Microarray QC Metrics and Thresholds:
Calibrated RNA Samples, Reference Datasets, and Standard Datasets

Leming Shi, PhD

National Center for Toxicological Research (NCTR)
U.S. Food and Drug Administration (FDA)
Jefferson, Arkansas 72079, U.S.A.

Leming.Shi@fda.hhs.gov
+1-870-543-7387
www.gene-chips.com

Outline

1. Overview of the project
2. Selection of RNA samples
3. Design of microarray experiments
4. Generation of “gold standards” (qRT-PCR)
5. QC metrics and thresholds
6. Data sharing, management, and analysis
7. Timetable
8. Logistics, challenges, and benefits
9. Expectations from the meeting

SPECIAL SECTION

GENES IN ACTION

NEWS

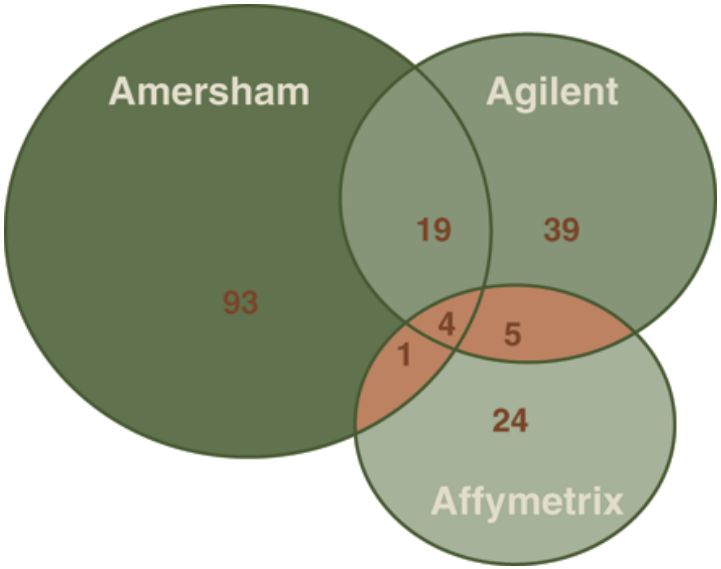
Getting the Noise Out of Gene Arrays

Thousands of papers have reported results obtained using gene arrays, which track the activity of multiple genes simultaneously. But are these results reproducible?

he gathered on kidney tumor cells, the less significant it seemed.

But those who have persevered with gene expression arrays attribute such problems to early growing pains. They claim

E. Marshall, *Science* 306, 630 (Oct 22, 2004).



"**Little overlap**."

"… the devices produced **wildly incompatible data**, largely because they were **measuring different things**."

