Harmonization and Standardization of Panel-Based Tumor Mutational Burden Measurement: Real-World Results and Recommendations of the Quality in Pathology Study

Albrecht Stenzinger, MD,[a,b,*] Volker Endris, PhD,[a] Jan Budczies, PhD, PD,[a] Sabine Merkelbach-Bruse, PhD,[c] Daniel Kazdal, PhD,[a,b] Wolfgang Dietmaier, PhD,[d] Nicole Pfarr, PhD,[e] Udo Siebolts, MD, PhD, PD,[f] Michael Hummel, PhD,[g] Sylvia Herold, PhD,[h] Johanna Andreas, BSc,[i] Martin Zoche, PhD,[j] Lars Tögel, PhD,[k] Eugen Rempel, PhD,[a] Jörg Maas, Dipl.-Pol.,[l] Diana Merino, PhD,[m] Mark Stewart, PhD,[m] Karim Zaoui, MD,[n] Matthias Schlesner, PhD,[o] Hanno Glimm, MD,[p] Stefan Fröhling, MD,[q] Jeff Allen, PhD,[m] David Horst, MD,[g] Gustavo Baretton, MD,[h] Claudia Wickenhauser, MD,[f] Markus Tiemann, MD,[i] Matthias Evert, MD,[d] Holger Moch, MD,[j] Thomas Kirchner, MD,[r] Reinhard Büttner, MD,[c] Peter Schirmacher, MD,[a] Andreas Jung, MD,[r] Florian Haller, MD,[k] Wilko Weichert, MD,[e] Manfred Dietel, MD[l]

[a]*Institute of Pathology, University Hospital Heidelberg, Heidelberg, Germany*
[b]*Translational Lung Research Center (TLRC) Heidelberg, German Center for Lung Research (DZL), Heidelberg, Germany*
[c]*Institute of Pathology, University Hospital Cologne, Cologne, Germany*
[d]*Institute of Pathology, University Regensburg, Regensburg, Germany*
[e]*Institute of Pathology, Technical University Munich (TUM), Munich, Germany*
[f]*Institute of Pathology, University Hospital Halle, Halle, Germany*
[g]*Institute of Pathology, Charité University Hospital, Berlin, Germany*
[h]*Institute of Pathology, University Hospital Dresden, Dresden, Germany*
[i]*Institute of Hematopathology, Hamburg, Germany*

***Corresponding author.**

Dr. Stenzinger, Dr. Endris, and Dr. Budczies contributed equally to this work.

