed biological insights into the workings of the vertebrate olfactory system gained by molecular, genomic, and computational approaches. He presented results of two studies carried out in his laboratory that utilized genome-wide approaches to identify genes involved in different aspects olfactory development.

Sunduz Keles pointed out that in a typical study with many phenotypes and many genetic markers (e.g. expression QTL (eQTL) studies), most markers are not associated with variation in most traits. This sparsity allows one to massively simplify the statistical problem by either filtering out or down-weighting the contributions of most loci. She presented a sparse partial least squares (sPLS) method based on a sparse prediction model.

Ru-Fang Yeh presented analysis issues for DNA methylation data from bead arrays. Aberrant cytosine methylation in CpG dinucleotides is associated with silencing of tumor suppressor genes in many cancers. Methylation status can be assessed using specialized microarrays. Yeh extends her methods developed for SNP array genotyping to make dichotomized methylation calls and derive associated confidence measures as an alternative for the manufacture-recommended metric, relative intensity ratio, and also developed a likelihood ratio test and a model-based clustering algorithm based on the underlying beta distribution for differential methylation and clustering analysis. These methods were illustrated on applications to cancer data.

Jenny Bryan presented several novel statistical approaches for both low-level and high-level analysis (normalization, clustering, growth curve modelling) of high-throughput phenotyping data from model organisms such as yeast.

Aseem Ansari presented results on the comprehensive binding preferences of polyamides against the entire sequence space of a typical 10bp binding site. He also described a new method for visualizing DNA binding data, called a Specificity Landscape. A Specificity Landscape displays the relative affinity of a

particular binding molecule for every DNA sequence assayed simultaneously. Specificity Landscapes were shown to accurately represent DNA sequence motifs with interdependent positions with high confidence.

Katherine Pollard noted that in many cases classical models for trait distributions (e.g. the normal model) are not appropriate with molecular phenotypes, which often have skewed or even discrete distributions, some times with point masses at zero. She presented several non-parametric methods for QTL mapping that avoid potentially incorrect distributional assumptions.

Karl Broman reviewed the traditional approach to multiple QTL mapping in which each genomic position is tested individually and then the family of tests is adjusted for multiple testing. He then argued that this problem is better viewed as one of model selection, and proposed a penalized likelihood method for simultaneously considering multiple loci. His method has better power than the traditional approach and also allows for the investigation of interactions.

Ingo Ruczinski discussed the problem of missing data and genotyping errors, which arise when the genotyping algorithms indicate that the confidence in certain genotype estimates is low. He presented several approaches to halding missing data and genotype uncertainties, and demonstrated that accounting for genotype uncertainty can be crucial when inferring possible copy number variants. The noveloty of this approach includes joint modeling of genotype calls and copy number, and in addition, integrating confidence estimates of the genotype calle and copy number estimates. The results presented in the talk demonstrated the superiority of the joint approach in terms of the accuracy of the genotype and copy number calls as shown on HapMap data. Although he focused on association studies, the methods may also be useful for QTL analysis.

Mark van der Laan presented a general maximum likelihood based approach targeting a user supplied parameter of the data generating distribution. This approach results in locally efficient estimators fully tailored for the parameter of interest, which have been shown to be more robust than maximum likelihood estimators. The method was illustrated in several applications: HIV drug resistance, detecting binding sites in the regulatory region of the yeast genome, breast cancer response to treatment and SNP association in case-control studies.

Adam Olshen presented the circular binary segmentation (CBS) technique for identifying regions of abnormal copy number. This has important ramifications in cancer, where progression often involves alterations in DNA sequence copy number. Multiple microarray platforms now facilitate high-resolution copy number assessment of entire genomes in single experiments. This technology is generally referred to as array comparative genomic hybridization (array CGH). The first published version of CBS was criticized for being slow. He present a methods for greatly speeding up the procedure. He also has shown approaches to recent copy number applications, including allele-specific copy number, clonality, and copy number variation. As in the talks of Ruczinski and Broman, the advantage of performing joint rather than unidimensional estimation of allele-specicifc copy number becomes apparent since join estimation allows making use of the summary constraint on the two alleles copy number number. Moreover, Adam discussed an interesting issue of clonality detection especially useful when one wants to distinguish between secondary primary and metastatic cancers. The clinical distinction is frequently vague; however, the ensuing treatment is dependent on the conclusion: a more aggressive treatment if metastasis, same treatment if secondary primary. Olshen has shown how copy number from paired samples could be used to distinguish these two situations.