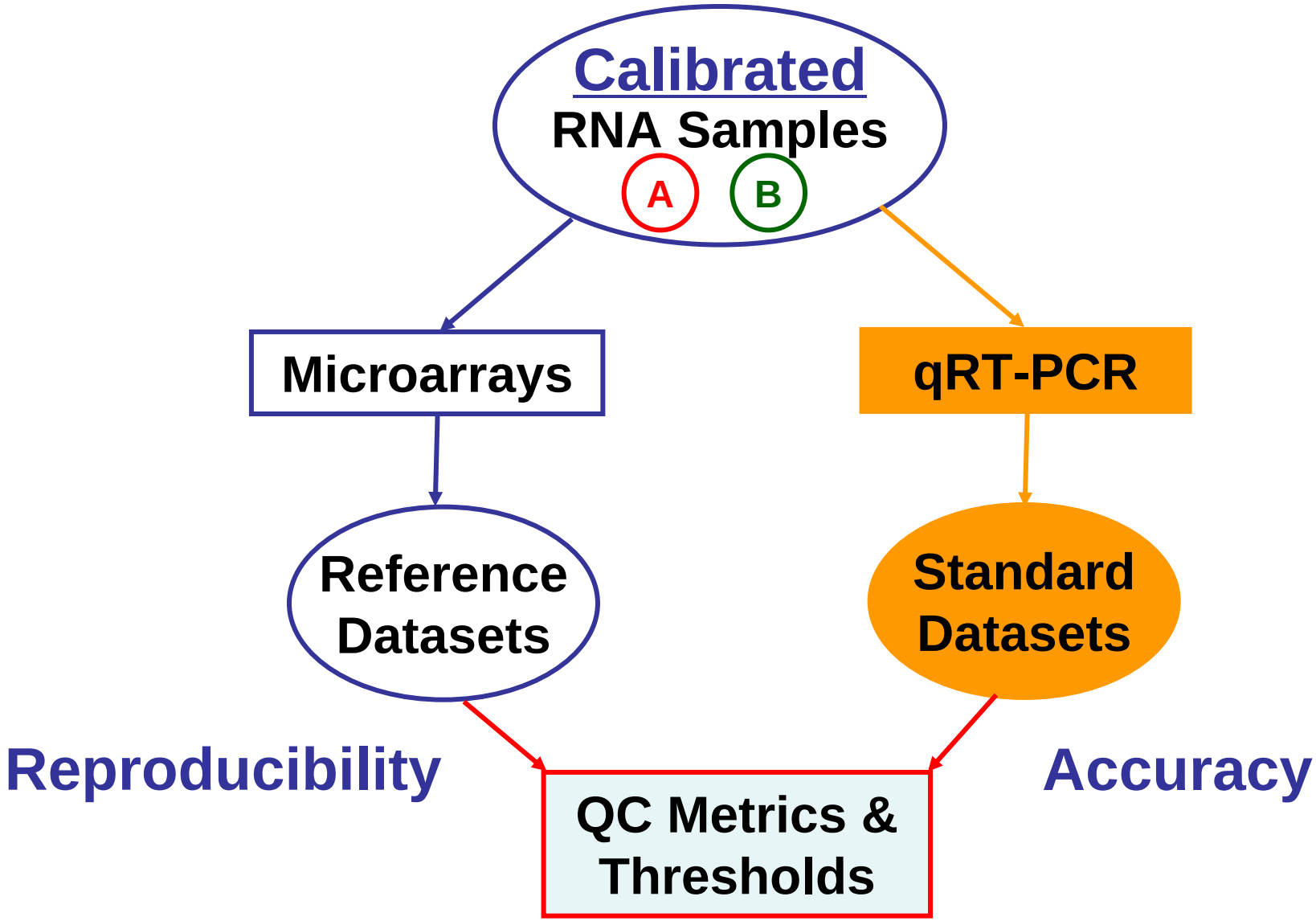
"… suggesting the need for establishing industrial **manufacturing standards**, and further independent and thorough validation of the technology."

P.K. Tan *et al.*, *Nucleic Acids Res* 31, 5676 (Oct 1, 2003).

???

- **Intra-platform/lab performance?**
- **Data analysis methods?**

Overview of Microarray QC Project



Selection of RNA Samples

A B

Two RNA samples for each species (Human, Mouse, and Rat).

Starting with one species (Human).

Criteria for RNA sample selection:

