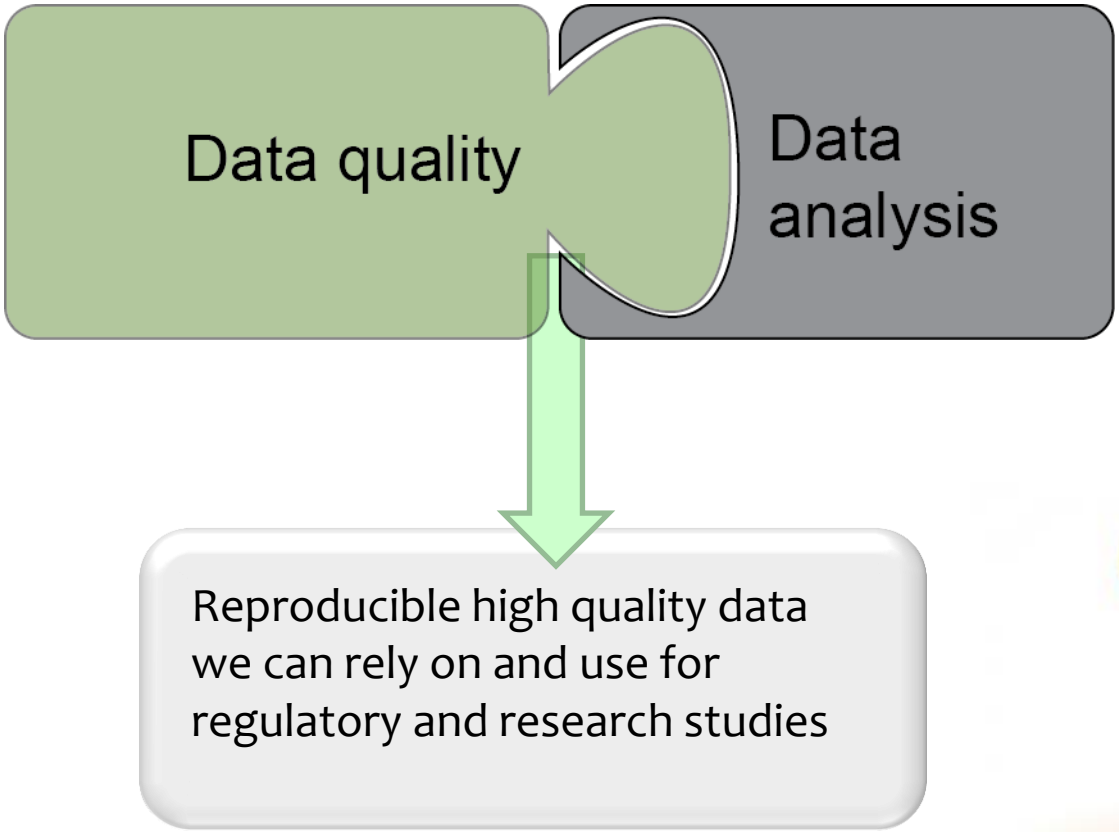
Public Health England
Protecting and improving the nation's health

Towards a framework for transcriptomics and other Big Data analysis for regulatory application

Timothy W Gant

Wenjun Bao; Remi Bars; Mohamed Benahmed; Alan Boobis; Timothy Ebbels, Karma Fussell, Lili Li; David Rouquie; Leming Shi; Kayo Sumida; Weida Tong; Shu-Dong Zhang; Madeleine Laffont (ECETOC); Alan Poole (ECETOC)

‘Omics Data quality and data analysis



Robust reproducible gene lists for regulatory guidance.

The caveat to ‘omics data

‘Omics has enjoyed a great deal of success in research.

Nevertheless there are serious challenges with high throughput data the have hindered wider adoption and too many studies have for example:

- Relied on too few replicates,
- Confusion between technical and experimental replicates
- Inadequately controlled variables
- Over normalization or filtering
- Inappropriate statistics

These challenges have hindered the use of ‘omics data in regulatory assessments

New applications such as epigenetics are going to amplify this challenge.

