Statistical Genomics in Biomedical 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)

July 18 – July 23, 2010

Overview and Recent Developments in Statistical Genomics in Biomedical Research

Genetic research has been transformed by technological developments, and has by necessity become extremely quantitative, as massive quantities of varying complex data types can now be generated very rapidly. High-throughput data are being used in a variety of basic science investigations that have implications in the diagnosis and treatment of human disease: identification and characterization of genetic variants associated with a particular disease within and across populations; discovery of gene expression signatures associated with disease phenotypes; identification and testing of potential disease biomarkers. Methodological advances in the statistical treatment and interpretation of these data are needed in order to meet the pressing need for powerful, efficient and robust analyses.

Biomedical progress is increasingly dominated by high-throughput technologies, presenting new statistical and computational hurdles to overcome in order to make sound quantitative inferences. These technological advances provide unprecedented opportunity for understanding the genetic basis and molecular mechanisms of disease, as well as normal biological function. At the same time, they have given rise to multiple and complex data types, posing serious modeling and analytic challenges.

Fostered by the development of new techniques in molecular biology, translational research has become an important aspect of clinical research. The aim of translational research is to translate knowledge derived from laboratory work (basic research) into clinical applications. Translational research occurs at the interface between quantitative methodology and clinical treatment. The field is highly multidisciplinary and team-based, encompassing researchers with backgrounds in lab-based basic science, clinical investigation, statistical methodology and computational/algorithmic development. Moving these very high-dimensional data into clinical practice will require biologically-inspired expansion of the existing statistical framework for dealing with these complex structures.

This workshop will focus on key areas of basic and clinical biological research that generate very high dimensional, complex data structures and which require further development of statistical and computational methodology for their efficient use.

The targeted areas are briefly described below.

Population and Quantitative Genomics

Patterns of genetic variation in a population can reveal the dynamics of that species' history, disease susceptability, and response to changing environmental conditions. The focus is on genome features that vary among individuals within a species. While population genomics does not necessarily involve measuring a phenotype, studying the association between this genetic variation and variability in traits of interest across a population can shed light on the underlying biology, for example, helping to identify genetic risk factors associated with disease. Genetic variation consists of single nucleotide polymorphisms, large-scale polymorphism, copy number variation, and insertions, deletions, and rearragements in the genome. The major statistical challenges in population genomics are appropriately modeling and quantifying uncertainty about the historical events (recombination, mutation, migration, drift) that shape genetic variation. Coalescent theory and diffusion models provide a sound basis for these studies, but require extension and algorithmic improvements to handle the high-dimensional data sets from emerging technologies.

Quantitative traits of an individual (*e.g.* molecular biomarkers, drug concentrations, physical properties) are typically controlled by a collection of genetic loci, called quantitative trait loci (QTLs). Genomic technologies are enabling the measurement of many such traits and the simultaneous study of their association with millions of genetic markers. Quantitative genomics is facilitated by controlled breeding and/or knowledge of population history and structure. Furthermore, studies are now being conducted on huge panels of organisms in which distinct (combinations of) genes have been knocked out, or knocked down. 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. Analysis of these complex data sets requires statistical methods for assessing power, modeling interactions, and accounting for multiple comparisons.

High-Throughput Sequencing Assays and Transcriptional Genomics

Advances in high-throughput sequencing capabilities have given rise to new, sequence-based versions of microarray-based assays. Common ones in current use include RNA-seq and ChIP-seq.

RNA-seq is a protocol for sequencing messenger RNA, and can be used as a tool to measure gene expression levels (e.g. for identifying differentially expressed genes) as well as for other aims requiring increased sensitivity compared to microarrays, notably to identify alternative splicing. This new technology requires a new generation of software for alignment to a reference genome. Software must be able to accommodate the reality that not all splice junctions are known, and the additional complications stemming from the large number of fragments with only very short overlaps. Some progress has already been made in this direction.

ChIP-seq is a sequencing-based alternative to (microarray-based) ChIP-chip that combines chromatin immunoprecipitation (ChIP) with DNA sequencing. This type of assay is used to study DNA-protein interactions. For example, gene expression is regulated by proteins known as transcription factors. Knowing how transcription factors and other proteins interact with DNA is crucial to understanding many types of biological functions.

ChIP-Seq has revolutionized experiments for genome-wide profiling of DNA-binding proteins, histone modifications, and nucleosome occupancy. As the cost of sequencing is decreasing, many researchers are switching from microarray-based technologies (ChIP-chip) to ChIP-Seq for genome-wide study of transcriptional regulation. Despite its increasing and well-deserved popularity, there is little work that investigates and accounts for sources of biases in the ChIP-Seq technology. These biases typically arise from both the standard pre-processing protocol and the underlying DNA sequence of the generated data.

Other difficulties arise in data analysis due to problems in peak identification/resolution (the precise DNA-protein binding location is not identified, only the ends of the ChIP fragments) and also due to regional biases such as sequencing and mapping biases. Model-based statistical approaches can be useful for resolving these difficulties.