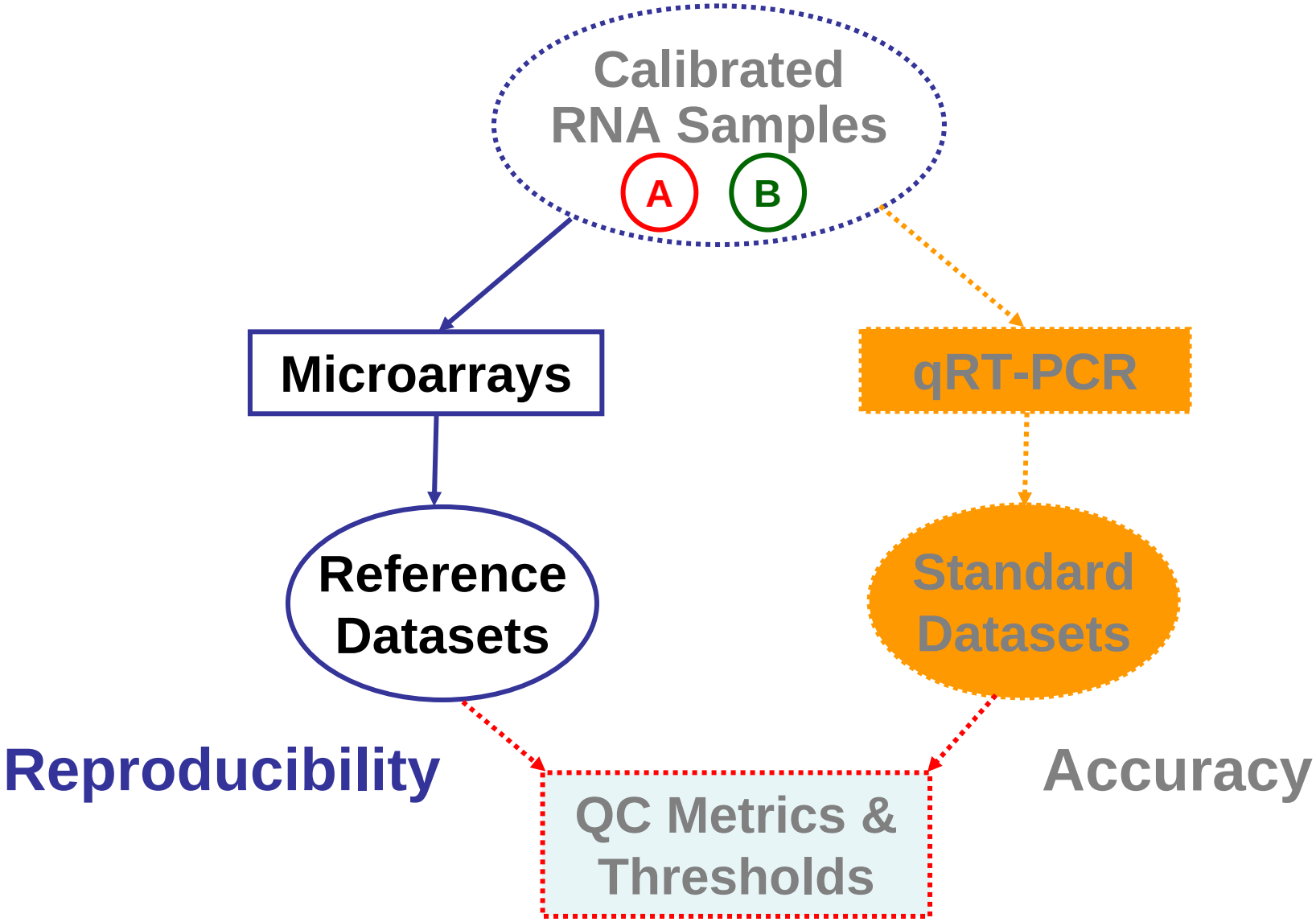
Available in large quantity
Reproducibility in production
High quality
Accessibility (commercial, e.g. Ambion and Stratagene)
Wide gene presence
Large fold changes for a number of genes

Options for RNA sample selection:

1. Two universal reference RNAs
2. Two tissue-specific RNAs
3. Two cell lines
4. Combination

Overview of Microarray QC Project:

Microarrays



Design of Microarray Experiments:

12 Technical Replicates

(4 replicates/day x 3 days)

A B

One-color platform: 24 hybs per lab

12 technical replicates for each RNA sample

Two-color platform: 30 hybs per lab

12 dye-swap pairs, plus

6 self-self hybs with sample A

(to investigate systematic dye-bias and false positives)

Design of Microarray Experiments:
Factors Not Addressed

Subjects:	Individual differences
Tissue dissection:	Sampling variability
RNA isolation:	Stability / integrity / purity

These are extremely important sources of variability, but are not specific to microarray technology. Therefore, they are not addressed in the current QC study.

