PROPOSED WORKSHOP "STATISTICAL DATA INTEGRATION CHALLENGES IN COMPUTATIONAL BIOLOGY: REGULATORY NETWORKS PERSONALIZED MEDICINE"

Objectives

The primary objectives of this workshop are;

- to identify statistical challenges that arise in joint, integrated analyses of biomedical and high-throughput genomic data;
- to present and critique solutions to these problems, with the goal of determining the strategies that will be most effective for yielding significant, reproducible biological discoveries;
- to bring together two communities that are strongly committed to tackling the data integration problem: the biologists generating the data and the statisticians developing analytical frameworks.

In recent years, several large-scale international genomics projects have been launched. They include The Cancer Genome Atlas (TCGA) consortium, the ENcyclopedia Of DNA Elements (ENCODE) and modENCODE (model organism) projects, and the 1000 Human Genomes project. These initiatives have now reached a peak in data collection, where massive amounts of high-quality clinical and genome-scale data have been made available to the research community. In addition, there exist several easily accessible WWW databases containing curated biological knowledge, such as gene functional annotation and gene-protein interactions, e.g. the Gene Ontology (GO) consortium, the Kyoto Encyclopedia of Genes and Genomes (KEGG), the Human Protein Reference Database (HPRD), and OncoMine, to name a few.

Joint integrative analysis of these rich data resources remains very difficult. The obstacles range from the technical (e.g. differences in the nomenclature used to uniquely identify a gene) to the conceptual (e.g. the scarcity of statistical methods that can reliably and efficiently accommodate highly disparate data types). In the proposed workshop, we intend to address topics across this wide spectrum. To ensure that the analytical tools are conceptually sound and produce results that are biologically interesting and experimentally verifiable, a collaborative approach is essential. The workshop is thus designed to foster deeper connections between "wet-lab" and "dry-lab" researchers and to be a forum for (1) the dissemination of cutting-edge developments, including new high-throughput biological assays and novel statistical methodologies, and (2) the identification of open problems in the analysis of these diverse data.

The main topic for the workshop is data integration and within that we target two specific application areas: the identification of regulatory networks through multiple data sources and the integration of clinical and genome-level data for personalized medicine. Finally, we place a special emphasis on studying data analysis strategies that promote reproducible research. We anticipate that a combination of seminar, poster, and brainstorming sessions, in which life scientists and quantitative scientists participate together

to flesh out the major issues and roadblocks, will help clear the way for advances in these targeted areas.

Relevance, importance, and timeliness

Relevance: High-throughput biology has changed both genomics and statistics research. Statistics has become an essential component in genomics research and analytical challenges from genomics have given rise to methodological advances. We are now facing a new challenge: integration of clinical data with multiple high-throughput and diverse data sources. Addressing the new statistical demands of data integration has considerable relevance for continued progress in biological research and presents exciting opportunities for further significant methodological advancement.

Importance: With the recent peak in data collection by the U.S. National Institutes of Health initiatives and international consortia (e.g. TCGA, 1000 Genomes, ENCODE), integration and statistical analysis of diverse data sources has become a bottleneck for progress. Yet, it is understood that these data contain information that can greatly improve our understanding of complex biological processes, as well as disease causes, prognoses, and treatments. However, it is no simple task to identify, extract, and integrate the relevant sets of data for a particular biological question. The aim of this workshop is for biologists and statisticians to work together to identify and formulate problems in data integration with the aim of deliverables that give biological insight and testable results.

Timeliness: While there are a number of conferences specializing in analytical challenges that arise in computational biology (e.g. ISMB and RECOMB), the primary quantitative discipline has been computer science rather than statistical science. These meetings are typically more focused on databases, algorithms, or software, rather than statistical modeling. Conversely, there are computational biology sessions at statistical conferences, but the number is small and these conferences are attended almost exclusively by statisticians. Opportunities for truly interdisciplinary interaction between statisticians and biologists are infrequent but crucially important.

Our aim is to organize an interdisciplinary workshop that addresses the enormous challenges of statistical data integration in biological research. Our goal is to focus on the statistical and computational methods development without losing sight of the underlying biology. Such a workshop would be an important development in the field and BIRS provides an ideal environment for such an endeavor.

Subject area overview

Recent large-scale initiatives, such as the TCGA consortium, the ENCODE and modENCODE projects, and the 1000 Human Genomes project, have produced massive amounts of diverse biological data. Examples of data sources include: genomic data, such as SNPs (Single Nucleotide Polymorphism) and CNV (Copy Number Variation); transcription data, such as mRNA and miRNA expression; and interaction data, such as information on transcription factor binding sites. These type of data come from microarray or, more recently, from next-generation sequencing technologies (e.g. Illumina, Pacific Biosciences).