The challenge

Inconsistency in large pharmacogenomic studies

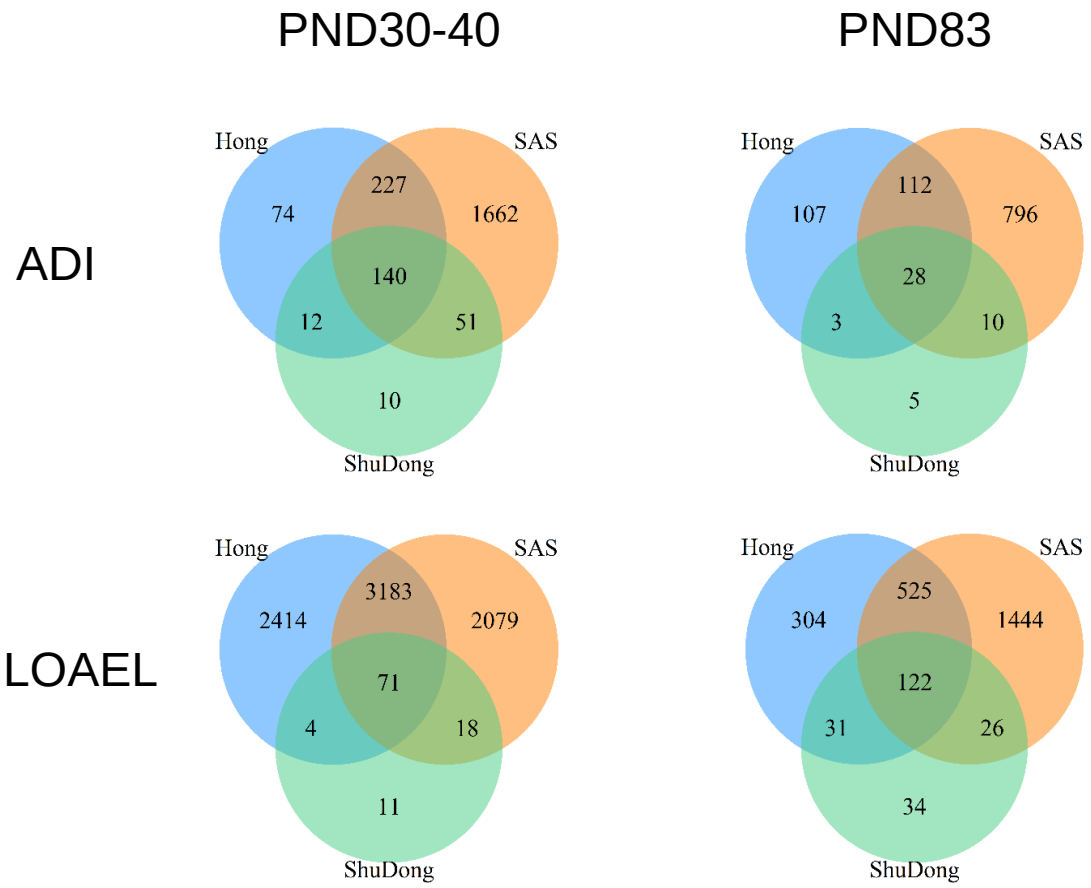
Benjamin Haibe-Kains[1,2], Nehme El-Hachem[1], Nicolai Juul Birkbak[3], Andrew C. Jin[4], Andrew H. Beck[4]*, Hugo J. W. L. Aerts[5,6,7]* & John Quackenbush[5,8]*

Two large-scale pharmacogenomic studies were published recently in this journal. Genomic data are well correlated between studies; however, the measured drug response data are highly discordant. Although the source of inconsistencies remains uncertain, it has potential implications for using these outcome measures to assess gene–drug associations or select potential anticancer drugs on the basis of their reported results.

nature

2013

Exemplar of the Challenge



There was one further analysis that was consistent with that of Hong – same mathematical analysis method used.

ECETOC team - What is the aim?

- To build on the work carried out and published by the MAQC consortium to ensure data quality in the analysis of ‘omics data
- To develop a framework of best practice in ‘omics data analysis to go from raw data to gene list
- To make the framework applicable to data from all ‘omics methods but with an initial focus on transcriptomics
- To render the analysis framework ‘opt out only with justification’ whereby deviations from the framework would have to be justified

Has this been done already?

- **OECD** does not have a guideline document for the univariate analysis of ‘omics data
- **MAQC** considered the quality of data generation from microarrays and from high throughput sequencing but did not consider the analysis as a separate issue
- **MAQC** did though develop guidelines for data analysis but did not incorporate these into a framework
- **So -** While there are many accepted and widely applied methods for the analysis of ‘omics data there is no clear guideline or framework available for application in a regulatory environment.

MAQC projects

2005-2006	2007-2010	**2008-2014**	**2012-??**
MAQC-I: Differentially Expressed Genes (DEGs)	MAQC-II: Classifiers and GWA	**MAQC-III (SEQC):** Next-Gen Sequencing Quality Control	**MAQC-IV (PADRE):** Predicting Adverse Drug Reactions and Efficacy

