Bioinformatics Opening Workshop
September 8-12, 2014

SPEAKER TITLES/ABSTRACTS

Veera Baladandayuthapani
UT MD Anderson Cancer Center

“Bayesian Models for ``Integromics''”

Due to rapid technological advances, various types of genomic, epigenomic, transcriptomic and proteomic data with different sizes, formats, and structures have become available. Each of these distinct data types provides a different, partly independent and complementary, high-resolution view of the whole genome. Modeling and inference in such studies is challenging, not only due to high dimensionality, but also due to presence of structured dependencies (e.g. regulatory mechanisms). Furthermore, analyzing data from multi-platform genomics experiments combined with patientsâ€™ clinical outcomes helps us understand the complex biological processes that characterize a disease, as well as how these processes relate to the development of the disease. I will discuss integrative frameworks for modeling such multi-dimensional data that include current state-of-the art methods and future directions.

Greg Buck
Virginia Commonwealth

“Changing Paradigms of the Vaginal Microbiome in Health, Disease and Pregnancy”

Over the past five years, we (the Vaginal Microbiome Consortium at VCU [VMC]) have applied so-called metagenomic technologies to study the role(s) of the microbes that colonize the female urogenital tract in women’s health and disease in a project entitled the Vaginal Human Microbiome Project (VaHMP). More recently, the VMC has launched the Multi-Omic Microbiome Study-Pregnancy Initiative (MOMS-PI), which was funded by NIH to apply so-called ‘multi-omic’ technologies to study the contribution of the vaginal and related microbiomes to adverse outcomes in pregnancy, focused on preterm birth and stillbirth. In the former cross-sectional community-based study, we have enrolled and analyzed the microbiome profiles of over 6,000 women visiting VCU’s multiple women’s clinics, including Labor & Delivery. This large cross-sectional sampling has provided new insights into the roles of the microbiome in bacterial vaginosis, sexually transmitted infections, maturation and aging, menopause, etc.. The demographic diversity of our population also provided further confirmation of the significant racial/ethnic differences in the microbiota populating the female urogenital tract. The vaginal microbiomes of over 1000 pregnant women in this initial study provided intriguing clues to the roles of the bacteria inhabiting the female urogenital tract in adverse outcomes in pregnancy. These preliminary data in turn led us to embark on the MOMS-PI study, in which we are longitudinally sampling ~2000 women during their pregnancies, and their neoates after birth, at multiple body sites including the vagina, cervix, rectum, skin, nares and mouth. These samples are processed using ‘multi-omics’ technologies, including 16S rRNA ‘metagenomic’ surveys, whole metagenome shotgun sequencing, whole metatranscriptome sequencing, lipidomics, cytokine profiles, and interactomics with the goal of developing an integrated data set.

This work was supported by National Institutes of Health Common Fund grants 4UH3AI083263 (The Vaginal Microbiome: Disease, Genetics and the Environment) and 8U54HD080784 Multi-omic Analysis of the Vaginal Microbiome During Pregnancy (MOMS-PI), funded as components of the National Institutes of Health Human Microbiome Project (see http://commonfund.nih.gov/hmp/index and http://hmp2.org).

Susmita Datta
University of Louisville

“An Integrative Exploratory Analysis of –Omics Data from the ICGC Lung Cancer Data”

We considered an exploratory analysis of multi-platform -omics data involving gene expression, microRNA expression, protein expression, somatic copy number variation, and methylation profiles for subjects with lung adenocarcinoma collected from the International Cancer Genomic Consortium (ICGC). The analysis provided interesting system biological view of the lung cancer progression status.

We present ways to integrate statistical analysis of these data sets of multiple molecular profiles in order to identify connections between important genes, miRNAs, proteins, chromosomal segments, and methylation patterns associated with the disease process rather than merely identifying those which are associated with the disease from each individual molecular profile. We used various clustering techniques for grouping the subjects according to their molecular profiles individually and in an integrative manner into different stages of clinical outcome. We have also used penalized regression techniques to find the predictive performances of their molecular profiles individually and collectively. In addition, we have used ensemble classifiers developed by our research team to classify the patient samples into two different classes of the disease outcomes.

