

Mathematical & Computational Biology Seminar

Organizer: Lior Pachter

Wednesday, 2:00–3:00pm, 939 Evans

Apr. 22 **Sandrine Dudoit**, University of California, Berkeley
Statistical Inference in mRNA-Seq

For the past decade, microarrays have been the assays of choice for high-throughput studies of gene expression. Recent improvements in the efficiency, quality, and cost of genome-wide sequencing are prompting biologists to rapidly abandon microarrays in favor of ultra high-throughput sequencing, a.k.a., second-generation or next-generation sequencing: e.g., Applied Biosystems' SOLiD, Helicos BioSciences' HeliScope, Illumina's Genome Analyzer, and Roche's 454 Life Sciences sequencing systems. These high-throughput sequencing technologies have already been applied to monitor genome-wide transcription levels (mRNA-Seq), DNA-protein interactions (ChIP-Seq), chromatin structure, and DNA methylation. While sequencing-based gene expression studies have been touted as overcoming longstanding limitations of microarray-based studies, these new biotechnologies raise similar as well as novel statistical and computational challenges, in areas such as image analysis, base-calling, read-mapping, and (differential) expression inference.

This talk will report on our investigation of two mRNA-Seq datasets obtained using Illumina's Genome Analyzer platform to measure transcript levels in reference samples from the MicroArray Quality Control (MAQC) Project. We focus on the analysis of mapped read counts and the following three main issues: (1) exploratory data analysis (EDA); (2) assessment of biological effects of interest (e.g., expression levels in Brain vs. UHR RNA) and nuisance experimental effects (e.g., library preparation, flow-cell, and lane effects); (3) identification of differentially expressed genes.

This is joint work with James H. Bullard, Steffen Durinck, Kasper D. Hansen, and Elizabeth A. Purdom.