Franck Picard considered joint analysis of multiple array CGH profiles. Most current segmentation methods can deal with one CGH profile only, and do not integrate multiple arrays, whereas array CGH microarray technology are becoming widely used to characterize chromosomal defaults at the cohort level. Picard presented a new statistical model to jointly segment multiple CGH profiles based on linear models. This strategy turns out to be very powerful for the joint segmentation of multiple profiles, as well as for the joint characterization of aberration types (status assignment of regions based on the cohort). The computational difficulties of simultaneous estimation are addressed using such tricks as model estimation with CART and linear programming. Overall, linear models offer a unified framework for the joint analysis of multiple CGH profiles.

Annette Molinaro presented a new experimental and analytical methodology to obtain enhanced estimates which better describe the true values of DNA methylation level throughout the genome, giving a model-based estimate of the absolute and relative DNA methylation levels. This model has been successfully applied to evaluate DNA methylation status of normal human melanocytes compared to a melanoma cell strain. Importantly, the model-derived DNA methylation estimates simplify the interpretation of the results both at single-loci and at chromosome-wide levels. This feature should make the method more accessible to

life scientists.

Mark Segal extends random forests (ensembles of decision trees) to multivariate responses and illustrates their use on several yeast microarray experiments, including cell cycle, and various stresses. Segal demonstrates that random forest derived covariate importance measures more reliably identify key regulators compared to relying on a single tree. Further, utilizing the proximity matrix from the forest output to cluster genes into homogeneous groups based on both motifs and expression values, Segal showed that the multivariate response random forest effectively reveals high-order motif combinations that influence gene expression patterns, thereby obviating the need for examining the entire combinatorial space of all motif pairs.

Jason Lieb described a number of projects, including identifying DNA-encoded regulatory elements and exploring how targeting and transcriptional output relate to each other in a simple developmental context for yeast.

Terry Speed reported on a case-study of qRT-PCR normalization, using principal components analysis for data quality assessment and normalization. He also considered the more general question of how to assess the effectiveness of a normalization method in the absence of other data (e.g. calibration data) and discussed a framework for quality assessment. This work has implications for other data types, such as microarrays.

Simon Tavaré described a high-throughput sequencing technology that replaces cloning and sequencing of bisulfite-treated DNA to identify DNA methylation patterns in single cells. The technology can be used to reconstruct ancestral information about stem cells and their lineages, and also applied to study tumour evolution.

Keith Baggerly Discussed the difficulties in predicting response to chemotherapy based on microarray data. The usual approach is to define a gene expression signature of drug sensitivity. In establishing the signatures, it would be preferred to use samples from cell lines, as these can be grown in abundance, tested with the agents under controlled conditions, and assayed without poisoning patients. Recent studies have suggested how this approach might work using a widely-used panel of cell lines, the NCI60, to assemble the response signatures for several drugs. Unfortunately, ambiguities associated with analyzing the data have made these results ambiguous and difficult to reproduce. Baggerly described methods to make the analyses more reproducible, so that progress can be made more steadily.

Neil Hayes discussed clinical experience in the genomic classification of lung cancer. He described an in-depth analysis of three independent lung cancer cohorts demostrating reproducibility of the gene-expression based signatures for known clinical subtypes and also for survival. The approaches presented were exemplary in terms of the study design, care with the classifier building and clear conclusions.

Pratyaksha (Asa) Wirapati looked at leveraging the accumulating public data to carry out combined analysis of data from multiple cancer studies by using hierarchical modeling for detection of differential gene expression, prediction, and cluster analysis. He presented a framework to modify standard single set microarray data analysis methods to accommodate datasets from multiple studies.