***Disclosure:** Dr. Stenzinger received advisory board honoraria from AstraZeneca, Bayer, Bristol-Myers Squibb (BMS), Illumina, Novartis, Seattle Genomics, Takeda, and ThermoFisher; speaker's honoraria from AstraZeneca, Bayer, BMS, Illumina, Merck Sharp & Dohme (MSD), Novartis, Roche, Seattle Genomics, Takeda, and ThermoFisher; and research funding from Chugai Pharmaceuticals, BMS, and Bayer. Dr. Endris received personal fees from Thermo Fisher and Astra Zeneca. Dr. Merkelbach-Bruse received support from Quality in Pathology (QuIP), during the conduct of the study; personal fees from BMS, Roche, Novartis, Pfizer, AstraZeneca, and Bayer; and nonfinancial support from Illumina and Janssen. Dr. Dietmaier received support from Stockholding Healthineers. Dr. Zoche received grants from F. Hoffmann-La Roche AG, Basel, CH, during the conduct of the study and grants from F. Hoffmann-La Roche AG, Basel, CH, outside the submitted work. Dr. Kazdal received personal fees from Pfizer Pharma GmbH. Mr. Maas received personal fees from BMS. Dr. Horst received personal fees from Bayer, BMS, and Roche. Dr. Tiemann received support from Merck Sharp & Dohme and BMS. Dr. Moch received grants from Roche, during the conduct of the study, and grants and personal fees from Roche, outside the submitted work. Dr. Schirmacher received grants from QuIP, during the conduct of the study; grants and personal fees from BMS, MSD, Roche, AstraZeneca, and Novartis; personal fees from Chugai Pharmaceuticals and AbbVie; grants from Sanofi-Aventis; personal fees from Ipsen; grants and personal fees from Pfizer; and grants from Illumina and Thermo Fisher, outside the submitted work. Dr. Jung received support from QuIP GmbH and personal fees and nonfinancial support from BMS, during the conduct of the study; and personal fees and nonfinancial support from Amgen, Boehringer Ingelheim, Novartis, Bayer, Merck Serono, Roche Pharma, Biocartis, MSD, and ThermoFisher, outside the submitted work. Dr. Haller received nonfinancial support from Illumina, personal fees and nonfinancial support from Qiagen, during the conduct of the study; grants, personal fees, and nonfinancial support from Illumina and Qiagen; and personal fees from BMS, Novartis, Merck, AstraZeneca, Pfizer, and Bayer, outside the submitted work. Dr. Tögel received nonfinancial support from Illumina, Inc., Qiagen, and Agilent Technologies, during the conduct of the study. Dr. Weichert received personal fees from Roche, MSD, BMS, AstraZeneca, Pfizer, Merck, Eli Lilly, Boehringer, Novartis, Takeda, Amgen, Astellas and grants from Roche, MSD, BMS, and Bruker, outside the submitted work. Dr. Fröhling received personal fees from Amgen, Eli Lilly, Roche, and Bayer, and grants from AstraZeneca, Pfizer, and PharmaMar, outside the submitted work. The remaining authors declare no conflict of interest.*

Address for correspondence: Albrecht Stenzinger, MD, Institute of Pathology, University Hospital Heidelberg, Im Neuenheimer Feld 224, 69120 Heidelberg, Germany. E-mail: albrecht.stenzinger@med.uni-heidelberg.de

[j]Institute of Pathology, University Hospital Zurich, Zurich, Switzerland
[k]Institute of Pathology, University Hospital Erlangen, Erlangen, Germany
[l]Quality in Pathology (QuIP), Berlin, Germany
[m]Friends of Cancer Research (FoCR), Washington, District of Columbia
[n]Department of Otorhinolaryngology, University Hospital Heidelberg, Heidelberg, Germany
[o]Bioinformatics and Omics Data Analytics, German Cancer Research Center (DKFZ), Heidelberg, Germany
[p]Department of Translational Medical Oncology, National Center for Tumor Diseases (NCT Dresden) and University Hospital Carl Gustav Carus, Dresden, and Translational Functional Cancer Genomics, German Cancer Research Center (DKFZ), Heidelberg, Germany and German Cancer Consortium (DKTK), Dresden, Germany
[q]Division of Translational Medical Oncology, National Center for Tumor Diseases (NCT) Heidelberg and German Cancer Research Center (DKFZ), Heidelberg, Germany
[r]Institute of Pathology, Ludwig-Maximilians University (LMU), Munich, Germany

ABSTRACT

Introduction: Tumor mutational burden (TMB) is a quantitative assessment of the number of somatic mutations within a tumor genome. Immunotherapy benefit has been associated with TMB assessed by whole-exome sequencing (wesTMB) and gene panel sequencing (psTMB). The initiatives of Quality in Pathology (QuIP) and Friends of Cancer Research have jointly addressed the need for harmonization among TMB testing options in tissues. This QuIP study identifies critical sources of variation in psTMB assessment.

Methods: A total of 20 samples from three tumor types (lung adenocarcinoma, head and neck squamous cell carcinoma, and colon adenocarcinoma) with available WES data were analyzed for psTMB using six panels across 15 testing centers. Interlaboratory and interplatform variation, including agreement on variant calling and TMB classification, were investigated. Bridging factors to transform psTMB to wesTMB values were empirically derived. The impact of germline filtering was evaluated.

