EPA
United States
Environmental Protection
Agency

Human Cell Line Derived Purified RNA and Cell Lysate Reference Materials for Transcriptomic Quality Control Assessments

Joshua A. Harrill

USEPA Center for Computational Toxicology and Exposure (CCTE)

Massive Analysis and Quality Control (MAQC) Society
April 29th, 2021

EPA United States Environmental Protection Agency

Disclaimer

The views expressed in this presentation are those of the author(s) and do not necessarily represent the views or policies of the U.S. Environmental Protection Agency, nor does mention of trade names or products represent endorsement for use.

EPA United States Environmental Protection Agency

Outline

- **Definitions**
- **MAQC / SEQC Projects**
  - Use of Transcriptomics Reference Samples
  - Characteristics of Transcriptomics Reference Samples
- **High-Throughput Transcriptomics (HTTr) at USEPA**
  - HTTr Experimental Design
  - Novel Reference Samples
  - Metrics for Assessing Transcriptomics Assay Reproducibility
- **Conclusions**

Background / Definitions

- **Transcriptomics** is the study of transcriptomes: i.e. all, or a portion of, the RNAs (typically mRNA) expressed from the genes of an organism, tissue or cell.

- **Reliability** is "*the extent of reproducibility of results from a test within and among laboratories over time, when performed using the same standardized protocol*" (OECD GD 34).

- Reliability has been identified as a key criteria for the acceptance and use of **New Approach Methodologies (NAMs)** data in regulatory setting (USEPA 2018).

- The **need** and **means** for demonstrating reliability of transcriptomics assays prior to use in regulatory settings has been recognized in the scientific literature:

  - **MACQ Consortium, 2006:** "*Concerns have been raised regarding the **reliability** and consistency and hence potential application of microarray ['omics] technology in clinical and regulatory settings…it follows that before this technology can be applied in clinical practice and **regulatory decision making**, standards, quality measures and consensus on data analysis methods need to be developed…['omics] studies need unified metrics and **standards** which can be used to identify suboptimal results and monitor performance…*"

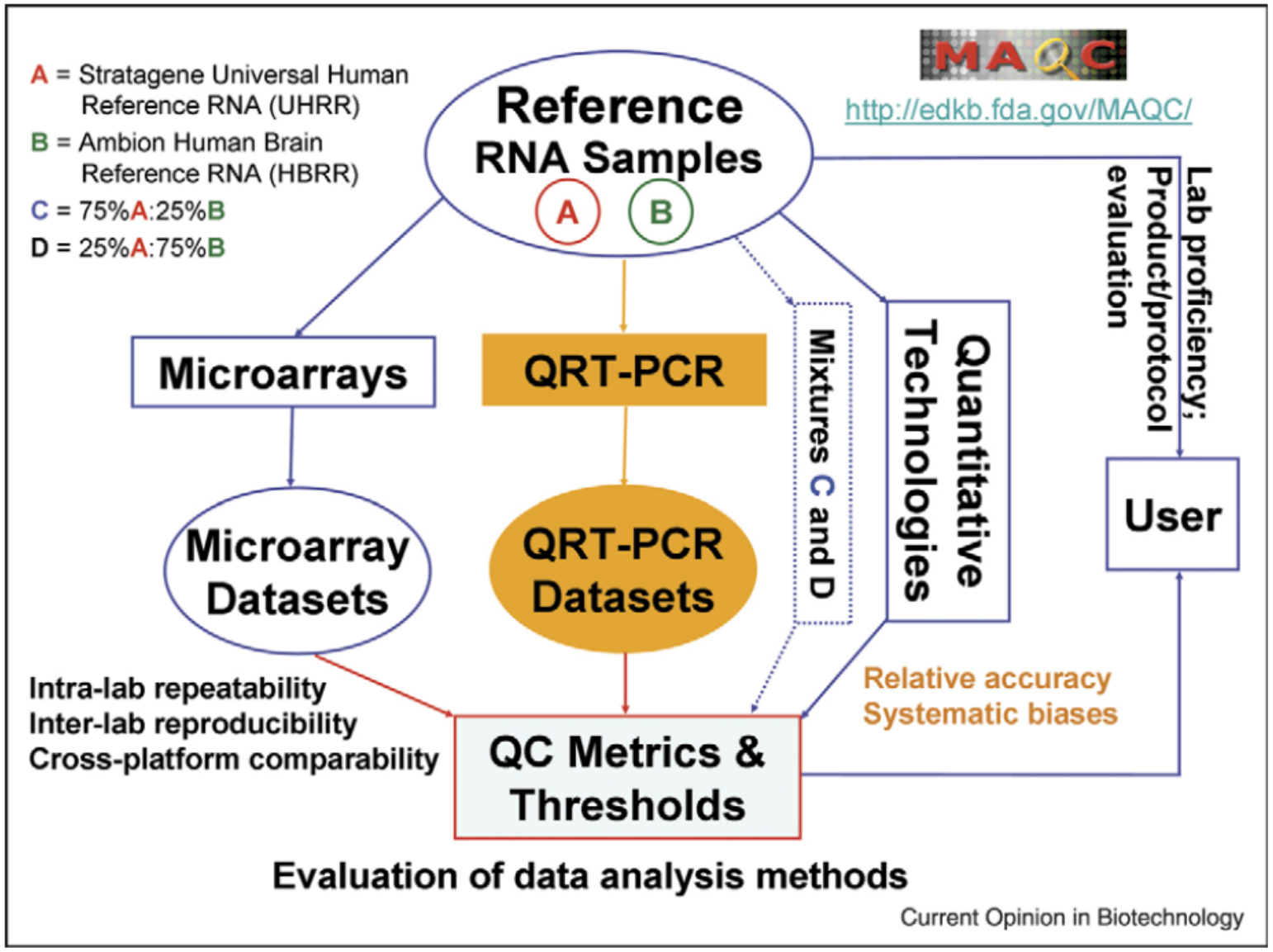
  - **ECETOC / Sauer et al., 2017:** "*The establishment of **calibrated RNA samples** and **reference datasets** were identified as crucial for an objective assessment of the performance of different microarray ['omics] platforms.*"

MAQC-I Project Outline

- **MAQC-I** project was initiated in 2005 to assess **reliability** and technical performance achievable using microarray technologies.

- Analysis of commercially-available **reference RNA samples**:
  - 40 test sites
  - 20 microarray platforms
  - 1329 individual microarrays

- The concordance of **differentially expressed genes (DEGs)** from paired reference samples was used as a metric to evaluate technical performance.



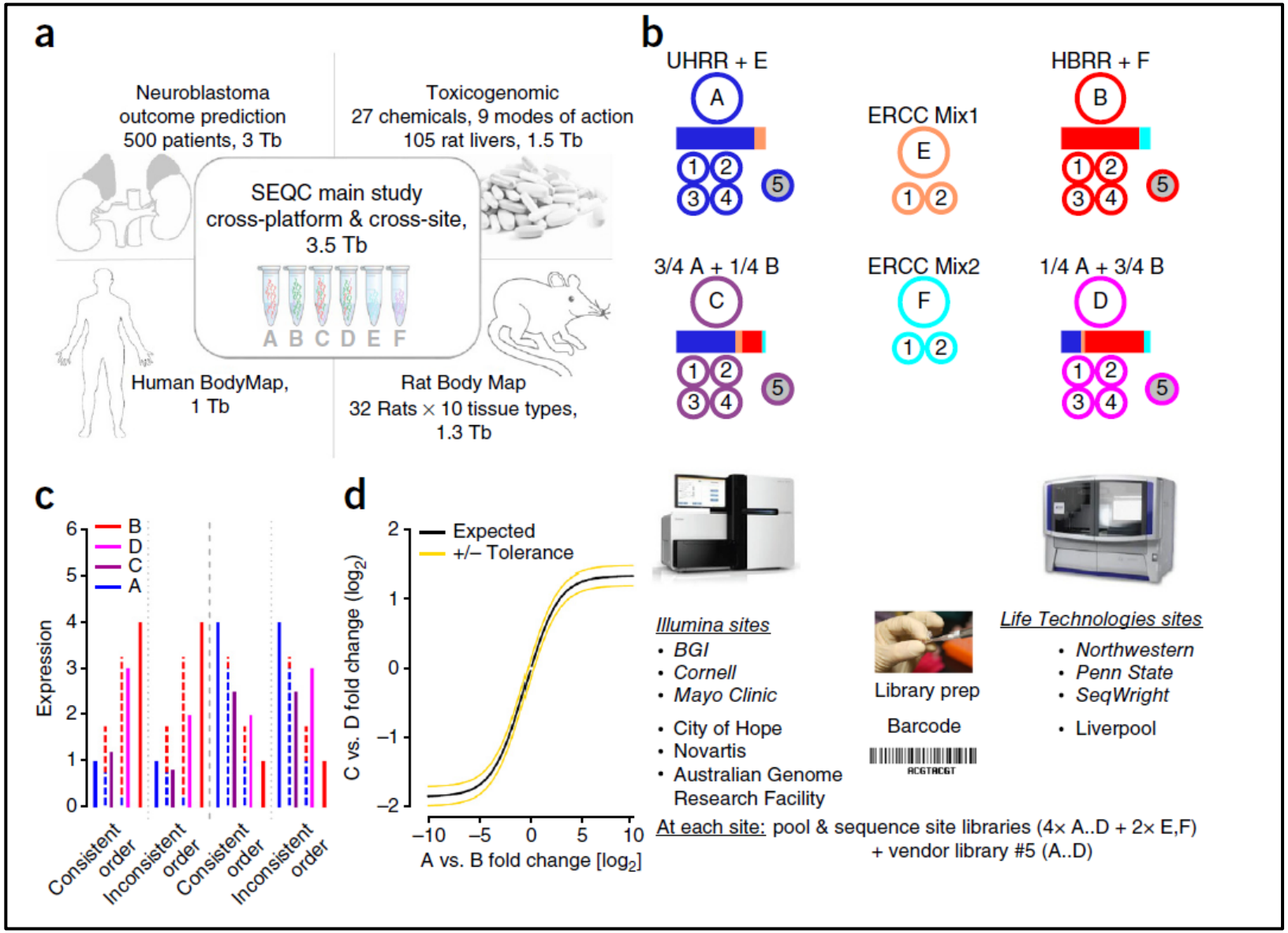
Shi et al. (2008). doi: 10.1016/j.copbio.2007.11.003

SEQC/MAQC-III Project Outline

- **SEQC/MAQC-III** project was initiated in 2013 to assess **reliability** and technical performance achievable using next generation sequencing (NGS) technologies.

