Analysing Data: Towards developing a framework for transcriptomics and other Big Data analysis for regulatory application.

Timothy W Gant

Centre for Radiation, Chemical and Environmental Effects, Public Health England, Oxfordshire, UK.

Big Data has been part of the landscape of toxicology for nearly two decades and has contributed much to our undertaking of modes and mechanisms of toxicity; but despite the extensive use of 'Big Data' and in particular 'omics data in toxicology research these data sets have yet to be routinely utilised in regulatory toxicity. This is partly because from the first generation of these data sets it was apparent that, even before considering interpretation, large data sets pose challenges. Some of these challenges have been quality control of data generation, normalization, recognition of outliers and univariate statistical analysis. Additionally there are challenges with the associated experimental meta data and last but not least data interpretation. There are biological and experimental variables revealed by these large data set that may not be seen, or be of consequence, when fewer measurements are taken.

The challenges of adequate meta data associated with the experiment and availability of the data were addressed with the standards set out in MIAME (Minimum information about a microarray experiment) (Brazma A et al **Nat Genet**. 2001 Dec;29(4):365-71). For data quality the MAQC consortium has led the way in addressing the issue of quality control in data generation both with microarrays (**Nat Biotechnol**. 2006 Sep;24(9):1151-61) and next generation sequencing methods (**Nat Biotechnol**. 2014 Sep;32(9):903-14). This consortium has also addressed to some extent best practice in the initial analysis of these data, but not to the point of recommendation of methods.

The adoption of standards for the univariate data analysis has been slower than the adoption of the standards for meta data collection and standardisation of methods for the generation of data. The causes for this are not clear but one possible reason is that there are many different ways of processing the data. Everyone has their favourite and can divide the data in the way that suits their experiment or hypothesis for example by changing the statistical parameters. While this can be acceptable for research where justification for the method used will be subject to peer review and likely replication, it is not acceptable for regulatory use where consistency is paramount.

While certain mathematical and statistical methods for the univariate have achieved a level of greater acceptability, a framework of best practice has not been developed that can be routinely applied to the primary analysis of data to the point of the generation of a gene list for subsequent interpretation. This presentation will outline this issue and present the initial thoughts from a group of experts convened to examine the issues under the auspices of ECETOC.