Transcriptional genomics is concerned with approaches for understanding transcriptional regulation, based on data from both gene expression studies (by high-throughput sequencing and/or microarrays), chromatin immunoprecipitation assays, and promoter sequence data. One aim is to catalog and gain an understanding of single transcription factors; another is to identify transcriptional modules, sets of genes that are co-regulated in a set of experimental or in vivo conditions.

Basic and Clinical Research: Predictive Diagnostics and Designing Clinical Trials

Translational research aims to bridge the disconnect between new basic science discoveries and the ability to translate those discoveries into effective, affordable and safe medical treatments for patients. There is a need for cross-disciplinary work of basic science researchers, clinical scientists, computational scientists and statisticians to develop biologically meaningful yet quantitatively sound approaches to integrating genomic, experimental and computational evidence during research and clinical drug development. The statistical challenges include: evaluating accuracy and precision of the technology used to measure biomarkers; evaluating prediction accuracy and developing appropriate statistical measures (along with uncertainty estimates) for the performance of these potential biomarkers relative to any established benchmarks; and assessing the reproducibility of experimental outcomes and the resulting inferences, both within and across different populations.

For progress in diagnosis, prognosis and treatment of human disease, associations between disease with the avalanche of genomic information (SNP genotypes, haplotype blocks, candidate genes/alleles, proteins, and metabolites) must be reliably quantified and assessed. There is a strong need for biologically relevant, powerful computational methods and models to integrate multi-level genome-wide evidence and to interpret the resulting high-dimensional outcomes so that strategies for clinical implementation can be developed. The major fundamental statistical challenges occur at the data analytic stage, where diverse data elements from all sources need to be incorporated into comprehensive models for prediction, risk assessment, and/or efficiency.

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.

Population and Quantitative Genomics

Jonathan Pritchard spoke about expression QTL mapping using RNA-Seq, a new, sequencing-based alternative to microarrays for measuring transcriptome activity. An important challenge of the post-genomic era is to make sense of how genome sequences control gene regulation. The talk discussed work using expression- and splicing-QTL (quantitative trait loci) in human lymphoblastoid cells as a model system for understanding how genetic variation can modify gene regulation. The focus was on the application of next-generation sequencing for measuring gene expression and splicing patterns and attempts to understand the mechanisms of action of eQTL SNPs.

Jeff Wall considered the problem of estimating human demographic parameters from sequence polymorphism data based on population genetic data. These data sets have the potential to inform about a species' demographic history, but most existing methods are not suitable for genomic-scale data. A composite-likelihood framework was presented for estimating demographic parameters such as split times and migration rates. The method was applied to the analysis of polymorphism data from different sub-Saharan African populations. His group has found evidence for population structure that likely predates the exodus of modern humans out of Africa. This finding has relevance with regard to current theories of human evolution.

Yoav Gilad has used next-generation sequencing to carry out comparative genomics in primates. Progress in evolutionary genomics is tightly coupled with the development of next-generation sequencing technologies, providing the ability to focus on a large number of outstanding questions that previously could not be addressed effectively. In the context of comparatives genomic studies in primates, new sequencing technologies have allowed collection of high resolution inter-individual and inter-species variation data from multiple dimensions of the regulatory landscape. These data are used to better understand the contribution of different regulatory mechanisms to overall inter-species differences in gene regulation, and allow identification of individual genes and entire pathways whose regulation evolves under natural selection in primates. These observations have the potential to help find functional genetic variation in humans. He provides an example where it was found that the set of genes previously associated with diseases that affect specific tissues is

enriched for genes whose regulation evolves under stabilizing selection in the same tissues.

John Ngai presented insights gained by transcriptome profiling on regulation of olfactory stem cell renewal and differentiation. The process of tissue regeneration is complex, requiring coordination of stem cell proliferation and differentiation to maintain or repair the structure. The olfactory epithelium (OE) is a sensory neuroepithelium whose constituent cell types – including the olfactory sensory neurons – are continuously replaced during the lifetime of the animal. Following severe injury that results in the loss of mature cell types, the OE is rebuilt by the proliferation and differentiation of adult tissue stem cells. The regenerative capacity and limited number of cell types make the OE an excellent model for investigating stem cell regulation in vivo. He discussed previous studies that have identified the horizontal basal cell (HBC) as the multipotent neural stem cell of the OE. However, the molecules and pathways regulating this adult tissue stem cell are unknown. He used whole genome expression profiling of FACS-purified HBCs to characterize the mRNA and miRNA transcriptomes of HBCs under conditions of quiescence and proliferation/differentiation. These studies allowed identification of groups of genes associated with different phases of the HBC life cycle.