Design of Microarray Experiments:
Factors Addressed

For each platform

Fixed:

Chip batch

Reagent batch

Protocol

Scanner settings
Same technician

Investigated:

"Day effect"

- For the same platform, different labs should use the same chip/reagent batch using vendor-recommended protocols.
- Each technical replicate should start from independent target synthesis and labeling reactions.
- Samples should be processed at 3 different days (4 replicates per days) by the same technician.
- Quality control metrics for individual steps should be recorded and provided.

Design of Microarray Experiments:
Reference Datasets: Task Assignments

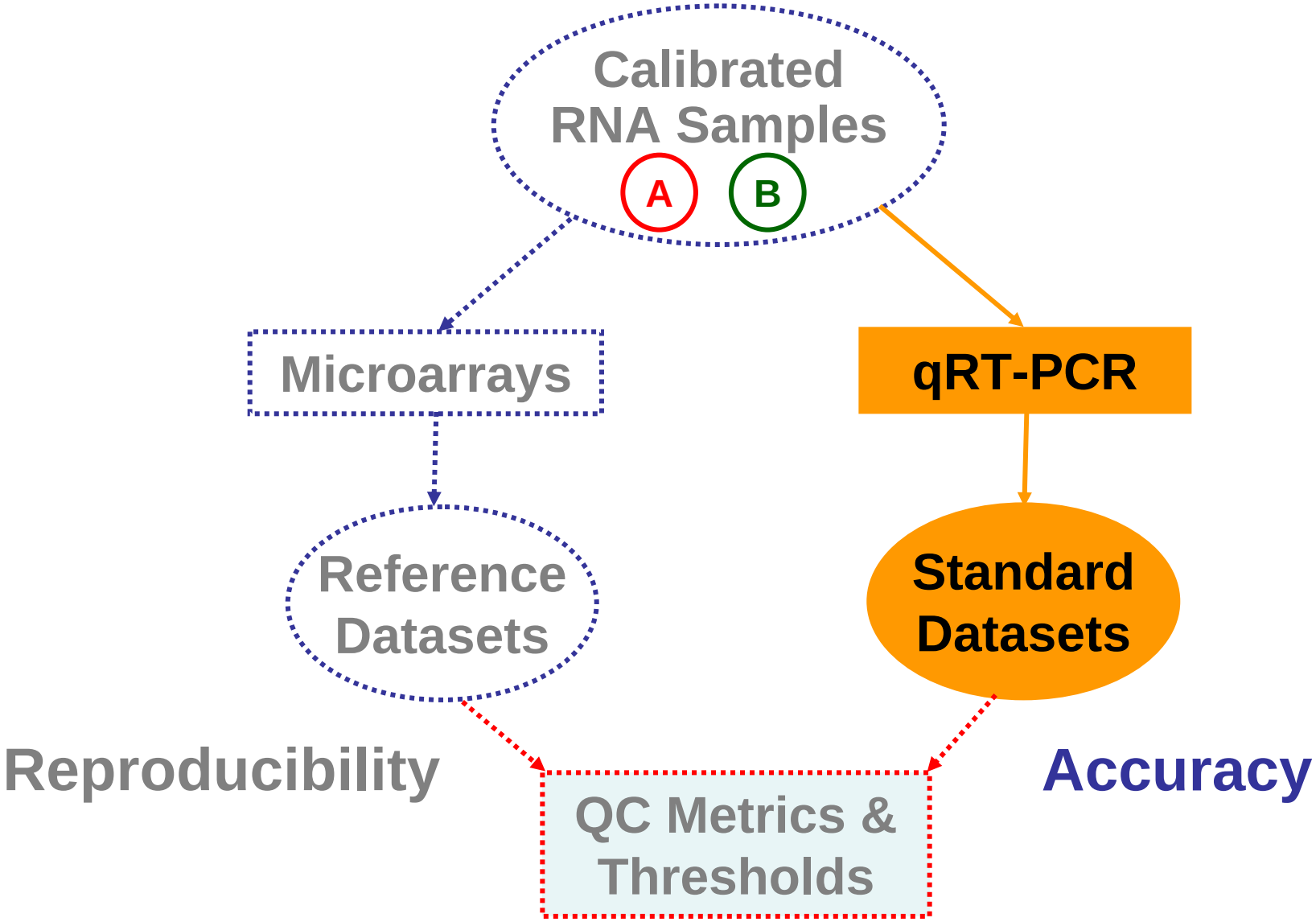
Provider	Array-type	Lab-1	Lab-2	Lab-3	Lab-n	Hybs/Lab
Affymetrix	U133v.2	Affymetrix	UCLA	Harvard?	CBER?	24
Agilent	G4112A	Agilent	NCTR	?		30
Applied Biosystems	Genome Survey V2.0	Applied Biosystems	?	Harvard?		24
GE Healthcare	UniSet 20K?	GE Healthcare	?	?		24
Illumina	Human-6 BeadChip?	Illumina	?	?		24
CBER	20K	CBER	NCTR	NCI		30
NCI	?	NCI	NCTR	CBER		30
NCTR	10K	NCTR	CBER	NCI		30