- Similar approach to MAQC-I using **reference RNA samples**:
  - 6 official test sites (4 non-official)
  - 2 NGS sequencing platforms
  - 120 libraries

- **Concordance of DEGs** and other metrics based on observed versus expected expression in reference sample dilution series were used to evaluate technical performance



SEQC/MAQC-III Consortium (2014). doi: 10.1038/nbt.2957

What are Reference Samples for 'Omics?

- **Reference samples** are either commercially manufactured or generated in bulk in a research laboratory and contain a mixture of biological molecules in varying amounts that can be measured using an 'omics platform.
  - **Transcriptomics → RNA Species**
  - Metabolomics → Small molecules
  - Proteomics → Proteins / peptides

- "*Expression values generated on different ['omics] platforms cannot be directly compared because unique labeling method and probe sequences will result in variable signals for probes that hybridize to [measure] that same target. Alternatively,* ***the relative expression between a pair of sample types should be maintained across platforms.***" [MAQC, 2006]

- *The combination of* ***biologically different RNA sources*** *and* ***known titration differences*** *provides a method for assessing the relative accuracy of an 'omics platform based on differential detection* [MAQC, 2006, adapted].

Selecting Reference Samples for Transcriptomics

Selection Criteria:

- A pair of samples from the same species that is being evaluated in experiment of interest.
- Substantial overlap in the identity of transcripts contained within each sample.
- Large range of differential expression levels upon comparison of samples.
- Qualitatively similar to test samples (i.e. RNA diluent, RNA purification method, etc.)
- Compatibility with 'omics platform.
- Suitability for use across 'omics technologies.
- Widespread availability.

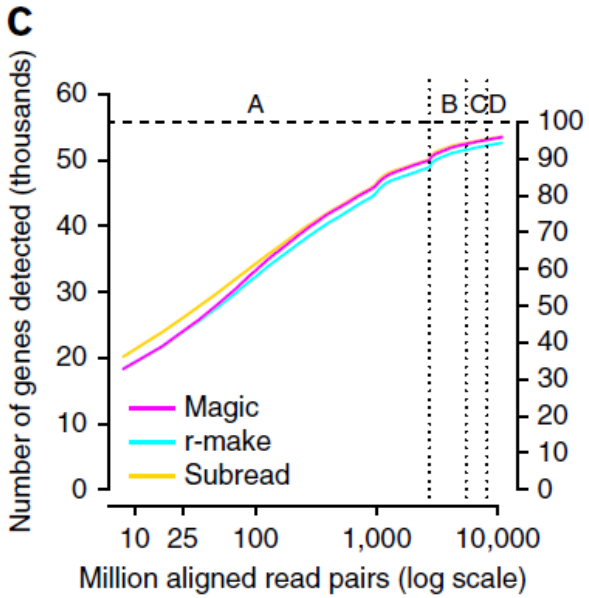
MAQC ID	Description
Sample A	**Stratagene Universal Human Reference RNA (UHRR)** A pool of RNA from 10 human cell lines Currently marketed by Agilent (Catalog #: 740000)
Sample B	**Ambion Human Brain Reference RNA (HBRR)** A pool of RNA from multiple human donors and brain regions Was marketed by ThermoFisher Invitrogen → Discontinued!
Sample C	75 % UHRR | 25 % HBRR
Sample D	25 % UHRR | 75 % HBRR

- The MAQC samples were chosen (not designed) based on commercially-available materials at the time.

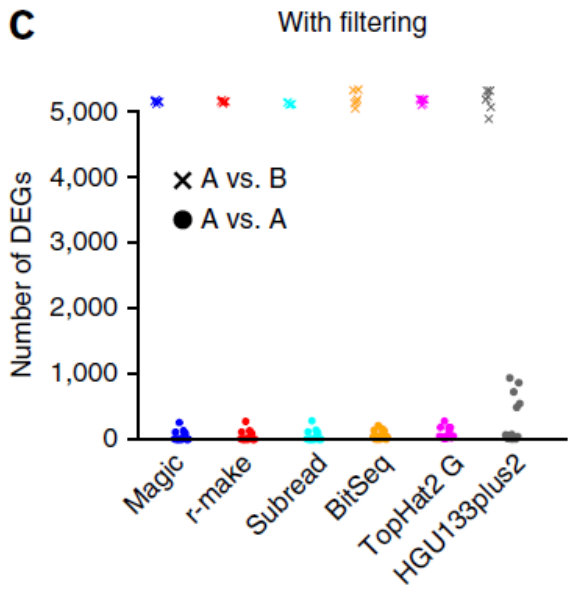
W.Tong - Personal communication

Characteristics of MAQC Reference Samples

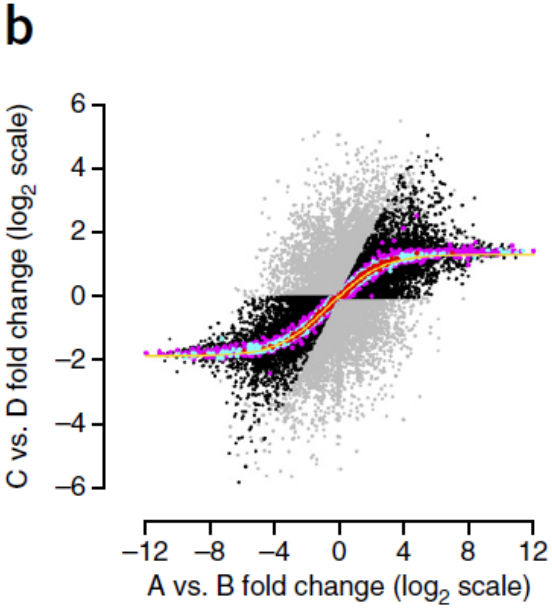
- Detectable expression of > 18,000 genes in each sample using RNA-Seq.
- Differential expression of > 5,000 genes upon comparison of Sample A & B.
- The range of differential expression spanned greater than 20 $\log_2$ units.



SEQC/MAQC-III Consortium (2014)
Figure 1C.



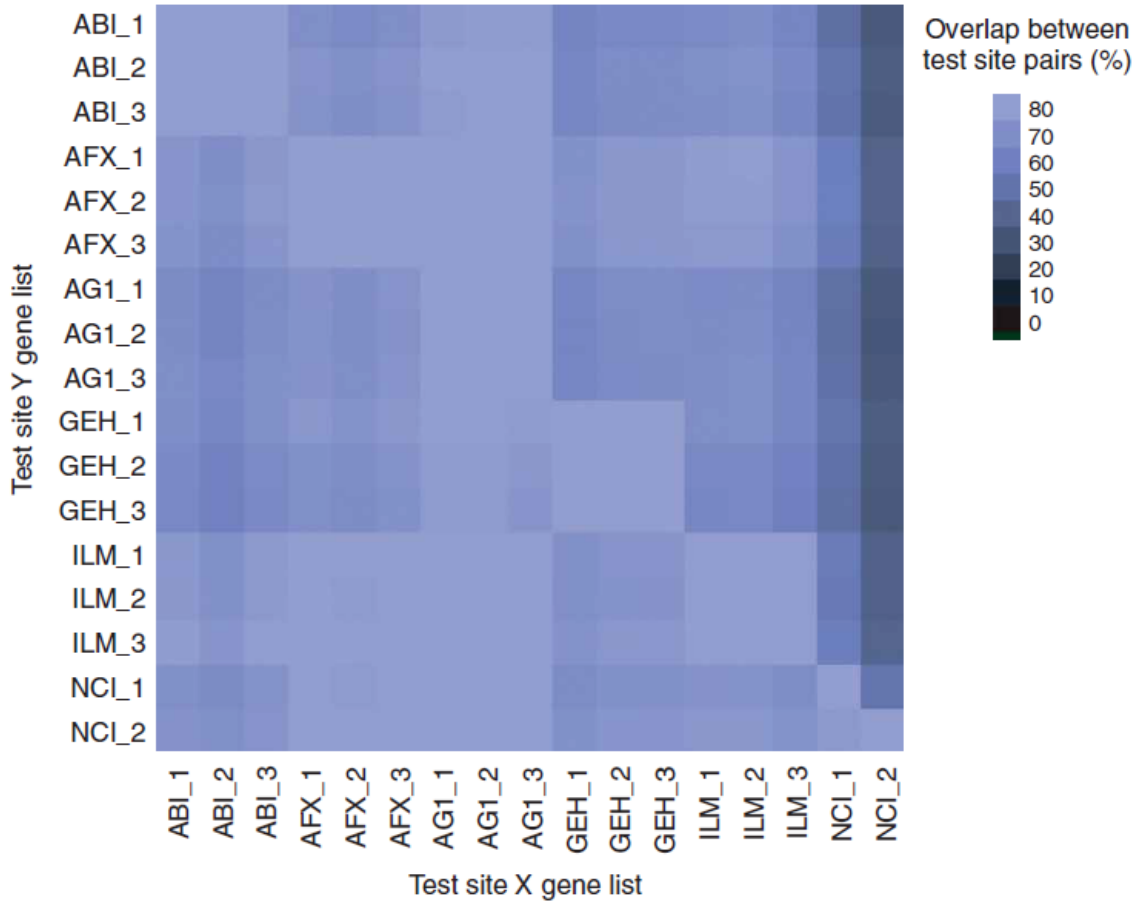
SEQC/MAQC-III Consortium (2014)
Figure 2C.



SEQC/MAQC-III Consortium (2014)
Figure 3B.