His group has found that p63, a member of the p53 tumor suppressor gene family, is highly enriched in quiescent HBCs. This finding is important because p63 is a key regulator of stem cell self-renewal and differentiation in all stratified epithelia investigated to date. Conditional inactivation of the p63 gene in HBCs results in the appearance of mature cells but loss of HBCs following regeneration. These results demonstrate a critical role of p63 in olfactory stem cell renewal and differentiation, and provide an entrée toward elucidating the downstream targets and interaction partners of this transcription factor. These studies provide the first molecular insights into the genetic network regulating stem cell dynamics in the OE and reveal an unexpected parallel between stem cell regulation in this sensory neuroepithelium and other epithelial tissues.

Transcriptional Genomics

Jason Lieb considered genome-wide measurement of transcription factor binding dynamics by competition ChIP. He presented a novel method applicable to a wide range of experimental next-generation sequencing datasets and signal patterns, including ChIP-seq, FAIRE-seq, and Histone Modification data. The method comprises a mixture regression-based framework that rigorously identifies, assesses and quantifies sets of factors that are relevant in explaining enriched and background signal in parallel. In addition, adjacent regions significantly enriched for signal are merged, allowing identification of both broad and short regions of activity. He provided a demonstration of how these factors play different roles across different data types, and showed how incorporating these factors into the modeling framework can lead to improved performance in the determination of biologically relevant loci. This method represents a significant shift away from earlier methods of peak calling/peak identification to a more flexible and unified modeling framework, applicable to many types of experimental situations.

Elodie Portales-Casamar discussed deciphering regulatory networks by transcription factor binding site analysis. She provided an introduction to regulation of gene expression, which can happen at multiple levels. These include chromatin modifications, initiation of transcription at gene promoters, alternative splicing and stability of RNA, protein modifications. The binding of transcription factors (TFs) to DNA sequences near or within genes is one of the primary mechanisms directing gene transcription. Understanding the interplay between TFs and their target genes is key to deciphering cellular regulatory networks that generate diverse types of cells and tissues within an organism.

She gave an overview of many of the common computational approaches to TF binding site analysis. Sets of known binding sequences are compared to construct TF binding models. Such necessary information is collected and disseminated through community-driven resources like PAZAR, a public database of transcription factor and regulatory sequence annotation, and the high-quality transcription factor binding profile database JASPAR. However, the compiled data still remains too sparse to cover the full spectrum of DNA-binding proteins. Genome-wide chromatin immunoprecipitation techniques (e.g. ChIP-Seq) are now providing larger data collections that allow for more accurate models and increase the quality of genome annotation. Such methods enable researchers to decipher entire regulatory networks in specific cellular contexts. The example included in the talk was for ChIP-Seq data analysis of the upregulation of detoxification

systems by the Nrf2 transcription factor in cells exposed to stress.

Sunduz Keles presented her work on the development of MOSAiCS: Model-based One & Two Sample Analysis and Inference for ChIP-Seq data. This model addresses a range of problems, from multi-reads to background adjustment to peak calling. She discussed various statistical aspects of ChIP-Seq data analysis, including handling of multi-reads and developing background models that adjust for apparent sources of biases due to ChIP-Seq experimental protocol.

The particular focus was on data from a naked DNA sequencing experiment, which sequences non-cross-linked DNA after deproteinizing and shearing, to understand factors affecting background distribution of data generated in a ChIP-Seq experiment. She outlined a background model that accounts for the observed sources of biases such as mappability and GC content. She then presented MOSAiCS, a flexible mixture modeling approach for detecting peaks in ChIP-Seq data. This model incorporates the background component derived from n aked DNA experiments and introduces a flexible model for the actual signal component. This model fits actual ChIP-Seq data very well, and also has the important practical advantage that one-sample analysis of ChIP-Seq data with MOSAiCS performs as well as the two-sample ChIP-Seq data analysis that utilizes sequenced naked DNA as control. A further extension of this model was developed for two-sample ChIP-Seq data analysis with Input DNA control.

Ting Wang presented his work on mapping the human DNA methylome (regions of methylated DNA) with MeDIP-Seq and MRE-Seq technologies. These represent two complementary approaches to detect methylated and unmethylated genomic DNA. The first, methyl DNA immunoprecipitation and sequencing (MeDIP-Seq), uses antibody-based immunoprecipitation of 5-methylcytosine and sequencing to map the methylated fraction of the genome. In the second method, unmethylated CpG sites are identified by sequencing size-selected fragments from parallel DNA digestions with the methyl-sensitive restriction enzymes (MRE-Seq). Using these technologies, he as able to generate data providing a genome-wide, high-resolution methylome map of human brain tissue, and a second map of human ES cell H1. These maps on average interrogate close to 90% of all CpGs (25 million of 28 million total) and 98% of CpG islands in the human genome, at the modest expense of relatively small amount specimen and a few lanes of Illumina flowcell.