The ENCODE project aims at identifying all functional elements in genome sequences. Its pilot phase started in 2003 and focused on 1% of the human genome. Since then, ENCODE has extended to model organisms such as mouse (mouse-ENCODE), fruitfly, and worm (modENCODE).

ENCODE and modENCODE generate genomewide binding profiles of multiple transcription factors, histone modifications, DNaseI sensitivity, DNA methylation, copy number variation, and gene expression utilizing sequencing and array-based technologies.

The Cancer Genome Atlas was launched in 2006. It is now a repository for thousands of tumor samples from some of the worst prognosis human cancers, such as glioblastoma and ovarian cancer. The goal of the TCGA is to use multiple data sources to gain insight into the molecular bases for human cancers. To this end, TCGA includes both clinical data (survival, treatment, grade of tumor) as well as genome-level data, including DNA mutation, copy number variation, alteration in methylation of the DNA, and mRNA and miRNA expression.

The 1000 Human Genomes project is the most recent of these initiatives, launched in 2008. In this massive effort, genetic variation between 2000 individuals from 20 population groups has been analyzed using next-generation sequencing technologies. In addition, the project is also a repository for the individual cell-lines, which opens up for follow-up studies on e.g. mRNA expression and drug-genome interactions.

All three projects outlined above revolve around the central theme that to understand a biological process or system we need to take a global approach and consider all its components. Below, we briefly summarize the central problem of data integration in high-throughput biology. We also describe two targeted areas where data integration is essential and on which the workshop will focus.

- *Data integration.*

  Disease phenotypes emerge from the joint effects of inherited and acquired genetic variation, as well as environmental factors. For instance, in cancer tumors it is natural to model the influence of point mutations and chromosomal aberrations on mRNA expression, and relate these effects to therapy response or patient outcome. We are now in a position where we can get data for most components of such a model. The goal is to jointly analyze the data sources to identify the disease drivers (the genetic variation), and the disease markers (expression), as well as the pathways and regulatory processes that are altered in the disease.

  Each data source provides a different view of the genome. It seems clear, though, that data integration is not very straightforwardly applied to 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.

  One common approach to data integration has therefore been to integrate *analysis results* (so-called "late data integration") rather than the data directly. For example, in analyzing mRNA and protein expression data, one can separately identify the mRNAs that are differentially expressed between two disease states, and similarly for the proteins. A biological pathway can now be ranked based on the over-representation of differentially expressed genes and proteins within it.

  However, integration of analysis results does not capitalize on the informative biological correlation between the different data types, results in data reduction (i.e. loss of information), and often involves manual steps that are not easily reproducible across labs. Thus, procedures that simultaneously integrate multiple data types into a probabilistic model are needed (so-called "early data integration"). We are in the early days of developing such

procedures, with some promising results appearing in the bioinformatics and biostatistics literature (e.g. Vaske et al., 2010; Shen et al., 2009).

Both the late and early data integration approaches will be well-represented at the workshop. One of the goals of the workshop is for life scientists and computational scientists to get an opportunity to discuss the relative merits and limitations of the above approaches to integration. In addition, we also plan to include a session on data integration as pertaining to biological knowledge, e.g. gene onthology and pathways. Similarly to the late and early integration of experimental data, biological knowledge can be used alternatively as post-processing and validation of analysis results or as prior knowledge in a biological systems model. The relative merits of these approaches will be debated at the workshop.

- *Regulatory networks in complex biological systems.*

  A current "hot" topic in systems biology research is the identification of regulatory networks and modules. Network reconstruction is also an area with considerable current statistical activity. Still, how to integrate several data sources for network modeling is a largely undeveloped area. Some efforts have been made using so-called late integration, i.e. separate estimation of network models from each data source and then weighted combination of results to produce a final network. Early integration of multiple data sources for network reconstruction has just begun to appear in the literature, e.g. modeling the mRNA concentration of a gene by a set of Ordinary Differential Equations, where synthesis and degradation rates are allowed to depend on the mRNA concentration of other genes and "perturbations" such as copy number variations.

  However, how to expand the scope of both late and early data integration to network reconstruction from several data sources remains an open problem. If we take cancer as an example, genetic variations or perturbations can be thought of as disease drivers. In contrast, the impact of these variations on expression, and through this on cell-function and disease progression, are comprised by both direct and indirect network effects of the disease drivers. By including data pertaining to both disease drivers (SNPs, CNV) and disease markers (mRNA, miRNA, protein expression), as well as clinical data (survival, treatment regime), in regulatory network construction, one aims to identify both (1) the main drivers of the disease and how they relate to disease progression, and (2) the regulatory processes that are most affected and can therefore be used as biomarkers or for disease diagnosis.