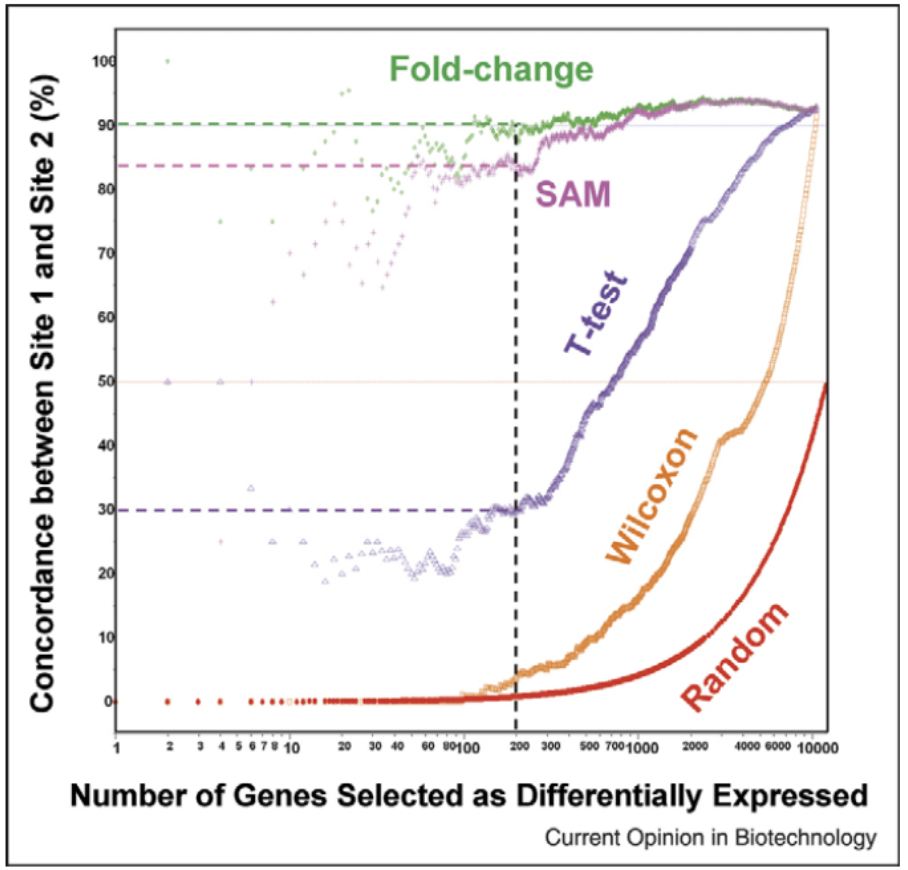
MAQC-I Project: Key Findings

Overlap in DEG Lists



MAQC Consortium (2006) doi: 10.1038/nbt1239

Concordance in DEGs Using Various Criteria

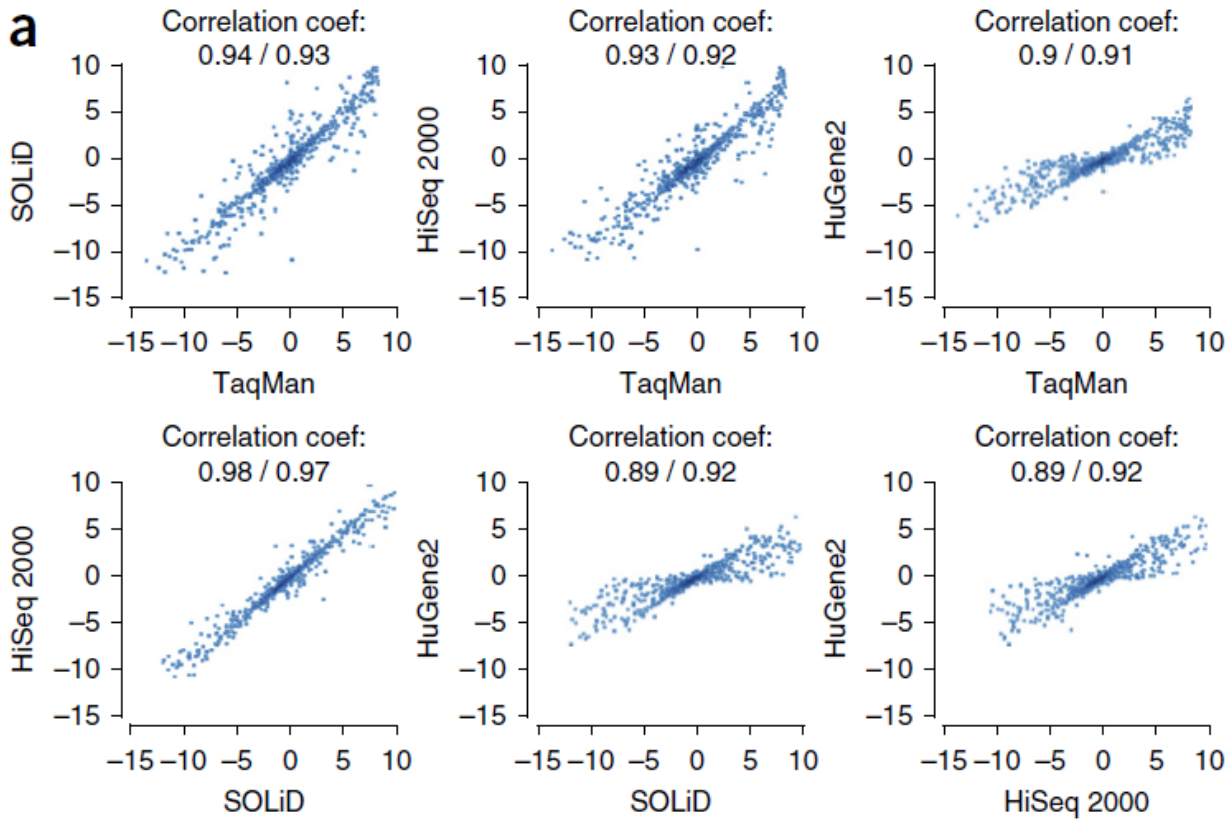


Shi et al. (2008). doi: 10.1016/j.copbio.2007.11.003

- Fold-change criteria for DEG determination results in higher concordance than *p*-value based criteria.

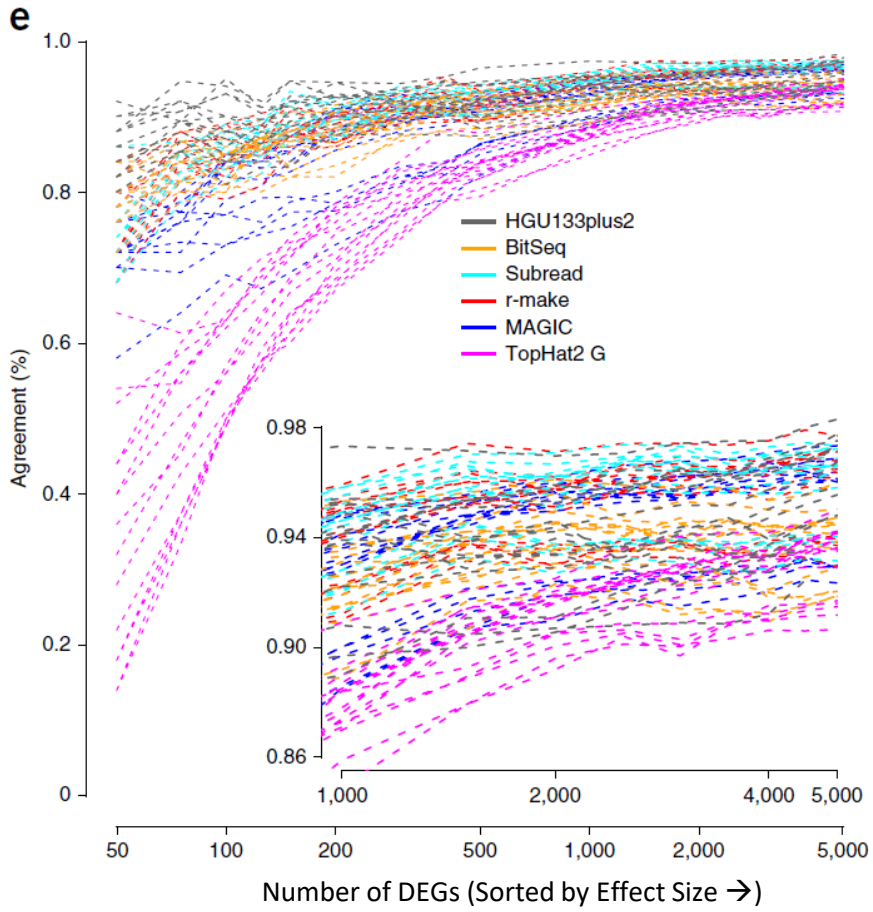
SEQC/MAQC-III Project: Key Findings

Correlation of Fold-Change Values



SEQC/MAQC-III Consortium (2014)
Figure 5a.

Agreement of DEGs



SEQC/MAQC-III Consortium (2014)
Figure 3e.

- *Reproducibility is highest when considering DEGs with large effect sizes (i.e. fold changes)*

EPA United States Environmental Protection Agency

Summary & Scientific Advancements

- The MAQC-I and SEQC/MAQC-III projects demonstrated that whole transcriptome technologies are **reliable** and can yield **reproducible** results across **laboratories and technology platforms**.

- Since these projects were completed, transcriptomics technologies have continued to evolve → ex. **targeted RNA-Seq**

- Increasing efficiency and declining costs associated with generating whole transcriptome profiles has made **high-throughput transcriptomics (HTTr)** a feasible option for chemical bioactivity screening in *in vitro* test systems.

