

My Experience with Emerging Genomics Technologies

Weida Tong, Ph.D

Division of Bioinformatics and Biostatistics, NCTR/FDA

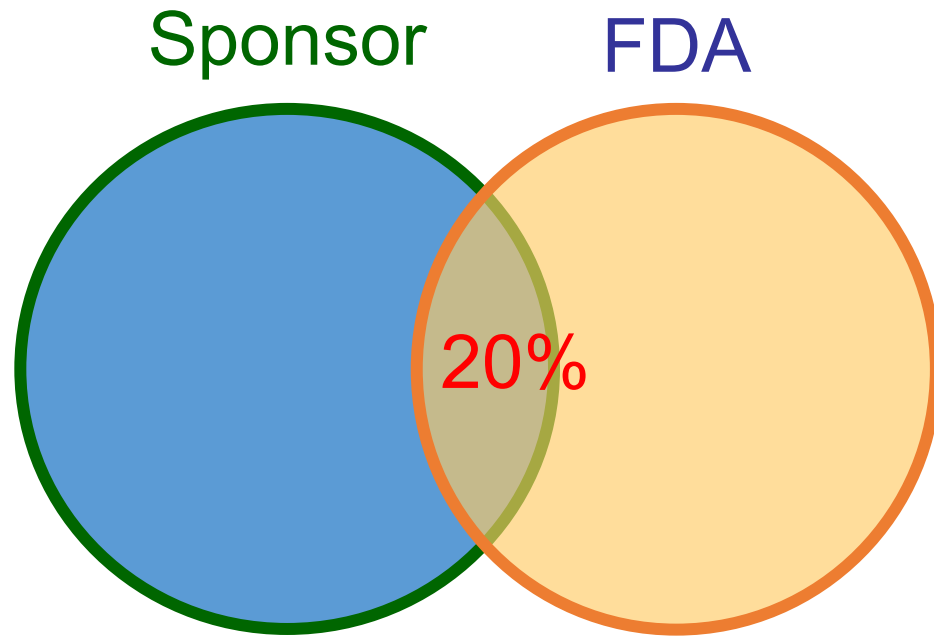
Weida.tong@fda.hhs.gov; 870-543-7142

FDA's Voluntary Genomics (eXploratory) Data Submission (VGDS/VXDS)

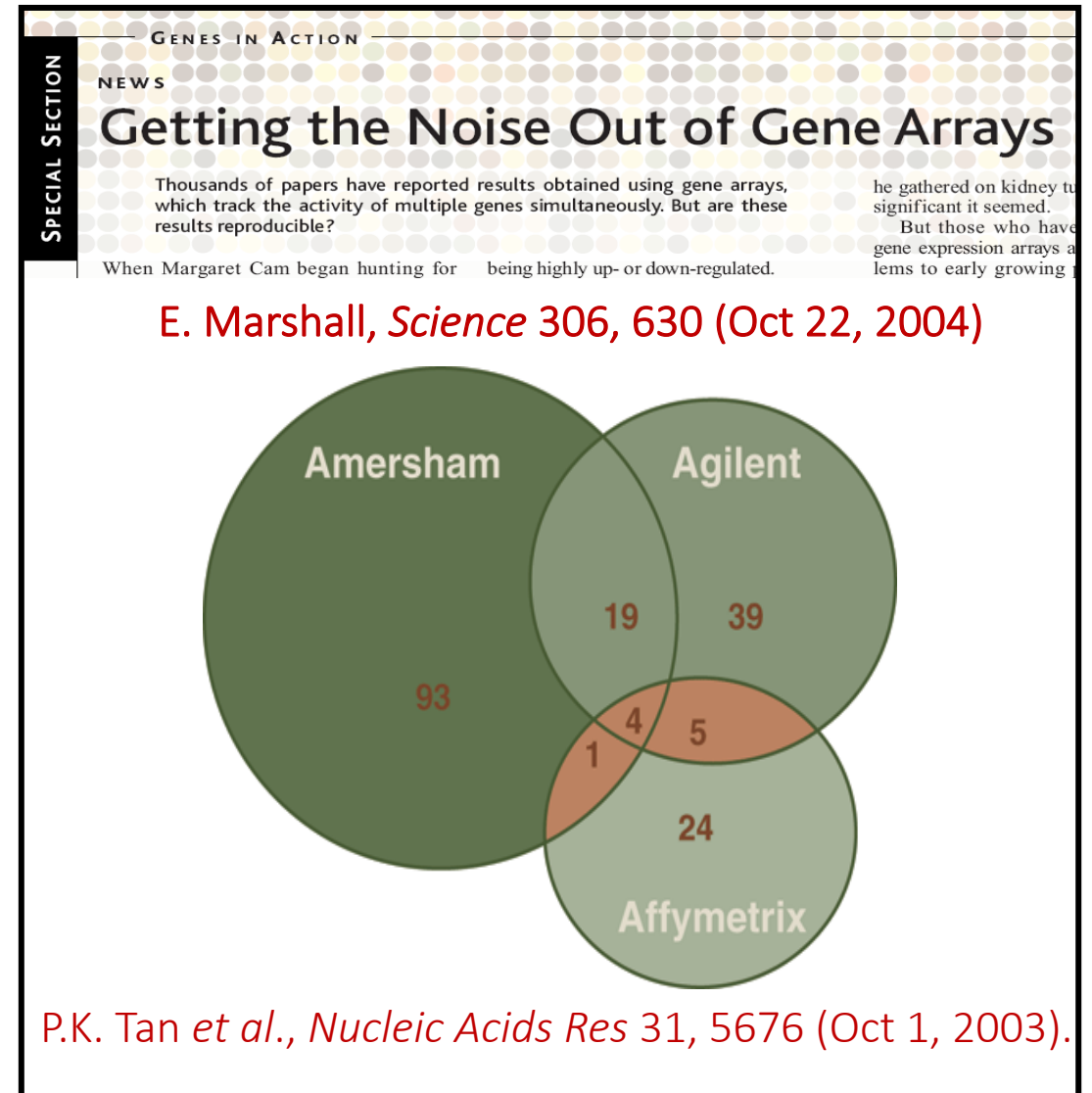
- Defined in Guidance for Industry on Pharmacogenomics (PGx) Data Submission (draft document released in 2003; final publication, 2005):
 - To encourage the sponsor to interact with FDA through submission of PGx data on a voluntary basis
 - To provide a forum for scientific discussions with the FDA outside of the application review process
 - To establish a regulatory environment (both the tools and expertise) within the FDA for receiving, analyzing and interpreting PGx data
- Set up a bioinformatics infrastructure to address three VGDS goals:
 - Data repository (developed ArrayTrack)
 - Reproduce the sponsor's results
 - Conduct alternative analysis

