Some Pertinent Findings from MAQC Related to Reproducibility of Gene Expression

Pierre R. Bushel, Ph.D., M.S.

Peer Review of Draft NTP Approach to Genomic Dose-Response Modeling Expert Panel Meeting
October 23rd - 25th, 2017
NIEHS, RTP, NC
Outline

- Brief overview of MAQC/SEQC
- How to assess reproducibility?
- A few ways MAQC explored reproducibility
  - Between sites
  - Data processing platform
  - Across platforms
  - Transcriptional response dependency
- Take home messages
MicroArray Quality Control (MAQC) Consortium

An FDA-led community wide crowd-sourced effort to assess technical performance and application of genomics technologies (microarrays, GWAS and next-gen sequencing) in clinic and safety evaluation.

**MAQC-I**

- **2005.2 – 2006.9**
- Assess reliability of microarrays
  - Repeatability
  - Reproducibility
- 137 participants
- 51 organizations
- 6 papers, 2006
  - In Nat Biotech

**MAQC-II**

- **2006.9 – 2010.10**
- Assess microarray based biomarkers
  - Clinical use
  - Safety evaluation
- 202 participants
- 97 organizations
- 13 papers, 2010
  - By Nat Publishing Group

**MAQC-III/SEQC**

- **2008.8 – 2014**
- Assess reliability of next-gen sequencing (RNA-seq) and compare it with microarrays
- >180 participants
- 73 organizations
- 10 manuscripts
  - 3 in Nat Biotech
  - 2 in Nat Comm
  - 2 Genome Bio
  - 3 SciData

**MAQC-IV/SEQC2**

- **2015 - Present**
- QC and reliable use of Whole Genome/Exome Sequencing (WG/ES) and Targeted Gene Sequencing (TGS) in clinical application and regulatory science research

Courtesy of W. Tong
Previously: Dr. Leming Shi, Professor Fudan University in Shanghai, China (formally with NCTR)  

Currently: Dr. Weida Tong, Director Division of Bioinformatics and Biostatistics NCTR
Javier Santoyo-Lopez, Laure Sambourg, Elia Stupka and Yiming Zhou

National Institutes of Health • U.S. Department of Health and Human Services

Courtesy of W. Tong
What Do We Mean by Reproducibility?

Under the same (or close to) conditions, study design, protocols and research tools, produce the same results from a previous experiment.
2016 Nature Survey on Reproducible Science

More than 70% of researchers have tried and failed to reproduce another scientist's experiments.
Gene Expression Reproducibility in the Context of MAQC/Sequence Quality Control

Experiment 1...N

Sample → RNA → Array or Seq → DEGs

Can we obtain the same DEGs?

Courtesy of W. Tong
Some Factors Evaluated in Relation to Reproducibility in a Toxicogenomics Study

- Study design
- Platform
- Between and within study sites
- Data processing/Normalization
- Treatment effect
MAQC-I: Rat Toxicogenomics Study

6-week-old Big Blue Fisher 344 male rats 12 weeks exposure

AA - Aristolochic acid; RDL - Riddelliine; CFY - Comfrey; CTR - Control

Kidney
- AA (10mg/kg) 0.9% NaCl
- CTL

Liver
- AA (10mg/kg) RDL 1mg/kg
- CFY 8% 0.9% NaCl
- CTL

Microarrays from Applied Biosystems, Affymetrix (2 sites), Agilent, and GE Healthcare.

Results are summarized in

Courtesy of W. Tong
Between Sites Reproducibility - Rat TGx Study

AA  CTL

Kidney

AA  RDL  CFY  CTL

Liver

Site 1

Site 2

Rank DEGs

t-test trt vs ctl

Courtey of W. Tong
Concordance of DEGs Between Two Study Sites

Percent of overlapping Genes (POG) = \( \frac{2 \times \text{intersect}(\text{DEG}_{\text{Site } 1}, \text{DEG}_{\text{Site } 2})}{\text{DEG}_{\text{Site } 1} + \text{DEG}_{\text{Site } 2}} \times 100 \)
Percentage of Overlapping Genes

- FC ranking
- FC + P < 0.05
- FC + P < 0.01
- P + FC > 2.0
- P + FC > 1.4
- P Ranking

MAQC-I: Coupling p-value with Fold Change
Improved Reproducibility

- Within site
- Between sites
- Between microarray platforms
MAQC-III/SEQC1: Let’s Up the Ante Regarding Between Platforms Reproducibility

Are the DEGs detected on the microarray platform detected on the mRNA-Seq platform?
MAQC-III/SEQC1: Toxicogenomics Study Design

**Samples:** Total RNA, then poly-A selection  
**Affy chip:** Rat 230_2.0, MAS5 and RMA normalizations  
**RNA-Seq:** Illumina HiScanSQ or HiSeq2000, 100bp PE  
Depth of 23-25 M reads  
6 Bioinformatics pipelines  

Results published in Wang et al. (2014) Nature Biotechnology
Root Mean Squared Distance

Measures the overall gene expression distance/deviation between pairs of samples $i$ and $j$

\[
RMSD_{ij} = \sqrt{\frac{\sum_g (I_{ig} - I_{jg})^2}{N_g}}
\]

$I_g$ is the log2 transformed expression level of gene $g$ in the corresponding sample and $N_g$ is the number of genes in the set

Compute the average RMSD for all pairs of replicates and compound treatments
Variability of Expressed Genes

High expressed genes have small variance.
qPCR Validation

DEG Validation by qPCR

- Microarray
- RNA-Seq

above median below median
The Chemicals Elicited a Wide Range of DEGs

*limma Treated vs control, FC > |1.5| and p-value < 0.05*

How does the strength of the perturbation affect the agreement between the two platforms?
Concordance Linearly Correlates with the Treatment Response

The stronger the system is perturbed, the higher the concordance

Courtesy of W. Tong
Concordance Increased at the Pathway Level

DEGs were mapped to GeneGo and IPA pathways. Concordance is the percentage of enriched pathways shared by the two platforms.

Courtesy of W. Tong
Take Home Messages

- Use fold change threshold coupled with a p-value cut off
- Filter out low expressed genes (primarily for microarray)
- Know your chemical’s transcriptional strength (if at all possible)
- Pathways perform better than genes individually
- It is of interest if some of these findings can be extended to other transcriptomics platforms such as Tempo-Seq
Acknowledgements

- MAQC/SEQC consortium participants
- Weida Tong for sharing some slides
- Jianying Li at NIEHS for informatics support
- Rick Paules and Scott at NIEHS for provision of samples and/or biological/toxicological insight