The role of DNA methylation in gene bodies was investigated with these methylome maps. From high-resolution coverage of CpG islands, the majority of methylated CpG islands were revealed to be in intragenic and intergenic regions, while less than 3% of CpG islands in $5\prime$ promoters were methylated. The CpG islands in all three locations overlapped with RNA markers of transcription initiation, and unmethylated CpG islands also overlapped significantly with trimethylation of H3K4, a histone mark enriched at active promoters. The general and CpG-island-specific patterns of methylation are conserved in mouse tissues. These and other results support a major role for intragenic methylation in regulating cell context-specific alternative promoters in gene bodies.

High-Throughput Sequencing

James Bullard presented an overview of a new proprietary third generation sequencing technology from Pacific Biosciences, the PacBio RS, scheduled for full commercial release this year. The focus of this talk was on describing the types of data available to analysts, the open source software being produced by the company, and the repositories where example data can be obtained.

Margaret Taub talked about detection of single-nucleotide variants with high throughput sequencing, including current practices and pitfalls. The talke included results on one targeted re-sequencing dataset as well as some of the publicly available data from the 1000 genomes project. She explored the impacts of technical and sequence-specific properties on accurate variant detection.

Kaspar Hansen provided an overview of results from an investigation of empirical features of RNA-Seq data, including methods for examining base-level effects and measuring goodness-of-fit of read count models. He also presented some graphics that explore detection as a function of annotation and additional exploratory analyses of RNA-Seq data.

Wolfgang Huber presented methodology and software for differential expression analysis of sequence count data. High-throughput nucleotide sequencing provides quantitative readouts in assays for RNA expression (RNA-Seq), protein-DNA binding (ChIP-Seq) or cell counting (barcode sequencing). Statistical inference of differential signal in such data requires estimation of their variability throughout the dynamic range, which is typically much larger than the dynamic range of microarrays. When the number of replicates is small, error modeling can be used to achieve greater statistical power. He proposed an error model that uses the negative binomial distribution, with variance and mean linked by local regression, to model the null distribution of the count data. The method controls Type I error and is shown to provide good detection power for detection of differentially expressed genes. A free open-source R software package, DESeq, is available from the Bioconductor project.

High-Throughput Biological Assays

Laurent Jacob discussed obtaining higher statistical power for identification of differentially expressed pathways using known gene networks. The problem of identifying sets of genes which are differentially expressed between two clinical groups is cast as a multivariate two-sample test. Under the assumption that the shift of expression is coherent with a known network structure, he has shown that integrating this structure in the test statistic leads to more powerful tests. He also discussed systematic testing of all sub-networks of a large network for de novo pathway identification. The behaviour of this new approach was illustrated on both synthetic data and on a breast cancer hormone therapy resistance expression dataset.

Jean-Philippe Vert talked about including prior knowledge in shrinkage classifiers for genomic data. Estimating predictive models from high-dimensional and structured genomic data, such as gene expression of comparative genomic hybridization (CGH) data, measured on a small number of samples is one of the most challenging statistical problems raised by current needs in post-genomics. Popular tools in the fields of statistics and machine learning to address this issue are shrinkage estimators, which minimize an empirical risk regularized by a penalty term, and which include for example support vector machines or the LASSO. He discussed new penalty functions for shrinkage estimators, including generalizations of the LASSO which lead to particular sparsity patterns, and which can be seen as a way to include problem-specific prior information in the estimator. Several examples illustrating the approach were included, such as the classification of gene expression data using gene networks as prior knowledge, and the classification and detection of frequent breakpoints in CGH profiles.

Pierre Neuvial presented work on targeted maximum likelihood estimation of the relationship between copy number and gene expression in cancer studies, a specific type of data integration problem. Identification of genes whose DNA copy number is "associated" with their expression level in a cancer study can help pinpoint candidates implied in the disease and improve understanding of its molecular bases. DNA methylation is an important player to account for in this setting, as it can down-regulate gene expression. He developed a method based on Targeted Maximum Likelihood to quantify the relationship between copy number and expression, accounting for DNA methylation. He explained the method and its statistical properties. Some preliminary results were shown from a simulation study as well as from a real data set from the Cancer Genome Atlas project (TCGA).

Robert Scharpf discussed his work on a multilevel model to address batch effects in copy number estimation for high-throughput SNP arrays. Submicroscopic changes in chromosomal DNA copy number dosage are common and have been implicated in many heritable diseases and cancers. Recent high-throughput technologies have a resolution that permits the detection of segmental changes in DNA copy number that span thousands of basepairs across the genome. Genome-wide association studies may simultaneously screen for copy number-phenotype and SNP-phenotype associations as part of the analytic strategy. However, genome-wide array analyses are particularly susceptible to batch effects as the logistics of preparing DNA and processing thousands of arrays often involves multiple laboratories and technicians, or changes over calendar time to the reagents and laboratory equipment. Failure to adjust for batch effects can lead to incorrect in-

ference and requires inefficient post-hoc quality control procedures that exclude regions that are associated with batch. His work extends previous model-based approaches for copy number estimation by explicitly modeling batch effects and using shrinkage to improve locus-specific estimates of copy number uncertainty. Key features of this approach include the use of diallelic genotype calls from experimental data to estimate batch- and locus-specific parameters of background and signal without the requirement of training data.

