

Public-private collaborations aiming to improve the reliability and consistency of clinical tests to guide cancer care

Lisa M. McShane, PhD

U.S. National Cancer Institute, National Institutes of Health

Joint Statistical Meetings
August 6, 2020

(Virtual presentation)

Precision medicine

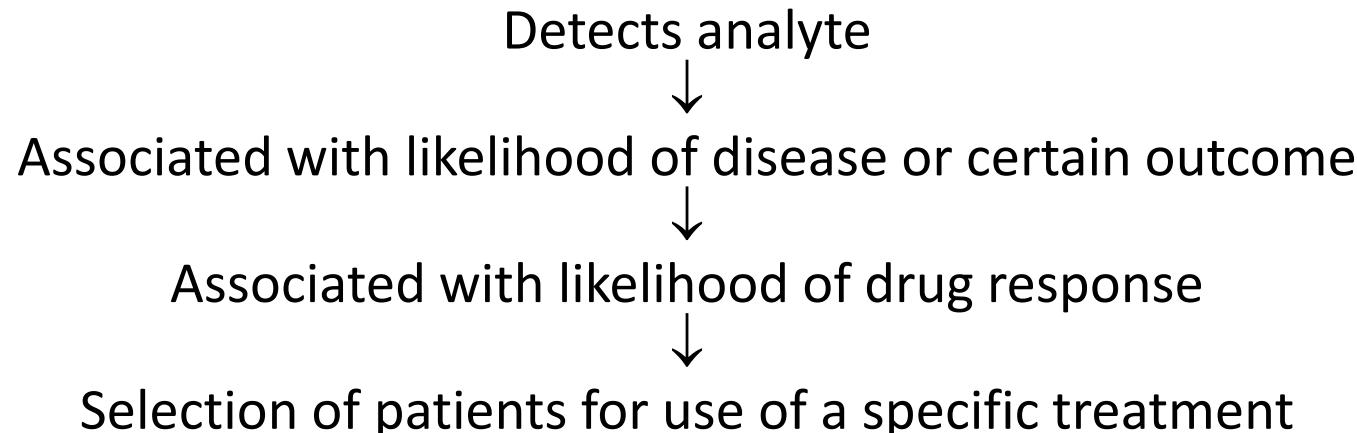
“An emerging approach for disease treatment and prevention that takes into account individual variability in genes, environment, and lifestyle for each person.”*

- Predict more accurately which treatment and prevention strategies for a particular disease will work in which groups of people
- Often reliant on biomarker-based clinical tests, e.g., those that measure single analytes or “signatures” derived from high-dimensional omics data (e.g., genomics, transcriptomics, proteomics, epigenomics, metabolomics, etc.)

*<https://ghr.nlm.nih.gov/primer/precisionmedicine/definition>

Regulatory oversight of biomarker tests compared to therapeutics

- New **therapies undergo rigorous evaluation** by the FDA to ensure safety and effectiveness before marketing
- Biomarker test regulatory oversight is **inconsistent**
 - CLIA regulations (administered by CMS) require laboratories follow **good laboratory practice**
 - Clinical performance of tests not directly evaluated by CMS
 - FDA review of tests focuses on those **marketed as kits or devices, varied claims**



Regulatory oversight of biomarker tests compared to therapeutics

- Large percentage of clinical tests are run as “laboratory developed tests” (LDTs, or “home brew” assays)
 - Subject to regulatory “enforcement discretion,” i.e., often no evaluation by the FDA or any other independent body
 - No independent clinical validation
 - No guarantee that tests intended to measure the same thing yield similar results
 - Impossible for FDA to review every LDT in existence
- Efforts to evaluate or ensure validity and reproducibility of biomarker tests often led by other interested parties, e.g., researchers, government, insurers, research funders, professional organizations, patient advocacy groups

Examples of collaborations between government, academia, industry, and advocacy groups to promote reliability, reproducibility, and harmonization of biomarker-based tests

- **Ki67:** Reproducibility assessment and standardization of Ki67 IHC proliferation biomarker evaluation in breast cancer
- **Omics predictors:** Criteria for the use of omics-based predictors in clinical trials
- **TMB:** Harmonization of tumor mutation burden (TMB) measurement for use in guiding immunotherapy for cancer

Ki67 immunohistochemistry (IHC) in breast cancer

- Assay method
 - Section of formalin fixed paraffin embedded tumors cut & mounted on glass slide
 - Stain with an antibody that detects the Ki67 proliferation protein in nucleus of cells
 - Record the percentage of *tumor cells* exhibiting Ki67 nuclear staining
 - Typically, the more proliferation, the more aggressive the cancer
- Ki67 sometimes used in combination with other clinico-pathologic factors to decide whether certain patients with early stage breast cancer need chemotherapy after surgery

Ki67 immunohistochemistry (IHC) in breast cancer

- Systematic review (Stuart-Harris et al, *The Breast* 2008; 17: 323-334)
 - 43 studies of Ki67 in early breast cancer, ≥ 100 patients
 - English publication, Jan. 1995 – Sept. 2004
 - Examine association of Ki67 with OS or PFS endpoint
 - Ki67 IHC typically performed by “home brew” assays
 - 7 different antibodies for Ki67, single or in combination
 - 19 different cut-points to define “Low” vs. “High,” ranging from 0-30%
 - Significant between-study heterogeneity and publication bias
- International Ki67 Working Group convened in London in March 2010 (Dowsett et al, *J Natl Cancer Inst* 2011; 103: 1656–1664)
 - Harmonization/standardization of Ki67 assays identified as a key need

Ki67 IHC reproducibility study #1

Design and variance components analysis

- 8 labs assessing different sections of same 100 breast tumors (tissue microarray, TMA)
- One set stained centrally; other set each laboratory stained its own by local method
- Model: $y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$

$y_{ij} = \log_2(\text{Ki67}+0.1)$ for the i^{th} patient ($i = 1, \dots, 100$) by the j^{th} lab ($j = 1, \dots, 8$)

(Ki67 = % tumor cells staining positive for Ki67)

μ = intercept (overall mean)

$\alpha_i \sim N(0, \sigma_P^2)$ (patient)

$\beta_j \sim N(0, \sigma_L^2)$ (lab)

$\varepsilon_{ij} \sim N(0, \sigma_E^2)$ (residual error)

Intraclass correlation:

$$ICC = \sigma_p^2 / (\sigma_p^2 + \sigma_L^2 + \sigma_E^2)$$

(Proportion of total variation explained by patient)

All random effects assumed independent.

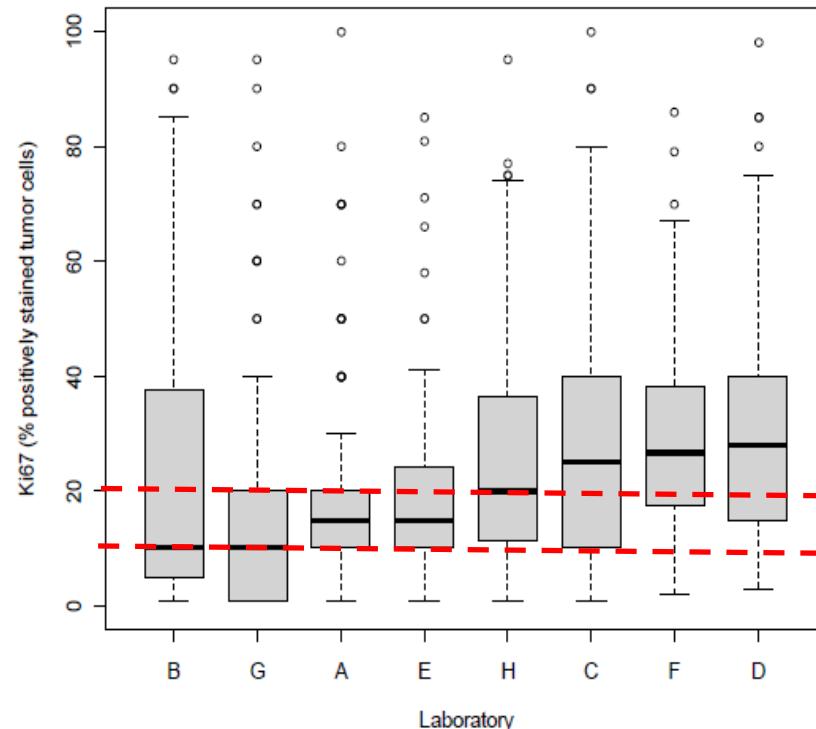
Estimates of variance components obtained using *lmer* in R package *lme4*.

Credible intervals for *ICC* computed using *HPDinterval* in R package *MCMCglmm*.