Rebecca Doerge
Purdue University

“Wanted: Large Complex `Omic Data Seeking Quantitative Reasoning and Analysis”

This is an exciting and influential time for the field of Statistics in science. Technological advances in genetic, genomic, and the other 'omic sciences are providing large amounts of complex data that are presenting a number of challenges for the biological community. Many of these challenges are deeply rooted in statistical issues that involve relatively small sample sizes with a large number of parameters (e.g., single cells, genes, exons, base pairs). Although there are many different computational tools for processing these data, there are still issues of data bias. Furthermore, there are a limited number of appropriate statistical methods, and even fewer that acknowledge the unique nature of these data (i.e., high dimensional discrete counts). After a discussion about the data and its structure, experimental design issues will be present. Next-generation sequencing experiments for both transcriptomic and epigenomic based questions will be presented with focus on dependence of high-dimensional data. This talk will be accessible to a broad scientific audience; an in depth understanding of statistics, biology and/or computing is not required.

Mark Gerstein
Yale University

"Human Genome Analysis"

The ENCODE and modENCODE consortia have generated a resource containing large amounts of transcriptomic data, extensive mapping of chromatin states, as well as the binding locations of over 300 transcription-regulatory factors for human, worm and fly. The consortium performed extensive data integration by constructing genome-wide co-expression networks and transcriptional regulatory models, revealing fundamental principles of transcription and network organization that are conserved across the three highly divergent animals.

I will give an overview of the data and some of the key analyses. In particular:

(1) A novel cross-species clustering algorithm to integrate the co-expression networks of the three species, resulted at conserved modules shared between the organisms. These modules are enriched in developmental genes and exhibited hourglass behavior.

(2) A global optimization algorithm to examine the hierarchical organization of the regulatory network. Despite extensive rewiring of binding targets, high-level organization principles such as a three-layer heirarchy are conserved across the three species.

(3) The gene expression levels in the organisms, both coding and non-coding, can be predicted consistently based on their upstream histone marks. In fact, a "universal model" with a single set of cross-organism parameters can predict expression level for both protein coding genes and ncRNAs.

(4) Their have been many analyses of pseudogenes.

(5) Finally, the extent of the non-coding, non-canonical transcription is consistent between worm, fly and human.

encodenets.gersteinlab.org
encodeproject.org/comparative

Debashis Ghosh
University of Colorado

"Kernel Machine Methods for Genomics Studies"

Increasingly in studies from genomics and imaging, there is consideration of high-throughput data. In this talk, we describe the use of kernel machine methods for the analysis of such data. These are techniques that have been popularized in the data mining community, and recently we have proposed and expanded upon statistical equivalences with these algorithms. This allows for a unified estimation and inferential framework. We discuss various appplications of kernel machines to the analysis of high-throughput data data as well as demonstrating how existing methodologies can be viewed as special cases of kernel machines. We will also describe an approach to power considerations with kernel machines.

Kasper Hansen
Johns Hopkins University

"Statistical Analysis of Epigenomewide Data"

Recent technological advances have made it possible to profile biochemical marks genomewide. We discuss challenges and solutions to the analysis of some of these data types, with a focus on DNA methylation and chromatin modifications.

Susan Holmes
Stanford University

"High Dimensional Data Integration"

Modern bioinformatics is challenged by the heterogeneity of the data available. Methods that integrate networks, trees, abundances and contiguous patient and environmental data are necessary. I will cover several approaches to these challenges and provide background in areas where progress is needed.

Laura Kubatko
Ohio State University

"Species-Level Phylogenetic Inference using Large-Scale Sequence Data"

The era of large-scale sequence data has led to unprecedented challenges in the inference of species-level phylogenies, involving both modeling and computation. I will review the issues associated with obtaining species-level phylogenetic estimates for large-scale data with realistic models subject to current computational limitations. I will then describe a new method for species tree estimation under the coalescent model that uses ideas from algebraic statistics to obtain accurate estimates from next-gen data sets in reasonable time. The utility of the method will be demonstrated via simulation and application to empirical data.

