PROPOSED WORKSHOP "***EMERGING STATISTICAL CHALLENGES IN GENOME-SCALE DATA ANALYSIS AND TRANSLATIONAL RESEARCH"

Objectives

The primary objectives of this workshop are

- to formulate and address emerging statistical problems in the analysis and combination of diverse types of biomedical and high-throughput genome-scale data;
- to facilitate meaningful interactions between the experimental biologists and research-oriented clinicans producing genome-scale data and the statisticians who will be addressing the issues that analysis of these data entails. Substantive collaborations between these groups are vital for transforming the massive amount of data produced by new technologies into important biological discoveries and translational research leading to the change of paradigm in patient management.

A secondary aim is to honor and celebrate the achievements and ongoing contributions to this field of Professor Terry Speed, who turns 65 in 2008. We believe that an appropriate recognition is to carry forward his first-rate example of forging productive statistical-biological hand-on collaborations, and that this workshop provides an effective means of doing so.

The workshop is intended to foster deeper connections between the biological and statistical research communities 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 data. The challenges include not only analyzing genotypic, gene expression and protein expression data, but also relating these to phenotypic data, such as biological and clinical outcomes, and further relating all of these to meta-data from WWW databases such as OncoMine, KEGG, GoldenPath, PubMed, the Gene Ontology (GO) Consortium, and the Pharmacogenetics and Pharmacogenomics Knowledge Base (PharmGKB).

We anticipate that this workshop will enable statisticians to articulate theoretically grounded statistical formulations of existing and emerging computational biological and clinical problems; create an exceptional opportunity for exchanging ideas between the communities; and help to shape the future of this dynamic field. Input from biologists and clinicans is absolutely crucial for identification of important questions and development of appropriate statistical methodologies. For this reason, we target areas that are relatively new to statisticians, as well as areas that have already been greatly influenced by statistical approaches.

The five targeted areas for the workshop include: patient classification, computational population genetics, pharmacogenomics, emerging technologies and data integration. We expect that the interaction between statisticians and biologists will lead to major advances in the analysis and integration of genome-scale data sets and in translational research. In addition, this relatively new and rapidly developing field enjoys an above-average representation of both young researchers and women.

Relevance, importance and timeliness

Relevance: It is now well accepted that the capacity to generate genome data has far outpaced the ability to analyze and interpret it. The rapid development of new high-throughput technologies allows investigation of biological processes on an ever-growing scale. Statistical genomics has adapted well to these changes, due to the great interest of statisticians in the methodological challenges inherent in a quickly evolving domain. Addressing the new statistical demands has much relevance for continued progress in biological and biomedical research predicated on genome-scale assays.

Importance: Genome-scale data are rising in prominence, and are rapidly becoming critical components of clinical import in human disease, most strikingly in cancer. Yet without sound methodology, accompanied by computationally feasible implementations, we risk missing, or misinterpreting, important information contained in these data. This workshop will help to enable transformation of the vast data resources emanating from multiple, diverse types of high-throughput assays into realizable health benefits, including: improved diagnostics, prognostics, risk assessments and treatments.

Timeliness: There are several well-established computational biology conferences (e.g., ISMB and RECOMB), where the primary quantitative discipline has historically been computer science, not statistical science. These meetings are often focused more narrowly on databases, algorithmic aspects, or specific software, and contain a rather weaker interdisciplinary component. Likewise, even though major statistical conferences often have sessions on computational biology, the number of those is still small, and the audience is almost exclusively statisticians. Opportunities for true interdisciplinary interaction are few. Yet in this field, rapid communication between biologists and statisticians is absolutely vital.

Our aim is to organize a more fully interdisciplinary workshop to address the challenges posed by the enormous need for quantitative data integration and modeling in biology. Although the field is very broad, our intent is to focus on the statistical, mathematical, and computing aspects without losing sight of the underlying biology. To this end, biologists have an intrinsic role to fulfill.

Achievement of this goal requires input from experts across the relevant scientific fields who have a broad vision for the global aims in advancing quantitative methods for genome-scale data. A workshop that specifically brings biologists and statisticians together would certainly be an important development in the field, and BIRS provides an unbeatable environment for this task.

Subject area overview

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. Cooperation among statistics, mathematics, computer science and biology is essential to the further advancement of both basic genome biology and high-level clinical applications of this new knowledge.

