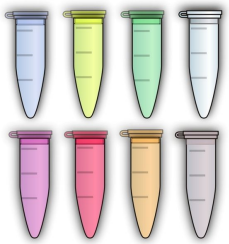
A general framework for evaluating cross-platform concordance in genomic studies

Tim Peters

BiocAsia 2018

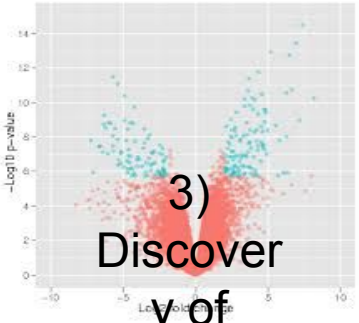
30/11/18

The path to glory…

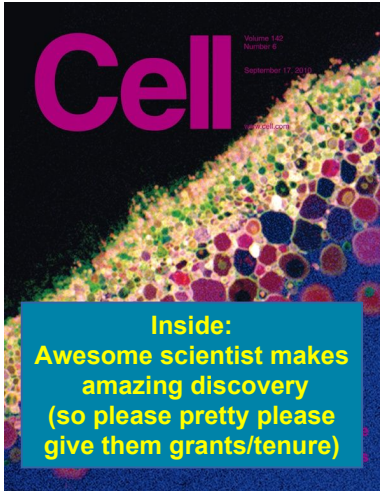


1) High-quality samples

2) Sequencing or other assay

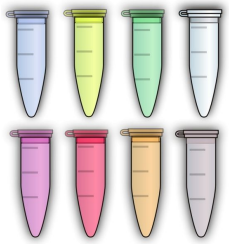


3) Discovery of novel panel of DE genes/variants/traits



4) Publicat

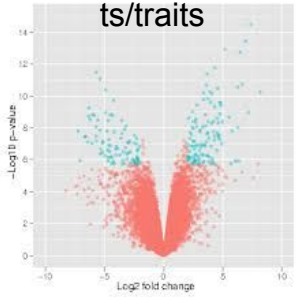
But in reality…



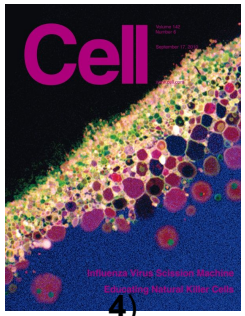
1) High-quality samples

2) Sequencing or other assay

3) Discovery of novel panel of DE genes/variants/traits



4) Submission to high-impact journal

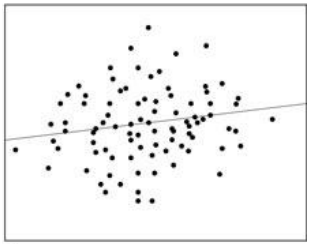


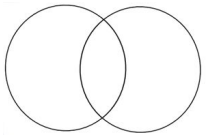
Reviewer:

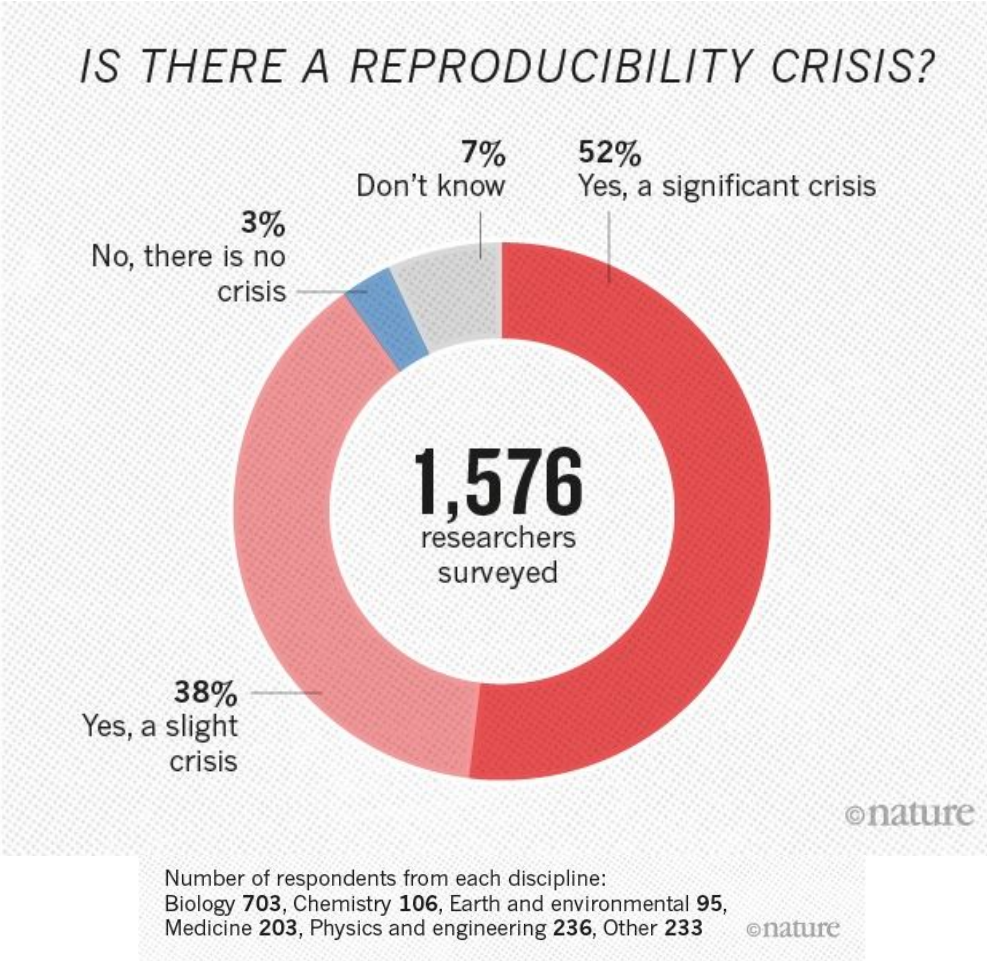
"These results need extra validation from an independent source!!!"

