THE UNIVERSITY OF CHICAGO

Department of Statistics

STATISTICS COLLOQUIUM

TERRY SPEED

Bioinformatics Division, WEHI, and Department of Statistics
University of California, Berkeley.

Removing Unwanted Variation from High-Throughput Omic Data*

MONDAY, November 26, 2012, at 4:00 PM

133 Eckhart Hall, 5734 S. University Avenue

Refreshments following the seminar in Eckhart 110

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 several other kinds of data. Large studies are often carried out over several years, and involve several hospitals or research centers. Unwanted variation (UV) can arise from technical elements 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. They can easily lead to spurious conclusions. For example, when doing clustering to identify new subgroups of the disease, one might identify one of the UV factors if its effect on gene expression is stronger than the subgroup effect. Note that similar problems arise when the objective is to combine several smaller studies. A very important objective is therefore to remove these UV factors without losing the factors of interest. The problem can be more or less difficult depending on what is actually observed and what is not. For example, when doing differential expression studies or supervised learning when the factor of interest is known and all the UV factors (say technical batches or different studies) are also known, the problem essentially boils down to a regression, and methods such as *Combat* generally give good results. When the UV factors are modeled as unknown, the problem becomes more difficult because one has to estimate UV factors along with their effects on the genes, and several estimates may explain the data equally well while leading to very different conclusions. This is partially addressed by methods like *SVA*. When neither the factors of interest nor the UV are observed, the problem is even more difficult. It can occur if one is interested in any kind of unsupervised analysis like clustering, or if one simply wants to "clean" a large dataset from its UV without knowing in advance what factors of interest will be studied. Some authors use *SVD* on the expression matrix to identify the UV factors. This approach may work well in some cases but relies on the strong assumption that all UV factors explain more variance than any factor of interest. Furthermore it will fail if the UV factors are too correlated with the factor of interest. Recently, we proposed a general framework to remove UV (called RUV) in microarray data using *control* genes. It showed very 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. Our objective here is to describe ways of doing similar things when carrying our supervised and unsupervised learning. We propose methods exploiting the existence of replicate arrays. The methods are illustrated on a gender study dataset, an MAQC dataset, some metabolomic data and some TCGA data.

* Joint with Johann Gagnon-Bartsch and Laurent Jacob.

For further information and inquiries about building access for persons with disabilities, please contact Dan Moreau at 773.702.8333 or send him an email at dmoreau@galton.uchicago.edu. If you wish to subscribe to our email list, please visit the following website:
https://lists.uchicago.edu/web/arc/statseminars.