2005

2012

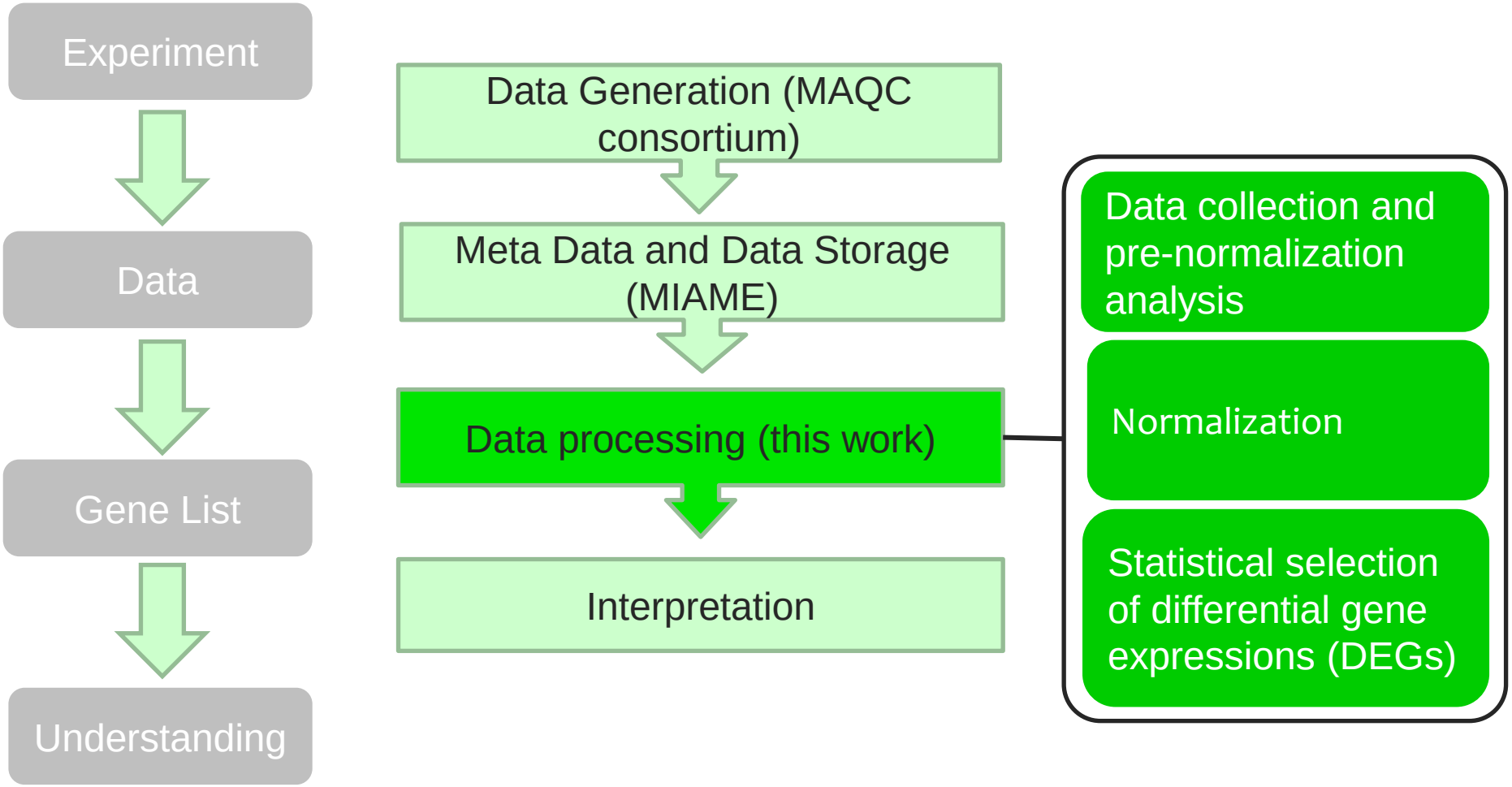
Combinations for analysing 'omics data

- Image processing
- Background handling
- Transformation
- Normalization
- Gene selection
- Classification
- Biological interpretation
- ……

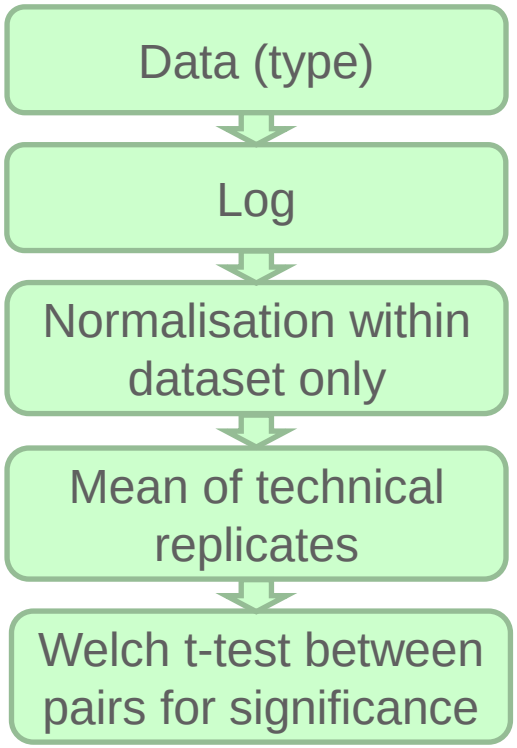
>10 million combinations

Based on estimation by Dr. Russ Wolfinger (SAS Institute Inc.)
The 4th MAQC Project Meeting, Feb. 3-4, 2006, Boston, MA

Processes from Experiment to Output



Output from the July 2015 workshop.



How to recognise outlier data sets?

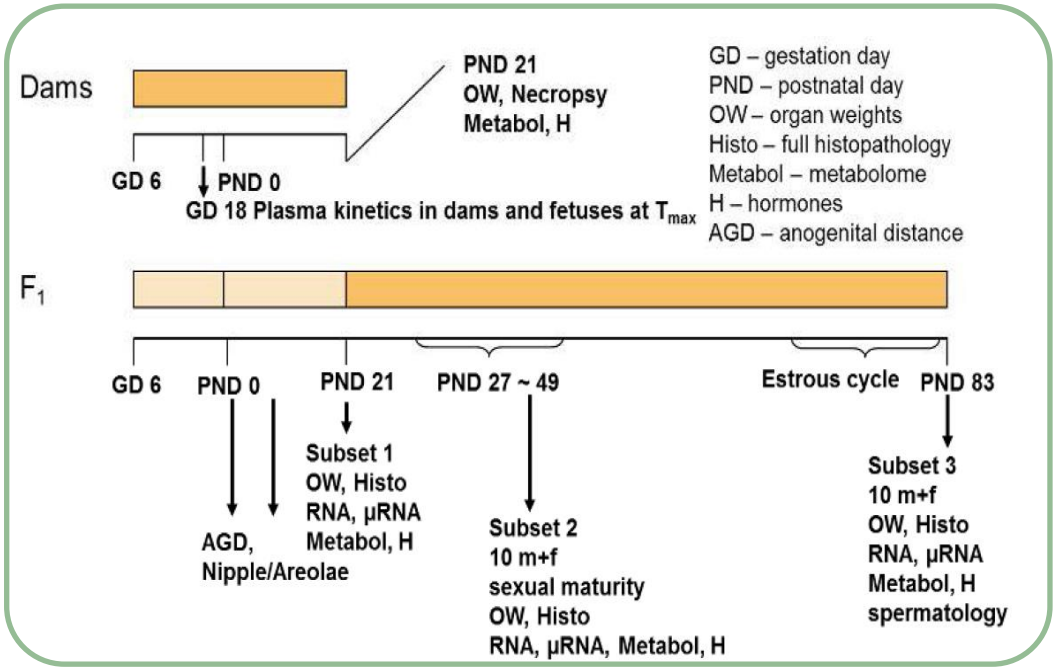
How to deal with low signal strength?

Type of normalisation and why not between data sets?