Poor reproducibility between different analysis methods and platforms

Different Data Analysis Methods, with the Same VGDS Data Set

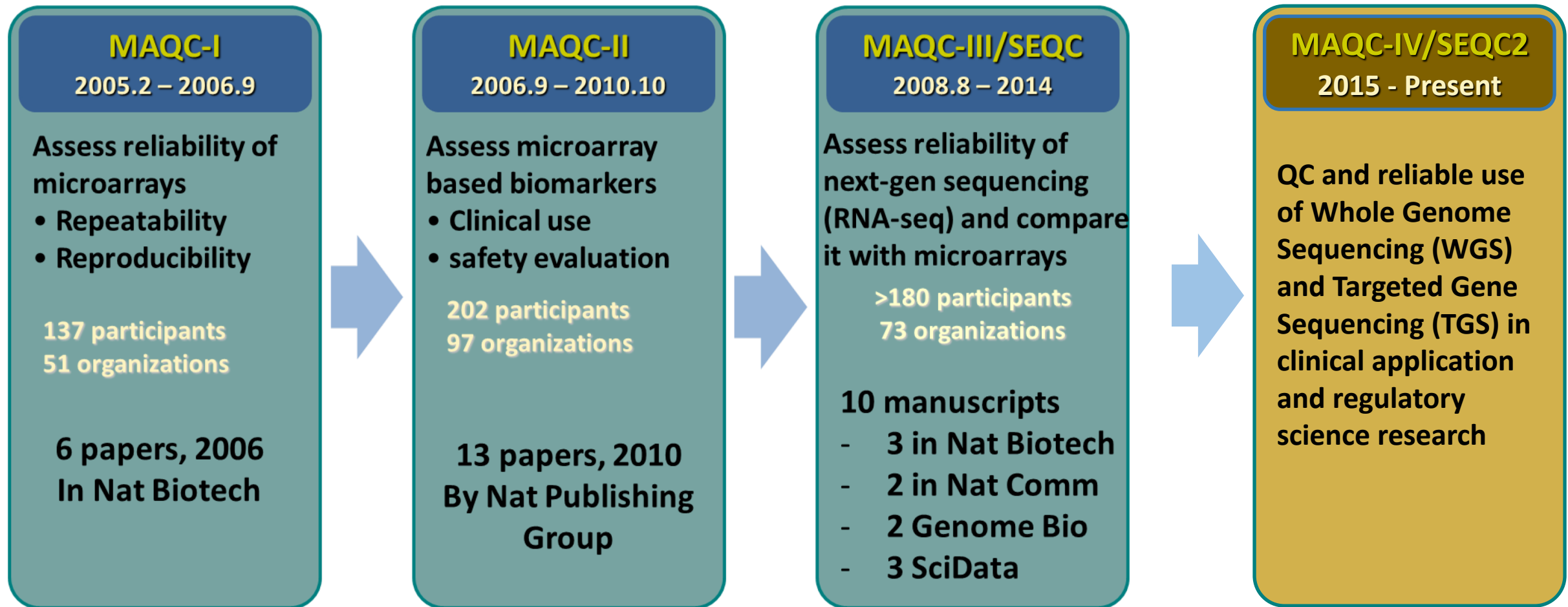


Updated from Federico Goodsaid, The 3rd MAQC Project Meeting, Palo Alto, CA, December 1-2, 2005