  The workshop's pre-invited speakers include both experts on statistical network reconstruction and life scientists involved in biological systems modeling, thus creating an ideal forum for discussing the challenges of identification of regulatory networks through multiple data sources.

- *Personalized medicine.*

  Above, we briefly touched on the open problem of how to include clinical information in integrative analysis. The ultimate goal is to gain sufficient understanding of how the clinical and molecular characteristics of an individual patient's disease are related and, consequently, to enable individual patient level-decisions for treatment and prognosis, i.e. personalized medicine.

  In cancer research and clinical oncology, some degree of personalized medicine has been utilized in patient management for decades. However, the variables that have informed these

early efforts have been crude (e.g. tumor's stage, grade, and anatomic location and demographic information such as patient's age, risk factors, and other basic clinical measures). In the last decade, specific molecular markers have been incorporated. They include PSA (prostate-specific antigen) for prostate cancer screening and detection and Her2/neu status for determining adjuvant therapy in breast cancer patients.

There is now a flood of new data being collected (discussed elsewhere in this proposal) that, if analyzed properly, will lead to new knowledge about clinically useful markers and the underpinnings of disease. This, in combination with new treatment modalities being developed in both the academic and pharmaceutical industry settings, has the potential to produce rapid improvements of personalized cancer care.

Two new and exciting trends currently driving personalized medicine will be explored at this workshop. The first is extending single genomic data type "signatures" to multiple genomic data types and integrative analyses. A recent example from The Cancer Genome Atlas (TCGA) is the identification of the CpG island methylator phenotype in glioblastoma (G-CIMP; Noushmehr et al., 2010). This integrated analysis of mRNA expression, methylation and clinical outcome identified a subgroup of patients with significantly longer survival. Additional work will be needed to match novel treatment strategies with this subgroup. We are in the early days of such findings, so a discussion of methods for discovery will be timely.

A second trend portends a move away from the paradigm of treatment determined by a tumor's tissue-of-origin and toward letting the mutational landscape drive treatment decisions. These genotype-directed approaches began with targeting of the "gene fusion" generated protein BCR-ABL in chronic myelogenous leukemias, the first cancer clearly linked to a specific genetic variation. Other examples include KIT mutations in gastrointestinal stromal tumors, and EGFR mutations in lung adenocarcinomas. Recently, (Palanisamy et al., 2010), identified rearrangements of the RAF kinase pathway across multiple cancers, thus suggesting an unexpected subset of patients for whom RAF inhibitors may be useful, even in tumor types that lack prototypical BRAF mutations.

These examples all highlight the possibilities inherent in a multi-faceted approach to data analysis. We have an ideal group to discuss how to identify additional cross-cancer events, as several of the participants are experts in cancer genomics. This interdisciplinary workshop is also an excellent setting for discussing the road blocks in personalized medicine, such as the design of clinical studies to validate the findings.

The proposed workshop centers on data integration in biomedical and genomic research. We are also addressing two specific focus areas; (1) the identification of regulatory modules and networks in complex biological processes and diseases and (2) personalized medicine. Our aim is to organize a workshop that addresses the interdisciplinary needs and challenges of these focus areas. **Additionally, a guiding principle of the workshop is to highlight reproducible research, a notion and practice which clearly deserve more attention and advocacy in the interdisciplinary field of statistical genomics.**

- *Reproducible research.*

  The complexity of the data analysis that underpins a "typical" genomics paper is extreme. While this complexity is hardly unique to computational biology, the cacophony that arises from the diverse disciplines and experimental platforms increases the degree of difficulty

substantially. Platform-specific standards have been developed for many assays (e.g. the MIAME standard for microarray data), but we are just beginning to conceive of similar standards for the conduct and reporting of an analysis strategy, especially considering the coming wave of publications pertaining to data integration.

There have been several highly-publicized instances where the analytical results of genomic studies have been impossible to reproduce, even with access to the underlying "raw data" and software implementations of the analysis methods. In 2004, Tibshirani was unable to replicate the results of Dave et al., who claimed to have identified two "immune response" gene clusters whose expression was predictive of survival in follicular lymphoma [see http://www-stat.stanford.edu/ tibs/FL/report/index.html for an account of Tibshirani's re-analysis of Dave et al., NEJM Nov 18, 2004 and the subsequent dialogue in NEJM]. More recently, Baggerly and Coombes have highlighted problems in a set of papers from investigators at Duke's Institute for Genome Sciences. This has led to the halt of clinical trials, an investigation by the National Academy's Institute of Medicine, and patient lawsuits. [Baggerly and Coombes, Ann. Appl. Stat. Volume 3, Number 4 (2009), 1309-1334].