Welch (deals with unequal variance better than Students t-test)

A fold change of 1.5 and p value of $p<0.05$ should be used as a cut-off

Exemplar data (Two generation study)



Agilent arrays
62976 data points
per array

gMedianData → 7.56 million data points

gProcessedData → 7.56 million data points

<u>Doses</u>
Flutamide – 0.0025, 0.25 and 2.5 mg/kg/day
Prochloraz – 0.01, 5 and 30 mg/kg/day
Vinclozolin – 0.005, 4 and 20 mg/kg/day
Controls – no compound

x

<u>Time points</u>
PND0
PND30-40
PND83

x

<u>Replicates</u>
4

=

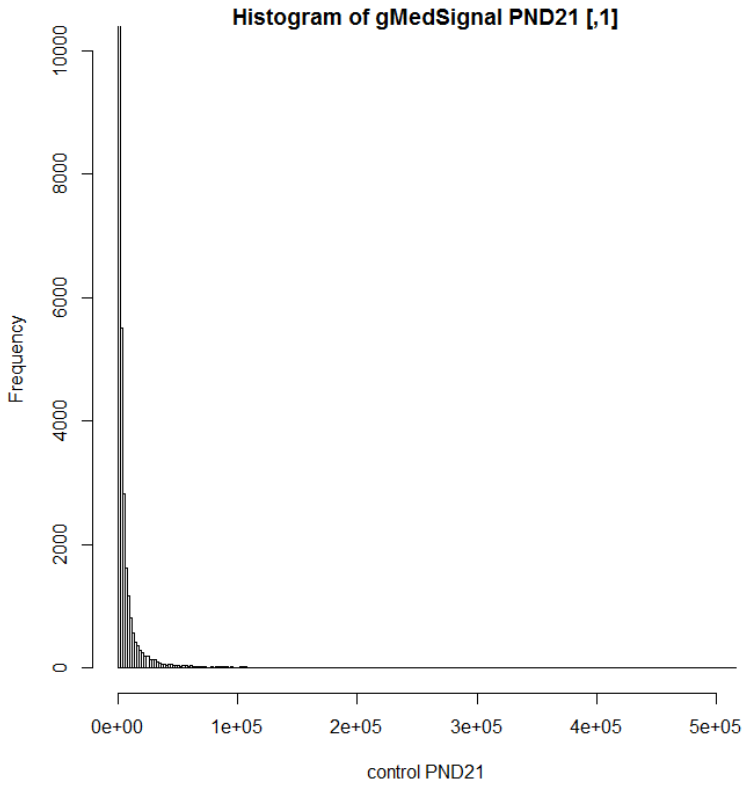
120 data sets

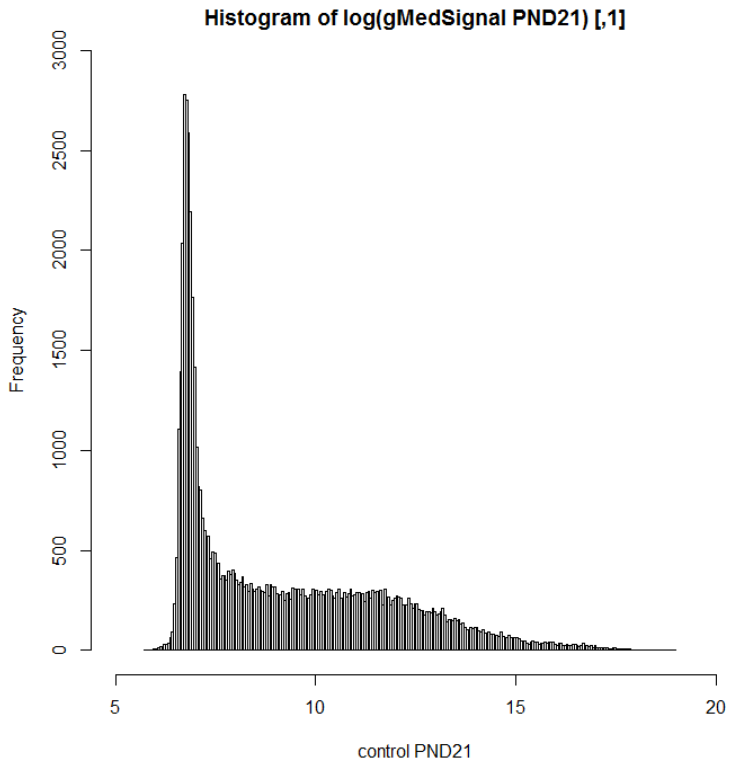
Data collection pre-normalization

- Data collection proceeds according to the manufacturers guidelines
- No prefiltering should be performed on the data except for the removal of spiked in standards.
- A normalization test should be performed on the raw data to ensure that it is normally distributed
- Outliers in the data can be identified using PCA plots. Data from the same sample types should group together.
- Not normal data and outliers should be removed.
- Justification should be made for any outlying data sets that are retained for analysis

Why log?

gMedianData





- Stabilises variance
- Spreads data over the range
- Produces a more normal distribution