The pace at which new technologies and data acquisition methods emerge makes computational and statistical genomics an extremely dynamic field. It is our goal to bring wet-lab biologists and research-oriented clinicians with interest in computational biology and statisticians working in several aspects of statistical genomics together in this workshop. This would serve as a great opportunity (1) to summarize new advances in biological technologies and state of the art statistical methodologies addressing relevant challenges, (2) to criticize and discuss limitations of the existing methodologies and formulations, (3) to explore ways to solve these issues, and (4) to discuss areas where more interaction among the two communities is needed.

We list below several areas of biological investigation that are fueled by technological advances and require rigorous statistical and computational analysis. There are no strict borders between topics, since most share high dimensional multivariate data that are similar in nature, and biological discoveries are often achieved through merging of various sources of data and perspectives. The workshop will focus around these five topics. For each topic, related statistical aspects such as parameter specification, estimation, inference and testing, model selection, and statistical computing issues will be addressed. We aim to have at least one well-known plenary biologist and one statistician to speak on each topic. Aside from regular talks, poster and software demonstration sessions will provide researchers opportunities to present current applications and results on these topics.

We have pre-invited some of the possible speakers from both communities including

should we specify the main area of work for each?* ***how many we should pre-invite and what if they don't reply?*** **we will need to cut ... since i am adding clinicians/large scale biologists for the translational part of the workshop, we will need to cut down on basic biologists*** ***do we need to focus on BIG names for pre-invitations (i think so) (and move smaller names to the list below or it does not matter?***

Mauro Delorenzi (Swiss Experimental Cancer Research Institute/Swiss Institute of Bioinformatics), Wolfgang Huber (EBI-EMBL) Simon Tavare (USC and Cambridge University, UK), Natalie Thorne (ambridge University, UK), Yee Hwa Yang (University of Sydney, Australia), Christos Sotiriou (Jules Bordet Institut, Brussels, Belgium) ***clinical, breast cancer, Pratyaksha Wirapati (Swiss Experimental Cancer Research Institute/Swiss Institute of Bioinformatics), Ruth Luthi-Carter (EPFL, Lausanne, Switzerland) ***biologist. Joe Gray (LBL) ** biologist/data generation*** Gordon Mills (MDA) *** clinican/biologist*** Keith Baggerly (MDA) ***statistician*** Arul Chinnaiyan (University of Michigan) ***biologistt/clinician/bioinformatics***

*** old names: Terry Speed (University of California, Berkeley), John Quackenbush (TIGR), David M. Rocke (University of California, Davis), Wyeth Wasserman (University of British Columbia, Vancouver), Tim Hughes (University of Toronto), Rafael Irizarry (Johns Hopkins), Karl Broman (John Hopkins) Robert Gentleman (Harvard University), Jason Lieb (University of North Carolina, Chapel Hill), Todd Lowe (University of California, Santa Cruz), Michael Newton (University of Wisconsin, Madison), Hao Li (University of California, San Francisco) and David Hinds (Perlegen Sciences). All of these researchers shared our enthusiasm in such a workshop and

showed great interest in participating.

- *Classification.*

  Translational aims are of paramount important in today's biomedical research. This year NCI has awarded a number of grants to begin generating atlas of genomic and genetics features in a variety of cancers. While the ultimate aim is to improve current management of cancer patients, the statistical problem is two fold. The first question is to understand how the high-dimensional data generated on the patients can be used to predict their response to standard or experimental therapies, estimate time to recurrence, assign probability of the progression under "wait and see" protocols as in prostate cancer or simply predict if an individual will develop a particular cancer. This problem falls into the class of *prediction statistical appoaches*. The second question involves identification of druggable markers of response to treatment, recurrence and progression and of early detection. This is known as *variable selection problem* and could lead to selecting the combinations of markers rather than indvidual ones. Note that the two questions are tightly linked.