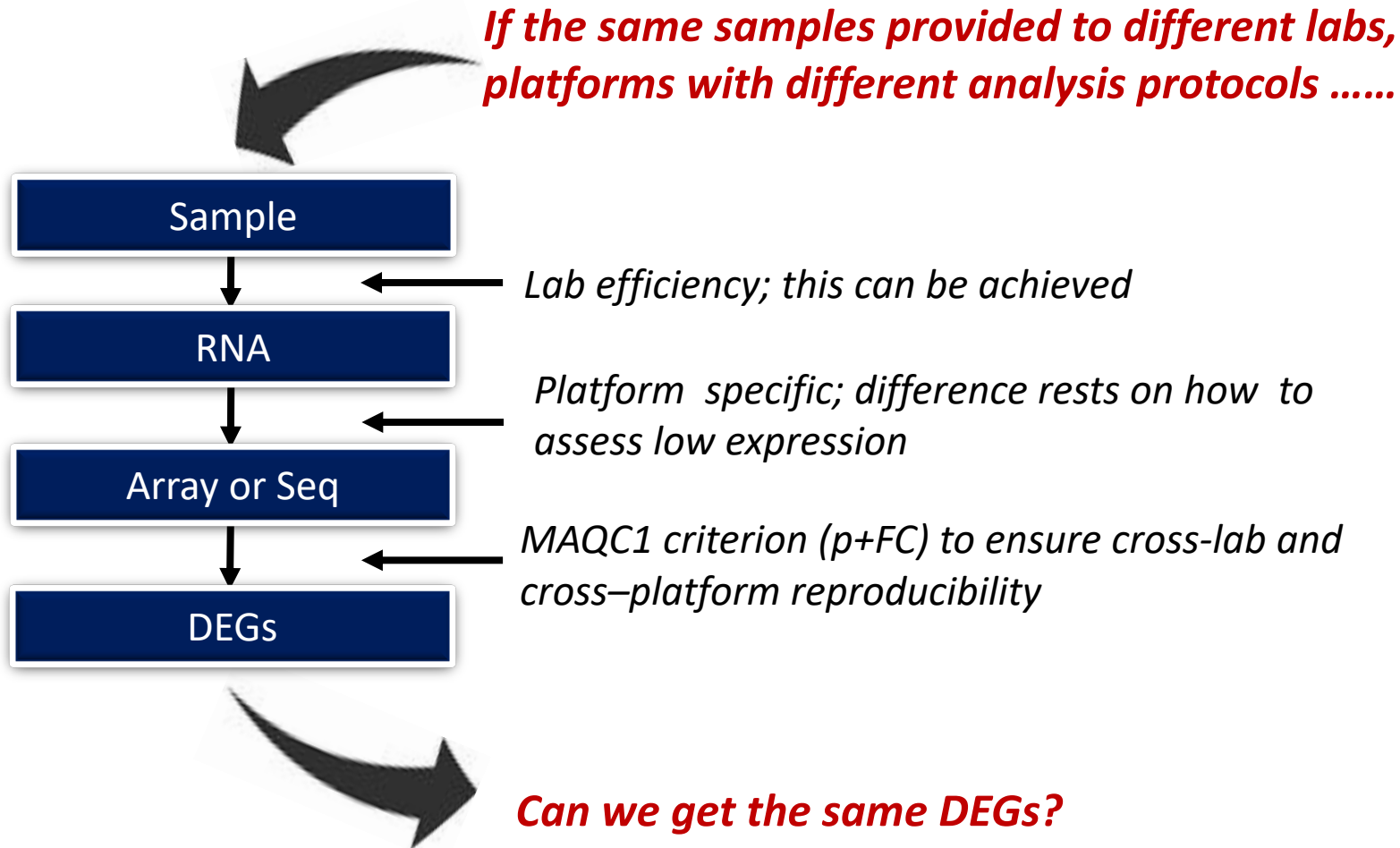


MicroArray Quality Control (MAQC)

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

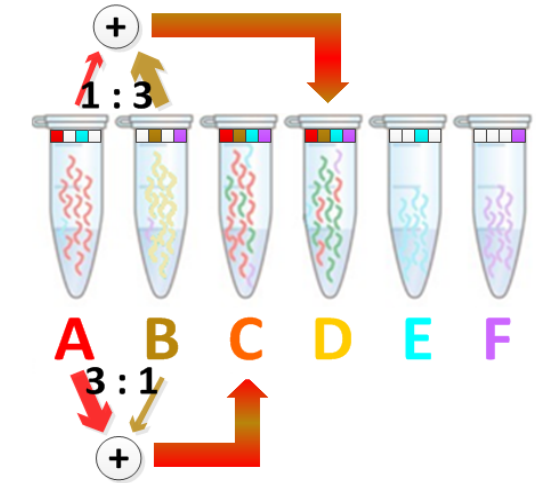


MAQC/SEQC Objectives:



MAQC-III/SEQC1 Study Design

6 reference samples



11 Labs

3 Platforms

- Illumina
- SOLiD
- 454

1. Reproducibility $\neq$ Statistical Significance

- Reproducibility is a third dimension beyond sensitivity and specificity:
 - It is not commonly considered by statisticians
 - Need a new performance metric to balance reproducibility and statistical significance
- How to generate a reproducible DEGs (**fold-change + p-value**):
 - Rank genes by FC and remove these with p-value smaller than a threshold