Recognising an Outlier

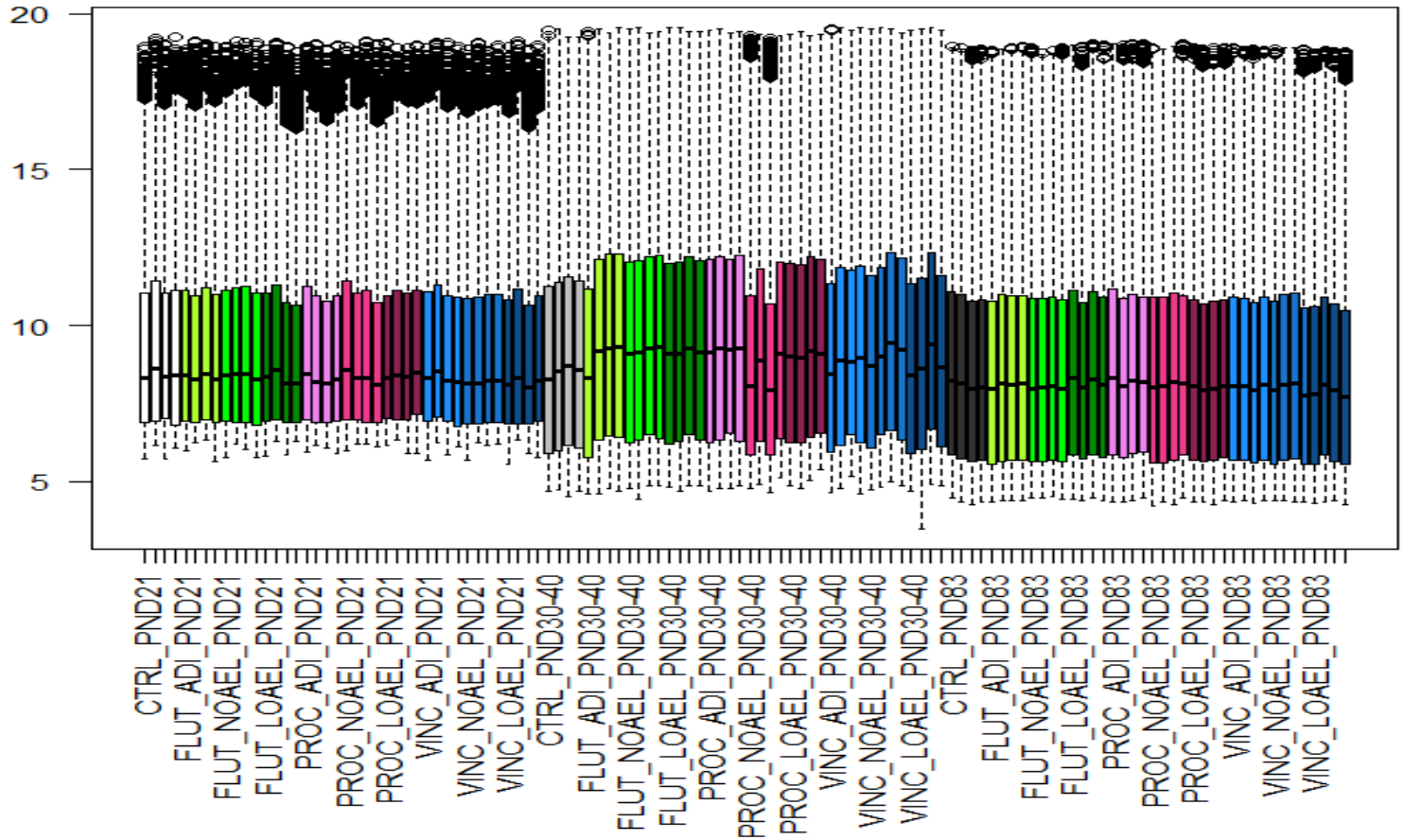
- For RNA was the RIN number low?
- For microarrays was there low dye incorporation?
- For RNA-Seq was the read depth low?
- For RNA-Seq – low % of mapped reads
- All methods – failure of manufacturers QC
- All methods – low signal to noise ratio
- All methods – Spiked in controls if present should pass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

Normalization

- The method of normalization used should be relevant to the type of data but in all cases should be the minimum necessary
- Normalization should be within sample only and generally not performed across the whole experiment (between array normalization). Between sample normalization methods, such as RMA (Robust Multi-array Average) would allow different samples to affect each other, such that the addition of new samples will result in changes of (normalized) expression values in the existing samples. This should be avoided.
- Test for normalization and outliers should be performed again as performed previously for the raw data