In both of these cases, the lack of standards for implementing, preserving, and disseminating the analysis of genome-scale data was a primary factor. To that end, various members of the community are rallying around this issue of "reproducible research". The solutions will be multi-level, ranging from software tools and practices to top-down prescriptions relating to the release and format for raw data, code, and results. General frameworks for reproducible research have been proposed in Gentleman and Temple Lang (2004). For instance, Sweave (Leisch, 2002), one such system applicable to R and LaTeX, allows the generation of integrated, dynamic, and reproducible statistical documents, intermixing text, code, and code output (textual and graphical). The document can be automatically regenerated whenever the data, code, or documentation text change. The reproducible research system (RRS) described in Mesirov et al. (2010) is an adaptation of Microsoft Word that links to the Broad Institute's GenePattern platform. While a number of statisticians have recently adopted the Sweave system for their research projects, the notion and practice of reproducible research clearly deserves more attention and advocacy in the interdisciplinary field of statistical genomics.

This workshop presents an excellent opportunity to construct guidelines and share solutions to this thorny problem early in the development of new approaches for the integrated analysis of diverse genome-scale data types. The planned meeting includes a wide range of expertise, with both statisticians and biologists exchanging ideas. We plan to have leading statisticians and biologists give plenary talks on each focus area. In addition, selected invited speakers will present their recent efforts on data integration within one of the focus areas. **Two informal evening sessions on software and databases will be included**, which will be the venue for highlighting software and practice that promote reproducible research. Following the successful structure of previous workshops, we plan to include several poster and brainstorming sessions where researchers can exchange ideas on how best to respond to the challenge of data integration and identify open problems and limitations with the available data and models in use.

Possible participants and their affiliations

In addition to the organizers, the following pre-invited participants have expressed interest and indicated their intention to attend this workshop. Several of our pre-invited participants are either life scientists generating one or more of the data sources mentioned in this proposal or statisticians/computational scientists working on these data sources.

- Wolfgang Huber, European Molecular Biology Laboratory, Heidelberg, Germany.
- Ingo Ruczinski, Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, USA.
- Mark Segal, Epidemiology and Biostatistics, University of California, San Francisco, USA.
- Dave Stephens, Department of Mathematics and Statistics, McGill University, Montreal, Canada.
- Raphael Gottardo, Fred Hutchinson Cancer Research Center, Seattle, USA
- Benjamin Haibe Kains, Dana-Farber Cancer Institute, Harvard School of Public Health, Boston, USA.
- Keith Baggerly, The University of Texas MD Anderson Cancer Center, Houston, TX, USA.
- Jeff Barrett, Sanger
- Carlos Bustamante, Stanford
- Goncalo Abecasis, Michigan
- David Haussler, UC Santa Cruz
- Peter Campbell, Sanger
- Nancy Zhang, Stanford
- Jeff Kidd, Michigan
- Terry Speed, Wehi
- Chad Creighton, Baylor
- Celia Greenwood, Toronto
- X. Shirley Liu, Harvard

Additional comments

The organizing committee has been selected to best represent the three targeted areas of the workshop as well as both communities that are strongly committed to tackling the data integration problem: the biologists generating the data and the statisticians developing analytical frameworks. Several of our members also have experience in organizing fruitful workshops in the areas of statistics and statistical genomics.

- Jenny Bryan is an Associate Professor at the University of British Columbia in the Michael Smith Laboratories and Statistics Department. She specializes in the analysis of high-throughput phenotypic data, such as the datasets being generated through whole genome RNAi or in large panels of knockouts. She has co-organized five previous successful Statistical Genomics meetings at BIRS in 2004, 2006, 2007 (2-day), 2008, and 2010. These workshops have brought together statisticians, biologists, and clinicians working on different aspects of genome-scale studies. They have been hugely successful, covering several broad areas of biological investigation that relied on statistical and computational methods.
- Aurelie Labbe is an assistant professor at McGill University in the department of Epidemiology, Biostatistics and Occupational Health. She works on methodological issue in data

integration and regulatory networks through the identification of expression quantitative trait loci. She has organized the very successful international workshop on "Computational statistical methods for genomics and system biology" that was held in Montreal in April 2011. This workshop brought together more than 100 participants from Europe and North America, as well as 25 invited speakers internationally recognized as leaders in the field.

- Stephen Montgomery
- Adam Olshen
- Ronglai Shen
- Paul Spellman

We would like to continue to build on our established track record of these extremely successful workshops, while adapting to address the emerging challenge of data integration in systems biology research.

Dates

Our preferred week

Preferred dates.

- June 3-8, 2012.
- August 13 - 18
- August 20 - 24
- July 23 - 27

Off-season dates.

- no strong preferences here?
-

Impossible dates.

- **March 10-13, 2013 (ENAR)**
- **??? (ICSA/Applied Statistics Symposium)**
- **August 3 - 8, 2013 (IMS/Joint Statistical Meeting)**
-