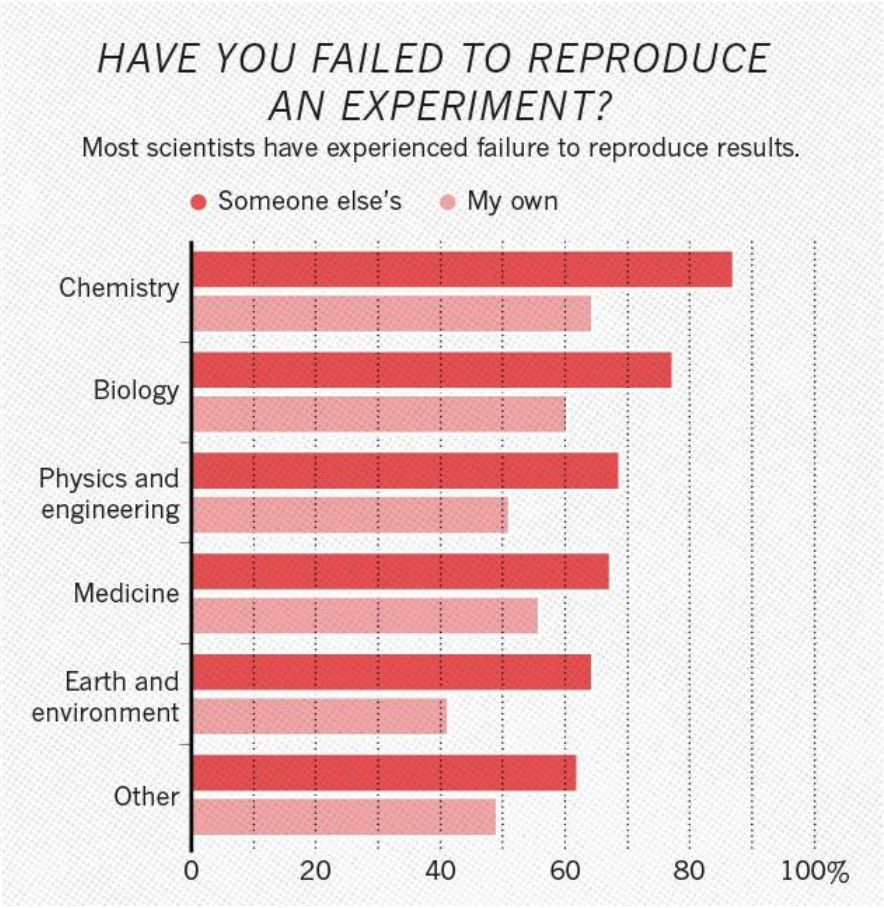
5) Validation via qPCR

6) **Poor concordance** between validation and original finding









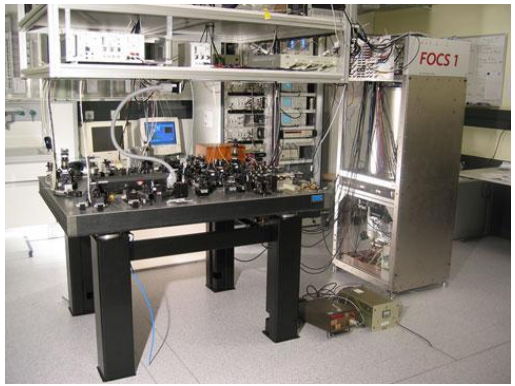
Nature 26/5/2016

The trust deficit in genomics: whence?

- **Incomplete account** of all molecules in assay
  - Stochastic sampling can amplify biases
- **Low replication** in individual studies
  - Non-robust
  - Unstable estimates of effect
- **Ambient variation** between timepoints/labs
  - i.e. batch effect
- **Systematic technical variation**
  - Fundamental difference in **measurement technique**

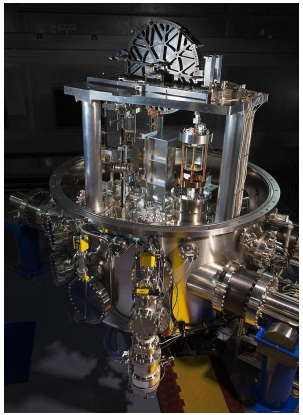
Measurement

- Measurement is what scientists do, more than anything
- Fundamental constants are critical to this
  - e.g. Planck's constant, Avogadro's constant
  - Ultra high-precision devices can approximate these