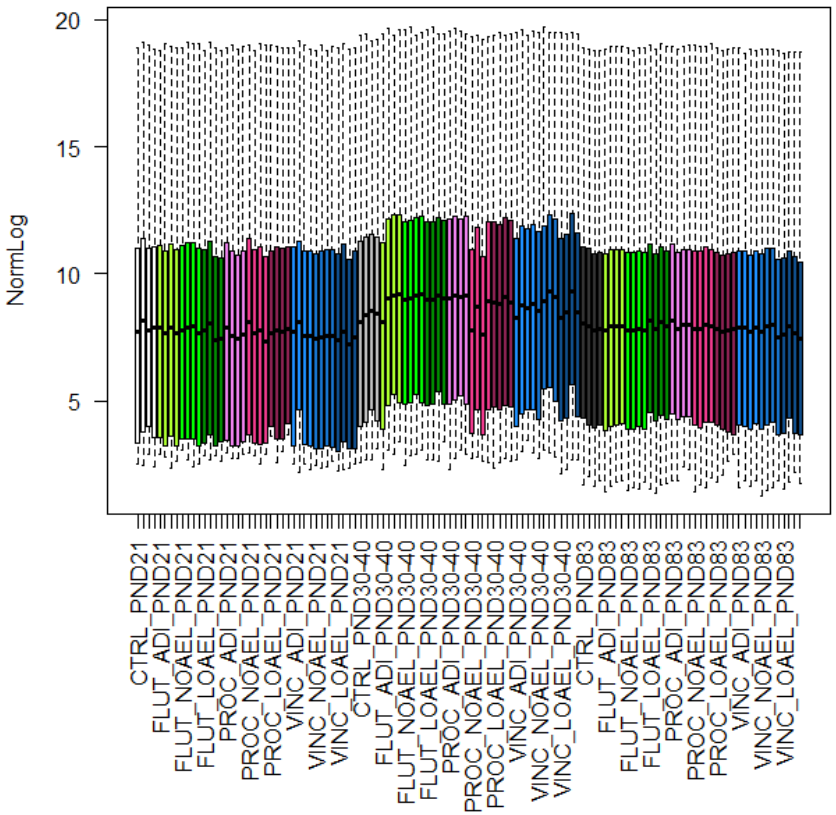
120 arrays, log; not normalised

gMedianData

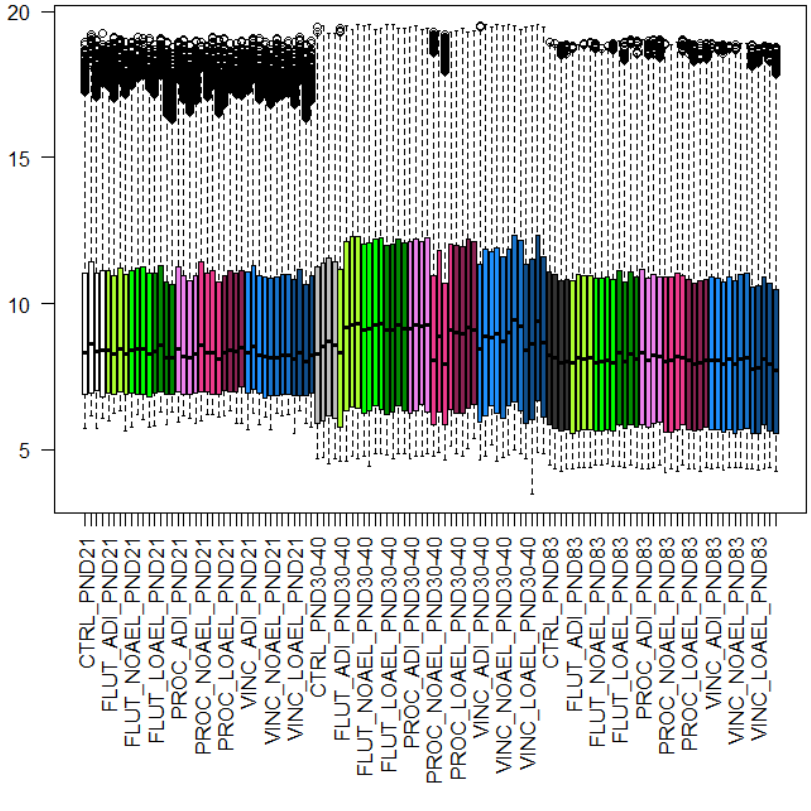


Starting data

gProcessedData

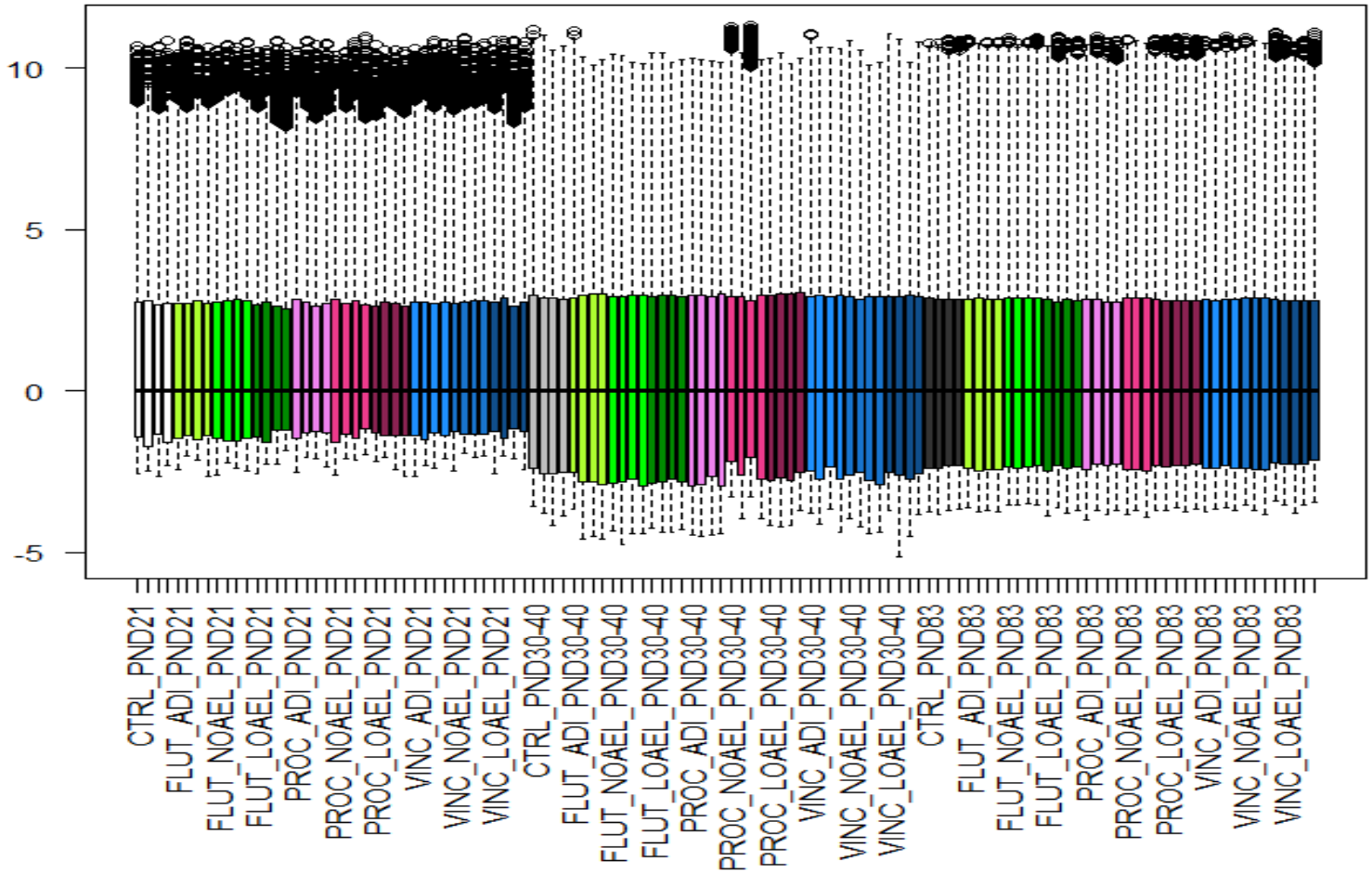


gMedianData



Median Centering Normalisation

gMedianData

