Normalization of Omic Data after 2007[a]

Terry Speed[b]

For over a decade now, normalization of transcriptomic, genomic and more recently metabolomic and proteomic data has been something you do to "raw" data to remove biases, technical artifacts and other systematic non-biological features. These features could be due to sample preparation and storage, reagents, equipment, people and so on. It was a "one-off" fix to what I'm going to call removing unwanted variation. Since around 2007, a more nuanced approach has been available, due to JT Leek and J Storey (SVA) and O Stegle et al (PEER). These new approaches do two things differently. The first is that they do not assume the sources of unwanted variation are known in advance, they are inferred from the data. And secondly, they deal with the unwanted variation in a model-based way, not "up front". That is, they do it in a problem-specific manner, where different inference problems warrant different model-based solutions. For example, the solution for removing unwanted variation in estimation not necessarily being the same as doing for prediction. Over the last few years, I have been working with Johann Gagnon-Bartsch and Laurent Jacob on these same problems through making use of positive and negative controls, a strategy which we think has some advantages. In this talk I'll review the area, and highlight some of the advantages of working with controls. Illustrations will be from microarray, mass spec and RNA-seq data.

[a] Joint with Johann Gagnon-Bartsch and Laurent Jacob

[b] Walter and Eliza Hall Institute of Medical Research