Jared Roach presented methodology and analysis of pedigree genome sequencing data for a small family with a rare disease. Full-genome sequences of a pedigree with p individuals can be represented as a series of genotype vectors. Consider a single chromosome with n positions. A genotype, $g_{i,p}$, is an observation of two alleles at a position i for individual p. For example, $g_{23145140,3}$ may be $\{A, C\}$. A genotype vector, $V_i = \{g_{i,1}, g_{i,2}, g_{i,3}, \ldots, g_{i,p}\}$, is an ordered list of genotypes for all individuals in the pedigree. The series of genotype vectors for a chromosome is thus $\{V_1, V_2, V_3, \ldots, V_n\}$. Binary inheritance vectors represent the not-directly-observed flow of alleles through the pedigree, and are parallel in structure to genotype vectors. These series of vectors can be regarded as emissions from Hidden Markov Models (HMMs) and illuminate underlying genetic features. He analyzed the whole-genome sequences of a family of four. HMMs enabled the precise identification of recombination sites and 70% of the sequencing errors. These analyses permit matching inheritance states and inheritance modes, and thus disease-gene identification.

Lei Sun presented work on a practical solution to the "winner's curse" in genome-wide scans. In genome-wide scans, the most significant variants detected in the original discovery study tend to have inflated effect size estimates due to the "winner's curse" phenomenon. The winner's curse has recently gained much attention in Genome-Wide Association Studies (GWAS), because it has been recognized as one of the major contributing factors to the failure of attempted replication studies. For example, five *Nature Genetics* publications in the first three months of 2009 acknowledged the effect of winner's curse in their discovery samples. However, none made statistical adjustments to the naive estimates.

Previous work (Sun and Bull, 2005) developed in the context of genome-wide linkage analyses has been extended to provide Bias-Reduced estimates via Bootstrap Re-sampling (BR-squared) for GWAS without collecting additional data. In contrast to the likelihood-based approaches, the proposed method adjusts for the effects of selection due to both stringent genome-wide thresholds and ranking of the association statistics over the genome. In addition, this method explicitly accounts for the effect of allele frequency because the expected bias is inversely related to power of the association test.

The method has been implemented to provide Bias-Reduced estimates via Bootstrap Re-sampling (BR-squared) for association studies of both disease status and quantitative traits, and applied in genome-wide association studies of Psoriasis and HbA1c. There is a greater than 50% reduction in the genetic-effect-size estimation for many associated SNPs, which translates into a greater than 4-fold increase in sample size requirements for replication studies. Thus, adjusting for the effects of the winner's curse is crucial for interpreting findings from genome-wide scans, and in planning replication studies, as well as attempts to translate findings into the clinical setting.

Mark Segal gave a talk on genomic applications of clustering with exclusion zones. Methods for formally evaluating the clustering of events in space or time, notably the scan statistic, have been richly developed and widely applied. In order to utilize the scan statistic and related approaches it is necessary to know the extent of the spatial or temporal domains wherein the events arise. Implicit in their usage is that these domains have no "holes" (or *exclusion zones*), regions in which events *a priori* cannot occur. However, in many contexts, this requirement is not met. When the exclusion zones are known it is straightforward to correct the scan statistic for their occurrence by simply adjusting the extent of the domain. He has tackled the more ambitious objective of formally evaluating clustering in the presence of *unknown* exclusion zones. By examining the behavior of *clumps* over the grid of putative cluster counts and lengths, he showed that the existence of exclusion zones manifests as a characteristic signature. This patterning is exploited to develop an algorithm for estimating total exclusion zone extent, the parameter needed to correct scan statistic based inference, with performance of the algorithm assessed via simulation study. Applications to genomic settings for differing marker (event) types are shown – *binding sites*, *housekeeping genes*, and *microRNAs* – wherein exclusion zones can arise through a variety of mechanisms. In several instances there are dramatic changes to unadjusted inference that does not accommodate exclusions.

Ingo Ruczinski discussed SNP association studies with case-parent trios. While at present most high-throughput SNP association studies are population-based, family-based designs also have some very attractive features. Case-parent trio designs in particular allow for the assessment of de-novo copy number variants, parent-of-origin effects, and transmission distortion. He discussed and demonstrated these via a genome-wide and a candidate gene association study that employ case-parent trios. The logic is also extended to regression methodology, originally developed for cohort and case-control studies, to detect SNP-SNP and SNP-environment interactions in studies of trios with affected probands. An efficient algorithm is derived to simulate case-parent trios where genetic risk is determined via epistatic interactions.