Gordon Mills gave an introduction to the topic of personalized medicine and a systems approach. Studies show that patients with the same type of cancer can have very different outcomes, even with the same treatment. Now physicians and researchers are developing personalized medicine treatment plans for each patient based on the molecular markers of their tumor. Systems biology is the study of the emergence of functional properties that are present in a biological system but that are not obvious from a study of its individual components. Systems biology is a data-driven process requiring comprehensive databases at the DNA, RNA, and protein level to integrate systems biology with cancer biology. Combining these patient and model-based databases with the ability to interrogate functional networks by a systematic analysis using siRNA libraries and chemical genomics provides an ability to link in silico modeling, computational biology, and interventional approaches to develop robust predictive models applicable to patient management. In describing the types of studies being carried out, he also emphasized the clinical needs for methodological development. He discussed some specific examples of utilization of diverse sources of data to identify specific genomic and genetic alterations which would make a cell susceptible to PI3K inhibitors. PI3K inhibitors are designed to attack a true heartland of cancer pathway and are being developed by nearly every company developing oncology drugs.

Jian-Bing Fan, from Illumina, discussed his company's development of technologies that address the scale of experimentation and the breadth of functional analysis required to achieve the goals of molecular medicine. There are array-based technologies for: SNP genotyping, copy number variation detection, DNA methylation studies, gene expression profiling, and low-multiplex analysis of DNA, RNA, and protein. These

serve as tools for disease research, drug development, and the development of molecular tests in the clinic.

Steffen Durrinck followed on with the recent technological advances in high-throughput transcriptome sequencing, For the last decade microarrays have been the major technology used to study gene expression. Despite their popularity, microarrays have known limitations such as cross-hybrization, probe affinity effects, availability for sequenced genomes only, and limited ability to study alternative transcription. Recent advances in sequencing technologies have significantly reduced the cost of sequencing, making it possible to now use sequencing for transcriptome studies. Because sequencing of transcriptomes on this scale is new there is a tremendous need for development of statistical and computational methods, for example to convert sequence data into exon and transcript-level expression measurements and to study differential transcript expression when comparing samples.

Hongyu Zhao gave an introduction to gene signaling pathways and showed how hierarchical models can be applied to the problem of signal transduction pathway analysis from single cell measurements. In contrast to measurements based on aggregated cells, e.g. gene expression analysis from microarrays, single cell-based measures provide much richer information on the cell states and signaling networks. The modeling framework allows pooling of information from different perturbation experiments, and network sparsity is explicitly modeled. Inference is based on Markov Chain Monte Carlo. Results from a simulation study demonstrate the effectiveness of this hierarchical approach, and the approach was also illustrated on experimentally-derived data.

Tim Hughes described efforts by his lab to determine the binding preferences of as many individual mouse transcription factors as possible, by determining binding specificity using a microarray technique. Mapping the complete spectrum of protein-DNA interactions is important for understanding global gene regulation and to fully decoding the genome and interpreting its evolution. The data accumulated thus far reveal a landscape of DNA sequence preferences, with many proteins exhibiting what appear to be multiple binding modes. Since the binding preferences correlate with conserved protein sequence features, the mouse data can be used to predict relative binding sequences in other species.

Rafael Irizarry presented a method that can accurately discriminate between expressed and unexpressed genes based on microarray data, thereby defining a unique "gene expression bar-code" for each tissue type. This method enables direct quantification of expression levels (rather than just relative expression between two samples) is also likely to contribute to better quantitative phenotyping for QTL studies. The method has been assessed using the vast amount of publicly available data sets, performing well in predicting normal versus diseased tissue for three cancer studies and one Alzheimer's disease study. The bar-code method also discovers new tumor subsets in previously published breast cancer studies that can be used for the prognosis of tumor recurrence and survival time. The bar-code approach to classification and discovery might also improved in various ways, for example by optimizing the simple detection method and distance calculations, or by expanding it to include microarray platforms in addition to the Affymetrix array types on which it has been developed.