Honzhe Li
University of Pennsylvania

"Microbiome, Metagenomics and High Dimensional Composiitonal Data Analysis"

The human microbiome is the totality of all microbes in and on the human body, and its importance in health and disease has been increasingly recognized. High-throughput sequencing technologies have recently enabled scientists to obtain an unbiased quantification of all microbes constituting the microbiome. Often, a single sample can produce hundreds of millions of short sequencing reads. However, drawing valid biological inferences from microbiome studies is difficult because of unique characteristics of the data produced by the new technologies, as well as the sheer magnitude of data. Analyzing these big data poses great statistical and computational challenges. Important issues include normalization and quantification of relative taxa, bacterial genes and metabolic abundances, incorporation of phylogenetic information into analysis of metagenomics data, and multivariate analysis of high dimensional compositional data. We review existing methods, point out their limitations, and outline future research directions.

Jun Liu
Harvard University

“Bayesian Inference of Three-Dimensional Chromosomal Organization from Hi-C Data”

Knowledge of spatial organization of the genome is critical for the study of transcription regulation and other nuclear processes in the cell. Recently, chromosome conformation capture (3C) based technologies such as Hi-C and TCC have been developed to provide a genome-wide, three-dimensional (3D) view of chromatin organization, but analysis algorithms for inferring chromosomal structures from such experiments are still under-developed. Here we describe a novel Bayesian probabilistic approach, denoted "Bayesian 3D constructor for Hi-C data" (BACH), to infer three-dimensional (3D) chromosomal structures. Applying BACH to a high resolution Hi-C dataset generated from mouse embryonic stem cells led to a model of the spatial arrangement of chromatin that revealed structural properties associated with euchromatic and heterochromatic regions in the genome. We also describe a BACH-MIX algorithm that can model the stability of chromatin structures, and found that variations of chromatin modeling are associated with genomic and epigenetic features. Our results demonstrate that BACH and BACH-MIX have the potential to provide new insights into the chromosomal architecture in mammalian cells.

Based on the joint with with Ming Hu, Steve Z Qin, Ke Deng, Jesse Dixon, and Bing Ren.

David Pollock
University of Colorado

“Epistasis, Convergence, and the Evolutionary Stokes Shift”

Adaptive convergence is a unique concept in evolutionary biology that provides information about the nature of the fitness landscape and the repeatability of evolution. It is extremely useful to evolutionary biologists because it can provide strong evidence for adaptation, but it can also vex them by confounding phylogenetic inference. However, to understand adaptive convergence properly, we also need to understand the patterns of non-adaptive convergence, also known as homoplasy. Although most scientists who have thought about it for a few moments believe that our evolutionary models are incorrect, we tend to assume, nevertheless, that they are “close enough” to correct, and that we can predict levels of non-adaptive convergence and their effect on phylogenetics (the “long-branch attraction” phenomenon). However, we have shown previously (Castoe et al., 2009) that they are not apparently good enough to accurately predict convergence levels in proteins. We wondered whether our “evolutionary Stokes shift” model of protein evolution (Pollock et al. 2012; Pollock and Goldstein 2014) could help explain these differences, and whether fluctuating epistasis or coevolution affected convergence, so we began a study on how non-adaptive convergence changed over time in mitochondrial and model proteins. Surprisingly, non-adaptive convergence is much more common than expected in closely related organisms, falling off as organisms diverge. The extent of the convergent drop-off in mitochondrial proteins is well predicted by epistatic or co-evolutionary effects in our evolutionary Stokes shift models, but much less so by conventional or site-specific evolutionary models. Excessive convergence early on is thus not necessarily adaptive, and likelihood values using conventional models are grossly in error in a way that makes them deeply suspect for phylogenetic inference as well as detecting adaptive convergence. We are hopeful that using our new understanding we can develop new, more realistic models to better discriminate adaptive from non-adaptive convergence and improve the validity of phylogenetic inference.

Kim Siegmund
University of Southern California

"Statistical Analysis of Illumina HumanMethylation450 BeadArrays"

DNA methylation is a commonly studied epigenetic mark, its importance well-established in human development and disease. Today, Illumina's HumanMethylation450 arrays provide the most cost-effective means of high-throughput DNA methylation analysis. As with other types of microarray platforms, technical artifacts are a concern. I will introduce the Illumina HumanMethylation450 array, discuss approaches to assess data quality, and methods for signal processing. Specifically, I will discuss how to leverage 'hidden' information on the array to correct for background fluorescence, and describe a method to correct for bias from using two fluorescent dyes. The methods will be illustrated using replicate controls and biological samples. I will conclude with a discussion of the variety of statistical approaches applied in DNA methylation association studies.