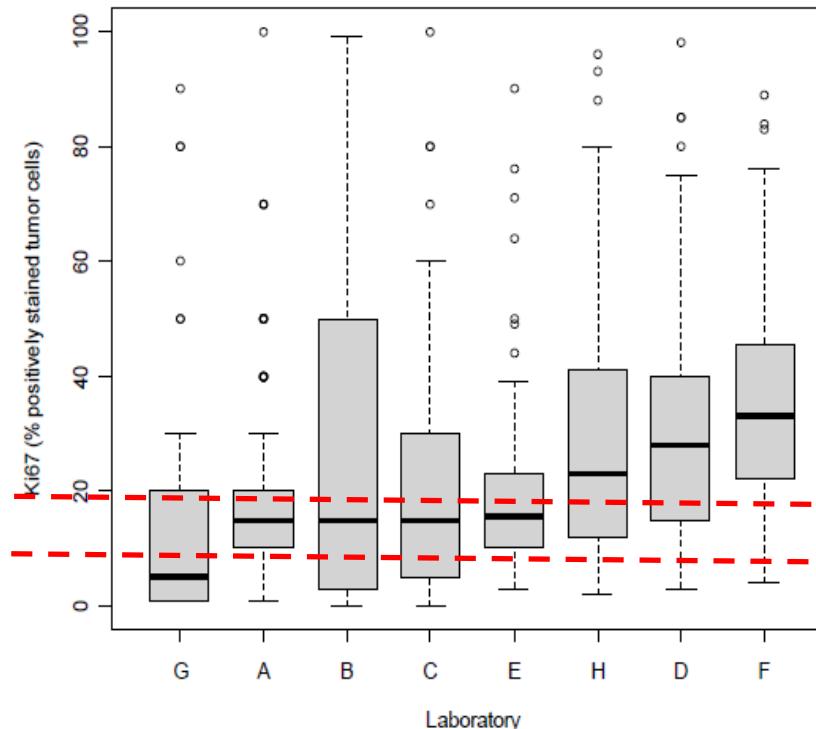
Ki67 IHC reproducibility study #1

Boxplots of Ki67 (% positive invasive tumor cells) with 8 labs assessing different TMA sections of same set of 100 breast tumors

Most cut-offs used to aid in clinical decisions about administering chemotherapy are in 10-20% range



Centrally stained, locally scored
Median range: 10% to 28%
ICC: 0.71, 95% CI=(0.47,0.78)



Locally stained, locally scored
Median range: 5% to 33%
ICC: 0.59, 95% CI=(0.37,0.68)

Developed Ki67 IHC standardized visual scoring methods

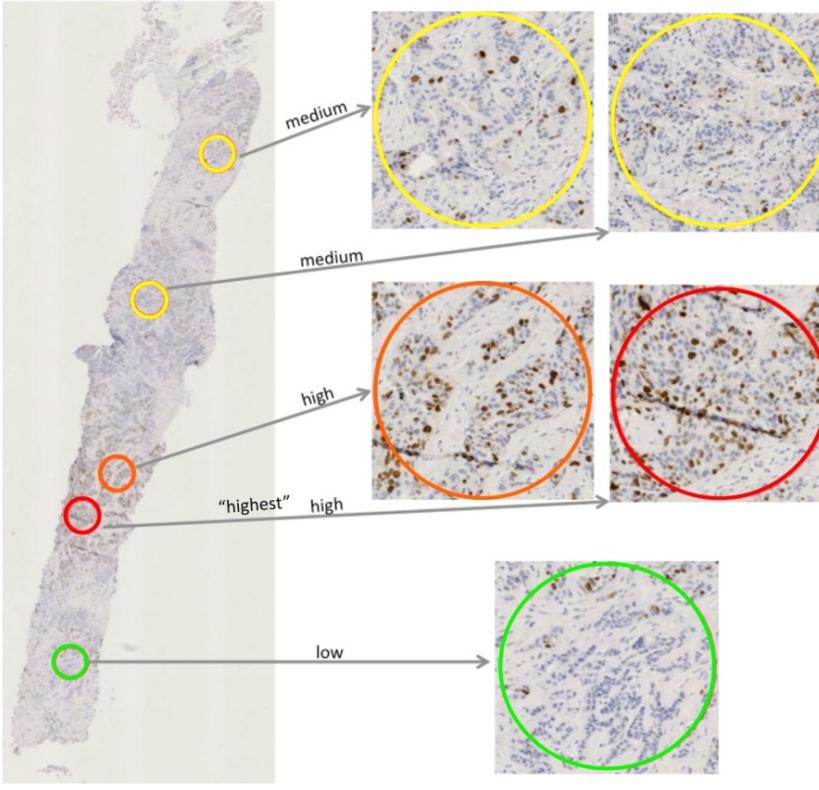


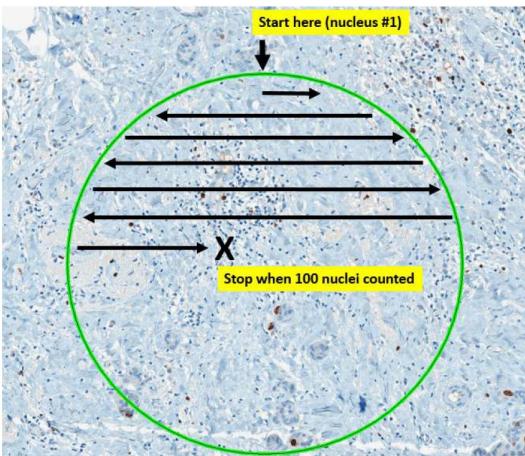
Image: 2 Ki67 -2013-08-29 16.18.49.ndpi

“Global” Scoring Method (2 versions):

- Select, mark and evaluate up to four fields of view (100 cells each) representative of regions with negative, low, medium, and high Ki67 staining rate.
- Estimate percentage of invasive tumor areas on slide represented by each region
- Estimate percent staining in each field of view (counting in typewriter pattern).
- Combine these estimates in both weighted and unweighted averages according to prespecified methods.

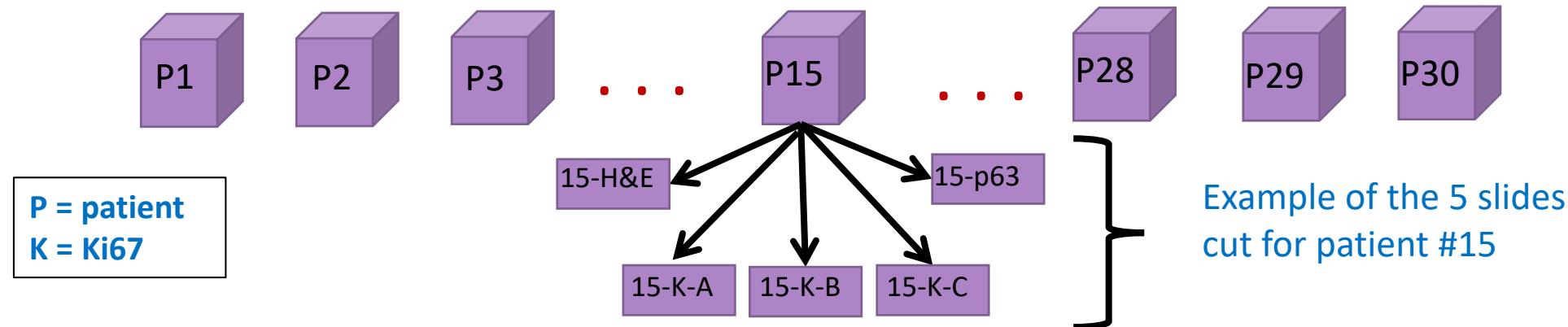
“Hotspot” Scoring Method:

- Select, mark and evaluate up to 500 tumor cells in a field of view that contains the region of highest Ki67 staining rate on the slide (the “hotspot”).
- Estimate percent staining, counting in typewriter pattern.



Ki67 IHC reproducibility study #3 schema

14-gauge core cut biopsies from 30 patients (range of Ki67 representative of ER+ breast cancer)



Cut 5 adjacent sections from *each block* & centrally stain:
1 H&E, 3 Ki67 (A-C labels randomly assigned), 1 p63 (myoepithelial marker)

Prespecified criterion for success: Lower 95% CI limit for ICC > 0.80 for any of 3 scoring methods

labs needed for adequate study power: 21 # recruited: 24 # actually completing study: 22

Each lab received:

- One of the three sets of 30 Ki67 stained glass slides, for visual scoring
- Access to a scanned image of each of their slides, for digital markup of the areas scored
- Access to scanned H&E and p63 images from each case, for reference if needed

Ki67 IHC reproducibility study #3

Spaghetti plots of Ki67 (% positive invasive tumor cells) with 22 labs (11 countries) assessing three sets of sections from a common set of 30 primary ER+ breast cancer core biopsies using three standardized scoring methods

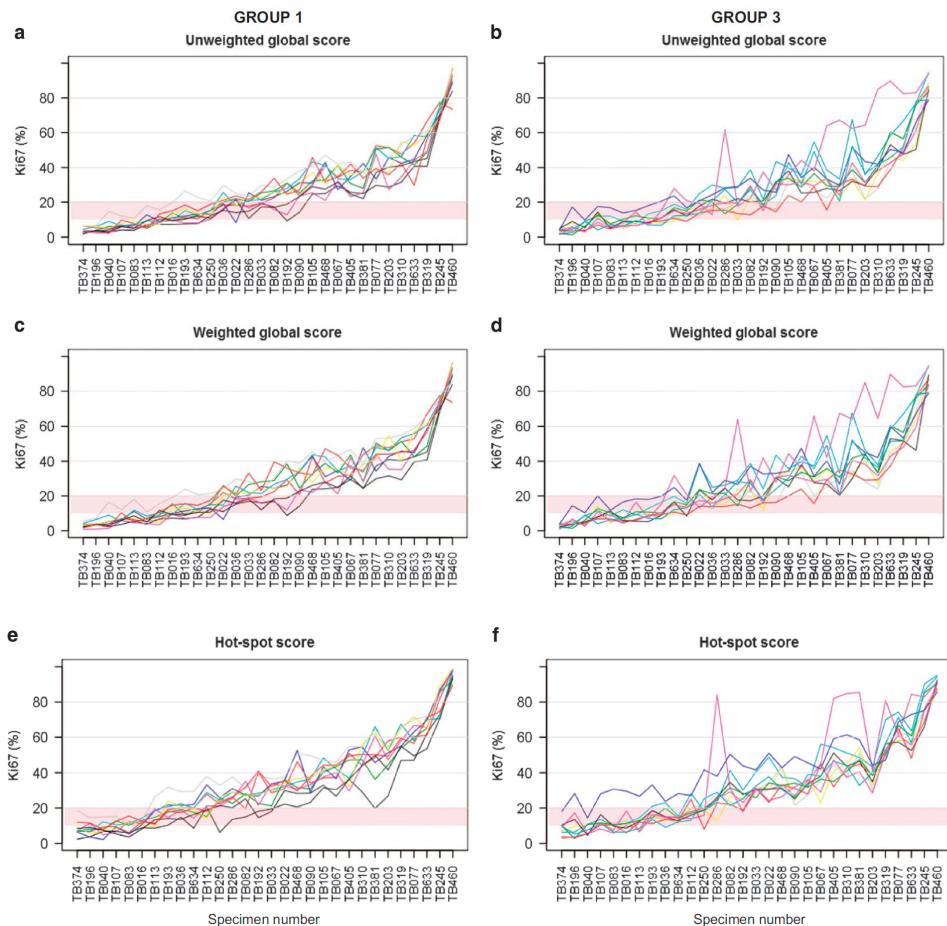


Figure 3. Variability in Ki67 scores (a, c and e correspond to Group 1; b, d and f correspond to Group 3). Each line represents Ki67 scores from one laboratory. Shaded region indicates Ki67 scores between 10 and 20%. Scores from Group 2 are not shown since there are only two laboratories in this group.