MAQC Consortium, Nat
Biotechnol, 2006

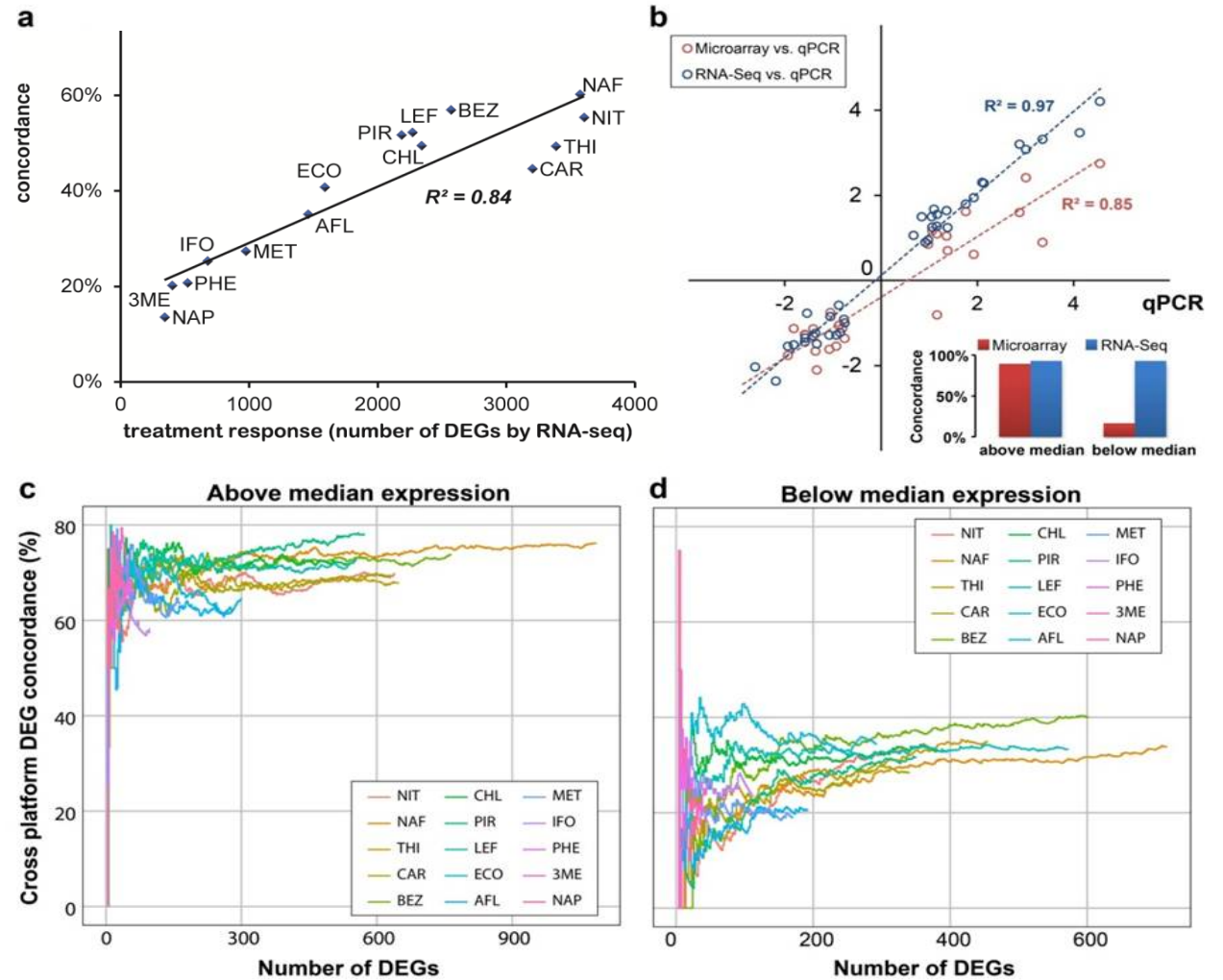
2. Not all the genes are created equal

The “treatment effect” dictates the concordance between two platforms in detecting DEGs

The concordance between the two platforms is high for the highly expressed genes, not for the weakly expressed genes

RNA-Seq agrees with qPCR for low expressed genes

RNA-Seq vs. microarrays



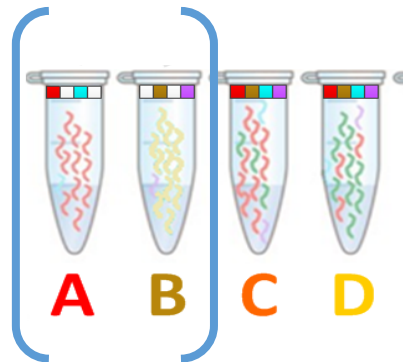
3. Not all the statistics are created equal

Calculation of Statistical Significance

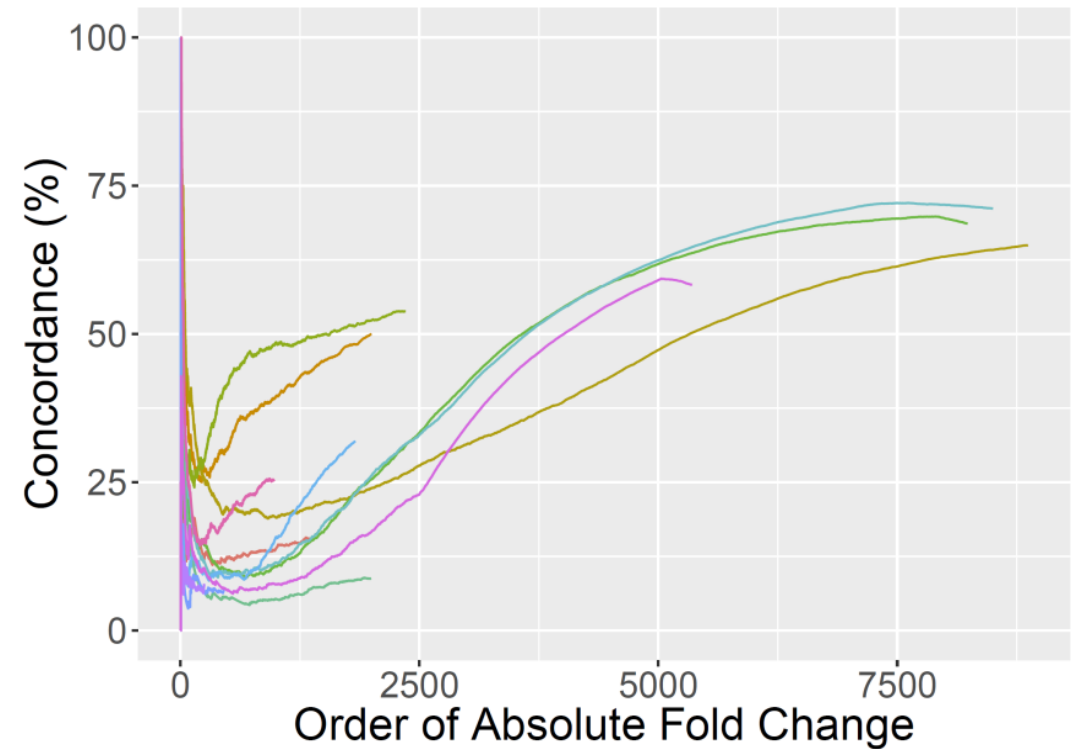
$$(P\text{-value}) = f\left(\frac{\text{Log2 (Fold Change of A/B)}}{\text{Standard Error}} \right)$$

If we calculate DEGs of A vs B in an experiment involving A, B, C and D groups

- **t-test** considers the variance only for A and B
- **ANOVA** considers the variance for all A-D



T-test vs ANOVA



My recommendation for reproducible transcriptomics data analysis

- Apply the MAQC reproducible DEGs rule (**fold-change + p-value**)
 - *Recommendation: Rank genes by FC and remove these with a lesser stringent p-value*
- Not all the genes are created equal:
 - The error structure varies among genes (e.g., high and low expression), which is technology (platform) dependent (e.g., RNA-seq vs AFX)
 - *Recommendation: remove these genes behaving badly*
- Not all statistical methods are created equal:
 - There is no “correct” method, and we should focus on the “less” wrong method.
 - *Recommendation: Limma is the “less” wrong method*

Reproducibility in Questions !!!