Terry Speed
Walter & Eliza Hall Institute of Medical Research and UC Berkeley

"Statistical and Mathematical Bioinformatics: Past, Present and Future (one person's view)"

What we now call Bioinformatics had no simple name before the mid-1990s, but was just thought of as computer methods for the analysis of protein and nucleic acid sequences. The journal *Computer Applications in the Biosciences (CABIOS*) was renamed *Bioinformatics* from 1998. Nevertheless, the field goes back to the early 1960s, perhaps earlier, and has always has lots of mathematics, computer science, probability and statistics in it, occasionally but usually not contributed by people who saw themselves as specialists in these areas. Computer scientists were quicker off the mark than statisticians, and became active in the area in the 1990s. Things changed around the year 2000, with the advent of microarrays and the completion of the Human Genome Project, after which many statisticians joined in. However, it is always wise to bear in mind that the majority of mathematical and statistical methods developed and analyses carried out in the field of Bioinformatics are, and will continue to be done by non-statisticians. So a broad view and open mind are essential.

In this talk will review the history of the field, highlighting mathematical and statistical notions as I go. I see six broad phases. In the 1950s the notion of proteins as amino acid sequences emerged, while the 1960s were dominated by molecular evolution. By the 1970s alignment and protein structure came to the fore, and the 1980s saw the emergence of molecular databases. Things started to expand rapidly in the 1990s, with the advent of the Human Genome Project, the rapid growth of DNA sequencing of model organisms, and the rise of microarrays. From the statistical viewpoint, the period since the year 2000 saw more microarrays and the era of high-throughput DNA sequencing. These generate more than enough work for all of us, and the field shows no sign of slowing down.

Marina Vannucci
Rice University

"Bayesian Models for Integrative Genomics"

Novel methodological questions are being generated in the biological sciences, requiring the integration of different concepts, methods, tools and data types. Bayesian methods that employ variable selection have been particularly successful for genomic applications, as they allow to handle situations where the amount of measured variables can be much greater than the number of observations. In this talk I will focus on models that integrate experimental data from different platforms together with prior knowledge. I will look in particular at hierarchical models that relate genotype data to mRNAs, for the selection of the markers that affect the gene expression. Specific sequence/structure information will be incorporated into the prior probability models. All modeling settings employ variable selection techniques and prior constructions that cleverly incorporate biological knowledge about structural dependencies among the variables. Applications will be to data from cancer studies.

Jean Yang
University of Sydney

“Joint Ranking of Micro-Rnas and Pathways from Matched Mirna and Mrna Data”

As microarray and other forms of high throughput technologies become more readily available there is a growing need to successfully integrate expression and other forms of high-throughput data with a focus on biological interpretation. In this talk, we will examine two types of data, gene (mRNA) and microRNAs (miRNAs) expression studies. We will describe a supervised approach, pMimCor, for integrating various databases of prior knowledge with experimental data to identify sets of genes that share a similar functional outcome and are potentially being regulated by a miRNA. This approach provides a joint ranking of regulatory miRNA and their potential functional impacts with respect to a condition of interest. We apply this approach to multiple experiments with various sample sizes and shown that pMimCor performs competitively across a range of different performance metrics when compared to other approaches.

Ziheng Yang
University College London

“Multispecies Coalescent and its use in Population Genetic Inference from Genomic Data”

The multispecies coalescent is a straightforward extension of the standard (single-population) coalescent to the case of multiple species or populations. It supplies the probabilistic distribution of the gene trees given the underlying species tree, which is the basis for several inference problems in population genomics, such as estimation of migration rates between populations, species tree estimation despite conflicting gene trees, and species delimitation using genetic data. Full-likelihood based inference under the multispecies coalescent model has to meet the computational challenge of averaging over the gene tree distributions, which can be achieved using Markov chain Monte Carlo. While computationally much more demanding, full likelihood methods make a more efficient use of information in the genomic sequence data and have superior statistical properties to heuristic methods, which make inference using reconstructed gene trees without accommodating their uncertainties. In this talk I will give an overview of the application areas of multispecies coalescent and discuss some new research results on our MCMC implementations of the model for species delimitation and species tree estimation.