Unweighted global method
ICC: 0.87, 95% CI=(0.81,0.93)

Weighted global method
ICC: 0.87, 95% CI=(0.7999,0.93)

Hot-spot method
ICC: 0.84, 95% CI=(0.77,0.92)

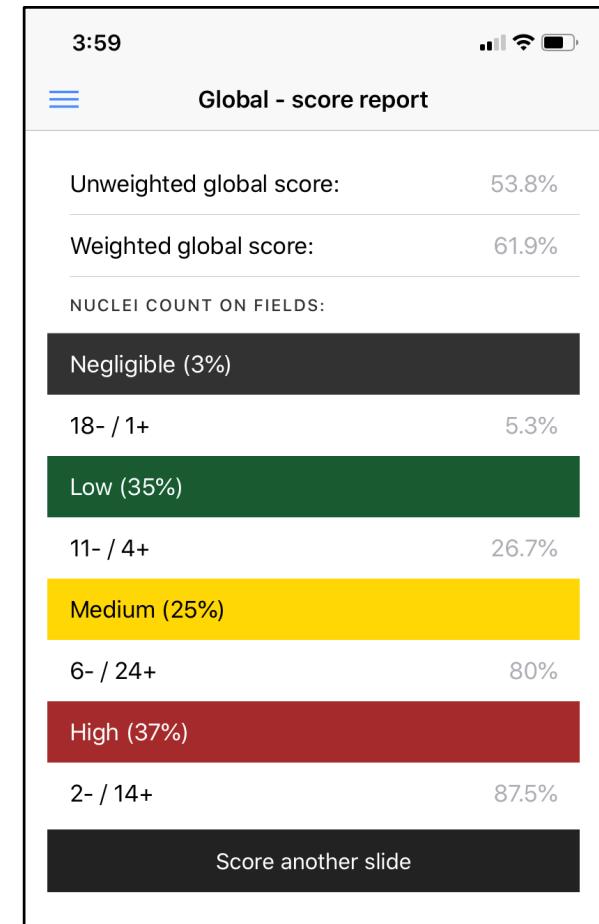
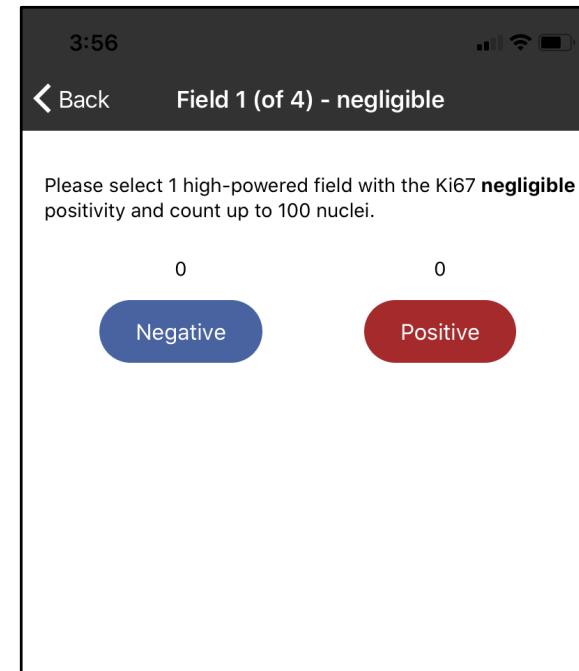
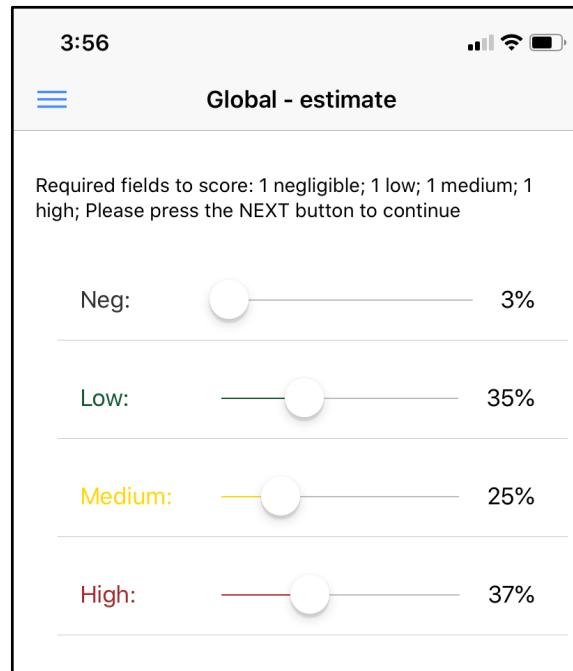
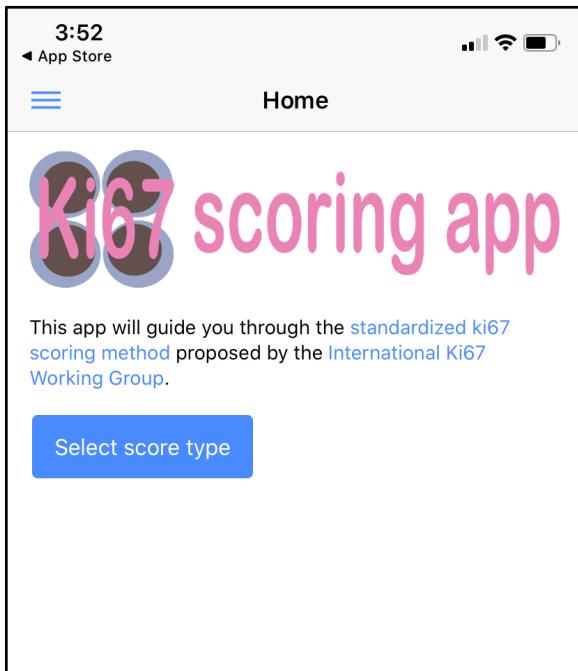
Slide set 2



Ki67 IHC standardized visual scoring methods

Ki67 scoring app for smartphones*

- Guides user through global and hotspot methods of standardized counting (including pdf of publication)
- Provides buttons for counting
- Calculates final summary results



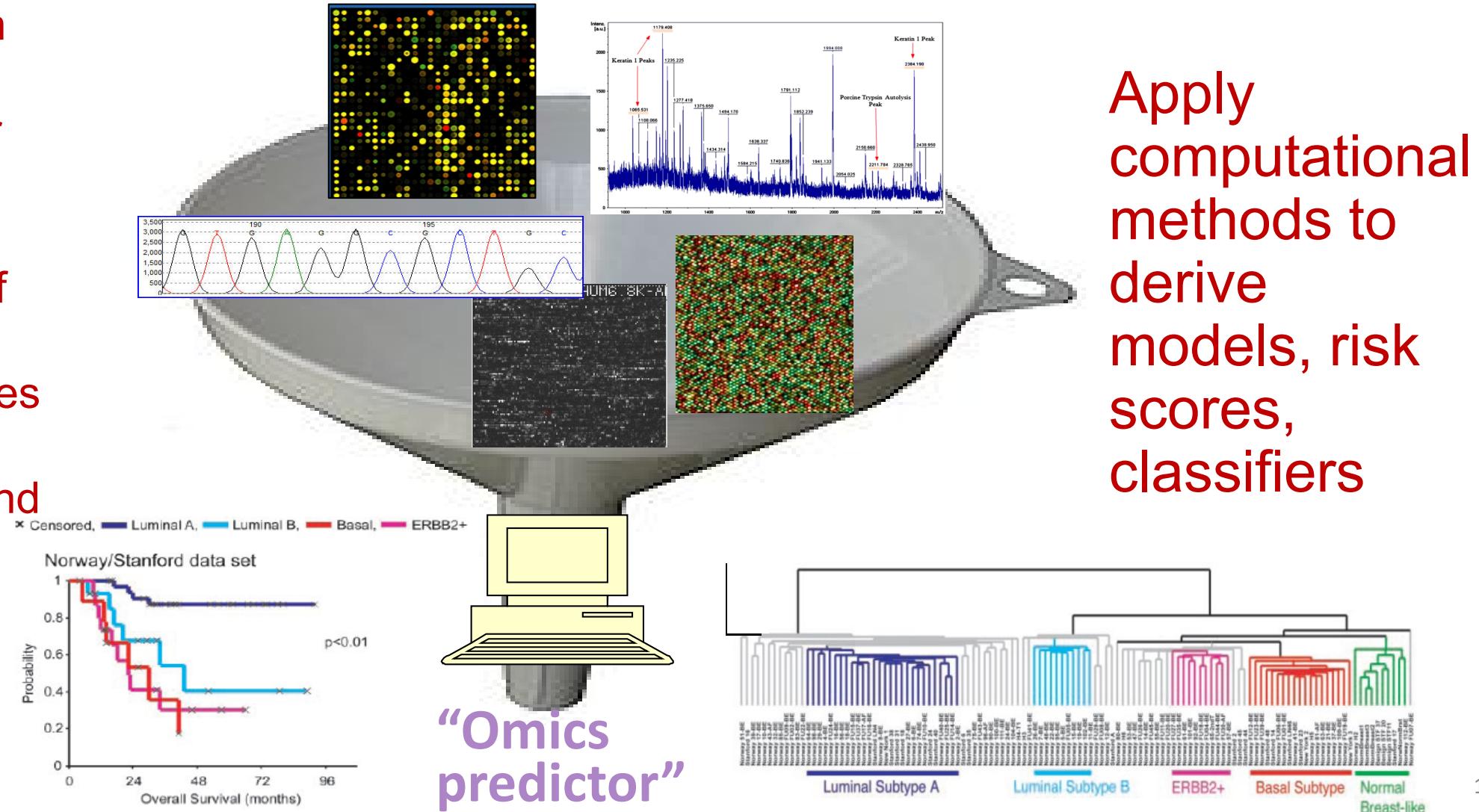
*Developed by Samuel Leung, Genetic Pathology Evaluation Centre, University of British Columbia; available in App Store