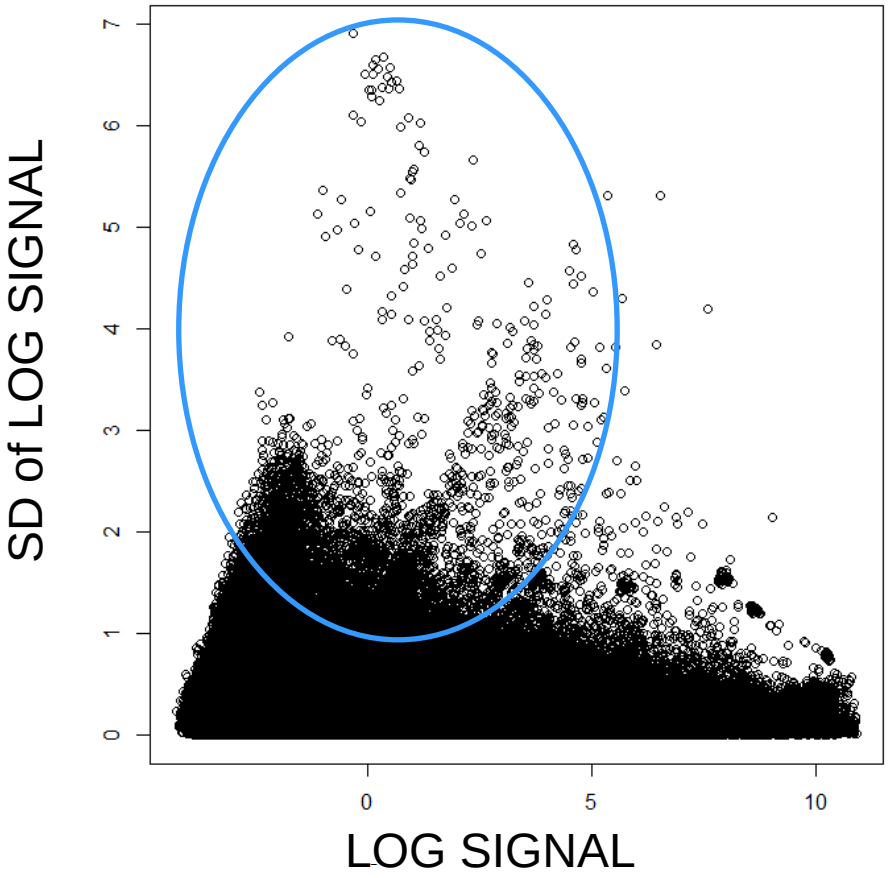


Statistics

- Panel considered that 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

Low signal strength

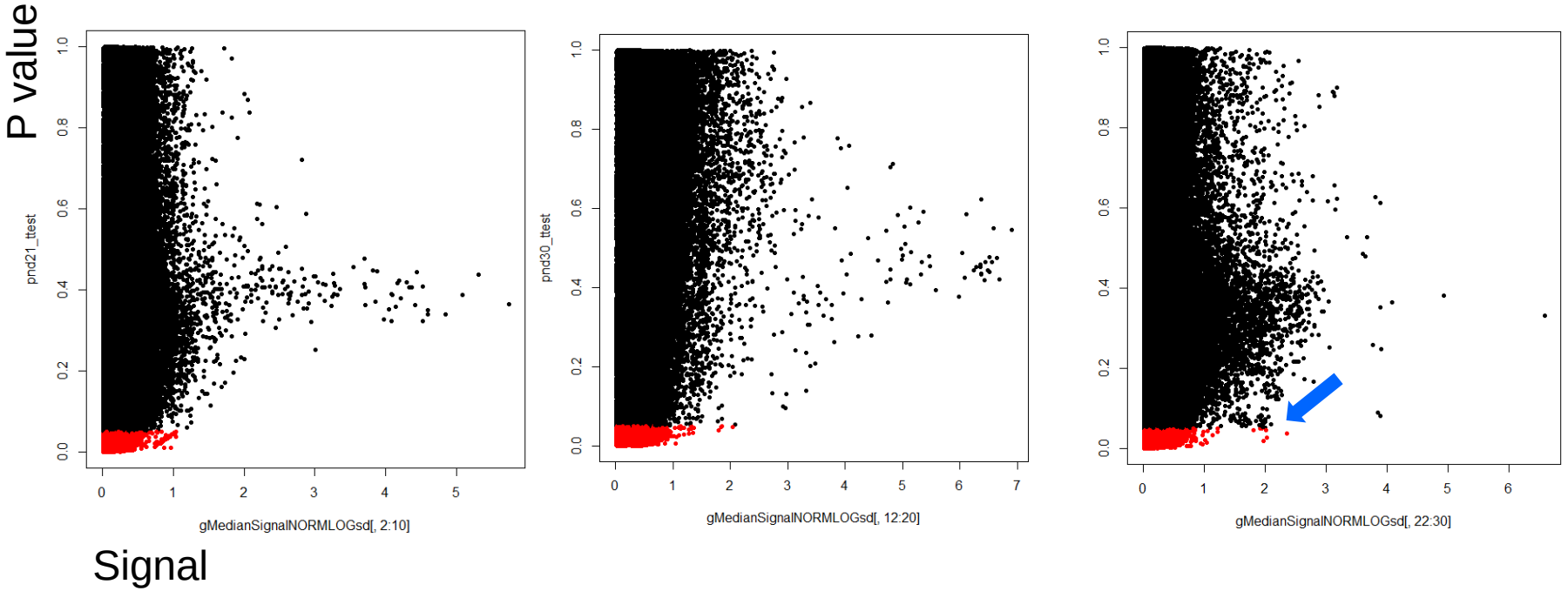
gMedianData



Values are the mean of the four biological replicates for each time and dose.