- “In 2012, OECD invested ~\$59 billion in biomedical research - ~89% of which may have been wasted on science that cannot be reproduced. Are we looking at a \$52.5 billion white elephant?” (by GEN)
- Up to 85% resources are wasted in irreproducible science
- The rate of retractions increased by a factor of ~11 between 2001 and 2010
- Not correlates with journal impact factor

THE WALL STREET JOURNAL.

The Breakdown in Biomedical Research

Contaminated samples, faulty studies and inadequate training have created a crisis in laboratories and new treatments and cures

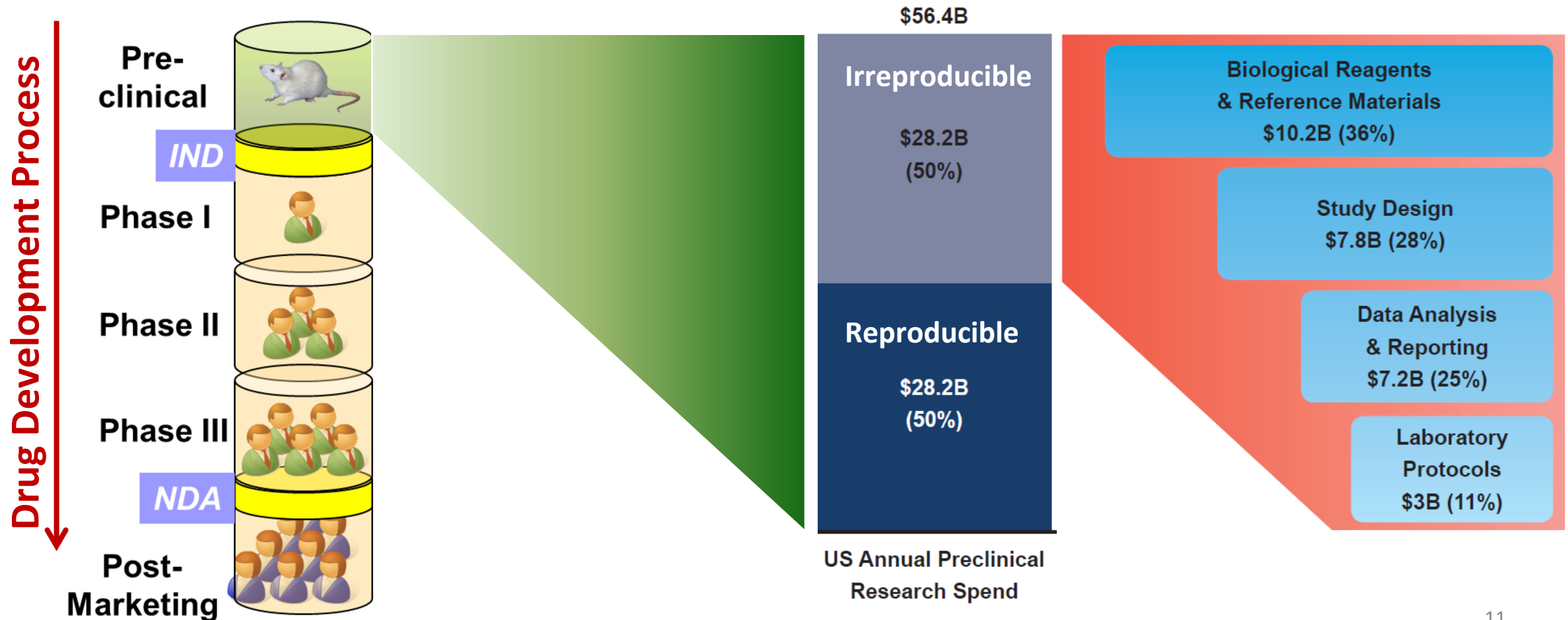


By **RICHARD HARRIS**

Updated April 7, 2017 2:05 p.m. ET

Irreproducibility in Preclinical Study

- Translational implication in drug development:
 - Only 20-25% of published data can be reproduced and translated for drug development (by Bayer Health)
 - Only 11% of cases can be reproduced (by Amgen)



Irreproducibility in Cell Based Assays

Systematic identification of genomic markers of drug sensitivity in cancer cells

Mathew J. Garnett^{1*}, Elena J. Edelman^{2*}, Sonja J. Heidorn^{1*}, Chris D. Greenman^{1†}, Anahita Dastur², King Wai Lau¹, Patricia Greninger², I. Richard Thompson¹, Xi Luo², Jorge Soares¹, Qingsong Liu^{3,4}, Francesco Iorio^{1,5}, Didier Surdez⁶, Li Chen², Randy J. Milano², Graham R. Bignell¹, Ah T. Tam², Helen Davies¹, Jesse A. Stevenson², Syd Barthorpe¹, Stephen R. Lutz², Fiona Kogera¹, Karl Lawrence¹, Anne McLaren-Douglas¹, Xenia Mitropoulos², Tatiana Mironenko¹, Helen Thif¹, Laura Richardson¹, Wenjun Zhou^{1,4}, Frances Jewitt¹, Tinghu Zhang^{2,4}, Patrick O'Brien¹, Jessica L. Boisvert¹, Stacey Price², Wooyoung Hur^{2,4}, Wanjun Yang¹, Xianming Deng^{2,4}, Adam Butler¹, Hwan Geun Choi^{1,4}, Jae Won Chang^{2,4}, Jose Baselga¹, Ivan Stamenkovic¹, Jeffrey A. Engelman², Sreenath V. Sharma^{1†}, Olivier Delattre⁶, Julio Saez-Rodriguez², Nathanael S. Gray^{1,4}, Jeffrey Settleman², P. Andrew Futreal¹, Daniel A. Haber^{2,8}, Michael R. Stratton¹, Sridhar Ramaswamy², Ultan McDermott¹ & Cyril H. Benes²