Joaquin Dopazo described the bioinformatic challenges of casting genomic data into biological concepts, with the ultimate goal of providing a functional interpretation of experimental results. This is often done now by using functional enrichment methods on a gene list resulting from the experiment. Because the gene list requirements may be too stringent, there is a loss of power to detect the relationships of interest. The assumption that modules of genes related by relevant biological properties, and not the genes alone, are the real actors of the cell biology dynamics, leads to the development of new procedures implicitly closer to systems biology concepts. Some advantages and difficulties with these systems approaches were described.

Yee Hwa (Jean) Yang presented work on identification of candidate microRNA using matched mRNA-miRNA time course data. This is an example of using data from multiple technologies to address a biological question. Here, integration of the different data types was used to reduce the number of candidate genes to follow up. Also discussed were some of the technical difficulties with matching diverse data types.

Darlene Goldstein discussed work using the relatively recent technology of glycan arrays, used to study the biological roles for oligosaccharides. Glycomics represents another strategy for biomarker discovery. Some applications in HIV and cancer using glycomics were also described. She closed the meeting with highlighting some of the common and recurring themes in high-throughput life science research.

4 Open Questions and Outlook for the Future

There remain several open areas of research in the domain of genome-scale data analysis; we outline some of these here.

Statistical Issues for Genome-scale Data

Major issues in all genome-scale data analyses are the high dimensionality (although sometimes sparse) and multiple testing problem. Although the sparse partial least squares approach shows promise, it (like all PLS methods) lacks a rigorous theoretical framework. Sunduz Keles presented some preliminary theoretical results, but more work is needed in this area. The use of nonparametric models for trait data can be appropriate, but the performance of these models versus the better understood parametric approaches requires further study.

Several of the methods presented in this meeting account for the effects of multiple loci on (single) phenotype (trait or disease) expression. Taking this approach to the next level, one could also consider multivariate phenotypes being jointly analyzed with respect to multiple loci. It would be interesting to evaluate whether additional power might be gained by combining information across traits. Katherine Pollard presented a Hidden Markov Model that attempts to do this type of pooling, but initial simulation studies indicated relatively poor performance in terms of identifying the location of QTLs. The underlying model needs some refinement to better capture the relevant system characteristics.

Genomics of Human Disease

One general area of great importance is the genomics of human disease. Some aspects for which statistical and computational problems remain to be addressed include: identification and analysis of appropriate intermediate and endpoint phenotypes, reliable systematic discovery of disease-associated polymorphisms and pathways, appropriate and powerful study designs in genome-wise linkage and association studies, models for mechanistic studies of disease-associated genetic variants, models incorporating gene-environment interactions, study designs and sample sizes which allow reliable detection of genetic effects, and the translational step between information obtained from these studies toward therapies of clinical usefulness at either the group or individual (personalized medicine) level.

The area of personalized medicine presents a number of statistical challenges. In searching for markers that distinguish people who will respond to a given treatment, the multiplicity problem comes greatly into play. With hundreds of thousands of tests being carried out, the potential for false positives is huge. Further statistical research is needed in model selection, particularly in development and characterization of procedures based on data-snooping, examination of the signal-noise patterns occurring in the various high-throughput technologies, and assessment of bias-variance tradeoff and overfitting. Evaluation of the stability and reproducibility of results should help to reduce the chance that false positive results are erroneously followed up in subsequent studies.

Models that leverage the dependence structures in genomic data will be particularly useful, since a dimensionality reduction should result in higher power. It might also be more interpretable and biologically relevant to focus on groups of genes rather than single genes. A common method for interpretation of results is a two-step approach in which genes of interest are initially selected based on analysis of experimental data, and then in a second, independent step the enrichment of these genes in biologically relevant terms is analyzed (e.g. using Gene Ontology data). It should be more powerful to consider groups of genes and functional knowledge in an initial modeling step so that the association of the sets of genes can be tested directly. There are many complications associated with this type of approach, but more work in this direction could prove to be fruitful.