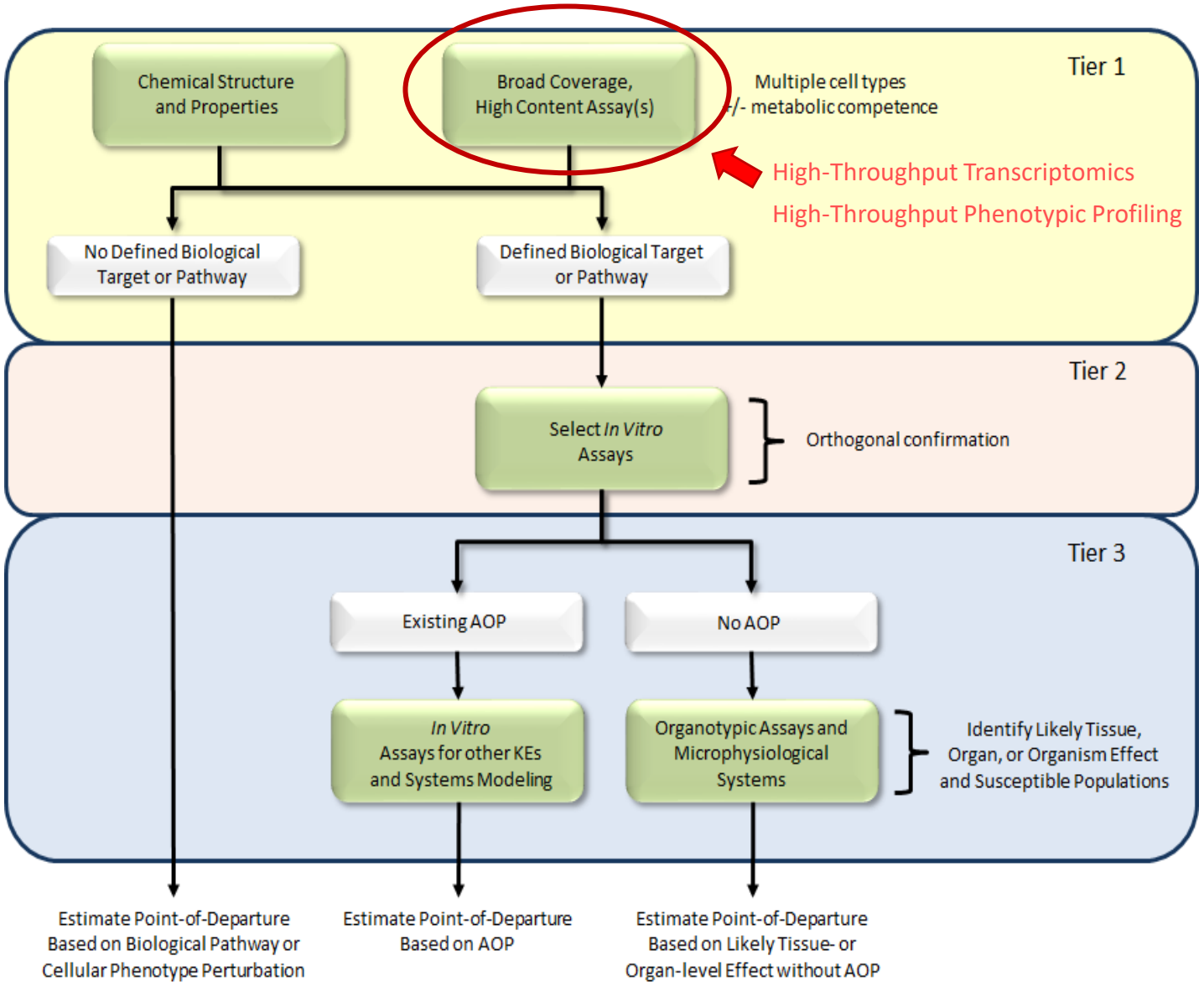
US EPA Tiered Hazard Evaluation Approach

- **New Approach Methodologies (NAMs)** are any technology, methodology, approach or combination thereof that can be used to provide information on chemical hazard and risk that avoids the use of intact animals.

- NAMs are a potential means to **reduce** the use of animals in toxicity testing and **accelerate** the pace of chemical risk assessment.

- US EPA CompTox Blueprint advocates the use of **high throughput profiling (HTP) assays** as the first tier in a NAMs-based hazard evaluation approach.

- **HTP assay criteria:**
  1. Yield bioactivity profiles that can be used for **potency estimation**, **mechanistic prediction** and evaluation of **chemical similarity**.
  2. Compatible with multiple human-derived culture models.
  3. Concentration-response screening mode.
  4. Cost-effective.

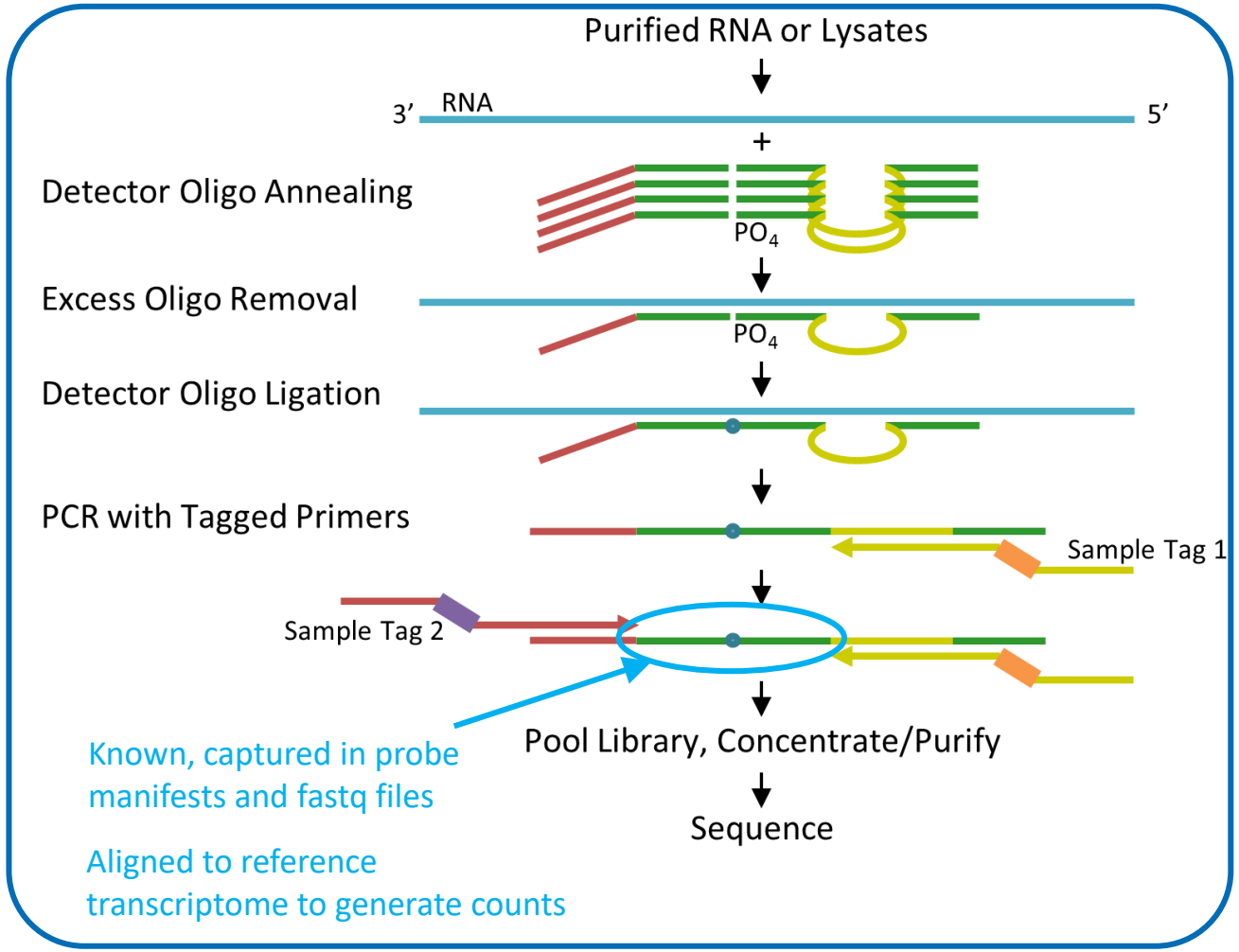


The NexGen Blueprint of CompTox as USEPA Tox. Sci. 2019; 169(2):317-322

Templated Oligo with Sequencing Readout (TempO-Seq)

- The **TempO-Seq** human whole transcriptome assay measures the expression of greater than 20,000 transcripts.
- Requires only picogram amounts of total RNA per sample.
- Compatible with purified RNA samples or **cell lysates**.
- Lysates are barcoded according to sample identity and combined in a single library for sequencing using industry standard instruments.
- Scalable, targeted assay:
  - 1) specifically measures transcripts of interest
  - 2) ~50-bp reads for all genes
  - 3) requires less flow cell capacity than RNA-Seq

TempO-Seq Assay Illustration



Yeakley, et al. PLoS One. 2017 May 25;12(5):e0178302.

Use of Reference Samples in HTTr Screening (1)

- Reference samples are intended to provide objective evaluation(s) of the **technical performance** of an 'omics assay…**NOT the biological response** of an *in vitro* test system.
  - Use reference treatments for this latter purpose.

- Processed in parallel with test samples → they should be subject to the same manipulations and assay conditions as test samples.

- Implemented in a manner that facilitates monitoring of consistency of transcriptomics assay results generated within studies, across studies, across laboratories and over time.

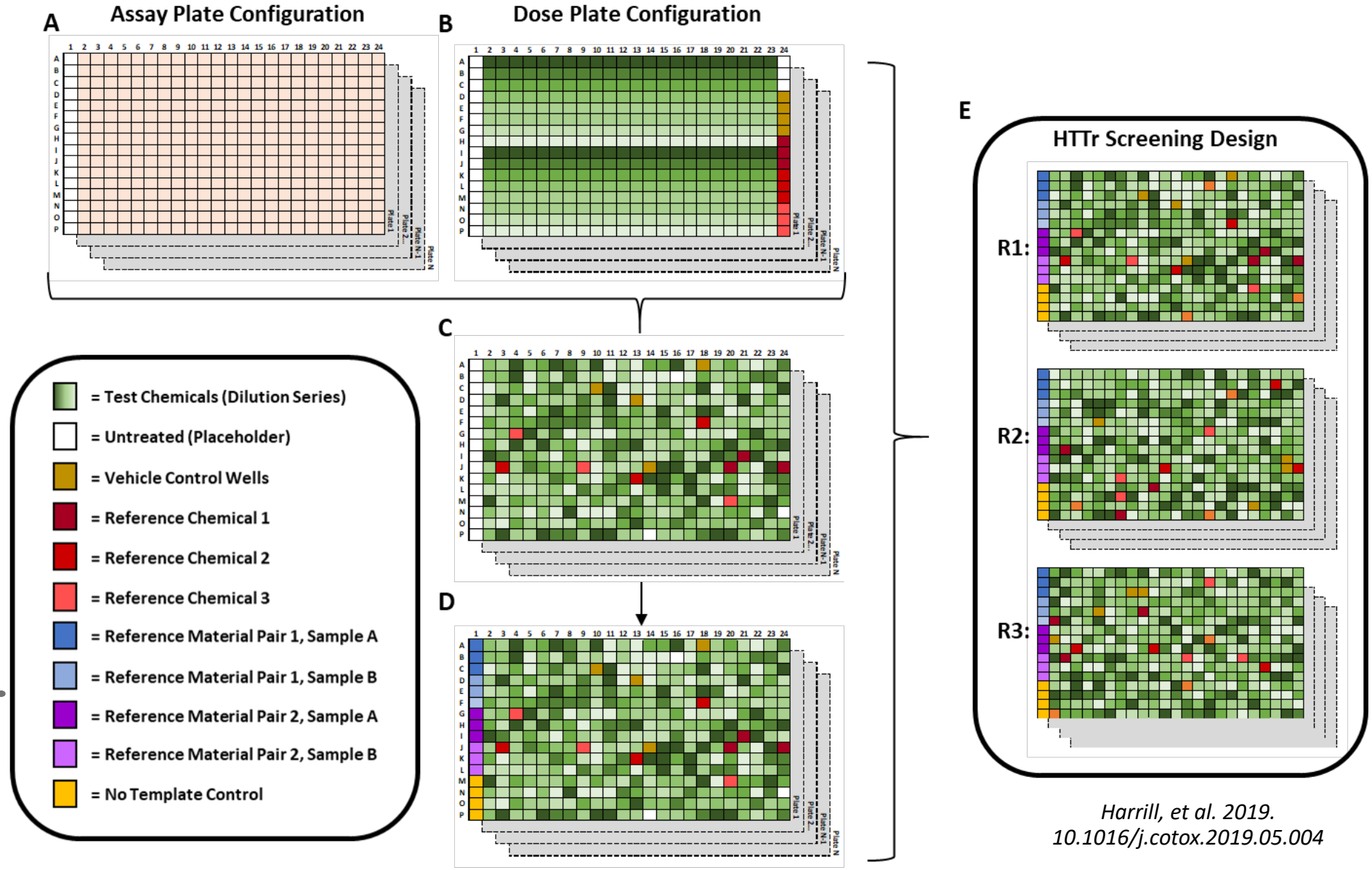
Use of Reference Samples in HTTr Screening (2)