Atomic clock

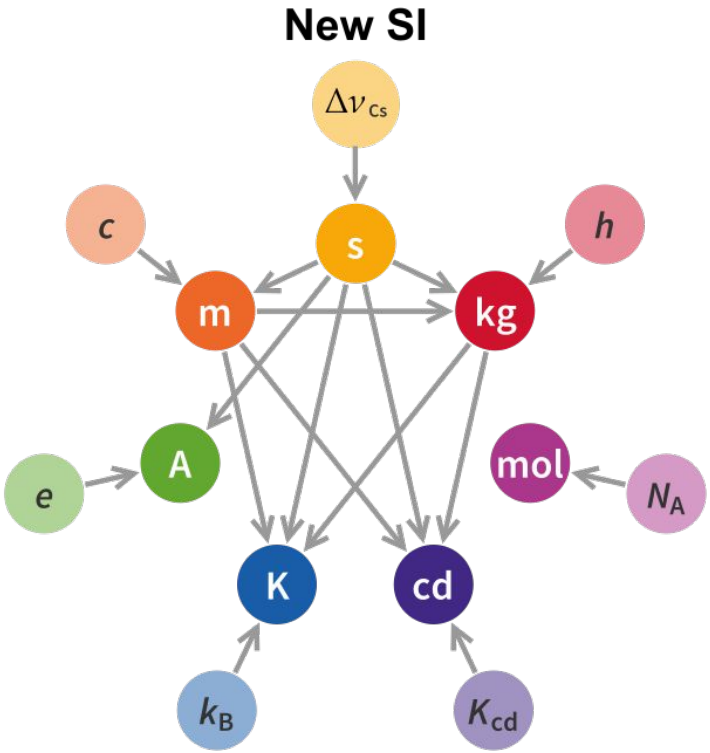
Kibble balance



No more artefacts… as of 16/11/18

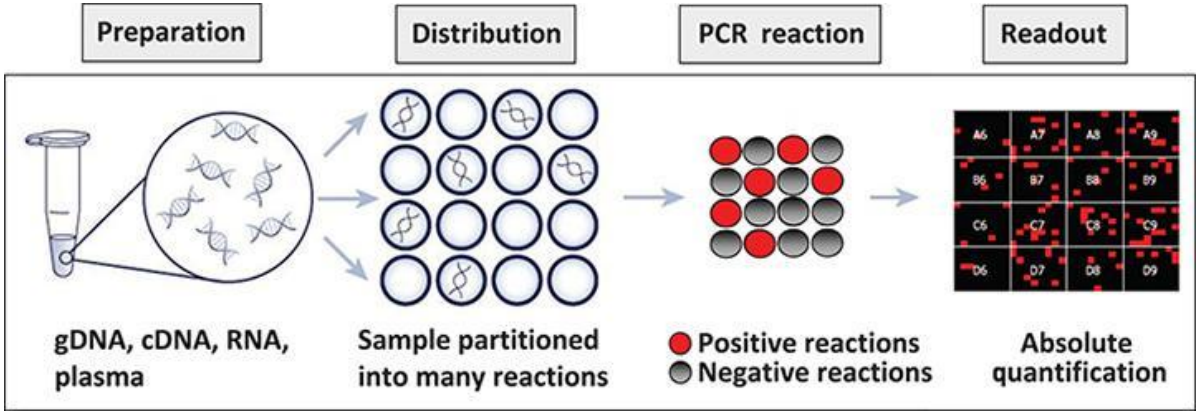
Versailles, France—A convocation of delegates representing 60 countries voted today in Versailles to implement the most significant change to the International System of Units (SI) in more than 130 years. For the first time, all measurement units will be defined by natural phenomena rather than by physical artifacts. The event was the 26th General Conference of Weights and Measures and was hosted by the International Bureau of Weights and Measures.

The U.S. delegation at the 26th General Conference of Weights and Measures where more than 55 countries voted to redefine four of the seven base units for the International System of Units (SI). L-to-R: Eric Lin, director, Material Measurement Lab, NIST; Claire Saundry, director of International and Academic Affairs Office, NIST; Willie May, U.S. member of the International Committee on Weights and



Constants for genomics?

- No real equivalent to physics/chemistry
- Spike-ins help with precision
    - Baker *et al.* (2005) *Nature Methods*
    - Deveson *et al.* (2016) *Nature Methods*
    - Hardwick *et al.* (2017) *Nature Reviews Genetics*
- Digital PCR (dPCR)
    - Sykes *et al*. (1992) *Biotechniques*



analytical chemistry

Article

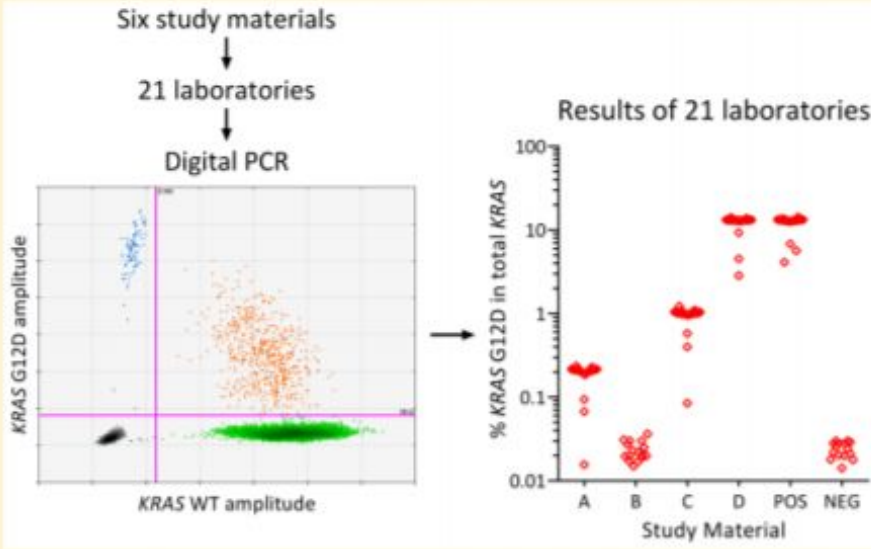
pubs.acs.org/ac

International Interlaboratory Digital PCR Study Demonstrating High Reproducibility for the Measurement of a Rare Sequence Variant