Results: Sixteen samples had low interlaboratory and interpanel psTMB variation, with 87.7% of pairwise comparisons revealing a Spearman's ρ greater than 0.6. A wesTMB cut point of 199 missense mutations projected to psTMB cut points between 7.8 and 12.6 mutations per megabase pair; the corresponding psTMB and wesTMB classifications agreed in 74.9% of cases. For three-tier classification with cut points of 100 and 300 mutations, agreement was observed in 76.7%, weak misclassification in 21.8%, and strong misclassification in 1.5% of cases. Confounders of psTMB estimation included fixation artifacts, DNA input, sequencing depth, genome coverage, and variant allele frequency cut points.

Conclusions: This study provides real-world evidence that all evaluated panels can be used to estimate TMB in a routine diagnostic setting and identifies important parameters for reliable tissue TMB assessment that require careful control. As complex or composite biomarkers beyond TMB are likely playing an increasing role in therapy prediction, the efforts by QuIP and Friends of Cancer Research also delineate a general framework and blueprint for the evaluation of such assays.

Introduction

Immune checkpoint inhibitors (ICIs) have greatly expanded therapeutic options in oncology.[1] Although many clinical trials have reported strong clinical responses across various tumor types, evidence is increasing that even in generally responsive tumor entities, many tumors are resistant at baseline or develop resistance to ICIs, for example, by immunoediting.[2] Moreover, adverse events associated with ICIs have been noted, particularly with combinatorial regimens that target cytotoxic T lymphocyte–associated protein 4 in addition to programmed cell death protein 1 or programmed death-ligand 1 (PD-L1).[3] Collectively, these observations argue for a sophisticated biomarker approach that reflects the interplay between the host's immune system and the cancer cells and is able to reliably separate likely responders from nonresponders.

To date, two predictive ICI-specific biomarkers have been approved in certain cancer types, which are as follows: (1) PD-L1, assessed by immunohistochemistry (IHC) with a wide range of different scoring systems and cut points depending on cancer type–specific trial results, and (2) high-level microsatellite instability or mismatch repair deficiency, assessed by either polymerase chain reaction (PCR) or IHC.[4,5] Whereas the former approach measures a continuous variable that serves as an approximation for T-cell anergy or tumor cells escaping immune response, the latter identifies a subgroup of cancers with a high

mutational burden and thus increased neoantigen load, which likely results in a higher propensity of immune cell–mediated tumor cell killing.

However, many cancer types, including NSCLC, do not harbor deleterious mutations in one of the DNA mismatch repair genes but have increased tumor mutational burden (TMB) associated with higher loads of neoantigens, which is caused by DNA damage through external noxae (e.g., ultraviolet light and smoking) or deleterious mutations affecting other DNA repair genes.[6]

Although clinical trials assessing the utility of TMB prospectively are ongoing, many retrospective analyses of individual patient cohorts and clinical trials have reported that TMB can be successfully used for patient stratification. Initial seminal studies employed whole-exome sequencing (WES) to measure TMB.[7-10] Because this approach has several limitations, including sample requirements, necessity for concurrent germline sequencing, extensive laboratory capacity for diagnostic application, and economic constraints in consideration of a diagnostic outreach setting, gene panels were designed and used to estimate TMB values, primarily in formalin-fixed and paraffin-embedded (FFPE) tissue and, more recently, in cell-free circulating tumor DNA.[11-13] Such assays have been successfully used under controlled trial conditions or at specific academic cancer centers. However, a detailed evaluation of the overall performance of commercially available sequencing panels that can be used as laboratory-developed tests and of the parameters affecting its diagnostic applicability is missing.