Translation from omics discoveries to clinically useful omics-based tests to guide clinical care

High-throughput omics assays (“big data” with $n \ll p$)

“Omics” is a term encompassing multiple molecular disciplines, which involve the characterization of global sets of biological molecules such as DNAs, RNAs, proteins, and metabolites.”

<http://www.iom.edu/Reports/2012/Evolution-of-Translational-Omics.aspx>



Omics tests: Skepticism, disappointment, and scandal

OvaSure test to detect ovarian cancer in women at high risk

FDA letter to LabCorp (Aug. 7, 2008):

“It has come to our attention that you are currently marketing the OvaSure Yale Ovarian Cancer Test . . .”

“Based on our review . . . we believe you are offering a high risk test that has not received adequate clinical validation, and may harm the public health”

<https://www.fda.gov/medical-devices/ivd-regulatory-assistance/ovasuretm-manufacturer-letter>

MISSING THE MARK

Why is it so hard to find a test to predict cancer?

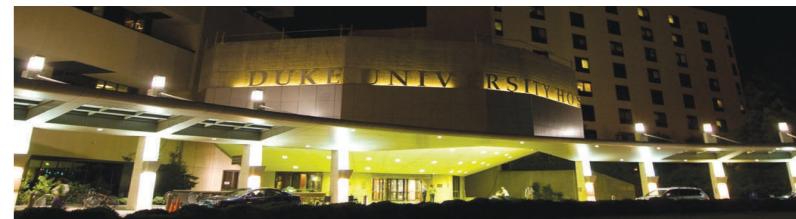
BY LIZZIE BUCHEN

In March, two studies appeared online that offered 19 pages of gloomy reading for anyone interested in cancer. They

women — to ask whether these seemingly breakthrough biomarkers were better at identifying women with early ovarian cancer than contrast, detected 63%). Mor’s panel already had a tortured history. A primary research paper behind it had been criticized by other

Lizzie Buchen, *Nature*, v. 471, March 24, 2011

Genomic chemo-sensitivity & prognostic tests to guide treatment decisions



Duke University allowed a controversial set of clinical trials to continue despite serious concerns about the validity of the data on which they were based.

ETHICS

Cancer trial errors revealed

University officials admit data withheld from review panel before misconduct charges arose.

BY EUGENIE SAMUEL REICH

It was a weekend that Michael Cuffe,

Yet Cuffe and Kornbluth had decided to restart them when a review panel seemed to validate Potti's method. The allegations that Potti,

Freedom of Information Act, Kornbluth and Cuffe have offered their account of the mistakes that led the trials to be restarted even after they



World politics

Business & finance

Economics

Science & technology

Culture

Blogs

Debate & discuss

Multimedia

Print edition

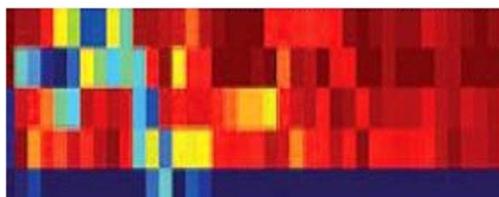
Misconduct in science

An array of errors

Investigations into a case of alleged scientific misconduct have revealed numerous holes in the oversight of science and scientific publishing

Sep 10th 2011 | from the print edition

ANIL POTTI, Joseph Nevins and their colleagues at Duke University in Durham, North Carolina, garnered widespread attention in 2006. They reported in the *New England Journal of Medicine* that they could predict the course of a patient's lung cancer using devices called expression arrays, which log the activity patterns of thousands of genes in a sample of tissue as a



Eugenie Samuel Reich,
Nature, v. 469, January 13, 2011

> 100 patients enrolled on clinical trials using faulty genomic tests

Rabiya S. Tuma, *The Economist*, September 10, 2011

Omics test scandal prompts Institute of Medicine study to examine field of translational omics

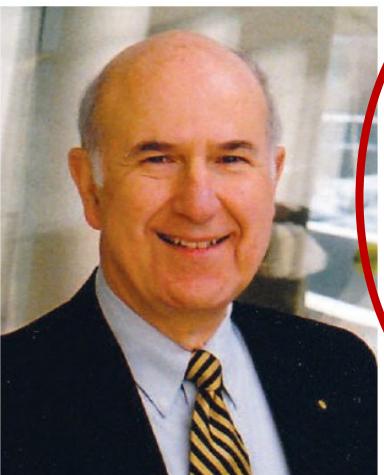
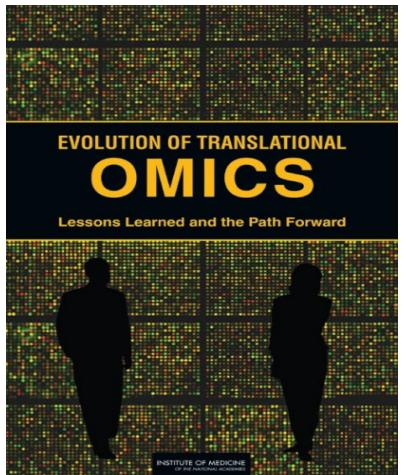
NEWS&ANALYSIS

Jocelyn Kaiser

CLINICAL MEDICINE

Science, v. 335, March 30, 2012

Biomarker Tests Need Closer Scrutiny, IOM Concludes



"There are a lot of lessons here that surely apply to other places."

—GILBERT S. OMENN,
UNIVERSITY OF
MICHIGAN,
ANN ARBOR

NCI Omics Workshop: Held in June 2011 to reinforce and operationalize the principles outlined in the IOM report

Development of checklist covering 5 domains: Specimens, assays, prediction model development & validation, clinical trial design, ethical, legal & regulatory issues

<http://www.iom.edu/Reports/2012/Evolution-of-Translational-Omics.aspx>

NCI criteria for the use of omics-based predictors in clinical trials:

McShane et al. *Nature* 2013;502:317-320 (checklist)

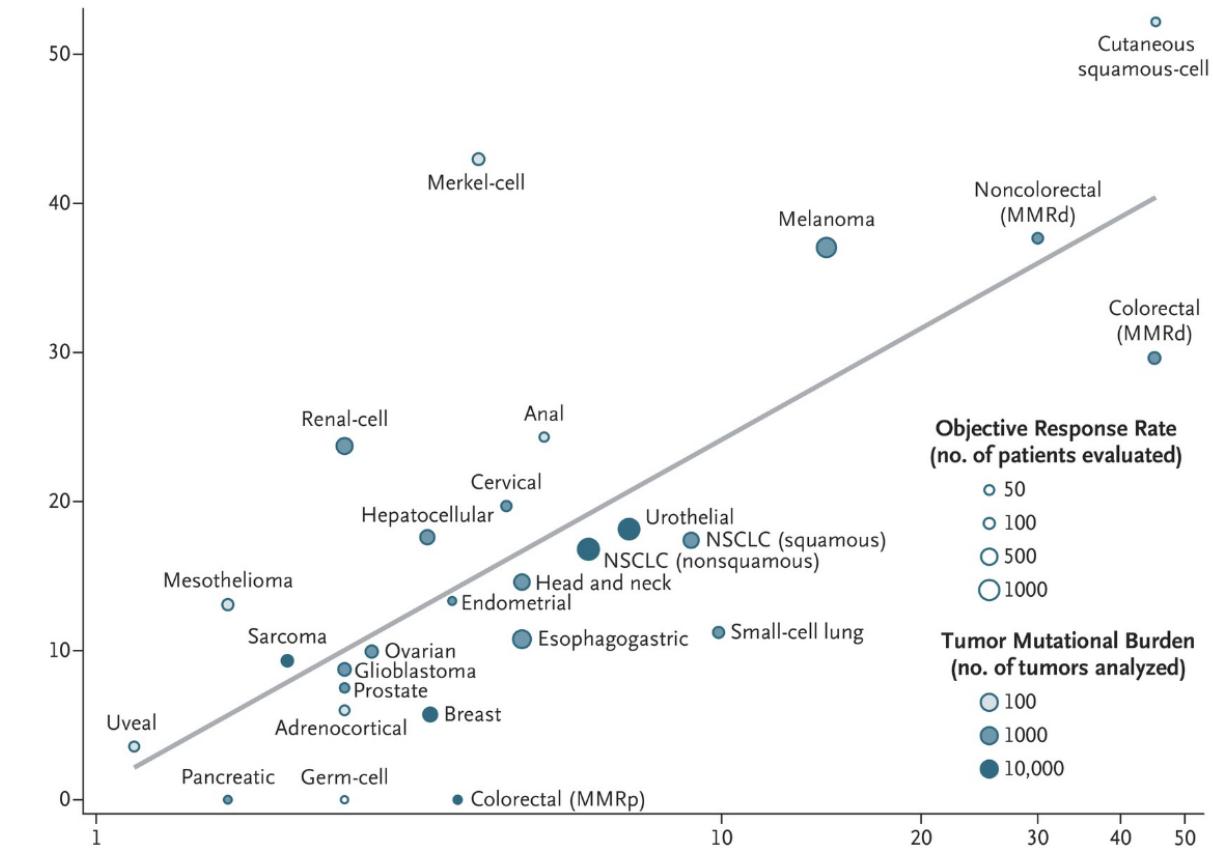
McShane et al. *BMC Medicine* 2013;11:220 (explanation & elaboration)

