Guideline Framework for the univariate analysis of 'Big Data' for regulatory use

:

Members of the ECETOC Data Analysis Framework Task Force

Background

Big data has become a common term to describe the large data sets that are often collected in biological and medical science. This is particularly true for molecular work where a data explosion is occurring driven by the 'omics technologies. However, in Toxicology the 'Big Data' collected from methods such as the 'omics has not been bioinformatically analysed in a consistent manner. For toxicology research that will be subject to peer review and replication the application of different analysis methods is less of an issue. For regulatory use however the absence of consistent analysis methods has hindered application in the regulatory toxicological risk assessment of chemicals.

Thus the challenge is in the consistent univariate bioinformatic analysis of 'Big Data' where data magnitude allows many possible approaches to be taken. The many ways by which the data can be 'sliced and diced' can lead to different outcomes. There are many methods all of which could be argued to be correct but the key question for regulatory purposes is which is most appropriate? This challenge has limited the uptake of high throughput molecular data in the regulatory arena.

Specific challenges in the analysis of 'Big Data'

- Formation of a reproducible list of relevant data from the output of a high throughput 'omics methods such as transcriptomics has been a challenge in regulatory toxicology due to the number of potentially applicable bioinformatic methods and statistical variables.
- In the univariate analysis the use of different normalisation methods, recognition of outliers and in particular statistical criteria can have a profound impact on the data included as the output of 'omics data
- Data subject to different analysis methods that can produce different outcomes depending on the methods employed. This is not acceptable for regulatory submissions.
- In research the use of different bioinformatic analysis methods can be judged in peer review but for regulatory use a formalised approach is required.

Objectives

- To produce a framework for 'Big data' univariate analysis where justifications are only necessary when deviating from the proposed framework.
- Formulate the framework into an OECD guidance document. A project submission form is in the process of being submitted to initiate this work.

Limits

- This project is only considering the transformation of data from the raw data to a statistically assessed processed gene list (univariate analysis).

The team

- Wenjun Bao
- Mohamed Bonahmed
- Tim Ebbels
- Karma Fussell
- Tim Gant
- Bruno Hubesch
- Madeleine Laffont
- Alan Poole
- David Rouquie
- Steffen Schneider
- Leming Shi
- Tokuo Sukata
- Kayo Sumida
- Weida Tong
- Shu-Dong Zhang

The process

- The group met for two days in July 2015 to formulate the framework (see next box)
- Three test teams are undertaking a trial of the framework on set of data developed by BASF under CEFIC LRI contract EMSG 56.
- The core group will meet again in January 2016 to review the output
- Telephone conferences have been held between these events to review progress and agree goals.

The draft framework

Initial Data collection

- Data collection proceeds according to the manufacturers guidelines
- No pre-filtering should be performed on the data except for the removal of spiked in standards.
- All of the data sets in the experiment should be examined to ensure consistency of quality according to the parameters below.
- Outlying data sets should be identified and justification should be made for any outlying data sets that are not removed before further analysis.

Outlying data sets

Outlying data sets should be removed from the data set before the commencement of further data processing. Some of the criteria for their identification are:

- For RNA was the RIN number low?
- For microarrays was there low dye incorporation?
- For RNA-Seq was the read depth low?
- For RNA-Seq – low % of mapped reads
- All methods – failure of manufacturers QC
- All methods – low signal to noise ratio
- All methods – Spiked in controls if present should pass manufacturers quality control
- All methods – data set does not conform to statistically assessed normality
- All methods – data from biological replicates does not cluster together on a PCA plot

Normalization

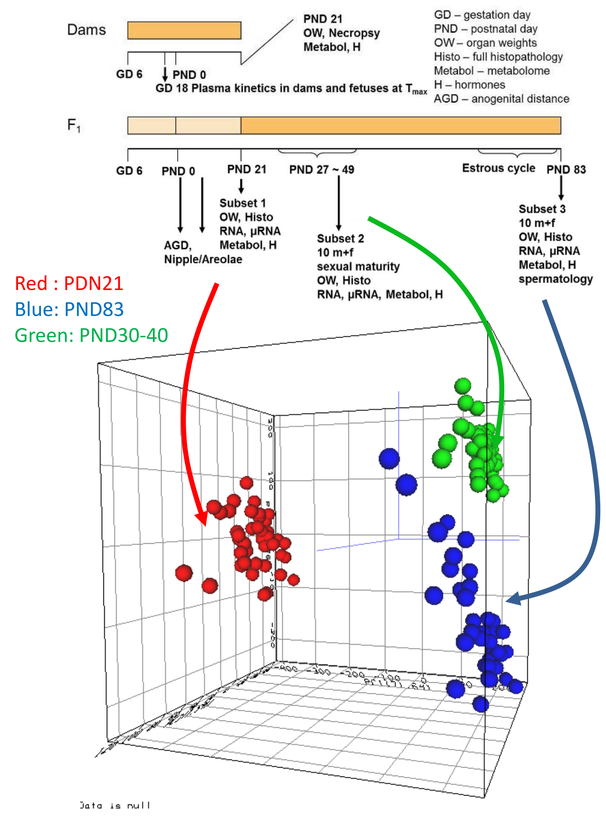
- The method of normalization to correct experimental and technical variables should be relevant to the type of data but in all cases should be the minimum necessary
- Normalization should be within sample only and not performed across the whole experiment (between sample normalization)[1]
- After normalization test again for outliers as performed previously for the raw data

1. Between sample normalization methods, such as RMA (Robust Multi-array Average) would allow different samples to affect each other, such that the addition new samples will result in changes of (normalized) expression values in existing samples. This contradicts statistical principals and physical reality, therefore should be avoided. Please see the paper at DOI:10.1109/BIBM.2014.6999142 for some discussions of the issue.

Statistics

- The combination of p value and fold change is the best way of recognising differential gene expression
- For calculating the p-value the Welch's test is recommended. This test is more robust to unequal variance and sample size than Student's t-test.
- A fold change of 1.5 and p value of $p<0.05$ should be used as a cut-off

Application – BASF data two generation study



The test data from BASF is for three compounds each at three dose levels:

- Flutamide
- Perchloraz
- Vinclozolin

Dosed to pregnant dams from gestation day 6 to 21 days post partum (PND 21).

Offspring dosed from PND 21 to PND 83

PCA separation of data processed using the framework of the top dose of flutamide. Some spreading of data can be seen in the PND 83 which could be due to inter-individual in sexual maturation or adaptation

Data was processed using the FDA arraytrack platform

From here:

- Test the framework on several data sets
- Publish as an ECETOC report
- Publish as a open access peer reviewed papers
- Progress to an OECD guidance document