Alexandra S. Whale,[†] Alison S. Devonshire,[†] George Karlin-Neumann,[‡] Jack Regan,[‡] Leanne Javier,[‡] Simon Cowen,[§] Ana Fernandez-Gonzalez,[†] Gerwyn M. Jones,[†] Nicholas Redshaw,[†] Julia Beck,[∥] Andreas W. Berger,[⊥] Valérie Combaret,[#] Nina Dahl Kjersgaard,[▽] Lisa Davis,[○] Frederic Fina,[◆] Tim Forshew,[¶] Rikke Fredslund Andersen,[◇] Silvia Galbiati,[♦] Álvaro González Hernández,[●] Charles A. Haynes,[■] Filip Janku,[□] Roger Lacave,[◇] Justin Lee,[◑] Vilas Mistry,[▱] Alexandra Pender,[△] Anne Pradines,[▲] Charlotte Proudhon,[▼] Lao H. Saal,[⬡] Elliot Stieglitz,[⬢] Bryan Ulrich,[⌀] Carole A. Foy,[†] Helen Parkes,[†] Svilen Tzonev,[‡] and Jim F. Huggett*[,†,◒]

[†]Molecular and Cell Biology Team, LGC, Queens Road, Teddington, Middlesex TW11 0LY, United Kingdom

ABSTRACT: This study tested the claim that digital PCR (dPCR) can offer highly reproducible quantitative measurements in disparate laboratories. Twenty-one laboratories measured four blinded samples containing different quantities of a *KRAS* fragment encoding G12D, an important genetic marker for guiding therapy of certain cancers. This marker is challenging to quantify reproducibly using quantitative PCR (qPCR) or next generation sequencing (NGS) due to the presence of competing wild type sequences and the need for calibration. Using dPCR, 18 laboratories were able to quantify the G12D marker within 12% of each other in all samples. Three laboratories appeared to measure consistently outlying results; however, proper application



SEQC/MAQC-III (*Nature Biotech.* 2014)

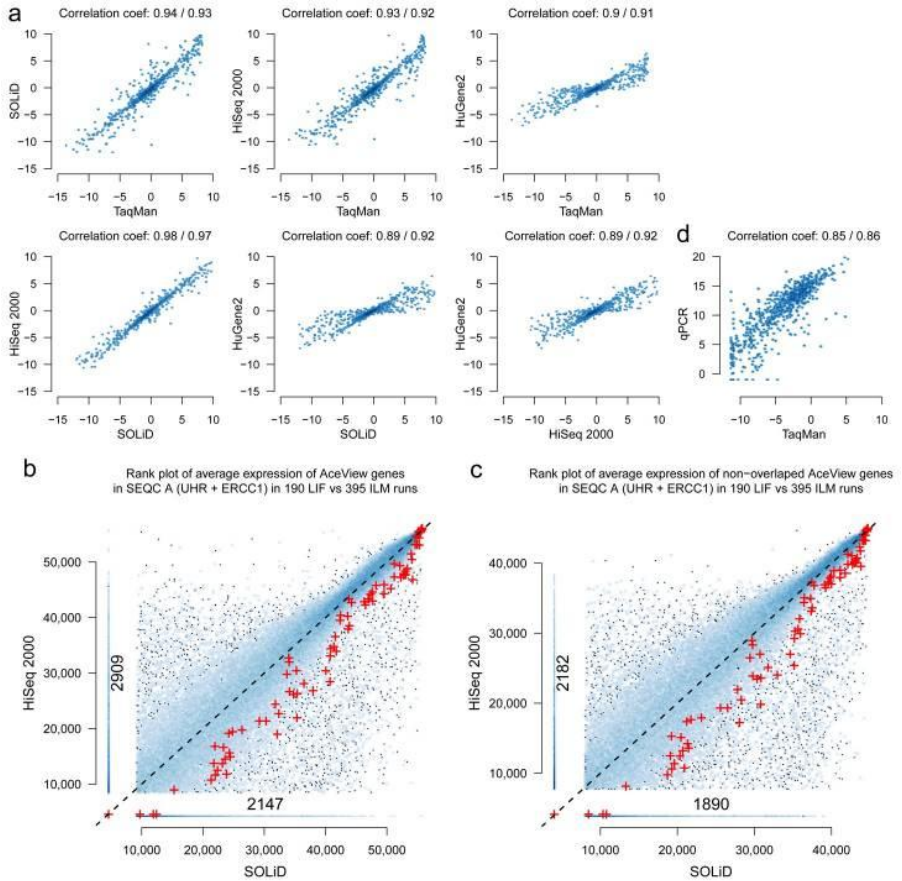


Figure 5

Cross-platform agreement of expression levels. (**a**) Comparison of log2 fold-change estimates for 843 selected genes. …(**b**) Comparison of absolute expression levels from HiSeq 2000 and SOLiD in a rank scatter density plot. Expression level ranks for sample A are shown on the *x*-axis for HiSeq 2000, and on the *y*-axis for SOLiD. Genes are represented by dots, and areas with several genes are shown in blue, with darker blue corresponding to a higher gene density in the area. **Large cross-platform deviations are seen even for highly expressed genes and these variations are systematic.** …. **ERCC spike-in signals are systematically lower in the HiSeq 2000 data, which may be explained by their shorter poly-A tails and differences in the library construction protocols.** (**c**) The same plot as (**b**) but removing the 11,066 genes that can be affected by the non-stranded nature of the applied Illumina protocol.