Wellcome Trust Sanger Institute
Massachusetts General Hospital

Sanger Institute:
138 drugs tested^d on
727 cell lines

nature

2012

The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity

Jordi Barretina^{1,2,3*}, Giordano Caponigro^{4*}, Nicolas Stransky^{1*}, Kavitha Venkatesan^{4*}, Adam A. Margolin^{1*}, Sungjoon Kim⁵, Christopher J. Wilson⁴, Joseph Lehar⁴, Gregory V. Kryukov¹, Dmitriy Sonkin⁴, Anupama Reddy⁴, Manway Liu⁴, Lauren Murray¹, Michael F. Berger^{1†}, John E. Monahan⁴, Paula Morais¹, Jodi Meltzer⁴, Adam Korejwa¹, Judit Jané-Valbuena^{1,2}, Felipa A. Mapa⁴, Joseph Thibault¹, Eva Bric-Furlong⁴, Pichai Raman¹, Aaron Shipway³, Ingo H. Engels⁵, Jill Cheng⁶, Guoying K. Yu⁶, Jianjun Yu⁶, Peter Aspesi Jr⁴, Melanie de Silva⁴, Kalpana Jagtap¹, Michael D. Jones⁴, Li Wang⁴, Charles Hattori³, Emanuele Palescandolo³, Supriya Gupta¹, Scott Mahan¹, Carrie Sougnez¹, Robert C. Onofrio¹, Ted Liefeld¹, Laura MacConaill³, Wendy Winckler¹, Michael Reich¹, Nanxin Li³, Jill P. Mesirov¹, Stacey B. Gabriel¹, Gad Getz¹, Kristin Ardlie¹, Vivien Chan⁶, Vic E. Myer⁴, Barbara L. Weber⁴, Jeff Porter⁴, Markus Warmuth⁴, Peter Finan¹, Jennifer L. Harris⁵, Matthew Meyerson^{1,2,3}, Todd R. Golub^{1,3,7,8}, Michael P. Morrissey^{4*}, William R. Sellers^{4*}, Robert Schlegel^{4*} & Levi A. Garraway^{1,2,3*}



Broad Institute
Novartis

Broad Institute:
24 drugs tested on
1,036 cell lines

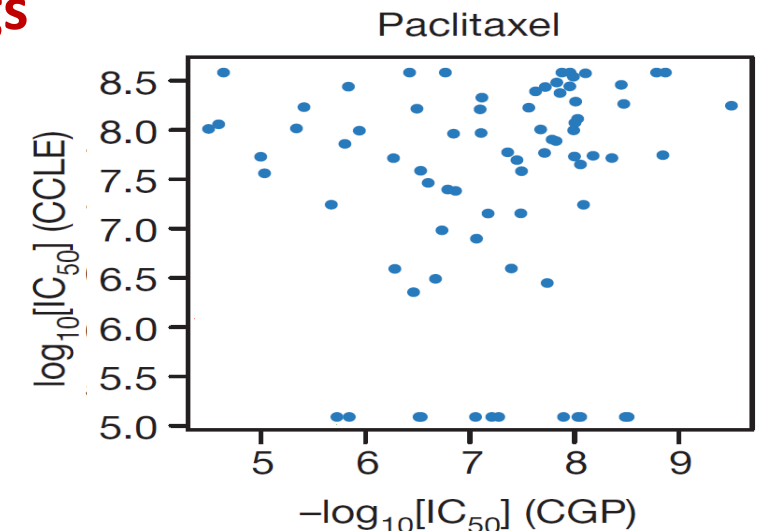
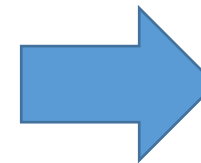
Common: 471 cell lines + 15 drugs

Inconsistency in large pharmacogenomic studies

Benjamin Haibe-Kains^{1,2}, Nehme El-Hachimi^{1,2}, Andrew C. Jin⁴, Andrew H. Beck^{4*}, Hugo J. W. L. Aerts^{5,6,7*} & John Quackenbush^{5,8*}