Reference Material Pair 1

- Purified RNAs

Reference Material Pair 2

- Bulk Lysates



Harrill, et al. 2019.
10.1016/j.cotox.2019.05.004

Reference Samples: History of Use for HTTr

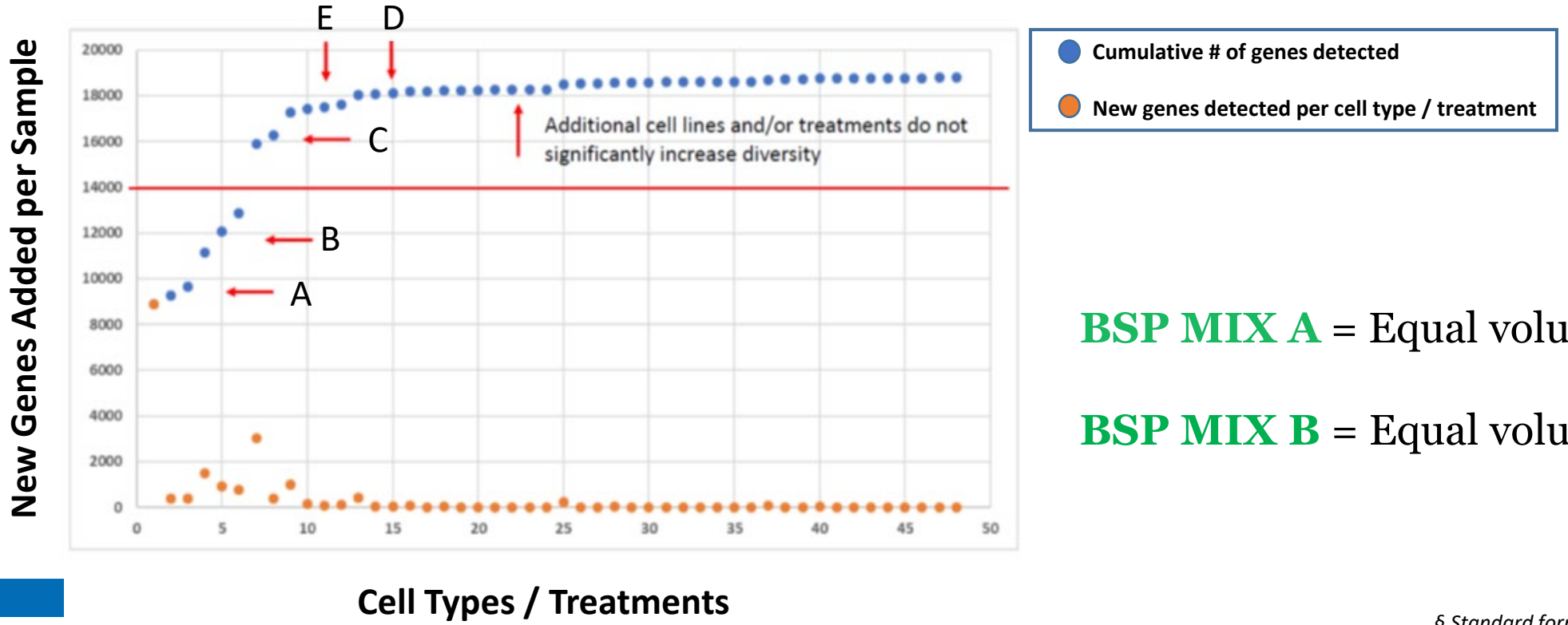
"Early days" (2017-2020) at US EPA:

Name	Description	Observations
Reference Pair #1 (purified RNA)	Takara UHRR (636690) Takara HBRR (636530)	• Not optimal for evaluating performance of cell-lysate compatible transcriptomics assays. • As with MACQ, were "selected" not "designed" • Finite resource.
Reference Pair #2 (bulk lysates)	**MCF-7 Cells** DMSO (0.5%) Treated TSA (1 µM) Treated	• Generated at US EPA • Fewer genes detected compared to Reference Pair #1. • Range of FC values smaller than Reference Pair #1. • Gene abundance was highly correlated within Reference Pair #2.

- US EPA perceived a need to develop **replenishable** human-derived transcriptomics reference samples that are:
  - Compatible with multiple assay technologies
  - Available as both purified RNA and cell lysates
  - Yield reproducible fold-change profiles across production batches

Engineering of Transcriptomics Reference Samples (1)

- Paired reference samples were prepared by combining the genetic material from different human-derived cell lines cultured under different conditions.
- Formulated to mimic the performance characteristics of MAQC samples.
- Prepared as both purified RNAs and cell lysates (*BioSpyder, Inc*).



BSP MIX A = Equal volume of A, B, C, D

BSP MIX B = Equal volume of A, B, C, E

§ Standard formulation, prep and analysis by BioSpyder, Inc.

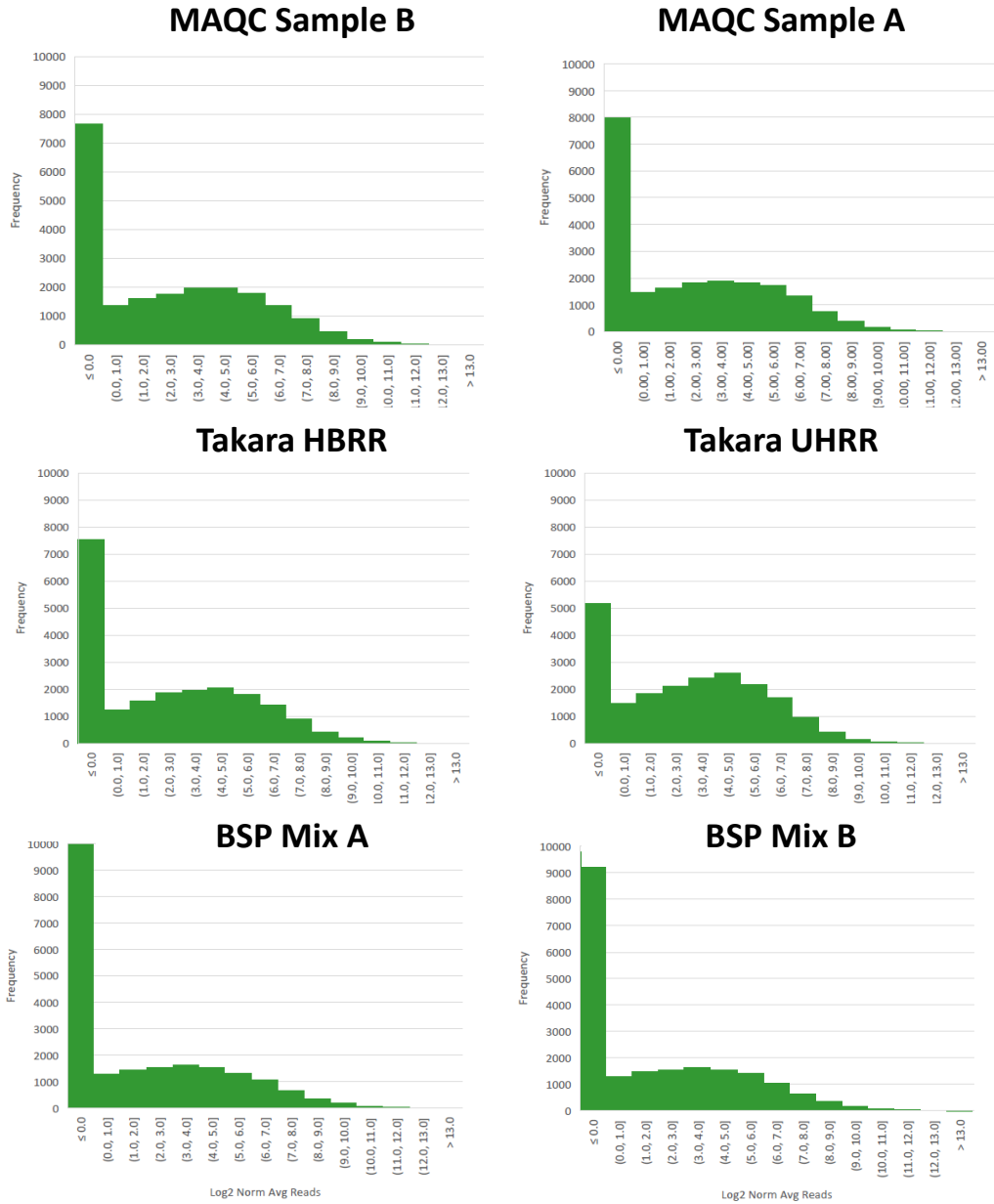
Engineering of Transcriptomics Reference Samples (2)

Sample	Number of Genes >5 counts at a depth of 8M (TempO-Seq)
BSP_RNA_A	13,962
BSP_RNA_B	13,779
BSP_LYSATE_A	14,919
BSP_LYSATE_B	14,565