The statistical challenges include but are not limited to study design and building predictors based on heterogeneous available cohorts, having relatively small ratio of sample size (in 100's, possibly in low 1000's by the time of the workshop) to the number of variables (100's of thousands of features at the moment for any given technology), multiple testing issues, development of computationally efficient classifiers able to explore interaction space of the variables and deal with the variety of the data types.

should this be combined with the Classification aim?*

- *Computational population genetics.* *** NOT CHANGED FROM PREVIOUS VERSION (should be changed???) *** Single nucleotide polymorphism (SNPs) are the most simple form and most prevalent source of genetic polymorphism in the human genome. The advent of SNP genotyping and haplotyping technologies are leading to accumulation of massive amounts of SNP data spanning a variety of species. One of the challenges faced by researchers in this field is how to relate such multimillion dimensional genotypic profiles to both biological and clinical phenotypes, such as disease and drug reaction. As more information accumulates, analysis of the emerging complex data requires comprehensive statistical methodologies capable of dealing with challenging issues such as censoring and causality.
- *Pharmacogenomics.* Genome-scale data are at the forefront of research into targeted therapeutics/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 therapeutic regime indicated for that particular individual. This multi-disciplinary field combines traditional pharmaceutical sciences with large scale data and meta-data on genes, proteins, and single nucleotide polymorphisms.

  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.
- *Emerging technologies.* There are several commonly used technologies for acquiring vast amounts of genomics and proteomics data, with further improvements and technological advances rapidly giving rise to newer assays. 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. Studies of disease processes in humans often include as additional information various types of patient clinical data and covariates.

Each technology involves computational, mathematical, and statistical issues regarding data acquisition, processing, analysis and subsequent interpretation. Statisticians have already contributed immensely in improving design and analysis of gene expression microarrays. Similar challenges are inevitably arising for newer platforms, such as the SNP chips used for high-throughput genome sequencing. Continued interdisciplinary research between biologists and quantitative scientists is crucial to achieving a high level of methodological success for analyzing these newer data types, which will only gain in importance.

- *Data integration.* The explosion of data is generating a number of new challenges in statistics, mathematics and computing. The data are of heterogeneous types measured across a number of biological organisms, very high dimensional, and typically exhibit substantial variability in addition to varying degrees of incompleteness. In order to use the abundance of information to significantly advance biological understanding, fundamentally sound quantitative methods for combination of the manifold data types are required. Appropriate data integration gives researchers power to uncover meaningful biological relationships, enabling further understanding, targeted follow-up, and efficient use of resources.

  Results and findings jointly learned 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 straighforward, mainly correlative approaches have been applied in gene expression and, lately, copy number data, for combining study results. It seems clear, though, that meta-analysis is not very straightforwardly applied to the problem of combining data of different types, the most obvious impediment being lack of a common parameter across data types and mix of letter-based (sequence,) categorical (SNP), ordinal (methylation, protein expression) and continuous (expression and copy number) data types. More sophisticated approaches include phylogenetic methods, hierarchical Bayesian models as well as variations on correlation-based approaches such as kernel, svd-type or distance-based methods. However, integrating multiple data types in an automated, quantitative manner remains a major challenge, where innovative approaches appear to be required. Note that the challenges are so novel that they include both identifying the biologically relevant questions arising from data integration as well as specifying the statistical models and corresponding parameters for estimation along with the required statistical methodologies.

A list of possible participants and their affiliation

should we specify the main area of work for each?* ***i assume we are not listing organizers, right?***

(1) Joe Felsenstein, University of Washington.
(2) Adam Seipel∗, University of California, Santa Cruz.
(3) Bret Larget, University of Wisconsin, Madison.
(4) Lior Pachter, University of California, Berkeley.
(5) Leonid Kruglyak, University of Washington, Fred Hutchinson.
(6) Charles Kooperberg, University of Washington, Fred Hutchinson.
(7) David Hinds, Perlegen Sciences.
(8) Jurg Ott, Rockefeller University.
(9) Hao Li, University of California, San Francisco