To address this important issue, we present the results of the multi-institutional Quality in Pathology (QuIP) study on a comparative assessment of TMB estimated by gene panel sequencing (psTMB) from 11 different institutes of pathology and four industrial laboratories. Analyzing 20 different FFPE cancer samples from routine diagnostics that reflect the full continuum of TMB, as measured by WES (wesTMB), provides real-world data on the following six different targeted gene panels designed for TMB estimation: Oncomine Tumor Mutational Load Assay (OTML; Thermo Fisher Scientific, Waltham, MA), QIAseq TMB panel (QIAseq; QIAGEN GmbH, Hilden, Germany), NEOplus RUO assay (NEOplus; NEO New Oncology, Cologne, Germany), TruSight Oncology 500 panel (TSO500; Illumina, San Diego, CA), a custom-designed academic panel (ACADEMIC; Agilent, Santa Clara, CA), and the FoundationOne assay (F1; Foundation Medicine, Cambridge, MA). Together with the efforts led by the Friends of Cancer Research,[14,15] this study sets the basis for harmonization of panel-based TMB measurement and supports implementation of TMB in routine diagnostic laboratories.

Materials and Methods

Samples

All patients provided written informed consent under an institutional review board–approved protocol, and the study was conducted in accordance with the Declaration of Helsinki. FFPE tissue specimens of 10 lung adenocarcinoma (LUAD), seven head and neck squamous cell carcinoma, and three colon adenocarcinoma were prepared and diagnosed at the Institute of Pathology Heidelberg, Germany. See Supplementary Table 1 for further detailed sample information. Only one block per tumor was selected, and consecutive sections were used for DNA extraction by the different laboratories. Tumor content was controlled using hematoxylin and eosin–stained slides on the first and last sections to ensure homogeneity throughout the slices.

Library Preparation and Sequencing

Protocols for the six applied panel-sequencing approaches (OTML, QIAseq, NEOplus, TSO500, ACADEMIC, and F1) and for WES are detailed in the Supplementary Materials and Methods (Supplementary Table 2). All assays were performed according to the manufacturers' protocols if not specified otherwise.

Data Analysis and Visualization

Data analysis and visualization were performed using the statistical programming language R (version 3.51).[16] Levels of psTMB were visualized as boxplots and as heatmaps, including hierarchical clustering of experiments (Manhattan distance, average linkage clustering). Spearman's correlations (ρ) and Pearson's correlations (R) of psTMB were calculated between pairs of experiments, clustered (Euclidean distance, average linkage clustering), and visualized as heatmaps. Error bars were plotted using the function plotCI from the R package gplots. Violin plots were generated using the R package vioplot.

Linear models without intercept were fitted to psTMB levels with wesTMB levels. Measurement of psTMB is influenced by different factors. Although misclassification of germline mutations as somatic mutations is independent of the TMB level, other factors, including the subsampling error caused by interrogation of only a limited part of the coding sequence, increase with a higher TMB.[17] Because the exact shape of the mathematical dependence of the TMB error on the level of the TMB is not known, linear models were fitted in the following two different ways: (1) standard linear regression (least square regression, LS) corresponding to constant error contributions, and (2) weighted linear regression (weighted least squares, WLS) with weights equal to the

reciprocal of TMB taking into account heteroscedasticity. The shape of the weights used in the WLS model reflects the mathematical law for the variation of psTMB that we recently uncovered and described—a linear increase of the variation of psTMB proportional to the level of TMB.[17]

Results

Study Outline

In this study (Fig. 1), FFPE tissue samples of 20 tumors (Supplementary Table 1) with existing matched WES data were analyzed using four commercial panel-sequencing TMB assays (Supplementary Table 2). Each assay was run by four different pathology laboratories and a reference laboratory of the panel provider on all samples. In addition, three pathology laboratories tested the ACADEMIC assay, and all samples were analyzed using the F1 assay. The analyzed study cohort was selected to represent the full spectrum of TMB values as characterized by The Cancer Genome Atlas for LUAD, head and neck squamous cell carcinoma, and colon adenocarcinoma, but it has a higher proportion of tumors with an intermediate TMB (100–300 mutations) (Supplementary Figure 1). In total, panel sequencing and psTMB measurement were successful in 467 of the 480 performed analyses (97.3%).