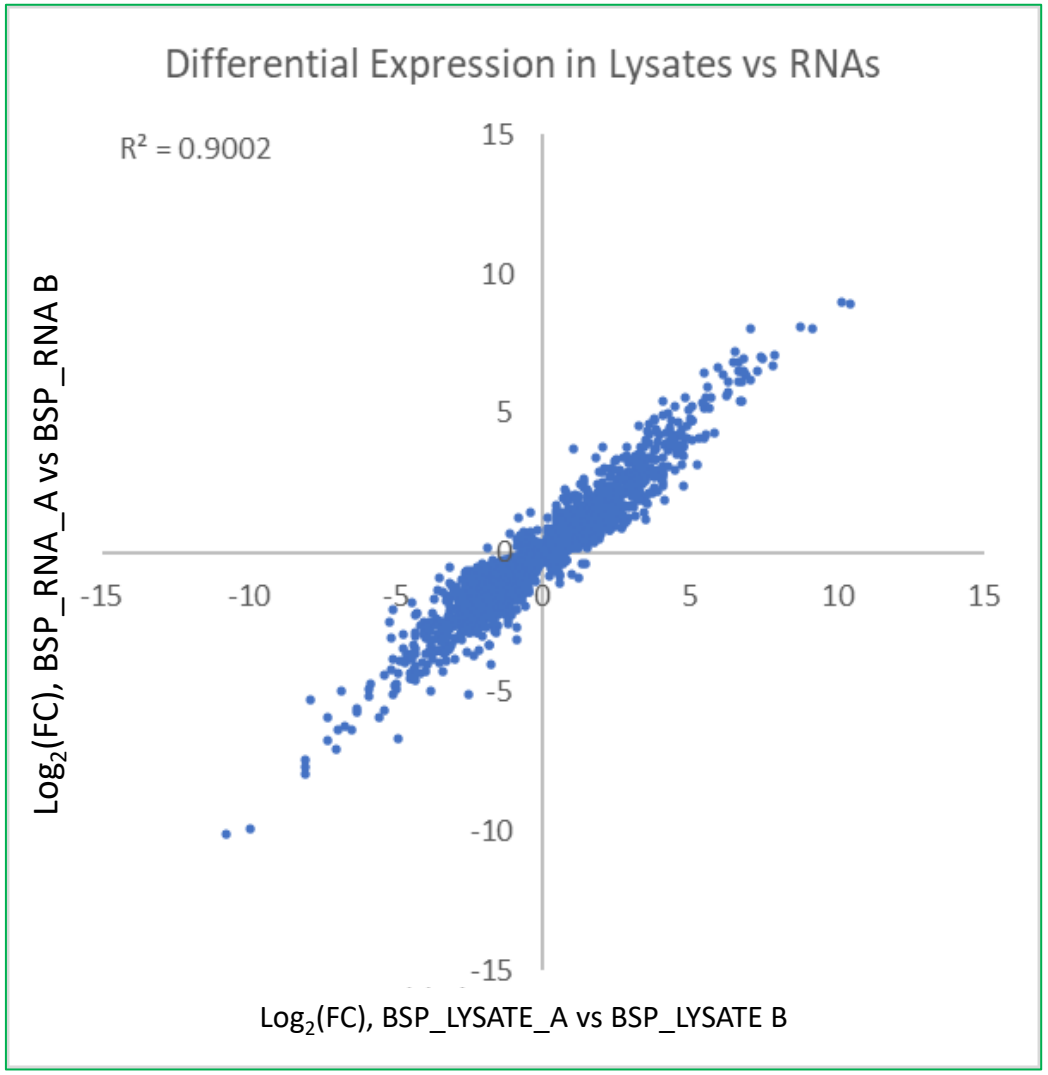
Samples	Number of Genes >5 counts in Common
RNA_A & RNA_B	12,881
LYSATE_A & LYSATE_B	13,546

§ Standard formulation, prep and analysis by BioSpyder, Inc.

- Similar numbers of detected genes in engineered reference samples versus MAQC or Takara samples.



Engineering of Transcriptomics Reference Samples (3)



** DESeq2-moderated fold changes*

	Raw Log_2 Fold Difference	
Comparison	**Min**	**Max**
MAQC	-16.43	13.30
Takara	-7.04	16.16
BioSpyder	-13.9	13.76

- Log_2(FC) profiles of BioSpyder RNA and lysates are highly correlated.
- Similar range of log_2(FC) values in BioSpyder reference samples versus MAQC samples.
- Takara had a slightly narrower range.

§ Standard formulation, prep and analysis by BioSpyder, Inc.

HTTr Screening: Experimental Designs

Parameter	Screen # 1	Screen # 2
Cell Type(s)	HepaRG (2-D)	
Screening Mode	Chemical Concentration Response	
Concentrations	8 Concentrations (3.5 $\log_{10}$ units)	
Time Points	24 hours	
Biological Replicates	3	
Assay Format	TempO-Seq Human Whole Transcriptome (v2)	
Target Read Depth	3 million mapped reads / sample	
Chemicals	1,218 ToxCast Chemicals	336 Reference Chemicals
HTTr Assay Plates	87	24
HTTr Samples	32,886	9,072
Reference Materials	Takara UHRR Takara HBRR DMSO Bulk Lysate Trichostatin A Bulk Lysate	BioSpyder Reference RNA A BioSpyder Reference RNA B BioSpyder Lysate A BioSpyder Lysate B

Metrics for Assessing Transcriptomics Assay Reproducibility

- **Sample-Based Metrics**
  - Gene abundance distribution
- **Comparison of Like Samples**
  - Correlation of expression values
  - D-statistic (House et al. 2017)
- **Comparison of Paired Samples**
  - Differential Gene List Overlap**
  - Dynamic range of FC
  - Correlation of L2FC profiles
  - Gene signature / pathway enrichment concordance (*if applicable*)

- Interpretation requires defining typical performance ranges.

Lesson Learned from MAQC*

A straightforward approach of FC ranking plus a non-stringent p-cutoff can be successful in identifying reproducible gene lists, whereas ranking and selecting differentially expressed genes solely by the t-test statistic **predestine a poor concordance in results**, in particular for shorter gene lists, due to the relatively unstable nature of the variance (noise) estimate in the t-statistic measure.

Furthermore, the impact of normalization methods on the reproducibility of gene lists becomes minimal when the fold change, instead of the p-value, is used as the ranking criterion for gene selection

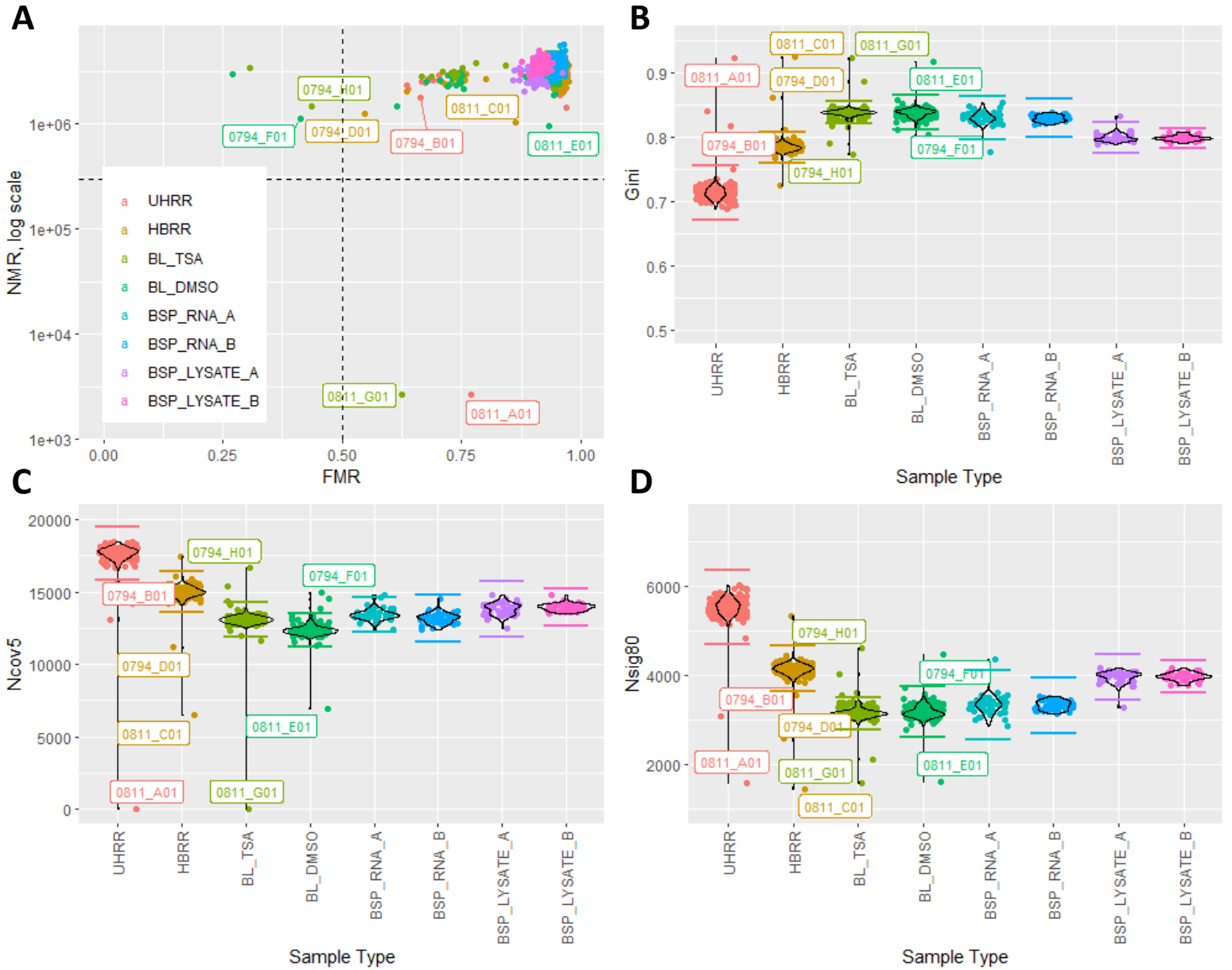
Sample Based Metrics for Transcriptomics Assay Performance

Sample-Based Metrics

Abbreviation	Description
NMR	Number of mapped reads, defined as a sum of total read counts over all detected probes.
FMR	Fraction of uniquely mapped reads
NCov5	The number of probes with at least 5 uniquely mapped reads
NSig80	The number of probes capturing the top 80% of total mapped counts in a sample
Gini	A measure of inequality in a distribution. Computed based on the distribution of raw counts for all probes including those with 0 aligned reads.