(10) Michael Eisen, University of California, Berkeley, LBL.
(11) Wing Wong, Harvard University.
(12) Todd Lowe∗, University of California, Santa Cruz.
(13) Jun Liu, Harvard University.
(14) Tim Hughes∗, University of Toronto.
(15) Jason Lieb∗, University of North Carolina, Chapel Hill.
(16) Joe Derisi, University of California, San Francisco.
(17) Terry Speed∗, University of California, Berkeley.
(18) Rafael Irizarry∗, Johns Hopkins University.
(19) Simon Cawley, Affymetrix.
(20) David M. Rocke∗, University of California, Davis.
(21) John Quackenbush∗, TIGR.
(22) Aad van der Vaart, Vrije Universiteit.
(23) Michael A. Newton, University of Wisconsin, Madison.
(24) Neil Clarke, Johns Hopkins University.
(25) Pat Brown, Stanford University.
(26) Wyeth Wasserman∗, University of British Columbia, Vancouver.
(27) Steven Brenner, University of California, Berkeley.
(28) Andrej Sali, University of California, San Francisco.
(29) David Baker, University of Washington.
(30) Ingo Ruczinski, Johns Hopkins University.
(31) Robert Gentleman∗, Harvard University.
(32) Tim Hubbard, The Wellcome Trust Sanger Institute.
(33) Rebecka Jörnsten, Rutgers University.
(34) Mark Segal, University of California, San Francisco.
(35) Ry Fang Yeh, University of California, San Francisco.
(36) Joe Gray, Lawrence Berkeley Laboratory
(37) Paul Spellman, Lawrence Berkeley Laboratory
(38) Keith Baggerly, MD Anderson
(39) Jeffrey Morris, MD Anderson
(40) Gordon Mills, MD Anderson
(41) Karl Broman, John Hopkins
(42) Shane Jensen, University of Pennsilvania
(43) Adam Olshen, Sloan-Kettering
(44) Natalie Thorne, Cambridge University, UK
(45) Mark Van Der Laan, UC Berkeley
(46) Hongyu Zhao, Yale University
(47) Annette Molinaro, Yale University
(48) Arul Chinnaiyan*, University of Michigan
(49) Debashishis Gnosh, University of Michigan

Additional comments

Our experience in organizing fruitful workshops in the general area of statistical genomics includes:

- *Statistics in Functional Genomics, Ascona (Switzerland) 2004.* This meeting, organized by Darlene Goldstein, Peter Bühlmann and Anthony Davison, included about 80 biologists and quantitative scientists (30% women) from all over the world. We were able to attract as invited speakers top experts from Berkeley, Harvard, Imperial College, and the Max Planck Institut (Berlin). The workshop was a great success, giving rise to some new collaborations, and was internationally very visible.
- *Statistical Science for Genome Biology, BIRS 2004.* This 5-day workshop, organized by Jennifer Bryan, Sandrine Dudoit and Mark J. van der Laan, brought together statisticians working in different genomics-related aspects of statistics. The workshop was a huge success and covered several broad areas of biological investigation that relied on statistical and computational methods. An outstanding aspect of the workshop was participation of many young researchers and women: 23 of the 39 participants were in the category of graduate student/postdoctoral researcher/assistant professor. Moreover, 13 among these were women.
- *Computational and Statistical Genomics, BIRS 2006.* Another 5-day workshop, organized by Jennifer Bryan, Sandrine Dudoit, Sunduz Keles, Katherine S. Pollard and Mark van der Laan, addressed computational and statistical problems associated with newer data types. This profitable meeting included a wide range of expertise, a large proportion of it from women.
- *Statistics for Biomolecular Data Integration and Modeling, Ascona 2007.* This workshop, with a focus on combining data and systems biology, is again organized by Darlene Goldstein, Peter Bühlmann and Anthony Davison. It is scheduled to take place in June 2007.

We would like to build on our track record of these extremely successful workshops, while adapting the scope to new emerging fields of statistical genomics. Enabling close communication between the biological and statistical communities will be conducive to creating statistical innovations that are relevant for advancement in the new and challenging problems that we plan to address.

Dates

Our preferred week (July 27-31) would be particularly advantageous, since the Joint Statistical Meetings (the largest statistics conference in North America) will be held in Denver, Colorado from August 3 - 7, 2008.

Preferred dates.

- July 27 - 31, 2008.
- June 1 - 5, 2008.
- June 22 - 26, 2008.
- June 29 - July 1, 2008.
- July 6 - 10, 2008.

Impossible dates.

- ????, 2008; ISMB. **could not find dates but small enough conference that should not really matter. remove?***
- June 13 - 19, 2008; IMS.
- August 3 - 7, 2008; JSM.
- July 13 - 18, 2008; IBS.

- ???, 2008; WNAR. **could not find dates but small enough conference that should not really matter. remove?***
- March 16 - 19, 2006; ENAR.