Statistical Modeling of Epigenomics and Gene Regulation
Workshop Program
Hosted by Harvard University and University of Edinburgh

Thursday, August 27 to Friday, August 28
Pierce Hall 209, Harvard University
29 Oxford Street
Cambridge, MA 02138, USA
# Schedule

All events are located in Pierce 209 unless otherwise noted. A map is located on page 7.

**Thursday, August 27**

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
</table>
| 9:15 am to 10:00 am | Workshop Check-in  
Continental Breakfast                                               |
| 10:00 am to 10:10 am | Welcome and Introduction  
*Edoardo Airoldi, Harvard University; and Guido Sanguinetti, University of Edinburgh* |
| 10:10 am to 11:10 am | Computational Methods for Large-scale Detection and Dissection of Human Regulatory Elements  
*Jason Ernst, University of California, Los Angeles*               |
| 11:10 am to 12:20 pm | First Contributed Session on Epigenomics  
*Andreas Kapouranis, Yuping Zhang, Zhengqing Ouyang, Tom Mayo, James Zou, and Gabriele Schweikert* |
| 12:20 pm to 2:00 pm | Lunch                                                                 |
| 2:00 pm to 3:00 pm | Scalable Bayesian Kernel Models with an Application to MAPK Pathway Addiction in BRAF Mutated Melanoma  
*Sayan Mukherjee, Duke University*                                  |
| 3:00 pm to 4:20 pm | Second Contributed Session on Gene Expression and Networks  
*Julien Gagneur, Kamrine Poels, Michael Tolstorukov, Feras Saad, Edward Wallace, Marieke Kuijjer, and John Santerre* |
| 4:20 pm to 4:50 pm | Poster Session Set-up  
Coffee Break                                                           |
| 4:50 pm to 7:00 pm | Poster Session and Reception  
*Maxwell Dworkin 119 and Lobby*                                      |
Friday, August 28

9:30 am to 10:30 am  Beyond DNA Methylation and Gene Expression: What Else to Learn from Bis-seq and RNA-seq Data  
*Michael Stadler, Friedrich-Miescher Institute, Basel*

10:30 am to 11:00 am  Coffee Break

11:00 am to 12:30 pm  Breakout Group Discussions on Challenges for SMEGR  
*Pierce Hall 100F, 209, and 320*

12:30 pm to 2:00 pm  Lunch

2:00 pm to 3:00 pm  Reports from Breakout Groups and Panel-Led Discussions  
*Panelists: Jason Ernst, Sayan Mukherjee, Uwe Ohler, and Michael Stadler*

3:00 pm to 4:00 pm  Predictive Computational Models: From Transcription Initiation to Enhancer-Promoter Interactions  
*Uwe Ohler, Max Delbrueck Center, Berlin*

4:00 pm to 4:10 pm  Workshop Closing  
*Edoardo Airoldi, Harvard University; and Guido Sanguinetti, University of Edinburgh*
Invited Speaker Abstracts

Computational Methods for Large-scale Detection and Dissection of Human Regulatory Elements
Jason Ernst, PhD
Assistant Professor, Biological Chemistry
Assistant Professor, Computer Science
University of California, Los Angeles

Understanding the human genome sequence and in particular the vast non-coding regions is a central challenge for modern molecular biology with profound implications towards understanding the genetic basis of disease. While understanding the genome by directly reading the primary DNA sequence is extremely challenging, the presence of epigenetic marks on top of the sequence holds great promise to aid our understanding of the genome. Technological advances in sequencing has made it possible in a single experiment to generate tens of millions of data points on the location of a epigenetic mark across the genome in a specific cell type, which is raising a number of computational challenges and opportunities. In this talk, I will first describe a method that I previously developed, ChromHMM, which learns de novo combinatorial and spatial patterns from maps of multiple epigenetic marks using a multivariate hidden Markov model. These patterns correspond to different classes of genomic elements, which I have then used to provide a cell type specific annotation of the human genome. I will then describe a combined computational modeling and experimental approach that in high-throughput can test putative regulatory elements of interest identified based on epigenomics patterns and identify within them at high resolution bases activating or repressing gene expression. Finally, I will describe a new method, ChromImpute, to impute maps of epigenetic marks that I have applied in the context of the Roadmap Epigenomics project to computationally predict over 4000 epigenomic datasets vastly accelerating the coverage of the human epigenome while providing overall more robust maps than have been obtained experimentally.