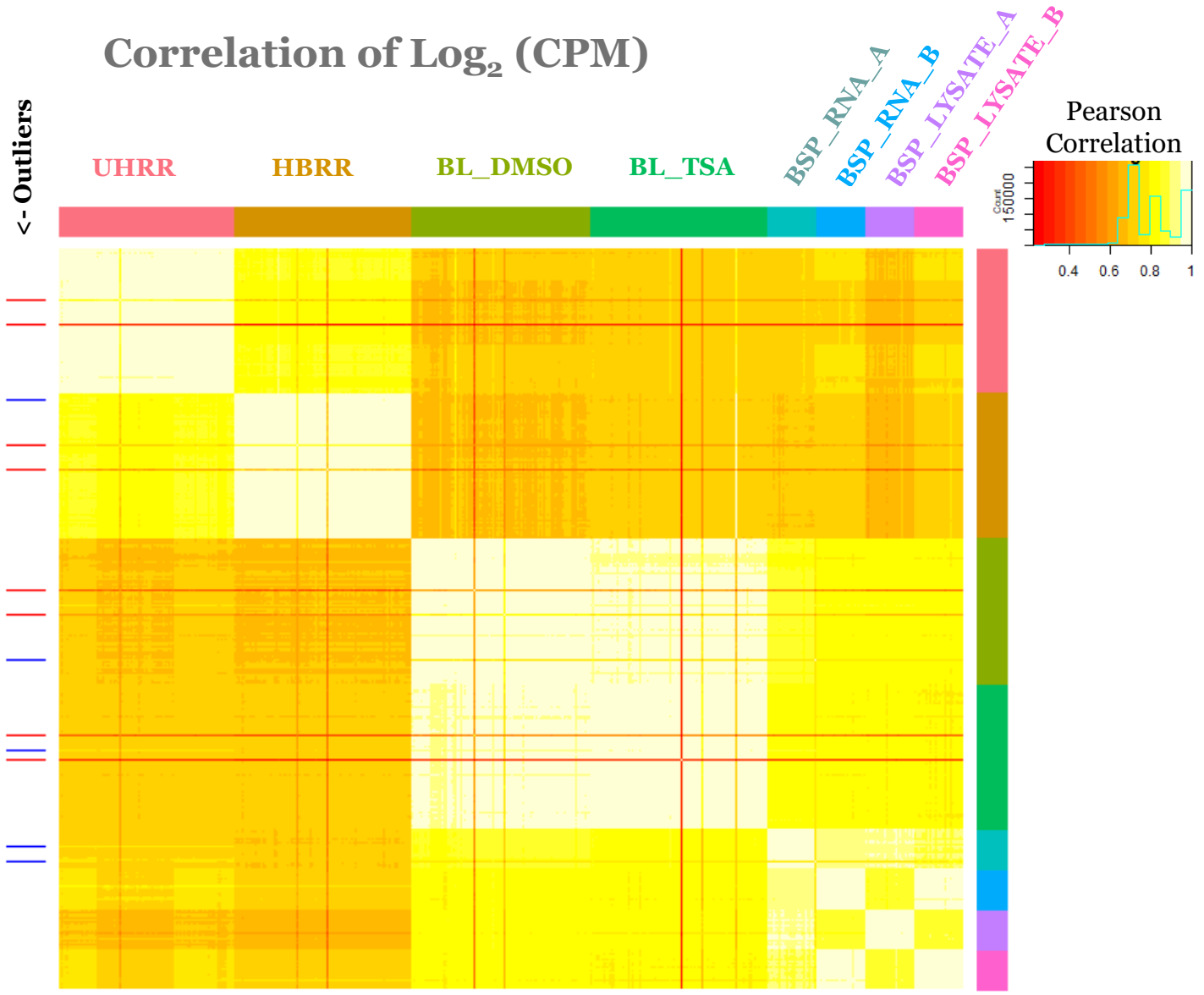
Harrill et al. (2020) doi: 10.1093/toxsci/kfab009

Flagging gates for gene abundance distribution metrics are based on Tukey's Outer Fence principle



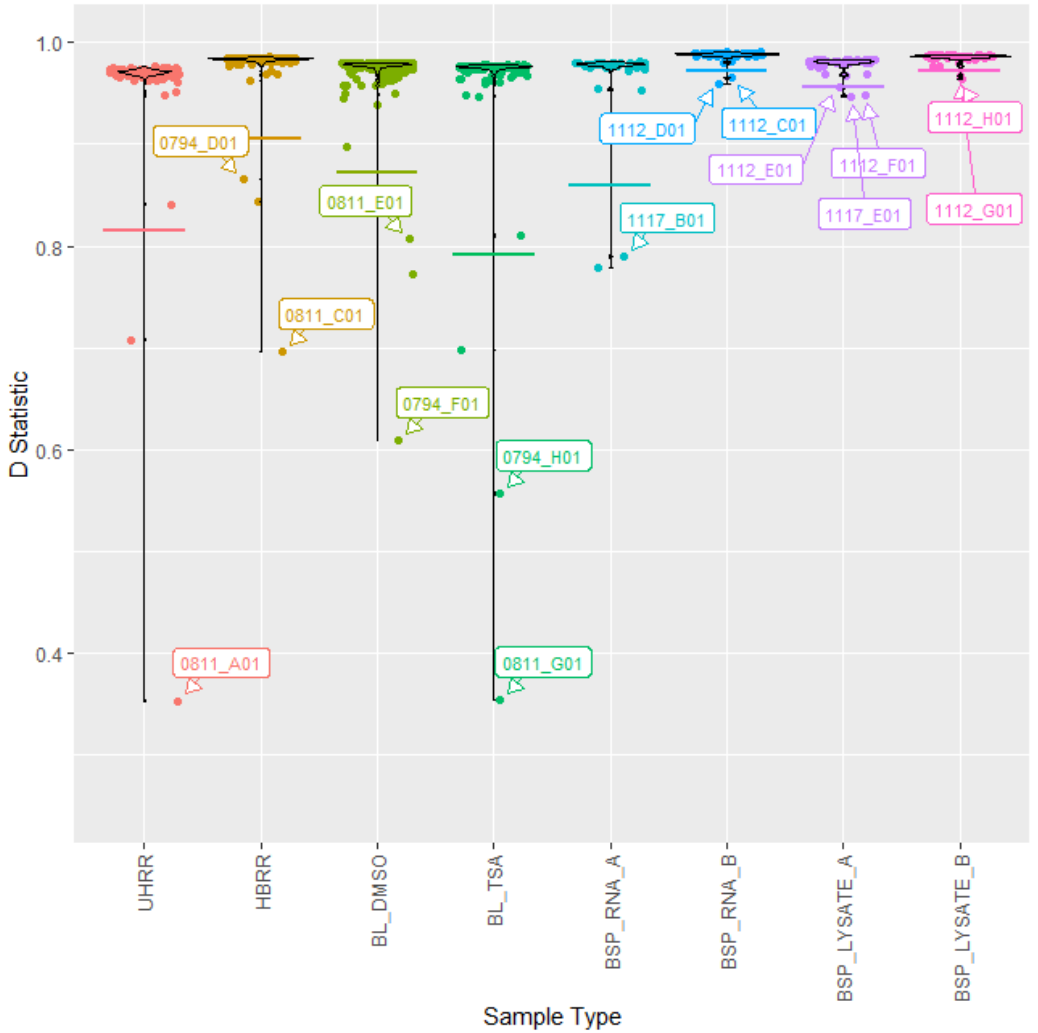
Comparison of Like Samples for Transcriptomics Assay Performance

Correlation of Log_2 (CPM)



D-Statistic (House et al. 2017)

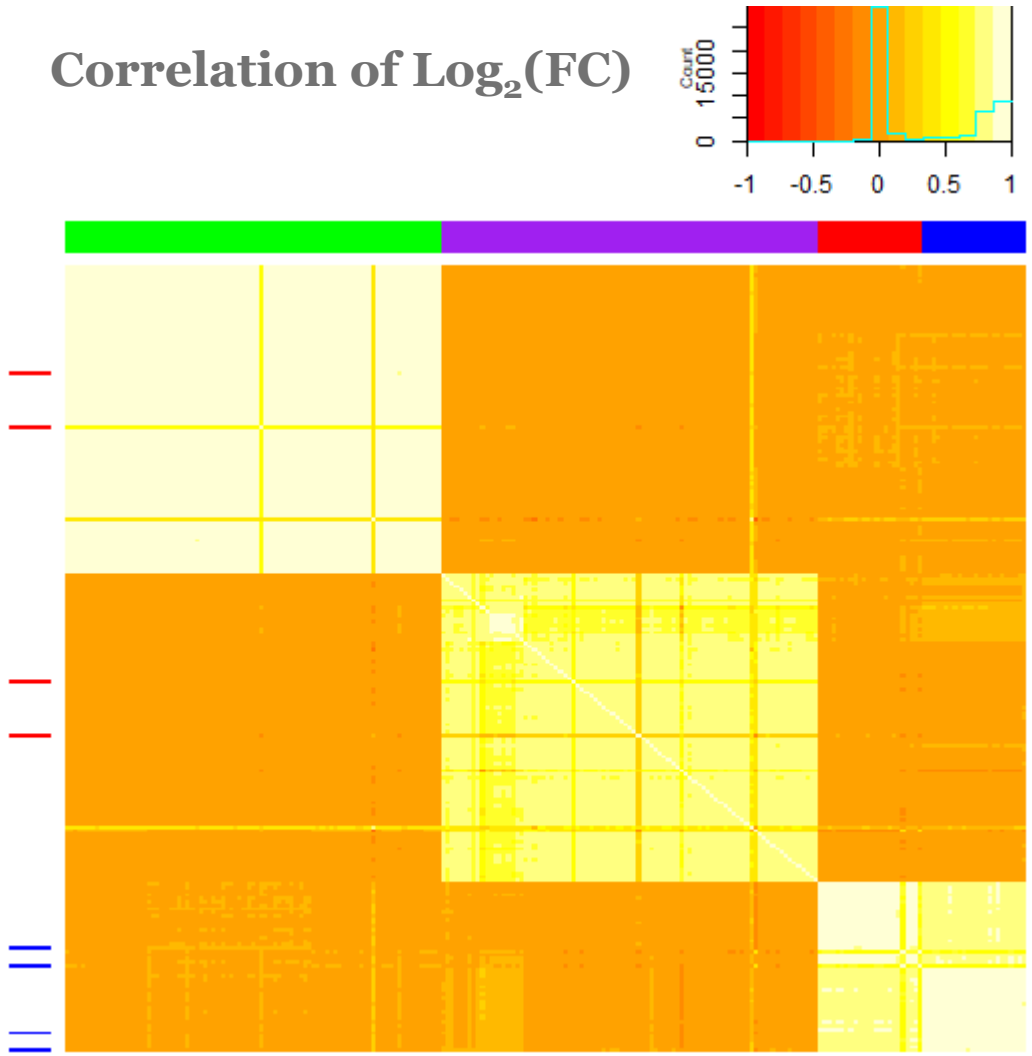
Mean correlation of Log_2(CPM) to like samples