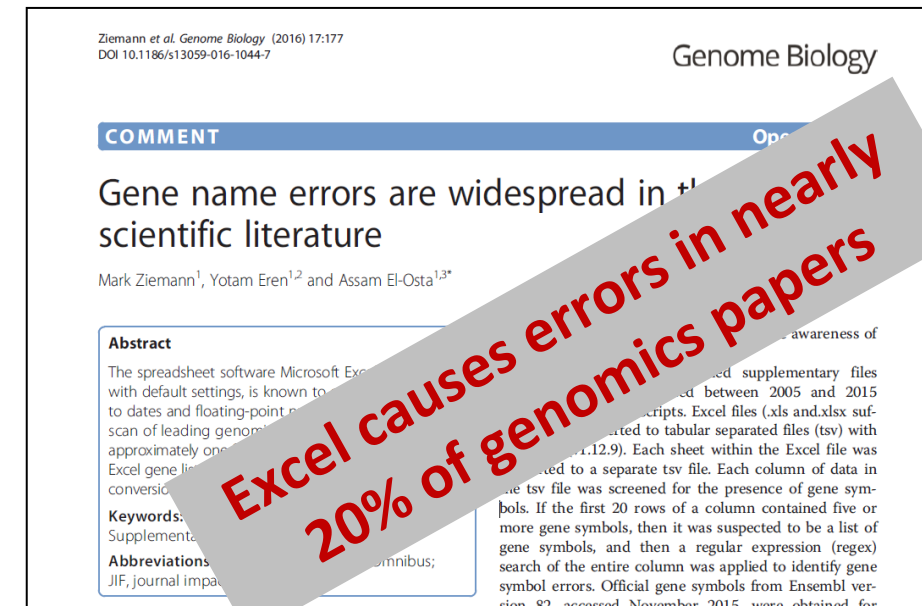
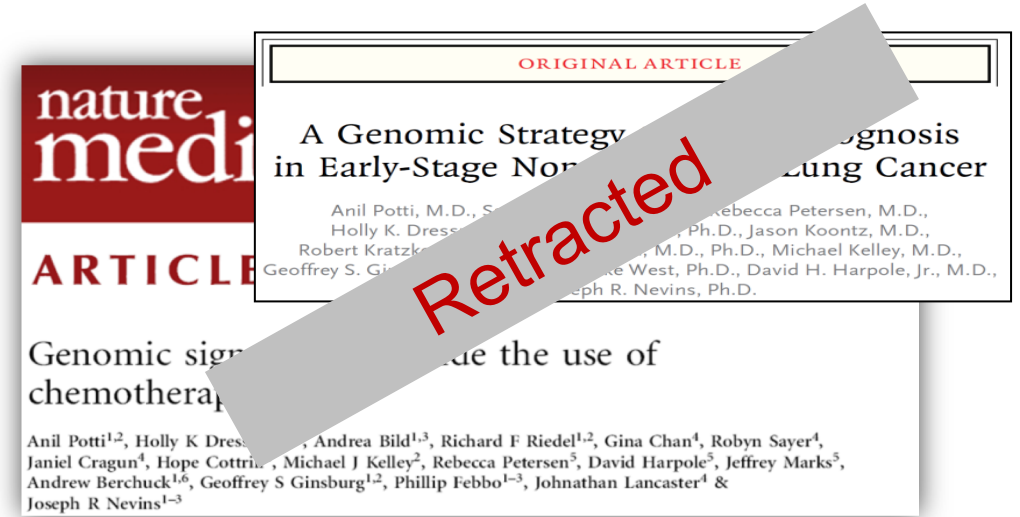
nature 2013

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.

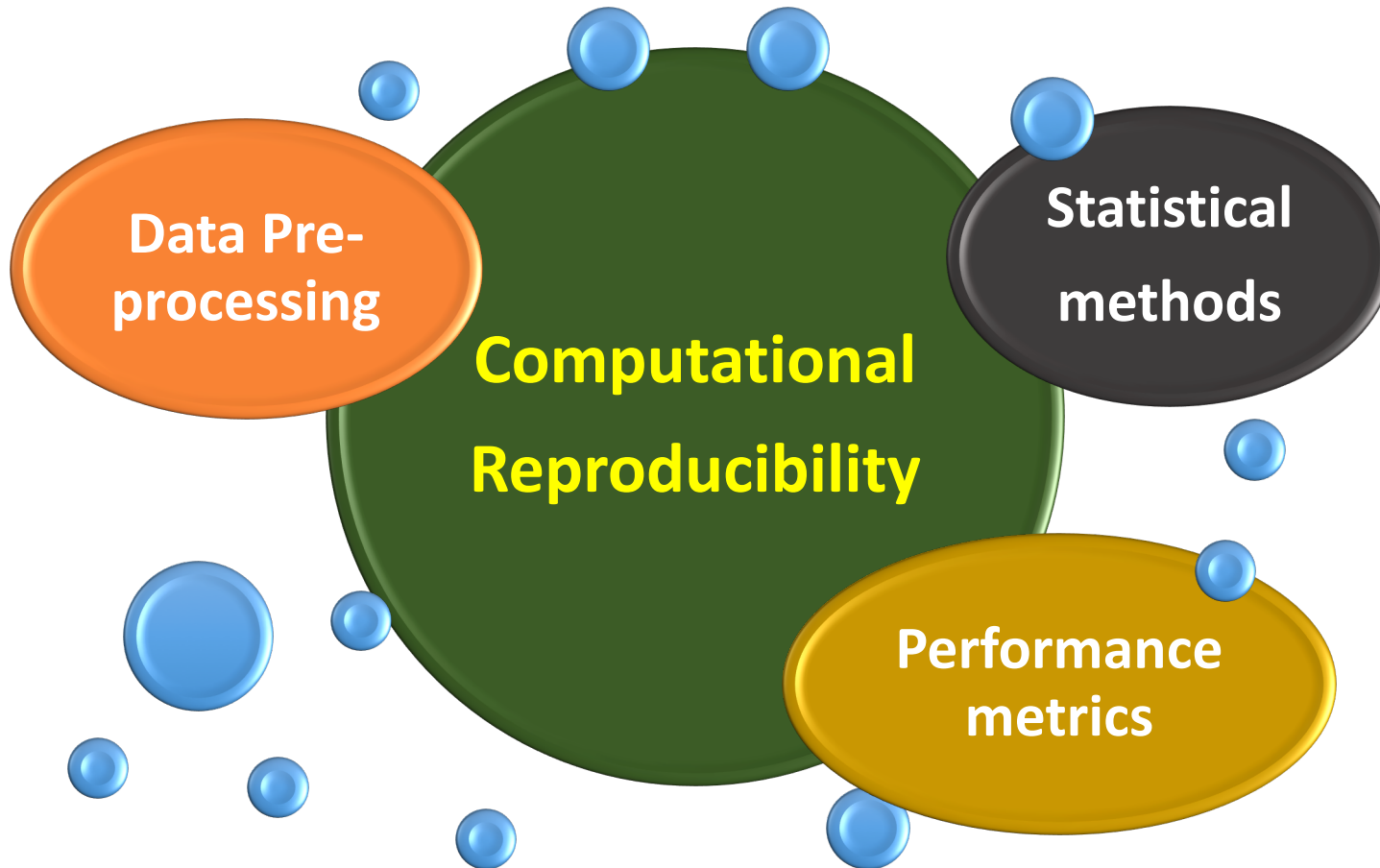


The cause of irreproducibility is diverse

- Some factors can be corrected/avoided:
 - Selective reporting: Victims of “publish and perish” paradigm
 - Sloppiness
 - QC of experimental parameters
 - Data fabrication – Evil! And hurting patients such as the Duke scandal
- Computational reproducibility – a distinct phenomena in the big data era:
 - In silico mistakes (algorithm errors): data too big to be examined individually
 - Inadequate statistics: #samples vs resources
 - Difference between analysis protocols



Computational reproducibility in analysis of high-throughput genomics data



- When come to the choice of analysis methods for genomics data, it is much to do with choosing a “less” wrong approach than a “correct” approach.