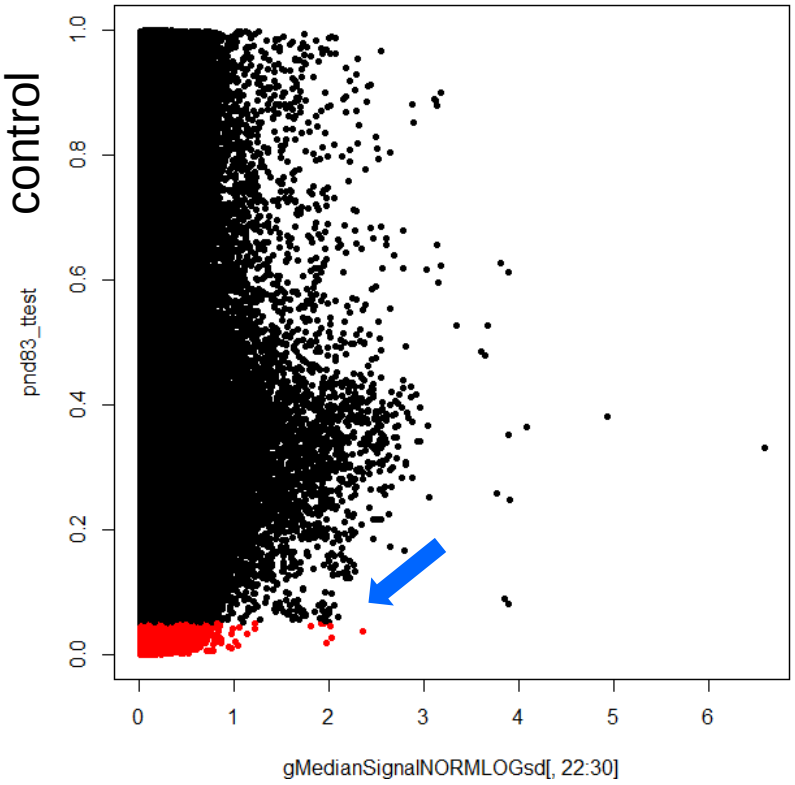
Greater variation in measurement is associated with a lower signal strength

Does the t-test deal with the high variance data?

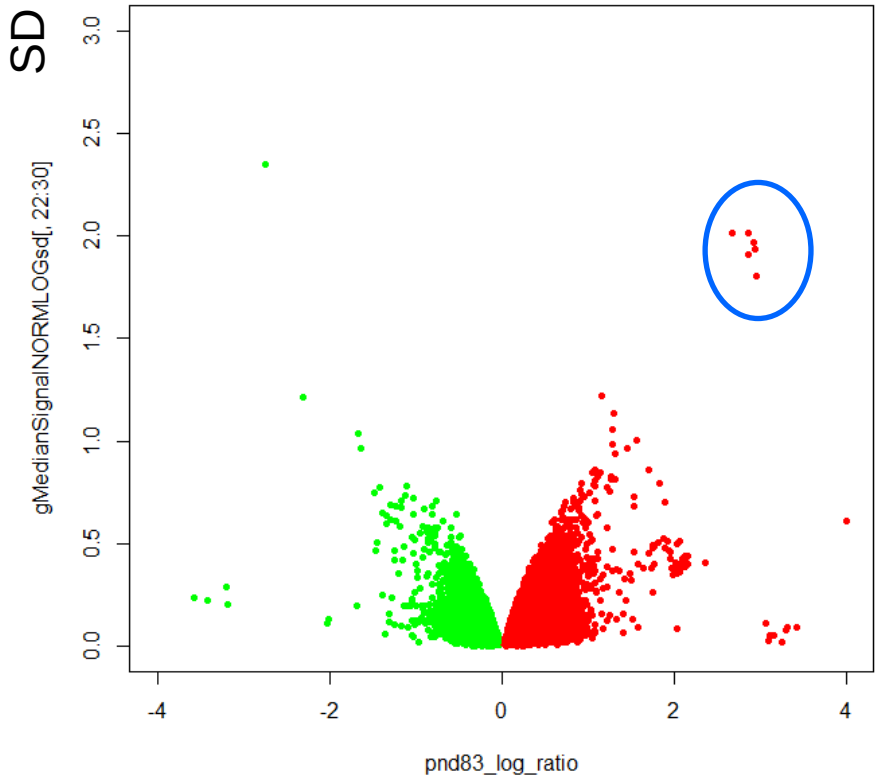


PND83

P value against control



SD



LogRatio

Where to from here

- The framework will be applied to some complex high throughput data generated as part of a project by ECETOC with BASF

- The framework will be developed written into a peer review paper and ECETOC report which will be finalised in 2016 for publication.

- OECD – A SPSF has been written and presented briefly to the Extended Advisory Group on Molecular Screening and
- Toxicogenomics (EAGMST). When the ECETOC written report is released it will be circulated to the EAGMST for comments and additionally may be presented at the June meeting.