Scalable Bayesian Kernel Models with an Application to MAPK Pathway Addiction in BRAF Mutated Melanoma
Sayan Mukherjee, PhD
Associate Professor of Statistical Science
Associate Professor of Computer Science and Mathematics
Duke University

Nonlinear kernels are used extensively in regression models in statistics and machine learning since from the perspective of predictive accuracy. Variable selection is a challenge in the context of kernel based regression models.

In linear regression the concepts effect size for the regression coefficients is very useful for variable selection. We will present a scalable Bayesian kernel model for which the analog for the effect size of each explanatory variable can be inferred. The key idea that allows to extract the effect size is a random Fourier expansion for shift-invariant kernel functions. We apply this idea to create a class of scalable Bayesian kernel regression models (SBKMs) for both nonparametric regression, binary classification, and non-linear mixed effects models. We outline how this model can be used to extraction of differential expression in cancer biology as well as non-linear mapping of complex traits, both eQTL analysis as well as GWAS.
Our work sought to explore the phenomenon of acquired resistance to MAP kinase (MAPK) inhibitors, drugs which are used as first-line therapies for the treatment of BRAF-mutant melanoma. While most patients respond to upfront treatment with these drugs, nearly all develop resistance in the span of months. Through a large number of studies, it is now clear that resistance can be caused by the activation of a diverse range of upstream cytoplasmic signaling pathways, a fact which makes the development of therapies to overcome resistance complex. Using a combination of statistical modeling of gene expression in drug-sensitive and drug-resistant melanoma cell lines and human tumors with experimental validation, we revealed a MYC-driven transcriptional program that serves as a nexus of these resistance pathways. MYC inhibition reverses resistance driven by upstream pathways in engineered, evolved, and intrinsically resistant cell lines, and chronic MYC inhibition in sensitive cells blocks the evolution of resistance.

Joint work with Lorin Crawford, Katie Singleton, and Kris Wood.

Predictive Computational Models: From Transcription Initiation to Enhancer-Promoter Interactions
Uwe Ohler, PhD
Professor
Max Delbrueck Center, Berlin

The transcription of metazoan genes is a complex process that includes regulatory regions close and far away from the genes they regulate (promoters resp. enhancers). My lab is interested in understanding the makeup and function of these different regions, and how their sequence and chromatin state work together to encode complex expression patterns in development and disease.

In the first part, I will explain our recent studies on transcription initiation and its directionality, in which we combined high-resolution nascent RNA initiation site mapping with underlying sequence and chromatin features.

In the second part, I will present a new predictive approach to link enhancers to target promoters that relies on the co-expression of genes rather than correlation of chromatin states with gene expression across conditions.

Beyond DNA Methylation and Gene Expression: What Else to Learn from Bis-seq and RNA-seq Data
Michael Stadler, PhD
Senior Computational Biologist
Friedrich-Miescher Institute, Basel

High-throughput sequencing has enabled efficient genome-wide measurement of biological variables such as transcribed genes of a cell and their chromatin state. These experiments yield rich data that contain various types of information, often beyond the scope of a given experiment or study. For example, RNA-seq is most often used to quantify RNA expression levels of genes, but also provides information about sequence polymorphisms, splicing, promoter usage and poly-adenylation. This talk will present two examples of additional information that can be obtained from sequencing datasets: In the first example, we study DNA methylation using whole-genome bisulfite sequencing and identified so called Low Methylated Regions (LMRs) that are formed by the binding of transcription factors and allow identification of enhancer regions [1,2]. In the second example, we demonstrate...
that an analysis of intronic and exonic reads in RNA-seq data (EISA, Exon-Intron Split Analysis) allows the characterization of transcriptional and post-transcriptional RNA regulation [3]. EISA can for example help to discriminate primary from secondary targets in miRNA transfection experiments. Finally, we have developed a set of freely available tools that implement the here described analyses [2,3,4].

Map

Workshop: Pierce Hall 209; 29 Oxford Street
Poster Session and Reception, Thursday Only: Maxwell Dworkin 119 and Lobby; 33 Oxford Street (Doors lock at 6:00 pm. Please do not prop the doors!)
Back-up Location: Science Center 300 H; 1 Oxford Street (You will be notified by email should we change locations. The Statistics Department offices are located on the 7th floor of the Science Center as well.)
Notes