The international MAQC Society launches to enhance reproducibility of high-throughput technologies

To the Editor: Reproducibility is a fundamental hallmark of good science. The US Food and Drug Administration (FDA)-led Microarray and Sequencing Quality Control (MAQC/SEQC) consortia conducted three projects¹⁻³ to assess the reliability and reproducibility of genomics technologies, including microarrays, genome-wide association studies, and next-generation sequencing. Here, we announce that this decade-long effort has led to the formation of a new international society, the Massive Analysis and Quality Control (MAQC) Society (<http://www.maqcsociety.org>), which is dedicated to quality control and analysis of massive data generated from high-throughput technologies for enhanced reproducibility.

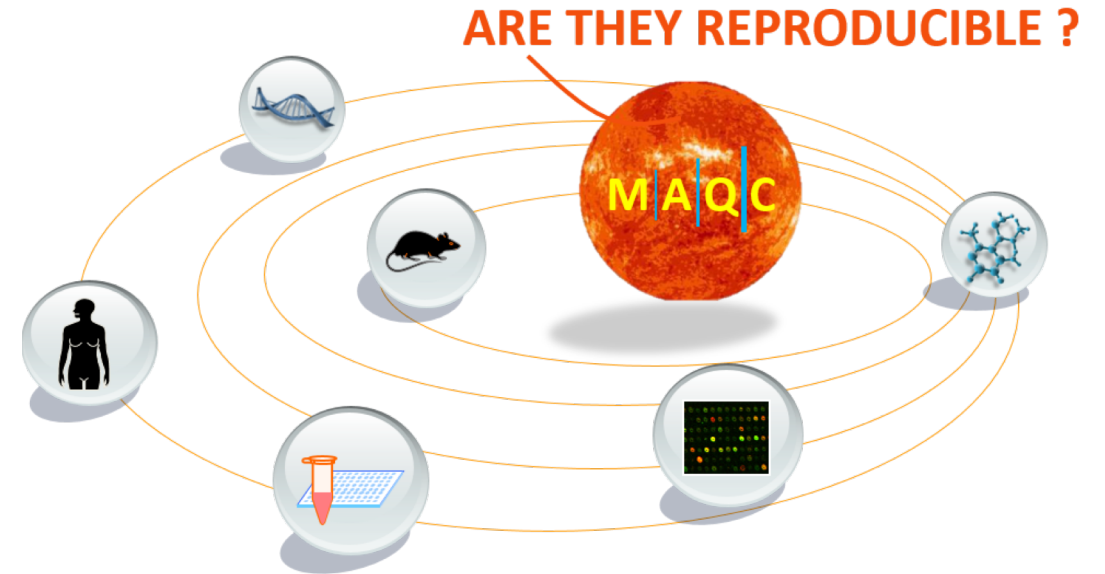
The goals of the MAQC Society are twofold: first, to advocate and facilitate enhanced reproducibility across multiple experiments, laboratories, and data analysis methods via the development and application of quality control practices and standard analysis protocols for biomedical data; and second, to advance understanding and best practices in the analysis of massive data from

across all biomedical fields who have an interest in enhancing reproducibility in the context of high-throughput approaches. The first Society meeting, which was held at SAS Institute (Cary, NC, USA) on April 12, 2017, and had a focus on 'Reproducible Genomics', gathered over 100 scientists (mostly MAQC/SEQC consortium members). The Society's membership currently comprises 35% academia, 31% government and regulatory agencies, 23% technology companies, 6% pharmaceutical companies, 4% clinicians, and 1% 'other' categories. Inspired by the productive and open discussion sessions at the first meeting, the Society seeks to foster a culture by which the members can cooperate to advance reproducibility by implementing standard protocols, pipelines, and best practices. We are promoting needed behaviors, such as the proper giving and receipt of feedback, to promote better practices in the life sciences.

To extend its mission, the MAQC Society plans to establish a series of case studies by which its members will share examples of published manuscripts with data and source

science for discovering effective therapies, identifying responder or adverse event-sensitive patient populations; indeed, clinical progress essentially depends on reliable and reproducible results. Error, whether it is human error, computational error, or technical error, and imprecise protocols could lead to irreproducible or inconsistent results that may contribute to patient risk or death. These concerns pervade high-throughput omics technologies⁴⁻¹⁶, such as microarrays^{1,2}, next-generation sequencing³, metabolomics, and proteomics for both preclinical and clinical studies.

Among various issues encountered, computational reproducibility becomes increasingly challenging in this field. This is simply due to the fact that the size of data is so massive that the manual inspection of data quality and analysis results is often impossible; thus, reproducibility remains largely at the mercy of whichever algorithm is used, which often lacks necessary benchmarking and metrics to assess reproducibility. Furthermore, a plethora of



MAQC Society Annual Conferences:

- April, 2017: NC, USA
- Feb, 2018, Shanghai, China
- April, 2019, Trento, Italy
- April, 2020, NC, USA
- April, 2021, Seville, Spain

AGE IS AN ISSUE OF
MIND OVER MATTER
IF YOU DON'T MIND
IT DOESN'T MATTER

Mark Twain



SCIENCE IS AN ISSUE OF
REPRODUCIBILITY OVER
INNOVATION, IF YOU
DON'T REPRODUCE IT
DOESN'T INNOVATE