SEQC/MAQC-III (*Nature Biotech.* 2014) Conclusions extract (**bolded** mine)

"… Specifically, a closer examination of the **varying amount of detected ERCCs** per sample indicated **substantial differences and inconsistencies** even across libraries prepared from the same sample at the same site and sequenced by the same machine. This implies **inherent limitations for the read-out of absolute expression level estimates and absolute quantification**.

"Considering the substantial disagreements even between different types of qPCR-based assays, we conclude that there is **no single 'gold standard.'** Although our cross-platform comparisons reveal common trends, drastic systemic differences remain..."

So, should we define a gold standard?

- Not with this level of observed interlaboratory variation.
- Defining one from a single observation will **bias** all subsequent observations towards the original
- *All* genomic measurements are estimates, and error is inherent in the measurement process

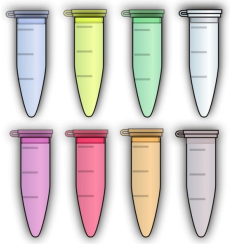
So we are left with...

- Empiricism (rather than rationalism)
- The interlaboratory test procedure is still useful – we have a lot of measurements of the same substances
- Could we leverage the information from them to
  - **characterise effects** (biases) of different technologies and laboratory conditions? (answer: yes)
  - What about **identifying deviant measurements** from the consensus (again: yes)

The row-linear model

- We don't have a reliable gold standard, but we do have a consensus (i.e. the mean)
- **What if we used a consensus "working truth" as the gold standard?**
- This approach is an official standard American Society for Testing and Materials (ASTM)
  - Standard E691: "Procedure for the design and analysis of interlaboratory tests"
  - Originally designed to benchmark performance of multiple US factories 50-60 years ago

Mandel, J. (1994). Analyzing Interlaboratory Data According to ASTM Standard E691. In *Quality and Statistics: Total Quality Management* (pp. 59-59–12).
Mandel, J., & Lashof, T. W. (1969). The Inter-laboratory Evaluation of Testing Methods. *Statistical Concepts and Procedures*, *1*, 170.



Microarray 1 Microarray 2 Sequencing qPCR

Row-linear model

Per-locus x *per-technology* estimates of **sensitivity** and **precision**

The concept (and some stats)

For **one** gene (say TP53):	LNCaP	HeSC	Liver	PBMC1	PBMC2	Row Mean a_i
Affy Microarray	2.26	3.59	2.73	4.19	3.06	3.17
Roche Microarray	8.35	8.92	8.09	8.77	8.57	8.54
PolyA+ RNASeq	7.38	8.05	7.01	7.91	7.53	7.58
RiboZero RNASeq	6.65	7.98	6.53	7.68	7.12	7.19
QPCR Lab 1	10.67	11.58	10.09	11.21	11.24	10.95
QPCR Lab 2	10.70	11.67	9.95	11.17	11.2	10.94
Column Mean x_j	7.67	8.63	7.40	8.49	8.12	

$$Z_{ij} = a_i + \boldsymbol{b_i}(x_j - \bar{x}) + \boldsymbol{d_{ij}}$$

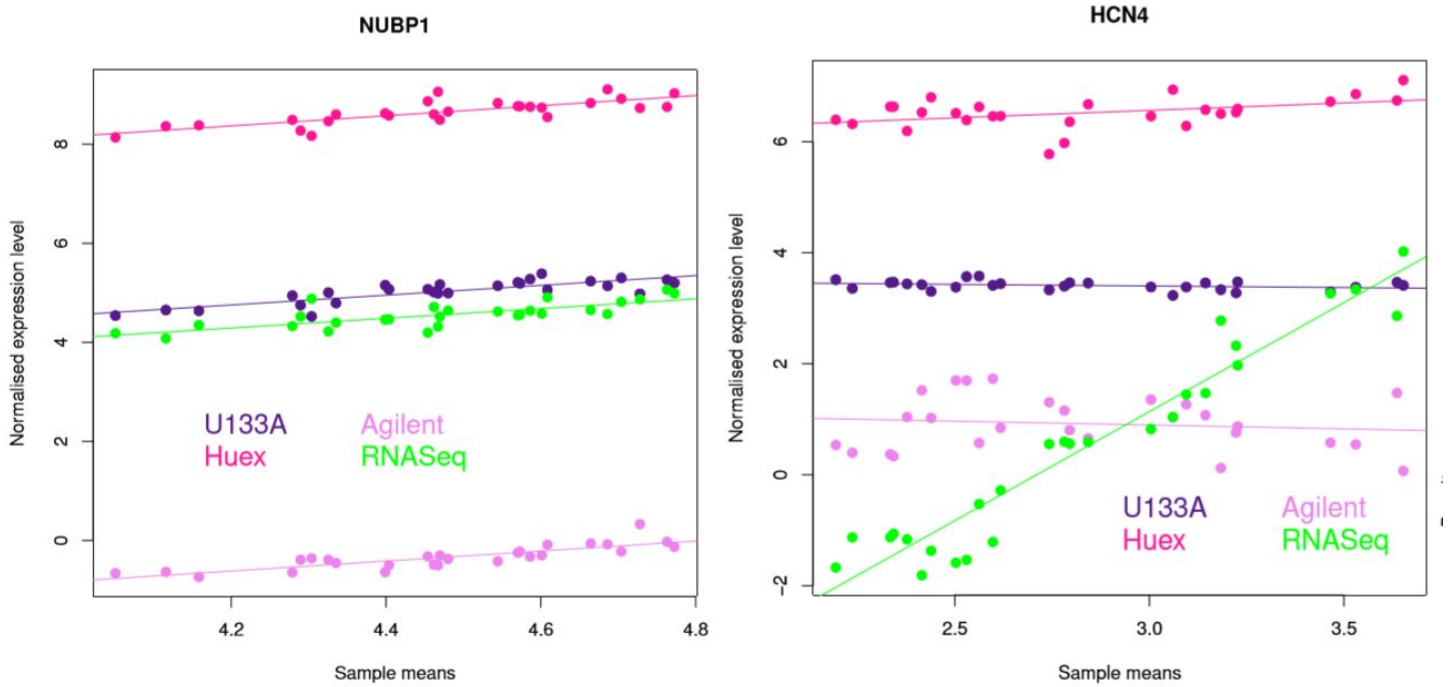
$\boldsymbol{b_i}$ - The slope of the fit for platform *i* i.e. **Sensitivity**