For each platform (array-type), data should be generated in at least 3 labs.

Total number of hybs: 648

Overview of Microarray QC Project:

qRT-PCR



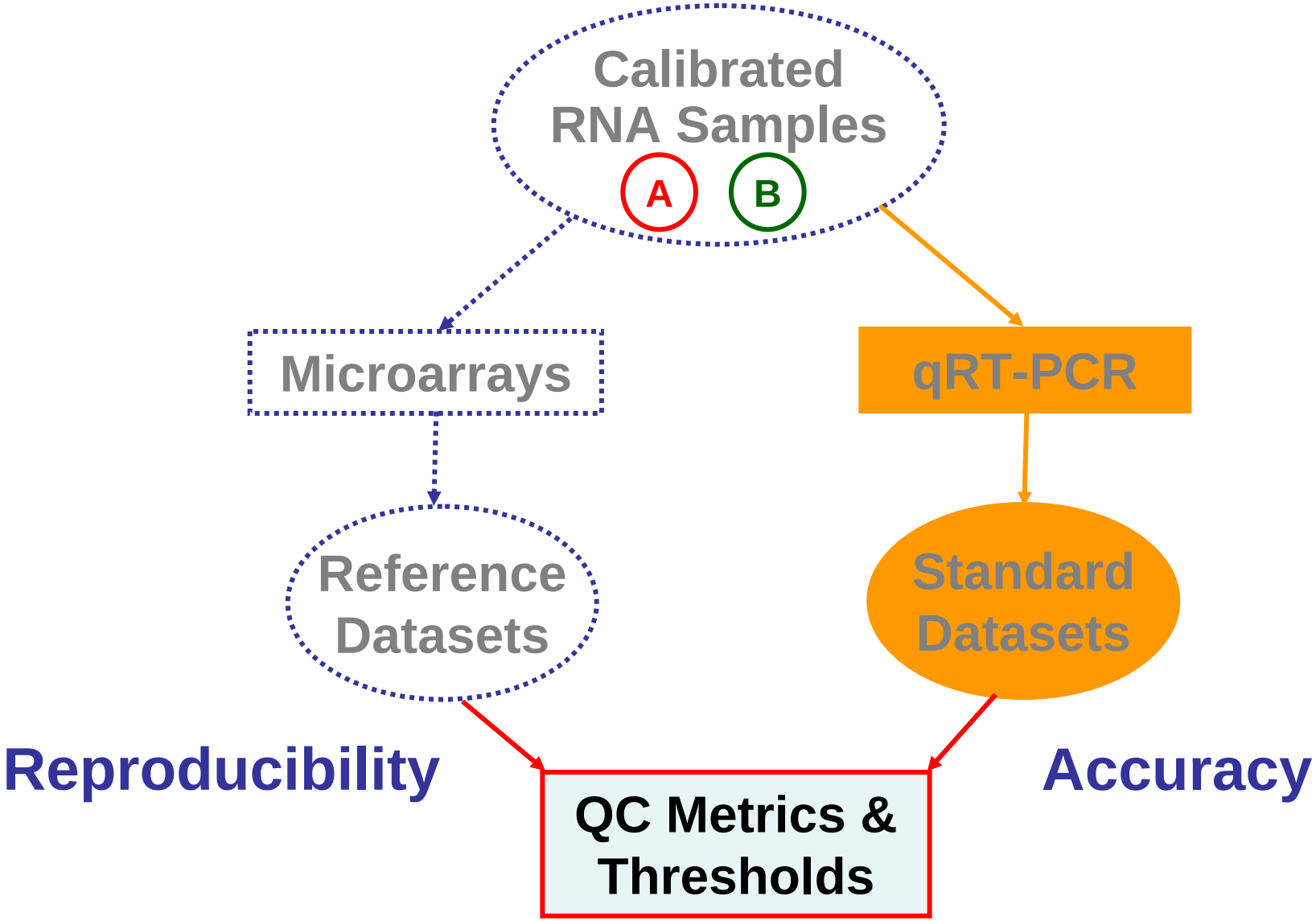
“Gold Standard” Datasets (qRT-PCR)