Houston Gilbert provided some results from a cross-platform evaluation study of reverse-phase protein microarray data. Reverse-phase protein microarrays (RPPMA) allow for the simultaneous detection of a single protein in complex analyte mixtures, such as those obtained from cell tissue culture or clinical sample protein lysate. To gain a better understanding of the RPPMA arena, he worked on an evaluation of three fee-for-service providers of this technology. Practical, statistical and biological results from the evaluation study have informed strategies for moving forward with RPPMA technology in research and development programs. The evaluation study has also highlighted areas for each of the companies to improve upon their own platforms.

From the Bench to the Clinic

Adam Olshen presented an overview of two projects involving high throughput data. One is more mature and concerns distinguishing primary tumors from metastases utilizing copy number data. The methodology was demonstrated on a lung cancer data set. The second is a work-in-progress involving methylation sequencing data. He discussed integrating multiple types of such data as well as methods for estimating copy number from it.

Mauro Delorenzi talked about translational studies for predictive and prognostic biomarkers in colon cancer, including of microsatellite instability by expression profiling in a clinical trial. Microsatellite instability (MSI) is the hallmark of a deficient mismatch repair system (MMR) in about 15% of colorectal cancers (CRC). Several studies confirm MSI as an independent prognostic marker associated with a better outcome in stage II and IIICRC. As MSI can be caused by different mechanisms, and dMMR leads to secondary oncogenic alterations, heterogeneity in the clinical and molecular features of MSI CRC is likely but not well understood. In this transcriptome expression study, his group explored which genes differentiate MSI from Microsatellite stable (MSS) tumors and looked for evidence of additional subclasses.

RNA extracted from formalin-fixed paraffin-embedded (FFPE) tissue blocks was used for expression profiling on the ALMAC platform SColorectal Cancer DSA Y T. Classifiers were constructed using AdaBoost and DLDA algorithms and assessed with area under the curve (AUC) by cross-validation. Survival analysis was based on Cox regression. Unsupervised methods allowed only weak separation of MSI and MSS specimen, despite 494 genes with significantly different expression (1% FDR), due to high variation in both groups. Gene expression differences were in agreement with results from reanalysis of three public datasets. Classifiers discriminated MSI and MSS with AUC of 0.96, using 40-80 selected genes. Prominent discriminatory genes include various pathways: Wnt (e.g. Axin2), MAPK (DUSP4); inflammation-immunity (REGs, STAT1), differentiation (TNNC2,mucins), metallothioneins. Association of these genes with RFS is heterogeneous.

Efficient discrimination of MSS and MSI in gene expression profiles can be obtained, with good quality FFPE material, using a multi-gene classifier. Inside both classes there is high residual heterogeneity. More samples are planned to be profiled in order to further define molecular subgroups and to search for prognostic genes. The ability to obtain reliable profiles form FFPE material implies that relevant information can be obtained from archival material stored in many biobanks.

Pete Haverty discussed the mutation spectrum revealed by paired genome sequences from a lung cancer patient. Previous studies have identified important common somatic mutations in lung cancers, but they

have focused primarily on a limited set of genes and have thus provided a constrained view of the mutational spectrum. He presented results from the complete sequences of a primary lung tumour (60x coverage) and adjacent normal tissue (46x coverage). Comparing the two genomes, a wide variety of somatic variations were identified, including >50,000 high-confidence single nucleotide variants. 530 somatic single nucleotide variants in this tumour were validated, including one in the KRAS proto-oncogene and 391 others in coding regions, as well as 43 large-scale structural variations. These constitute a large set of new somatic mutations and yield an estimated 17.7 per megabase genome-wide somatic mutation rate. Notably, there is a distinct pattern of selection against mutations within expressed genes compared to non-expressed genes and in promoter regions up to 5 kilobases upstream of all protein-coding genes. Furthermore, a higher rate of amino acid-changing mutations is observed in kinase genes. He presented a comprehensive view of somatic alterations in a single lung tumour, and provided evidence of distinct selective pressures present within the tumour environment.

Donald Geman talked about several projects in expression-based biomarker discovery and pathway regulation, mainly focused on cancer. The driving application is translational medicine. He argued that rank-based statistics can account for combinatorial interactions among genes and gene products; accommodate variations in data normalization and limited sample sizes; and avoid the "black box" representations and decision rules generated by standard methods in computational learning. Applications using single or pairs of top-ranked genes were presented for several cancer studies.

Predictive Diagnostics and Designing Clinical Trials

Jane Fridlyand gave a talk on the design of proof of concept trials in oncology, focusing on speed, cost and trial success. With the cost of bringing a drug to market in the range of nearly a billion dollars, pharmaceutical companies are concerned with achieving proof of concept as early as possible in the clinical development process, so that decisions on further development can be made with minimal loss. The bulk of the talk was devoted to the explanation of industry constraints to academic researchers focused on prediction optimality.