$\boldsymbol{d_i}$ - The scatter (residual) around the fit for platform *i* i.e. the inverse of **Precision**

Prerequisites

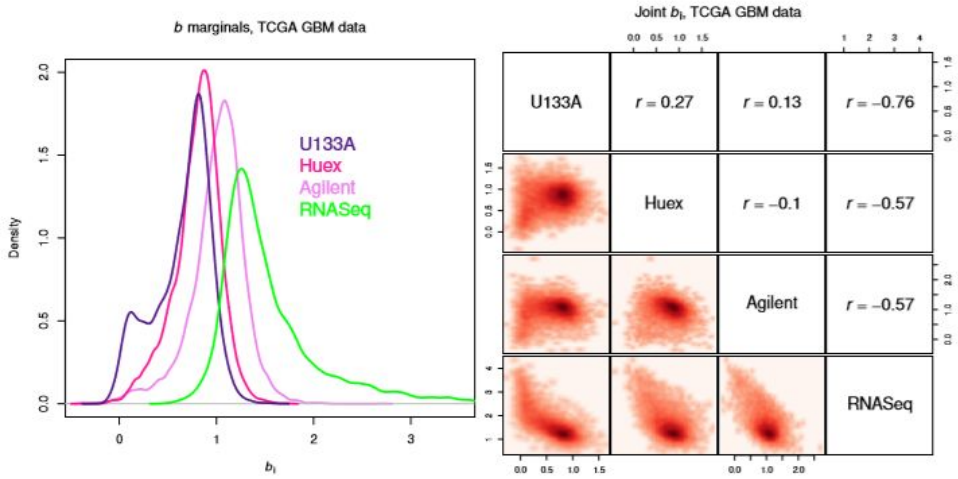
- A decent number (preferably **> 10**) of biological samples with either:
  - **Real biological variation** (i.e. gene expression not ~0; DNA methylation not all 1 or 0) OR
  - **Matched sets of aliquots** with variable NA concentrations across sets
- At least **3** different measurement technologies / strategies / lab conditions
  - Model needs 2 degrees of freedom to be fit properly

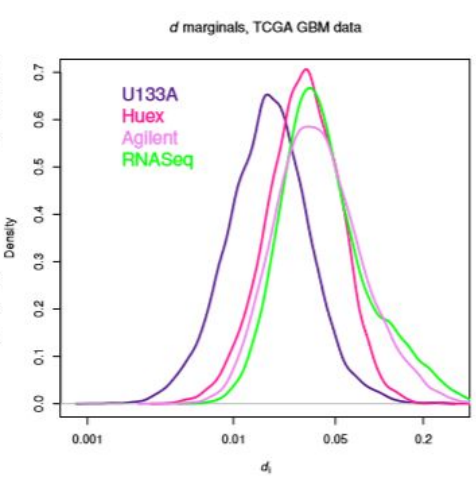
Examples – TCGA GBM data (Verhaak *et al.* 2010)

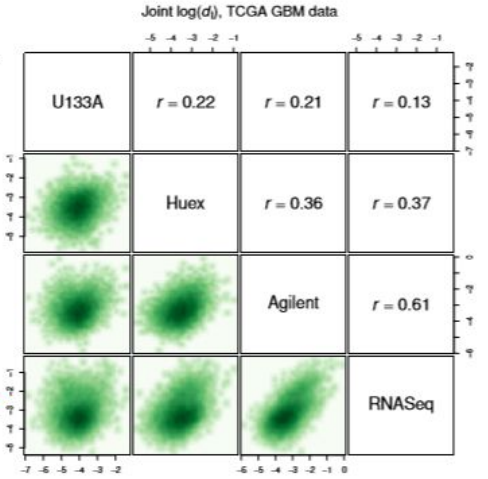


b_i - The slope of the fit for platform *i* i.e. **Sensitivity**
d_i - The scatter (residual) around the fit for platform *i*
i.e. the inverse of **Precision**