- Selection of a subset of genes (>1000? 5000?) covered by all microarray platforms
- Distributed, community-wide efforts
- Task assignments: 100-200 genes per lab for all participating labs?

qRT-PCR Working Group

Overview of Microarray QC Project:

QC Metrics and Thresholds

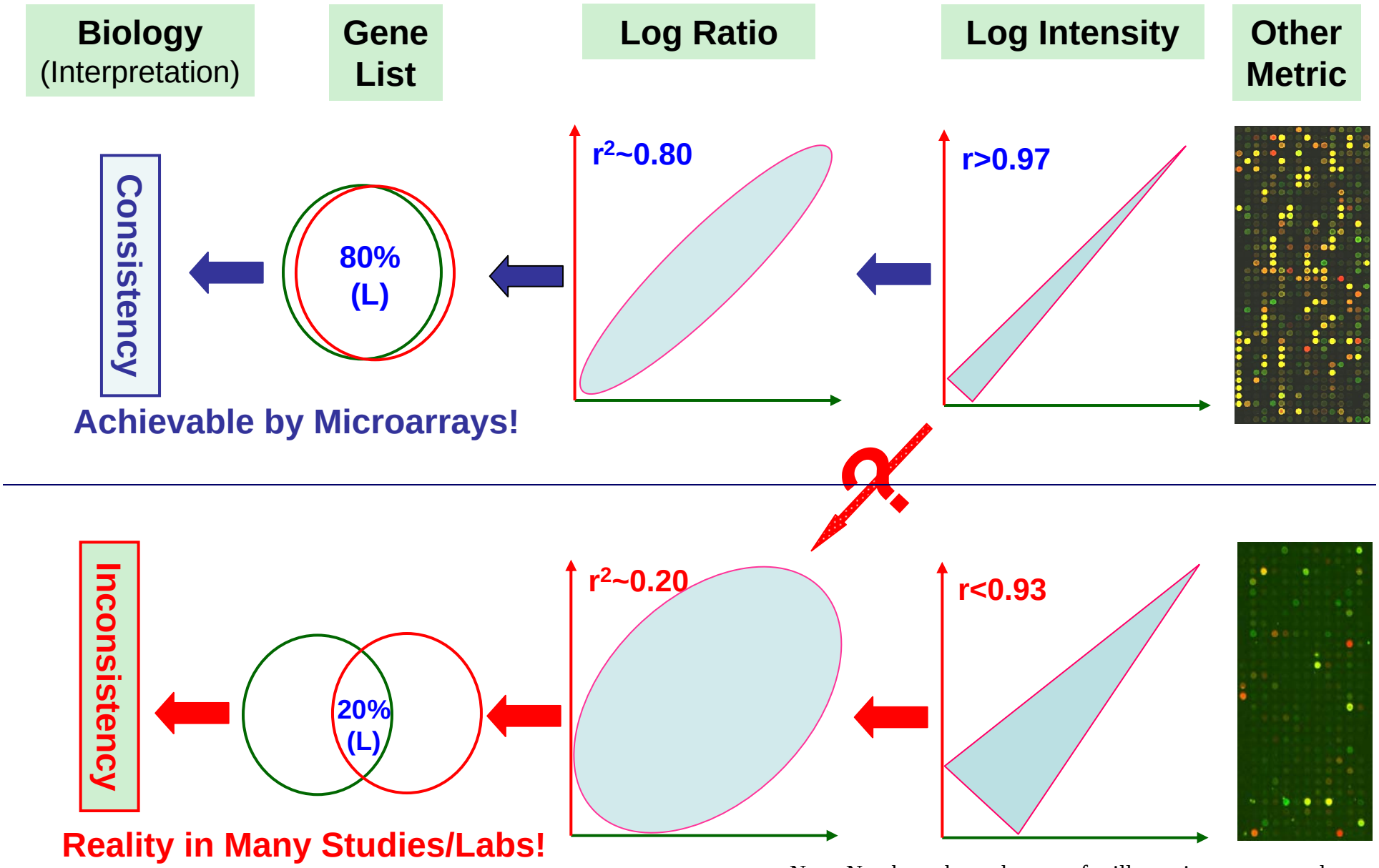


QC Metrics and Thresholds

- **Individual steps**
  - ✓ **RNA quality**: 260/280 ratio, Agilent Bioanalyzer, Gel electrophoresis
  - ✓ **cDNA/cRNA**: Synthesis yield
  - ✓ **Labeling**: Efficiency
  - ✓ **Hybridization**: Image defects, SNR, % present calls, 3'/5' GAPDH ratios, global 3'/5' ratio, …
- **Overall**
  - ✓ **LIr^2**: log intensity correlation squared
  - ✓ **LRr^2**: log ratio correlation squared
  - ✓ **POG**: Percentage of overlapping genes
  - ✓ **Statistic**: *p*, ANOVA, FDR, etc.
  - ✓ **GO?**
  - ✓ **Pathways?**
  - ✓ ……

Reproducibility ≠ Accuracy

Different QC Metrics



Note: Numbers shown here are for illustration purpose only.

Data Sharing, Management, and Analysis

1. Each participant should provide **image** files and **non-normalized** data along with its own **analysis report** (methods and results).
2. Data are expected to be centralized initially at FDA/NCTR with **ArrayTrack** (http://edkb.fda.gov/webstart/arraytrack/).
3. Data will be independently analyzed by at least three non-commercial groups (e.g., **FDA/NCTR**, **NIST**, and **Harvard**). Independent reports (results) will be provided to each participant.
4. Each lab will gain **access to other labs' data** and conduct its analysis, if desired.
5. A **project meeting** will be held to discuss the results.
6. **Manuscripts** will be written.
7. Data will be made **available to the public** (e.g., ArrayTrack, GEO, chip.org) for additional analysis by the public.
8. Guidance on QC metrics/thresholds and data analysis.

Timeline

• Experimental design:	**February 11, 2005**
• RNA sample selection:	**February 11**
• Task assignments:	**February 11**
• MTAs:	March 15
• RNA and array distribution:	March 31
• Array hybridization:	May 31
• Microarray data transfer to FDA/NCTR:	**June 30**
• Data sharing among participants:	July 10
• qRT-PCR:	**July 31**
• Initial, independent analysis reports:	August 31
• Project meeting on data analysis (DC, CA?):	**October**
• Manuscripts:	November 30
• Public data deposition/access:	November 30
• Populating datasets (human/mouse/rat):	May 31, 2006
• Public meeting on data analysis:	November, 2006
• (FDA) Report/guideline:	December 31, 2006