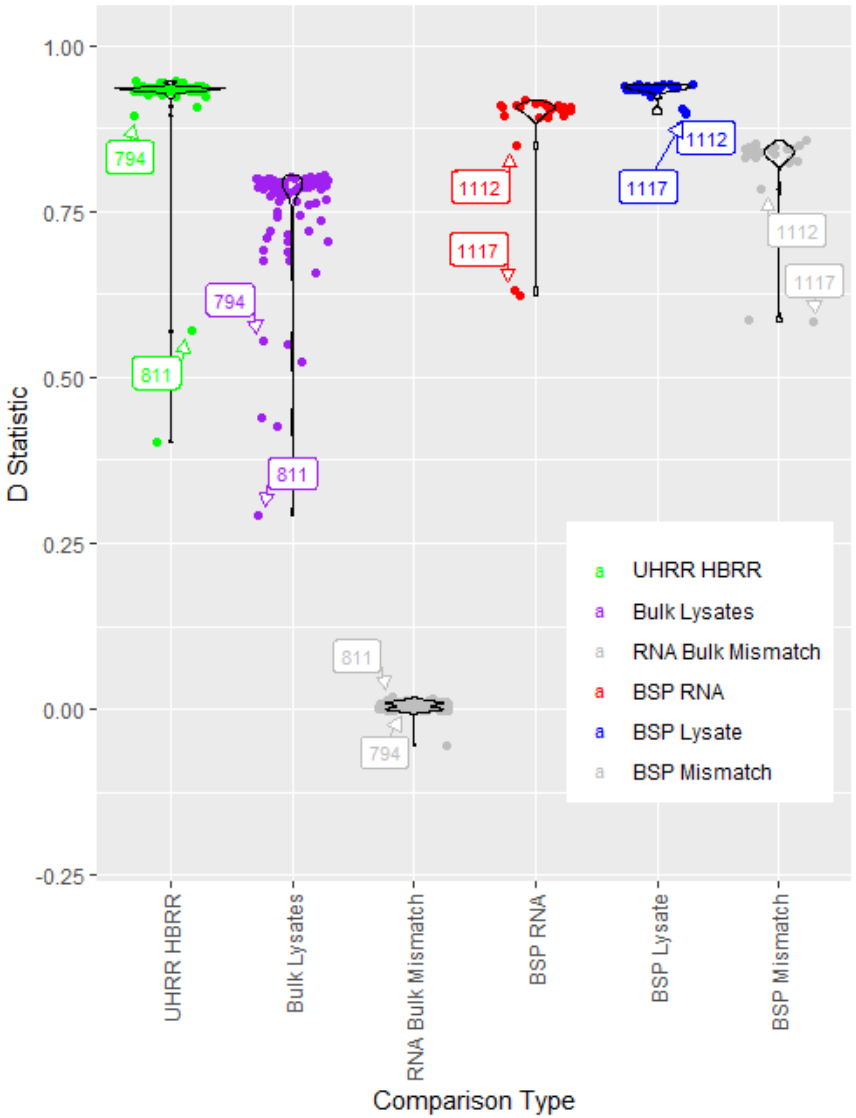
Comparison of Fold-Change Profiles for Transcriptomics Assay Performance

United States Environmental Protection Agency

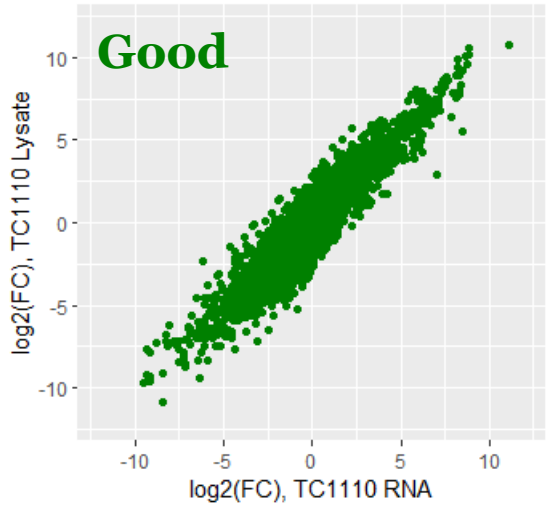
Correlation of Log_2(FC)

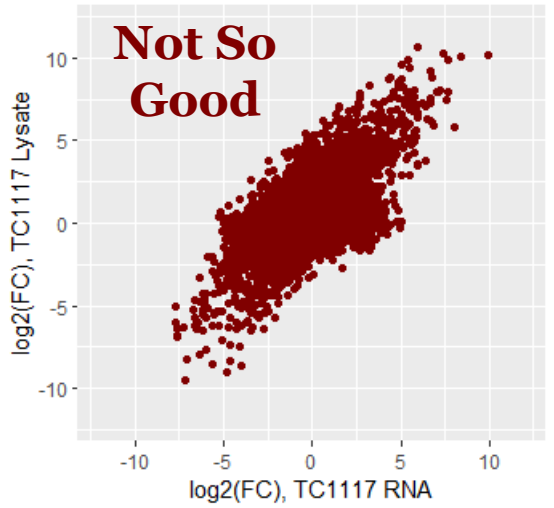


D-Statistic (FC)



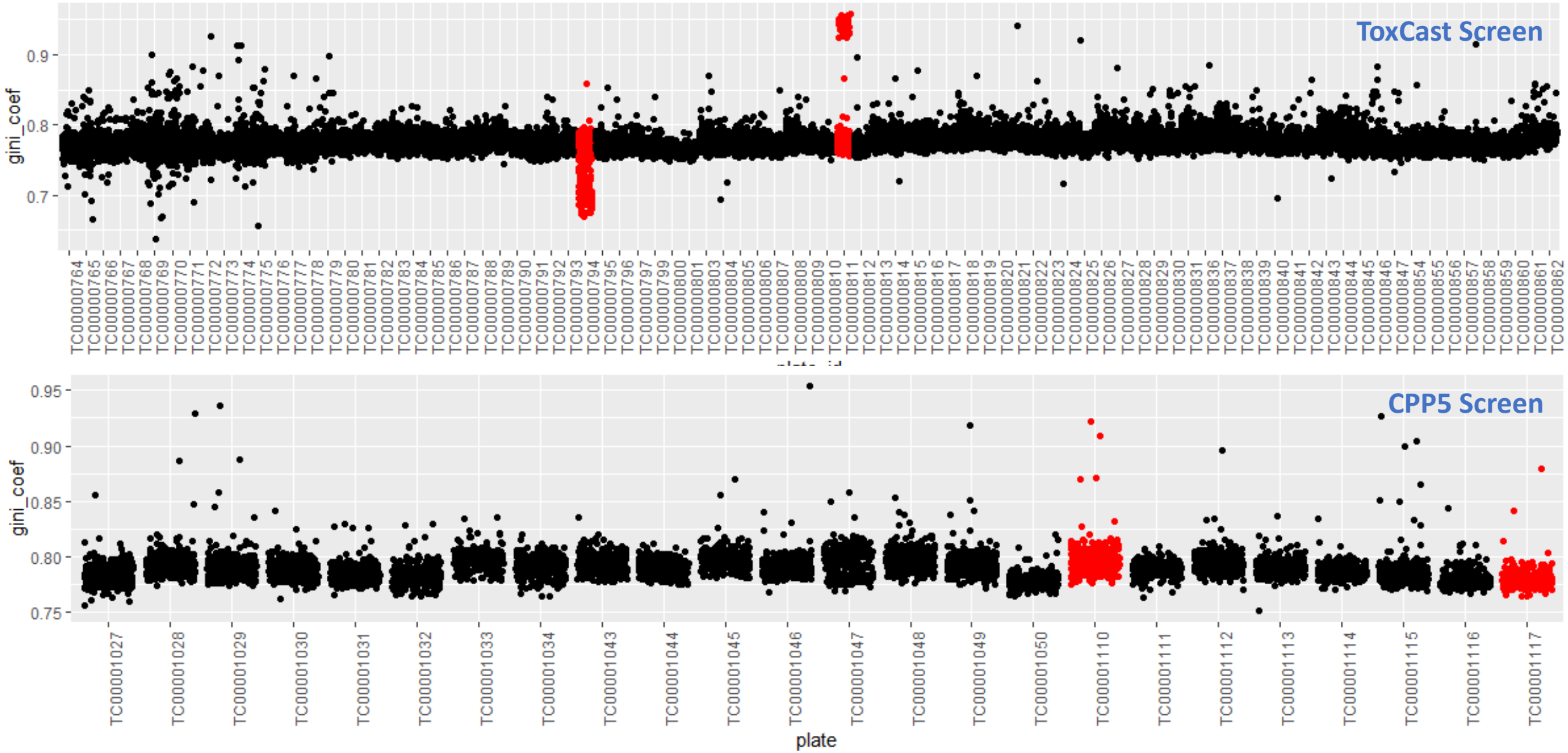
FC Correlation BSP RNA vs. Lysate





Analysis of Reference Samples Can Flag Underperforming Plates

= plate was flagged during analysis of reference samples



EPA United States Environmental Protection Agency

Summary

- Incorporation of **reference samples** into high-throughput transcriptomics study designs facilitates evaluation of assay performance within and across screening studies.
- **Performance metrics** may be calculated at different levels of data aggregation (per sample, comparison of like samples, comparison of differential expression in paired samples).
- We have engineered pairs of human-derived transcriptomics reference samples in both **purified RNA** and **cell lysate** format.
- Differential expression patterns are similar across engineered reference sample formats.

Acknowledgements

Office of Research and Development (ORD)
Center for Computational Toxicology and Exposure (CCTE)

- Johanna Nyffeler
- Clinton Willis
- Rick Brockway
- Megan Culbreth
- Dan Hallinger
- Terri Fairley
- Woody Setzer (ret) **
- Nisha Sipes **
- Russell Thomas **
- Logan Everett
- Imran Shah
- Richard Judson
- Derik Haggard

** = Co-authors on MAQC 2021 abstract

- Sreenivasa Ramaiahgari **
- Rick Paules (ret) **

- Megha Raghunathan **
- Jo Yeakley **
- Bruce Seligmann **
- Joel McComb **
- Pete Shepherd
- Milos Babic
- Dalia Gonzalez
- Kyle LeBlanc
- Garrett McComb