Ru-Fang Yeh discussed some of the statistical challenges in the development of predictive biomarkers. It has been increasingly important to incorporate diagnostics in the drug development process to improve response to treatment and help Several statistical issues arise during the development of predictive biomarkers that aim to identify patients who will benefit from a particular treatment. Examples that highlight statistical challenges in biomarker discovery and clinical applications were presented, including threshold selection for continuous biomarkers and implmentation of complex predictors.

Venkat Seshan concluded the session with a talk on two-stage designs for gene-disease association studies. Gene-disease association studies based on case-control designs may often be used to identify candidate SNPs (markers) conferring disease risk. If a large number of markers are studied, genotyping all markers on all samples is inefficient in resource utilization. He proposed an alternative two-stage method to identify disease-susceptibility markers. In the first stage, all markers are evaluated on a fraction of the available subjects. The most promising markers are then evaluated on the remaining individuals in the second stage. This approach can be cost effective since markers unlikely to be associated with the disease can be eliminated in the first stage. He presented tables showing optimal allocations and cost savings/increases in power and efficiency.

Open Questions and Outlook for the Future

There remain several open areas of research in the domain of statistical problems in high-throughput biomedical studies; we outline some of these here.

Statistical Challenges of New High-Throughput Technologies

DNA sequence data are becoming more prominent in high-throughput biomedical studies, and will only gain in importance as sequencing technologies become more reliable and less expensive. The aim of sequencing is to determine the order of the nucleotide bases in a DNA molecule, even up to an entire genome. This knowledge can be used in a wide variety of important applications such as identification of new genes and alternative gene splicing, mutation mapping, polymorphism discovery, DNA-protein interaction, and personalized medicine.

The number of applications of new sequencing technologies is large and growing. Sequencing can be used to provide genome-wide measures of: transcription levels (mRNA-Chip/mRNA-Seq), alternative splicing, protein-nucleic acid interactions, e.g., transcription factor, binding sites (ChIP-Chip/ChIP-Seq), DNA methylation (methyl-Chip/methyl-Seq), DNA copy numbers (aCGH), genotypes, etc.

With the new sequencing technologies come new statistical issues at nearly every step of the experimental pipeline: experimental design; exploratory data analysis and quality assessment/control; pre-processing steps such as image analysis, base-calling, read-alignment/mapping, normalization and expression quantitation; downstream analyses such as identification of differential expression; and the overarching issue of reliable software implementing the best developed methodologies.

Many of the methods developed for analyzing microarray data do not carry over to sequence data. There are new issues: different technology-dependent biases, and at an even more fundamental level the data have a completely different quality (continuous measures of fluorescence intensity for microarray data versus discrete counts for sequence data).

In the realm of experimental design, some of the new problems are: choice of sequencing depth (sample size – number of input samples/lanes); allocation of input samples to library preparations/flow-cells/lanes; control lane for calibration in base-calling; type of read (Strandedness, length, single-end, paired-end, or strobed reads); library preparation (priming, fragmentation protocols).

As with microarray experiments, low-level analysis/pre-processing is required for any type of study and is highly-dependent on the sequencing platform. Preprocessing steps include image analysis, base-calling, and read mapping/alignment. Unlike microarray probe sequences, sequence clusters do not lie on a grid, complicating image analysis. The base-calling relies on a deconvolution of the base sequence from measured fluorescence intensities of the four nucleotides at each cycle, to yield base-level and read-level quality scores. Challenges include existence of cross-talk and machine cycle effect. The resulting reads must be assigned to positions in the genome, transcriptome, or other reference sequence.

Even older algorithms developed for earlier generations of sequencing data cannot be used “as is”. Existing algorithms for alignment (read mapping) and sequence assembly of older (Sanger) sequence data cannot be used as-is, because of the much shorter reads generated by pyrosequencing (a few hundred nucleotides) – the algorithms do not scale to the increased data volume, and with shorter fragments there are in addition combinatorial complications. However, users are adopting the newer technologies due to their lower cost and higher throughput compared to Sanger sequencing. There is thus a need for new statistical models/frameworks and modified or novel scalable algorithms for processing and analyzing the vast amounts of deep sequencing data generated in a study.

Given base-level read counts, we need to derive expression measures for genomic regions of interest (e.g. exon, intron, splice junction, single-isoform gene, multiple isoforms from a given gene). Normalization requires adjustment of raw expression measures to ensure that observed differences in expression measures between lanes and/or between regions of interest are truly due to differential expression and not experimental artifacts (e.g., library preparation/flow-cell/lane effects, nucleotide composition). Other challenges include zero counts, heterogeneity of base-level counts, uncertainties in annotation, and alternative splicing.

Further investigation into a statistical framework is also needed. Of particular interest is the use of generalized linear models (GLM) and especially Poisson/log-linear regression, to evaluate and adjust for a variety of technical effects and to detect differential expression between regions of interest and/or input samples. Assessment of method performance on calibration/benchmark data sets, as has been done for microarrays, would provide researchers with valuable information on the reliability of existing and new methods.