Logistics

- **MTAs**
  - ✓ NIH/FDA/CDC template
- **RNA sample distribution**
  - ✓ 400? ug total RNA per sample per lab
- **Microarray chip distribution**
  - ✓ 24 (one-color) or 30 (two-color) chips per lab
- **Reagents**
- **Cost-sharing?**

Challenges

- **RNA samples:**
  - ✓ To meet the criteria required for a “reference material”
- **Project participants:**
  - ✓ To remain committed
  - ✓ To keep on schedule
- **Microarray community:**
  - ✓ To adopt the calibrated RNA samples
  - ✓ To participate in generating and analyzing data
  - ✓ To promote coordination (e.g., with ERCC and NIST)

Benefits of Calibrated RNA Samples, Reference Datasets, and Standard Datasets

Advantages of Calibrated RNA Samples:

Consistent profiling of identical pair of samples
Data share and comparison made possible
Establishment of Standard Operating Procedures (SOPs)
Quality control/assurance (avoid procedural failures)
……

Benefits of Reference/Standard Datasets:

Achievable performance on each platform
Reproducibility, accuracy, and robustness
Cross-laboratory and cross-platform concordance
QC metrics and thresholds
SOPs for data analysis
Regulatory guidelines
Public confidence
……

Expectations from the Meeting

1. To decide whether this study should be done
2. To finalize **experimental design**
3. To finalize **RNA sample selections**
4. To agree on each participant's **task assignments**
5. To agree on a plan for **data sharing, management, and analysis**
6. To agree on a **timetable**
7. To enumerate **QC metrics**
8. To form four **working groups**:
   - RNA samples
   - Microarrays
   - qRT-PCR
   - Data analysis

The Mission of FDA's National Center for Toxicological Research (NCTR) Jefferson, Arkansas

"To conduct peer-reviewed scientific research that supports and anticipates the FDA's current and future regulatory needs."

Acknowledgments

Toxicoinformatics
Megan Cao
Hong Fang
Weigong Ge
Li Guo
Steve Harris
Huixiao Hong
Roger Perkins
Feng Qian
Hongmei Sun
Weida Tong
Qian Xie

Biometry
Jim Chen
Bob Delongchamp
Ralph Kodell

Functional Genomics
Bill Branham
Varsha Desai
Jim Fuscoe
Tao Han
Cathy Melvin

Neurotoxicology
John Bowyer
Tucker Patterson
Bill Slikker
Z. Alex Xu (now at Gene Logic)

Systems Toxicology
Yvonne Dragan
Lei Guo

FDA/CBER
Jing Han
Raj Puri

FDA/CDER
Felix Frueh
Federico Goodsaid

FDA/CDRH
Gene Pennello
Uwe Scherf

FDA/CFSAN
Greg Akerman (now at EPA)
Tom Cebula
Scott Jackson
Joseph LeClerc

FDA/CVM
Heather Harbottle

Dan Casciano, Director, FDA/NCTR

UCLA/Cedars-Sinai
Yongxi Tan
Charles Wang

Non-FDA Collaborators
many…

Confirmed Project Participants

FDA Centers

CBER, CDER, CDRH, CFSAN, CVM, and NCTR

NIST

Marc Salit

Microarray Platform Providers

Affymetrix:	Janet Warrington, Jacques Retief
Agilent:	Jim Collins
Applied Biosystems:	Gary Schroth, Lu Zhang, Jack Zhai
GE Healthcare :	Timothy Sendera, Richard Shippy
Illumina:	Shawn Baker

RNA Sample Providers

Ambion:	Bob Setterquist, David Dorris
Stratagene:	Gavin Fisher, Natalia Novoradovskaya

Others

Cedars-Sinai/UCLA:	Charles Wang
Children's Hospital/Harvard:	Zoltan Szallasi
NIH/NCI:	Ernest Kawasaki

……

Thank you!

Thank you!

A free software for the management, analysis, and interpretation of microarray data.

www.gene-chips.com

http://www.fda.gov/nctr/science/centers/toxicoinformatics/ArrayTrack/