Tumor Mutational Burden (TMB)

A biomarker to predict benefit from immunotherapy

- A measure of the number of somatic mutations per area of the tumor's genome (mut/Mb)
- High TMB occurs in numerous tumor types and has been shown to correlate with clinical benefit from cancer immunotherapies¹ (predictive biomarker)
- TMB can be calculated using whole exome sequencing (WES), but TMB estimates from targeted NGS panels are highly correlated to WES-TMB and available at a fraction of the cost and turn-around time

Correlation between TMB and ORR with anti-PD-L1/PD-1 therapy in 27 tumor types²



Clinical use of TMB guided by a gene panel-based assay

FDA approves pembrolizumab for adults and children with TMB-H solid tumors

On June 16, 2020, the Food and Drug Administration granted accelerated approval to pembrolizumab (KEYTRUDA, Merck & Co., Inc.) for the treatment of adult and pediatric patients with unresectable or metastatic tumor mutational burden-high (TMB-H) [≥ 10 mutations/megabase (mut/Mb)] solid tumors, as determined by an FDA-approved test, that have progressed following prior treatment and who have no satisfactory alternative treatment options.

Today, the FDA also approved the FoundationOneCDx assay (Foundation Medicine, Inc.) as a companion diagnostic for pembrolizumab.

Efficacy was investigated in a prospectively-planned retrospective analysis of 10 cohorts of patients with various previously treated unresectable or metastatic TMB-H solid tumors enrolled in a multicenter, non-randomized, open-label trial, KEYNOTE-158 (NCT02628067). Patients received pembrolizumab 200 mg intravenously every 3 weeks until unacceptable toxicity or documented disease progression.

The main efficacy outcome measures were overall response rate (ORR) and duration of response (DoR) in patients who have received at least one dose of pembrolizumab as assessed by blinded independent central review according to RECIST v1.1, modified to follow a maximum of 10 target lesions and a maximum of 5 target lesions per organ.

A total of 102 patients (13%) had tumors identified as TMB-H, defined as TMB ≥ 10 mut/Mb. The ORR for these patients was 29% (95% CI: 21,39), with a 4% complete response rate and 25% partial response rate. The median DoR was not reached, with 57% of patients having response durations ≥ 12 months and 50% of patients having response durations ≥ 24 months.

Adverse reactions occurring in patients with TMB-H cancer enrolled in KEYNOTE-158 were similar to those occurring in patients with other solid tumors who received pembrolizumab as a single agent. The most common adverse reactions to pembrolizumab are fatigue, musculoskeletal pain, decreased appetite, pruritus, diarrhea, nausea, rash, pyrexia, cough, dyspnea, constipation, pain, and abdominal pain. Pembrolizumab is associated with immune-mediated side effects, including pneumonitis, colitis, hepatitis, endocrinopathies, nephritis, and skin adverse reactions.

<https://www.fda.gov/drugs/drug-approvals-and-databases/fda-approves-pembrolizumab-adults-and-children-tmb-h-solid-tumors>

June 16, 2020

FoundationOneCDx assay approved as companion diagnostic

https://assets.ctfassets.net/w98cd481qyp0/YqqKHaqQmFeqc5ueQk48w/0a34fcdaa3a71dbe460cdcb01cebe8ad/F1CDx_Technical_Specifications_072020.pdf

TMB-H defined as TMB ≥ 10 mut/Mb

1/3

<https://www.fda.gov/drugs/drug-approvals-and-databases/fda-approves-pembrolizumab-adults-and-children-tmb-h-solid-tumors>

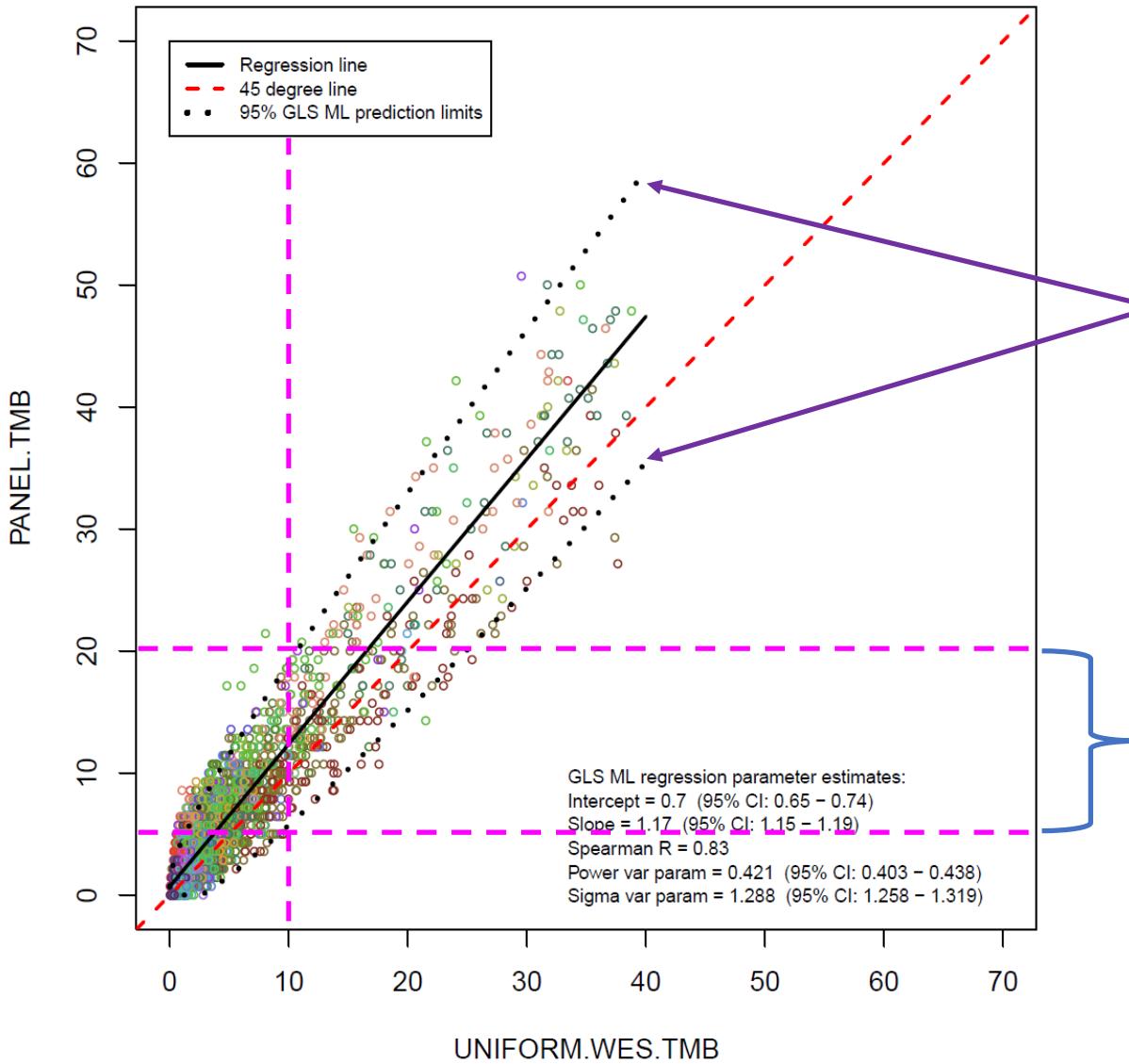
Friends of Cancer Research TMB Harmonization Project

Multi-stakeholder working group to align on and publish universal best practices for defining TMB, and analytic validation approaches including alignment against reference standards.