Integration of High-Dimensional Heterogeneous Data

Meta-analytic methods have been applied in the genomic context for combining study results, often one gene at a time via estimated regression coefficients or p-values. Such cases typically combine results for the same data type (e.g. gene expression data), but perhaps generated by different technologies (e.g. single channel or dual channel microarrays). It seems clear, though, that meta-analysis is not straightforwardly applied to the problem of combining data of different types, the most obvious impediment being lack of a common parameter across data types. It may also be desired to combine data types that are not gene-based (e.g. gene expression and glycomic data). Approaches other than meta-analysis include Bayesian models as well as correlation-based, kernel, svd-type or distance-based methods. However, integrating multiple data types in an automated, quantitative manner remains a major challenge.

Moving beyond the single gene at a time framework would also be valuable. For example, it would be quite useful to be able to integrate data on sets of genes (rather than only individual genes), or for multi-dimensional/multi-type molecular "signataures" in human disease, such as cancer. Development of a comprehensive catalog of signatures for different disease processes would spur method development, leading to the potential to elucidate molecular mechanisms important in disease pathogenesis and progression.

Translational Research

Translation of genome findings in complex disease is a complex and challenging aim. Genomic discoveries for monogenic diseases have led to clinical tests but there are fewer applications for complex diseases. Assessing the validity and utility of genomic tests for specific diseases can be difficult.

There is substantial scope for statistical thinking and methodology to help achieve translational aims. There is already a vast body of basic research that needs to be harnessed in a reliable and efficient manner so that the important findings impact on treatment options to improve human health. With a massive expenditure of resources already consumed, it is vital to take advantage of the existing Genome Wide Association Studies (GWAS) by exploiting previously generated initial GWAS scans. Problems still exist in addressing the replication and continuation or combination of GWAS findings, and epidemiologic data must also be integrated. Biological studies carried out as complementary studies must also be integrated in a fundamentally sound way.

It would seem that, at least initially, very specific models will need to be created, reflecting the conditions of a particular set of studies (or study types). A major hope is that a flexible, encompassing framework will emerge that will facilitate translation of basic findings to clinical relevance.

A medical topic of increasing influence is the importance of *structural variation in human diseases* such as cancer and HIV/AIDS. Until now, most genome-wide association studies have focused only on SNPs, or single nucleotide polymorphisms. However, larger structural variation (for example, copy number variation, or CNV) is also highly abundant and is increasingly recognized as a substantial contributor to disease/phenotype variation. Deep sequencing technologies can be used to detect structural variation, allowing for systematic study on a whole genome level of genetic alterations. Such studies should lead to greater understanding of fundamental disease processes, and result in important implications in the prevention, detection, diagnosis, prognosis, and treatment of cancer and of HIV/AIDS.

List of Participants

Bryan, Jennifer, University of British Columbia
Bullard, James, University of California, Berkeley
Delorenzi, Mauro, Swiss Institute of Bioinformatics
Dudoit, Sandrine,University of California, Berkeley
Eng, Kevin, University of Wisconsin-Madison
Erwin, Genevieve, UCSF/Gladstone Institute of Cardiovascular Disease
Fridlyand, Jane, Genentech, Inc.
Gilad, Yoav, University of Chicago
Gilbert, Houston, Genentech, Inc.
Goldstein, Darlene, Ecole Polytechnique Federale de Lausanne

Hansen, Kasper, University of John Hopkins
Haverty, Peter, Genentech, Inc.
Holloway, Alisha, Gladstone Institutes, University of California, San Francisco
Huber, Wolfgang, EMBL
Jacob, Laurent, University of California, Berkeley
Keles, Sunduz, University of Wisconsin, Madison
Leek, Jeff, Johns Hopkins Bloomberg School of Public Health
Lieb, Jason, University of North Carolina, Chapel Hill
Melsted, Pall, University of Chicago
Neuvial, Pierre, University of California, Berkeley
Ngai, John, University of California Berkeley
Olshen, Adam, University of California, San Francisco
Pollard, Katherine, Gladstone Institutes, University of California, San Francisco
Portales-Casmar, Elodie, University of British Columbia
Pritchard, Jonathan, University of Chicago
Roach, Jared, Institute for Systems Biology
Ruczinski, Ingo, Johns Hopkins University
Scharpf, Rob, Johns Hopkins University
Segal, Mark, University of California, San Francisco
Seshan, Venkatraman, Memorial Sloan-Kettering Cancer Center
Sun, Lei, University of Toronto
Taub, Margaret, Johns Hopkins University
Vert, Jean-Philippe, Mines ParisTech
Wall, Jeff, University of California San Francisco
Wang, Ting, Washington University, St. Louis
Wirapati, Pratyaksha, Swiss Institute of Bioinformatics
Yeh, Ru-Fang, Genentech, Inc.