In the area of biomarker discovery and translational research, the variety and ready availability of very high-dimensional data types is still waiting to be exploited effectively. It may also be the case that different diseases will require different modeling approaches to address specific problems. As an example, there are very different issues when considering breast cancer and ovarian cancer. Breast cancer is a relatively common disease while ovarian cancer is much more rare, so that sample size issues will be different. There is no screening available for ovarian cancer, and cases are most often detected only when the disease is in

the advanced stages. Thus, unlike in the case of breast cancer, there is little material available to detect some of the early changes associated with the disease. It therefore seems that more targeted approaches will be required to search for biomarkers for diseases like ovarian cancer.

Data Integration

There remain several open statistical problems in the joint analysis of data across different studies. One primary problem is as basic as getting such a project started. If we consider the spectrum of analyses outline above, it would seem preferable to combine more informative data (i.e. closer to raw than highly processed data). However, obtaining the appropriate raw data is not always very straightforward, even with databases containing publicly available data. For example, clinical data are usually excluded from these, due to legal and/or patient privacy concerns. However, without such clinical information, it is not possible to analyze data for association with these outcomes. Improved sharing mechanisms for primary data are needed. Until this situation improves, though, it might be possible to use a missing data-EM framework for inference based on processed information that is more readily available.

Enhanced networking and sharing databases will also benefit the advancement of personalized medicine, as there will be more data to mine and hence more reliable inference should be possible. Methods for integrated analysis of heterogeneous data will depend on the types of data available, and whether it is raw or summarized. This area remains wide open for innovative applications of statistics.

List of Participants

Ansari, Aseem (University of Wisconsin Madison)
Baggerly, Keith (M. D. Anderson Cancer Center)
Bengtsson, Henrik (University of Carlifornia, Berkeley)
Broman, Karl (John Hopkins University)
Bryan, Jennifer (University of British Columbia)
Bullard, James (University of California, Berkeley)
Collin, Francois (Genomic Health)
Conibear, Elizabeth (University of British Columbia)
Culhane, Aedin (Dana-Farber Cancer Institute)
Delorenzi, Mauro (Swiss Institute of Bioinformatics)
Dopazo, Joaquin (Centro de Investigacion Principe Felipe)
Dudoit, Sandrine (University of California, Berkeley)
Durinck, Steffen (Lawrence Berkeley National Laboratory/UC Berkeley)
Fan, Jian-Bing (Illumina)
Fodor, Imola (Genentech, Inc.)
Fridlyand, Jane (Genentech Inc.)
Goldstein, Darlene (Ecole Polytechnique Fédérale de Lausanne)
Gottardo, Raphael (University of British Columbia)
Hansen, Kasper (University of California, Berkeley)
Hayes, D. Neil (University of North Carolina)
Hughes, Tim (University of Toronto)
Irizarry, Rafael (Johns Hopkins University)
Keles, Sunduz (University of Wisconsin, Madison)
Kostka, Dennis (University of California, Davis)
Lieb, Jason (University of North Carolina, Chapel Hill)
Mills, Gordon (M. D. Anderson Cancer Center)
Molinaro, Annette (Yale University)
Ngai, John (University of California Berkeley)
Olshen, Adam (Memorial Sloan-Kettering Cancer Center)
Picard, Franck (Centre National de la Recherche Scientifique)
Pollard, Katherine S. (University of California, Davis)

Ruczinski, Ingo (Johns Hopkins University)
Segal, Mark (University of California, San Francisco)
Speed, Terry (Walter & Eliza Hall Institute of Medical Research)
Taub, Margaret (University of California Berkeley)
Tavaré, Simon (University of Southern California and Cambridge University)
Thorne, Natalie (Cambridge University)
van der Laan, Mark (University of California Berkeley)
Wirapati, Pratyaksha (Swiss Institute of Bioinformatics)
Yang, Yee Hwa Jean (University of Sydney)
Yeh, Ru-Fang (University of Carlifornia, San Francisco)
Zhao, Hongyu (Yale University)