TCGA GBM data: all 9,519 genes





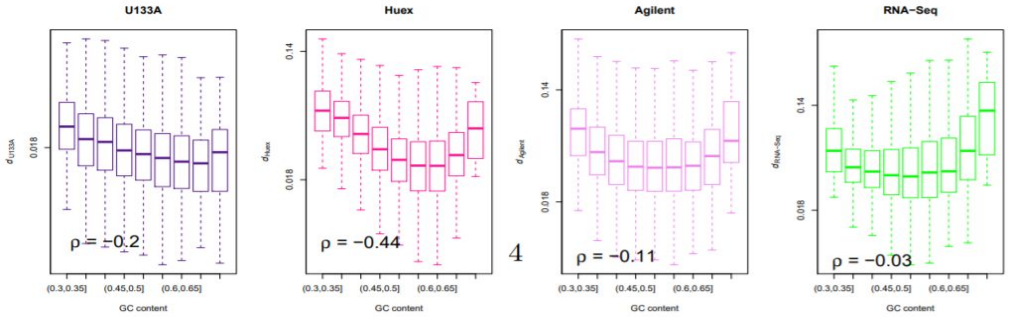


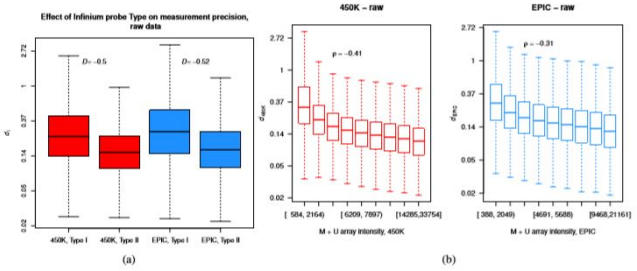
b_i peaks to the **right** -> (e.g. RNA-Seq) mean that platform is **more sensitive** to change in gene expression

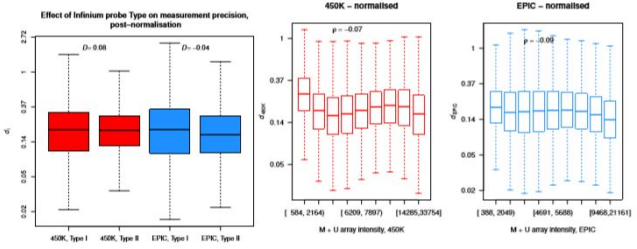
d_i peaks to the **left <-** (e.g. U133A) mean that platform's measurements are **more precise**

Gene-specific biases and interfering traits

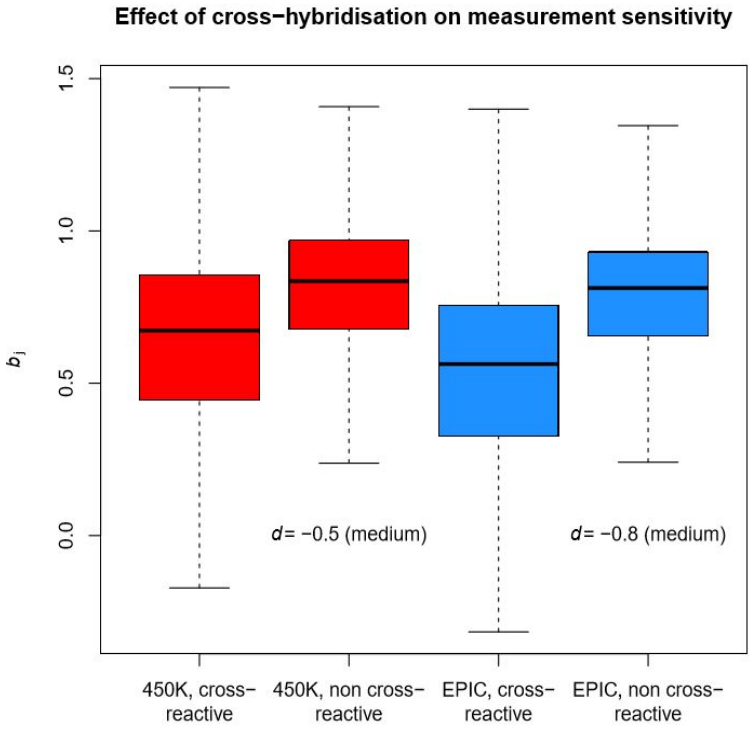
- Locus-wise sensitivities and precisions can be plotted against
  - GC content
  - CDS length
  - Coding vs. non-coding loci
  - Potential cross hybridisation events
- The row-linear model can also be used to evaluate normalisation strategies:

