ABSTRACTS: AMSI-SSAI LECTURER 2014 – PROFESSOR TERRY SPEED

Statistics lecture 1
Removing Unwanted Variation from high-throughput omic data
(joint with Johann Gagnon-Bartsch and Laurent Jacob)

Abstract: Over the last few years, many microarray-based gene expression studies involving a large number of samples have been carried out, with the hope of understanding, predicting or discovering factors of interest such as prognosis or the subtypes of a cancer. The same applies to proteomic and metabolomic data, and to other kinds of data. Such large studies are often carried out over several years, and may involve several hospitals or research centers. Unwanted variation (UV) can arise from technical aspects such as batches, different platforms or laboratories, or from biological signals such as heterogeneity in ages or different ethnic groups which are unrelated to the factor of interest in the study. This can easily lead to poor results. Recently, we proposed a general framework to remove UV (called *RUV*) in microarray data using *control* genes. It showed good behavior for differential expression analysis (i.e., with a known factor of interest) when applied to several datasets, in particular better performance than state of the art methods such as *Combat* or *SVA*. This suggests that controls can indeed be used to estimate and efficiently remove sources of unwanted variation. The methods are illustrated on a variety of kinds of omic datasets.

Statistics Lecture 2
Normalization of RNA-Seq Data: Are the ERCC Spike-In Controls Reliable?
(joint with Sandrine Dudoit, Davide Risso and John Ngai)

The External RNA Control Consortium (ERCC) developed a set of 92 synthetic polyadenylated RNA standards that mimic natural eukaryotic mRNA. The standards are designed to have a wide range of lengths (250-2,000 nucleotides) and GC-contents (5-51%). The ERCC standards can be spiked into RNA at various concentrations prior to the library preparation step and serve as negative and positive controls in RNA-Seq. Ambion commercializes spike-in control mixes, ERCC ExFold RNA Spike-in Control Mix 1 and 2, each containing the same set of 92 standards, but at different concentrations. We investigate the use of the ERCC spike-in controls for two main purposes: (a) Quality assessment/quality control (QA/QC) of RNA-Seq data and benchmarking of normalization and differential expression (DE) methods, and (b) Direct inclusion in between-sample normalization procedures. We have two RNA-seq data sets which make use of the ERCC controls: a local one concerning treated and untreated zebrafish tissue, and some of the SEQC samples. A variety of normalization methods will be compared, both using and not using the ERCC controls. One of the methods we discuss is a variant on our recently published RUV-2 method, which uses SVD on negative controls.

Bioinformatics lecture
Comparing and combining mutation callers
(joint with Su Yeon Kim and Laurent Jacob)

Somatic mutation-calling based on DNA from matched tumor-normal patient samples is one of the key tasks carried by many cancer genome projects. One such large-scale project is The Cancer Genome Atlas (TCGA), which is now routinely compiling catalogs of somatic mutations from hundreds of paired tumor-normal DNA exome-sequence datasets. Several mutation-callers are publicly available and more are likely to appear. Nonetheless, mutation-calling is still challenging and there is unlikely to be one established caller that systematically outperforms all others. Evaluation of the mutation callers or understanding the sources of discrepancies is not straightforward, since for most tumor studies, validation data based on independent whole exome DNA sequencing is not available, only partial validation data for a selected (ascertained) subset of sites.

We have analyzed several sets of mutation calling data from TCGA benchmark studies and their partial validation data. To assess the performances of multiple callers, we introduce approaches utilizing the external sequence data to varying degrees, ranging from having independent DNA-seq pairs, RNA-seq for tumor samples only, the original exome-seq pairs only, or none of those. Utilizing multiple callers can be a powerful way to construct a list of final calls for one's research. Using a set of mutations from multiple callers that are impartially validated, we present a statistical approach for building a combined caller, which can be applied to combine calls in a wider dataset generated using a similar protocol. The approach allows us to build a combined caller across the full range of stringency levels, which outperforms all of the individual callers.

Public lecture
A New Frontier: understanding epigenetics through mathematics
Scientists have now mapped the human genome - the next frontier is understanding human epigenomes; the 'instructions' which tell the DNA whether to make skin cells or blood cells or other body parts. Apart from a few exceptions, the DNA sequence of an organism is the same whatever cell is considered. So why are the blood, nerve, skin and muscle cells so different and what mechanism is employed to create this difference? The answer lies in epigenetics. If we compare the genome sequence to text, the epigenome is the punctuation and shows how the DNA should be read. Advances in DNA sequencing in the last five years have allowed large amounts of DNA sequence data to be compiled. For every single reference human genome, there will be literally hundreds of reference epigenomes, and their analysis could occupy biologists, bioinformaticians and biostatisticians for some time to come.