Workflow	Phase 1: In silico analysis	Phase 2: Empirical analysis	Phase 3: Clinical analysis
Samples	Publicly available TCGA data	Cells derived from human tumors (A: cell lines; B: clinical FFPE tumors)	Clinical Samples (correlation of TMB with outcome)
Goals	Identify sources of variability between TMB calculated using whole exome sequencing (WES) & various targeted panels used in the clinic	Agree upon creation of a universal reference standard using WES Identify sources of variability after alignment of TMB scores from targeted panels to the reference standard	Propose standards for defining clinical application of TMB and inform clinical use

Scatterplot of panel TMB vs. WES TMB for an example laboratory in Phase 1



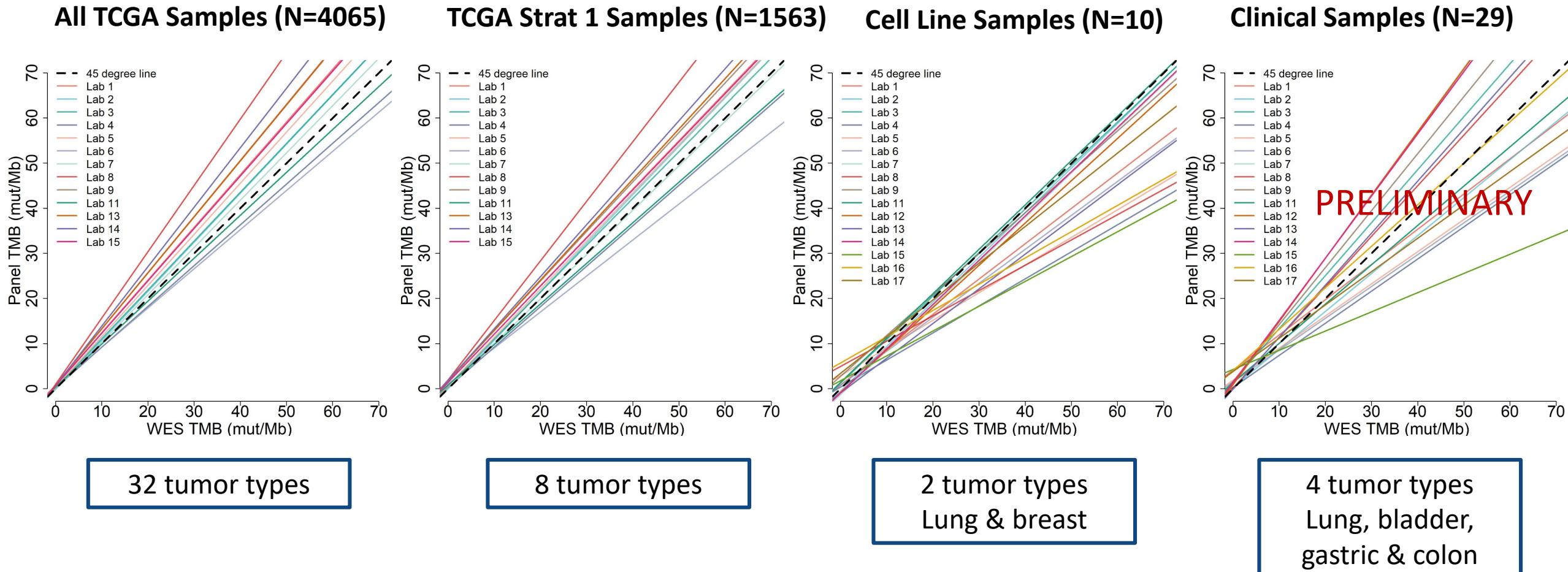
(*In silico* experiment using TCGA data from 32 tumor types)

95% prediction limits

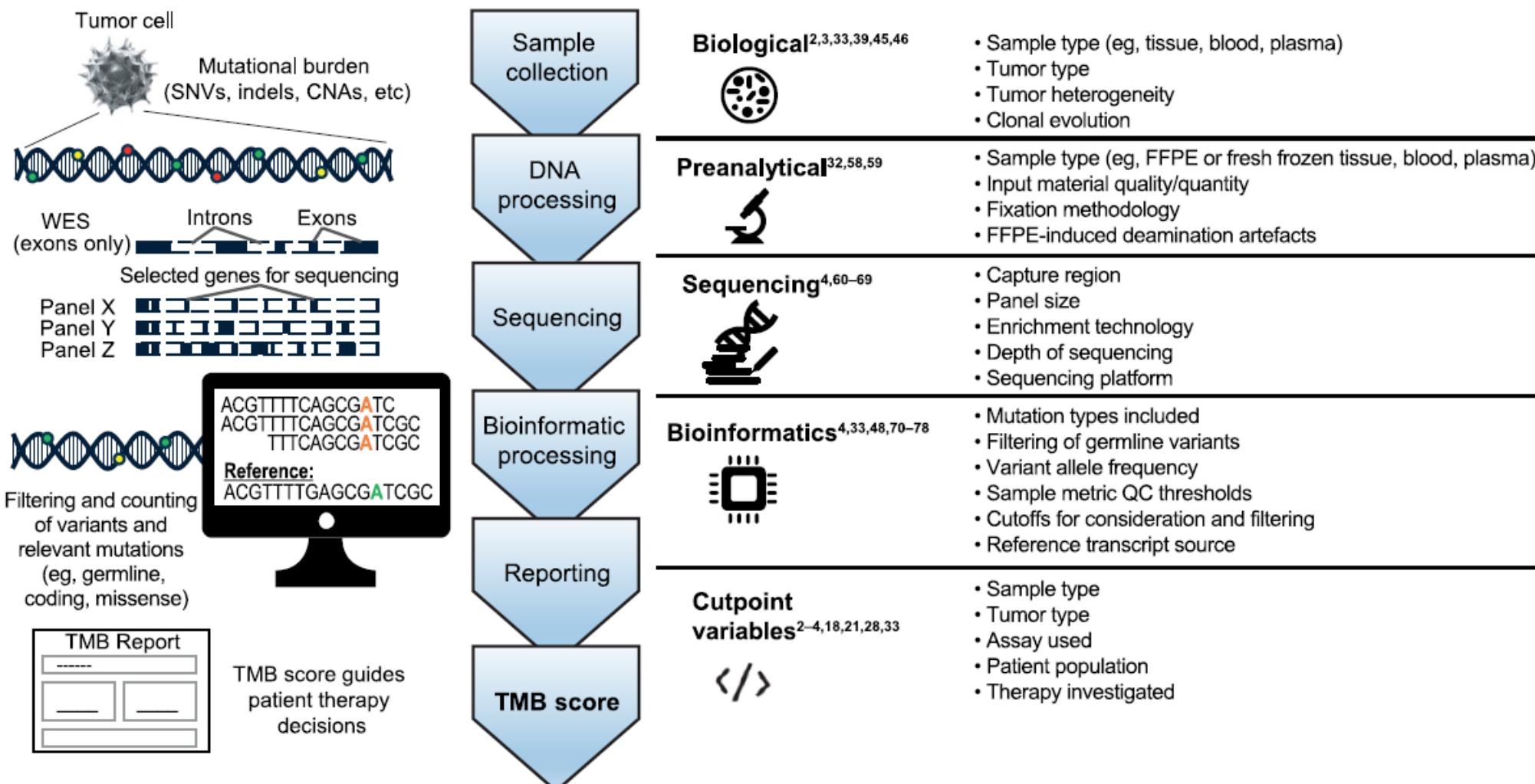
For tumors with true WES TMB = 10, we expect TMB estimated using this panel assay to fall between 5 and 20 about 95% of the time

Variation in Estimated Regression Lines

Panel TMB vs. WES TMB



Factors that affect TMB estimation



Calibration Approach Using All TCGA Samples

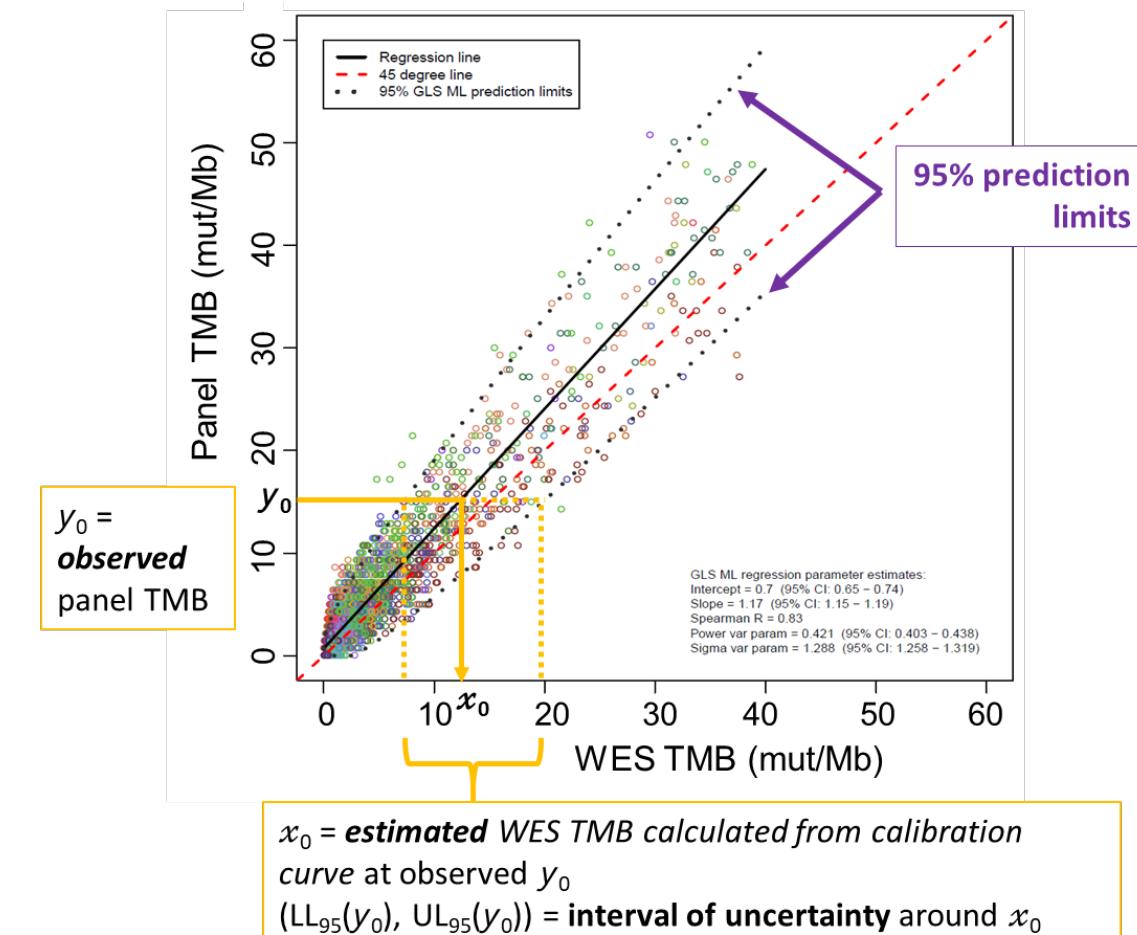
Observe panel TMB value y_0

Invert regression line to estimate WES TMB value x_0

Interval of uncertainty ($LL_{95}(y_0), UL_{95}(y_0)$):
Find x values where horizontal line $y=y_0$ intersects with 95% prediction limits