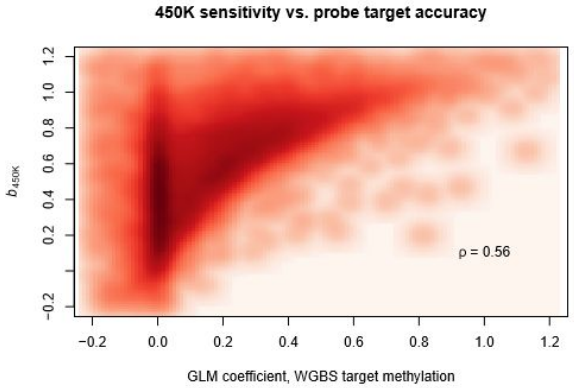


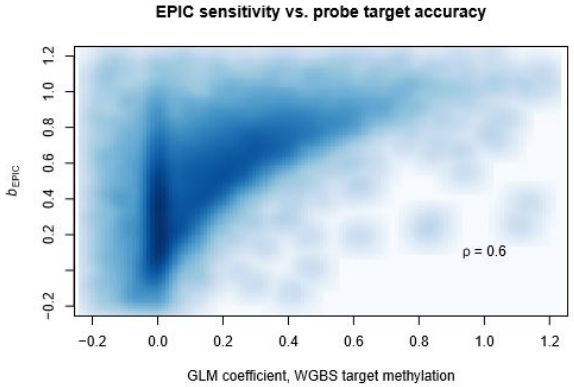




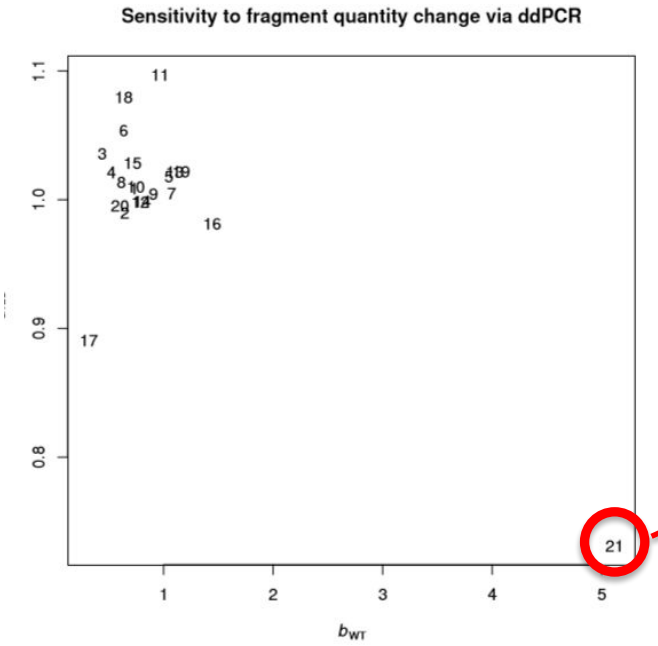
Cross-hybridisation blunts methylation array sensitivity



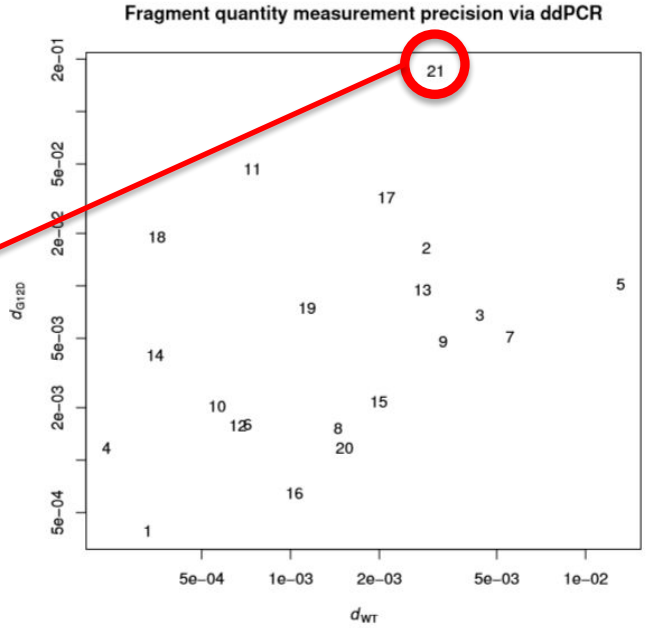




A true interlaboratory test - dPCR

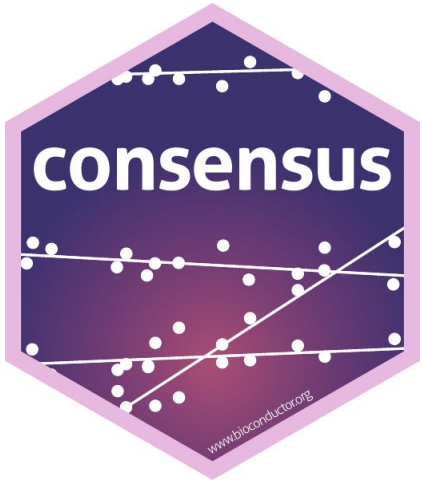


Laboratory 21 had the greatest number of deviant analysis parameters of all testing sites, including unique models of PCR plate sealer and thermal cycler



Package / Paper

- Bioconductor package *consensus*
  - Performs row-linear models
  - Plots marginal distributions in density
  - Plots joint distributions in smoothScatter
  - Plots least concordant genes/features in heatmap




```
tcga_mm <- MultiMeasure(names=c("U133A", "Huex", "Agilent", "RNA-Seq"),
                        data=list(U133A, Huex, Agilent, RNASeq))
tcga_mm

## MultiMeasure object with 4 platforms/conditions, 27 samples and 1000 measured loci.
```

- Peters, T. J. *et al*. Evaluation of cross-platform and interlaboratory concordance via consensus modelling of genomic measurements. *Bioinformatics*. In press.

Conclusions

- The row-linear consensus method can empirically assess the measurement quality of a suite of genomic platforms and laboratory conditions, **independent of the biological variation in the data** (provided that variation > 0)
- It can be used as a **screening procedure** to identify platforms with deviant measurements, and subsequently remove them

Thank you

Walter+Eliza Hall
Institute of Medical Research

DISCOVERIES FOR HUMANITY

Terry Speed

Clark Lab

Hugh French
Ruth Pidsley
Shalima Nair
Wenjia Qu
Elena Zotenko
Clare Stirzaker
Jenny Song
Aaron Statham
Kate Giles
Susan Clark

Stephen Bradford
Hilal Varinli
Peter Molloy
Firoz Anwar

SCIENCE AND
INDUSTRY
ENDOWMENT
FUND

Australian Government
National Health and Medical Research Council

UNSW SYDNEY

Ramaciotti Centre for Genomics