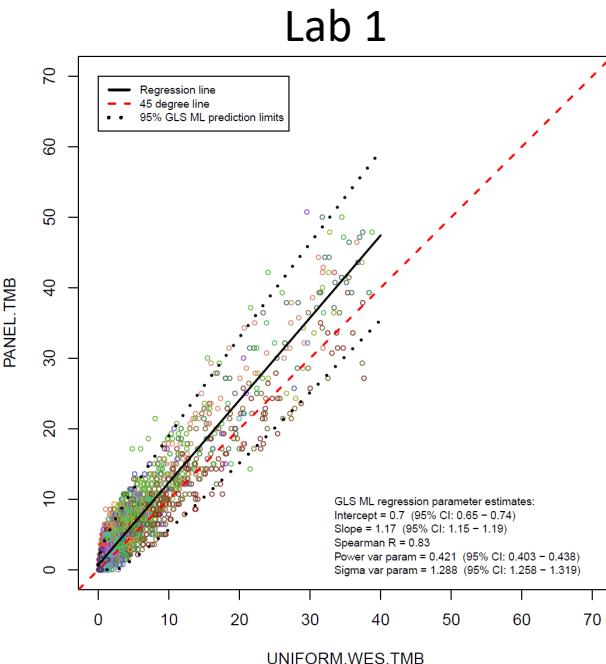
NCI team developed R software package with functions to estimate and apply the calibration model that will be made freely available along with source code

Calibration for individual laboratory informed by fitted regression line as well as scatter of points around the line (quantified by 95% prediction limits)

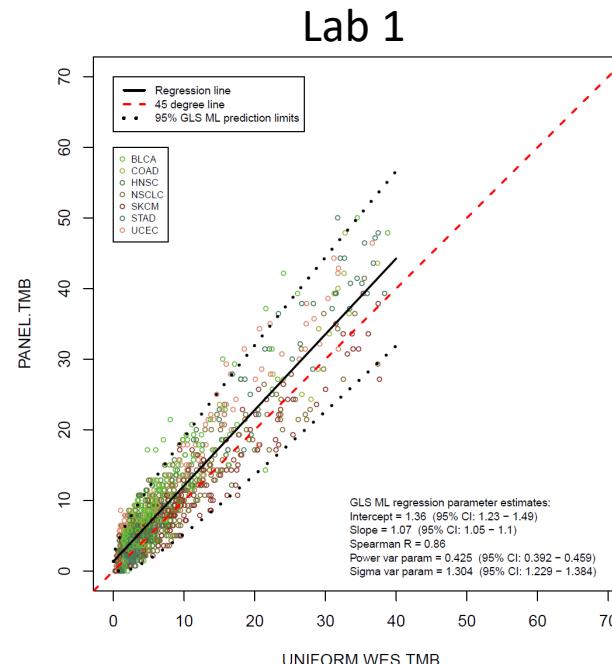


Calibration models using TCGA data and cell lines as reference standards

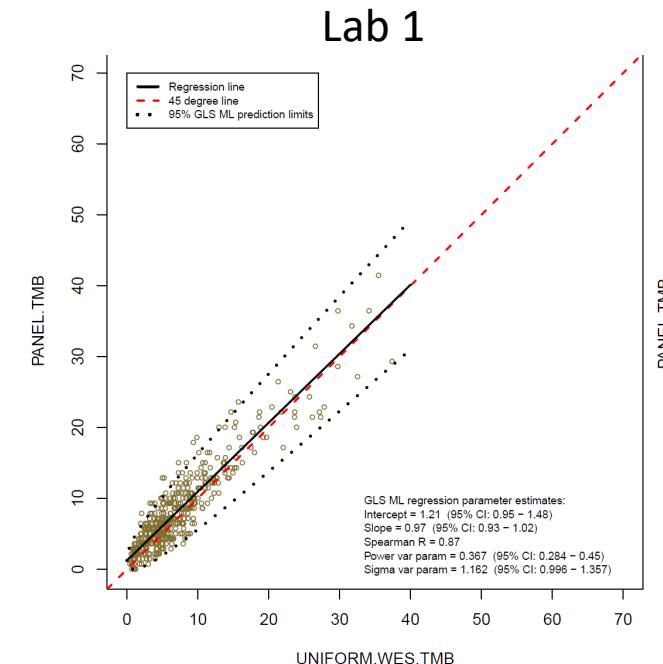
All TCGA Samples (N=4065)



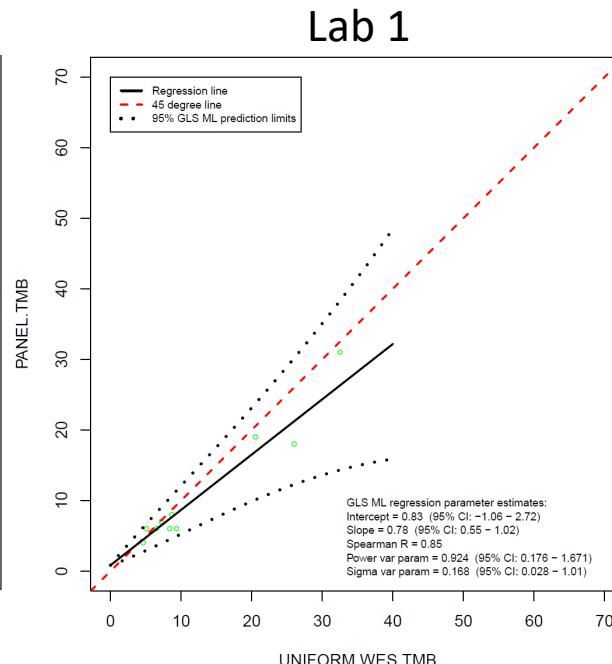
TCGA Strat 1 Samples (N=1563)



TCGA NSCLC Samples (N= 456)

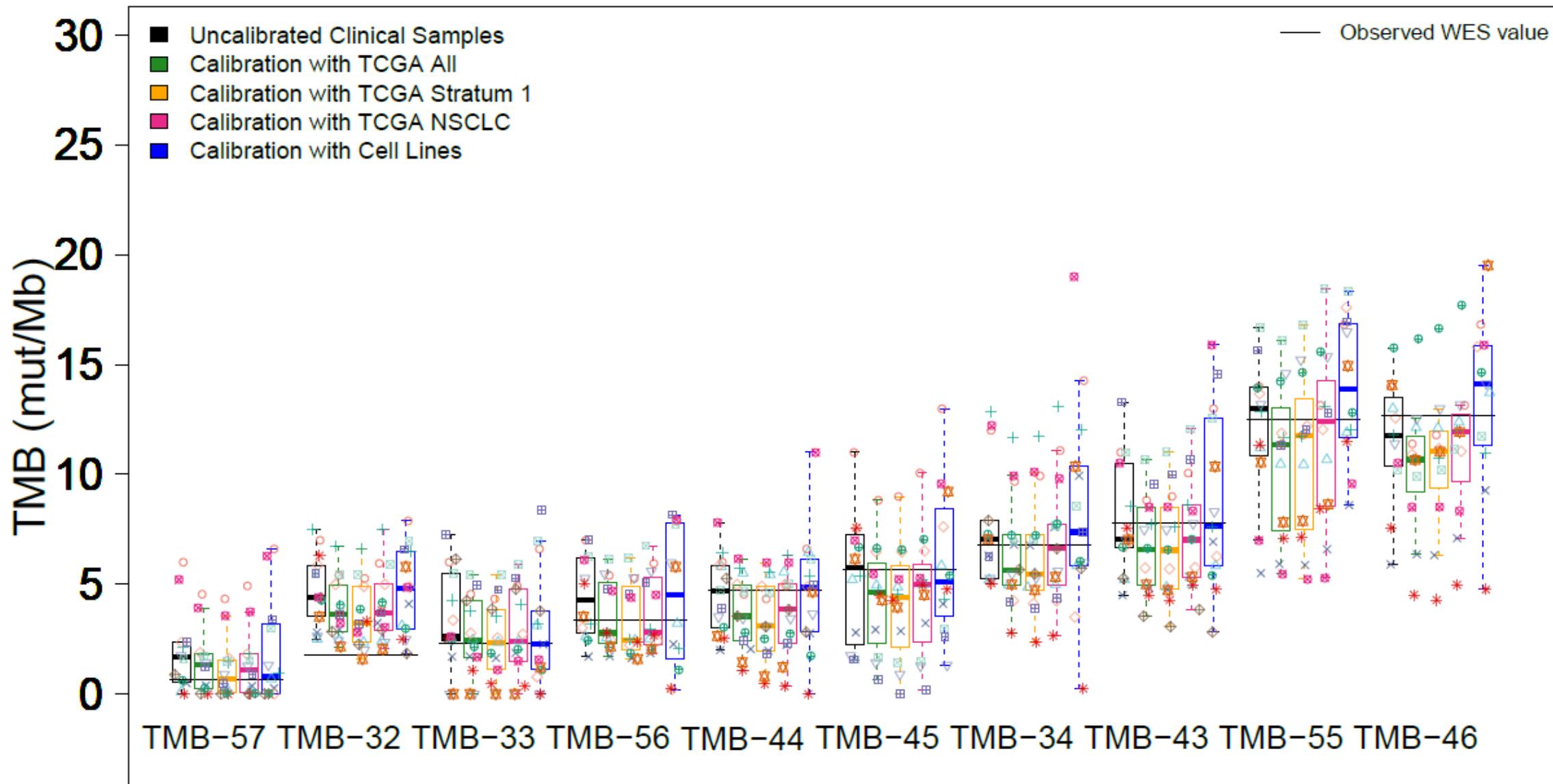


Cell Line Samples (N=10)

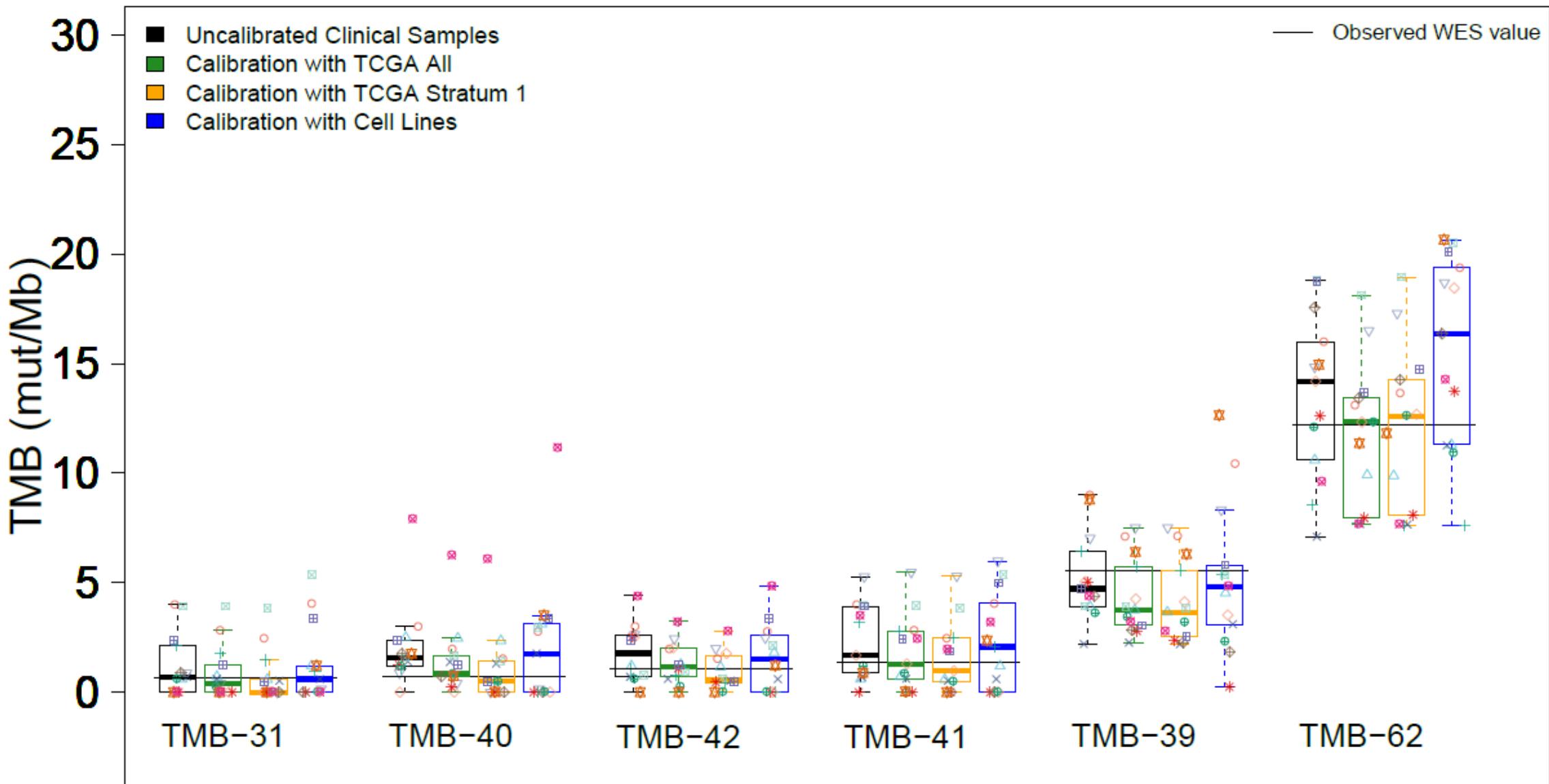


Given that TCGA calibration curves were estimated on a range restricted to WES TMB ≤ 40 , the calibration may not work well above that range. But, it is probably not too important to precisely estimate TMB values above 40 for most tumors.

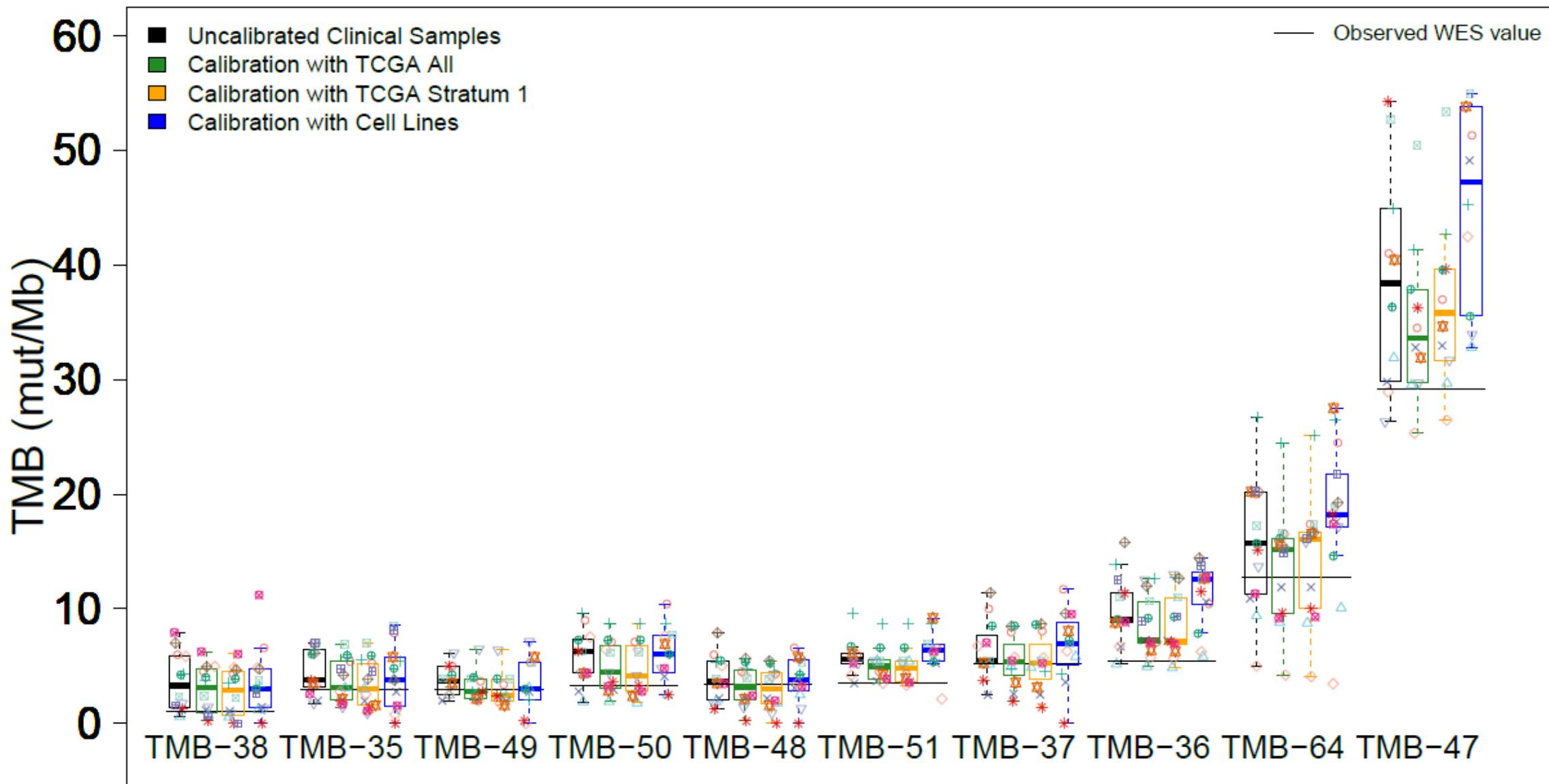
Application of four calibration approaches to lung clinical samples (n=10)



Application of three calibration approaches to gastric clinical samples (n=6)



Application of three calibration approaches to bladder clinical samples (n=10)



Potential applications for the calibration tool

1. Adjust panel TMB values to be more comparable to those generated by a reference standard method
 - Facilitate combining panel TMB data across studies to assess clinical correlation and determine cut-points
 - Evaluate impact of panel assay versioning and adjust calibration if needed
2. Eventual clinical implementation
 - Assist in tailoring TMB values or cut-points for clinical decision making to specific gene panel assay
 - Quantify imprecision relative to WES, especially near clinical cut-points

SPECIAL THANKS

- International Ki67 Working Group
 - Especially Mitch Dowsett, Dan Hayes, Torsten Nielsen, Mei Polley, Samuel Leung,
 - Breast Cancer Research Foundation (funding to academic collaborators)
- Institute of Medicine “Translational Omics” Committee
(<http://www.iom.edu/Reports/2012/Evolution-of-Translational-Omics.aspx>)
- Omics Workshop participants and checklist coauthors
- TMB Harmonization Project (www.focr.org/tmb)
 - Friends of Cancer Research (logistical and financial support for meetings and materials)
 - Especially project lead Diana Merino Vega
 - Entire TMB Harmonization Consortium
 - BRP statisticians and computational scientists Laura Yee, Qian Xie, Ming-Chung Li, Yingdong Zhao
 - MoCha Lab team: Mickey Williams, Tomas Vilimas, Lily Chen
- NCI leadership