TMB Levels and Correlations

Measurements of psTMB in the 20 tumor tissue samples ranged between 0 and 244 mutations per megabase pair (muts/Mbp) with a median of 9.2 muts/Mbp (Fig. 2A). With respect to interlaboratory and interpanel variance, four of the tumor samples (T4, T7, T13, and T15) stood out by having a larger interquartile range of psTMB compared with the remaining samples. This was mainly owing to unfavorable preanalytic quality parameters (degraded DNA or low tumor cellularity) (Fig. 2A). Two samples (T4 and T15) had a large interlaboratory variance of psTMB when each of the panels was analyzed separately, whereas this was not

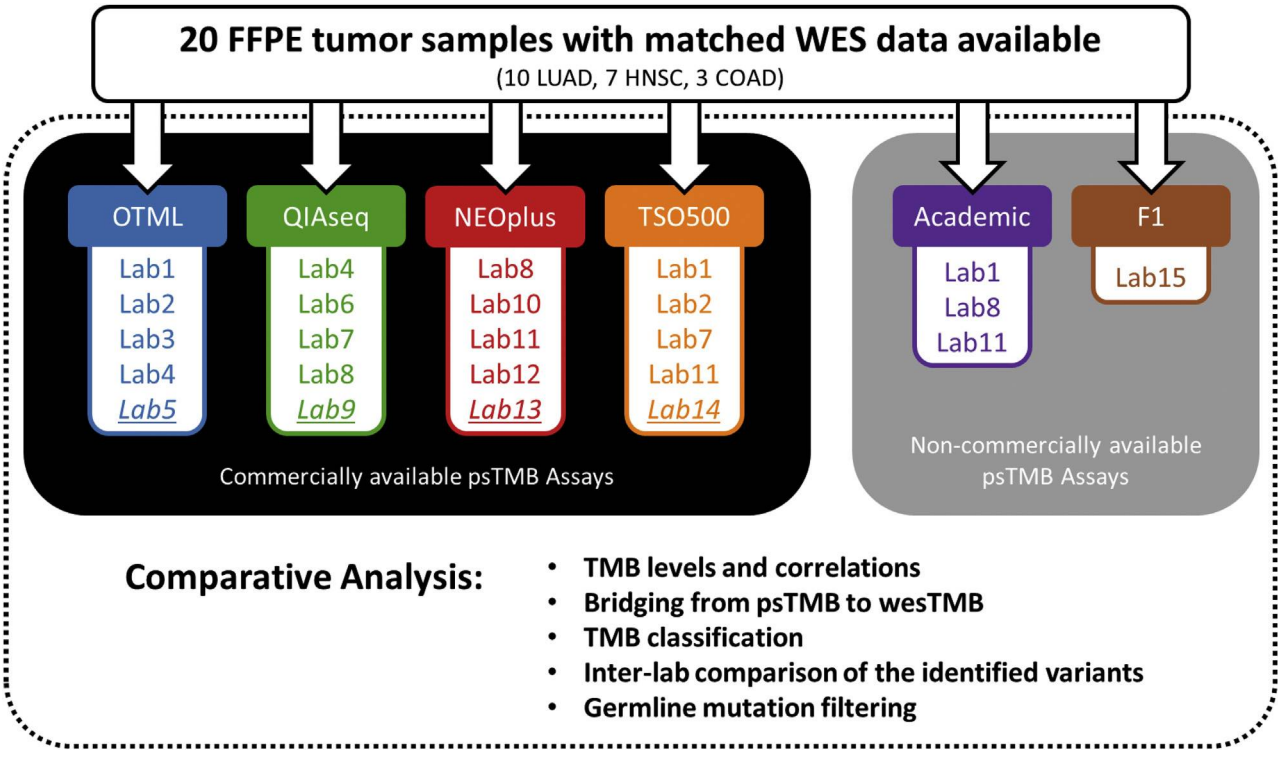

Figure 1. Outline of the QuIP TMB harmonization study. In this comparative study, FFPE tissue samples from 20 tumors were analyzed using four commercial panel-sequencing TMB assays. Each assay was tested by four independent pathology laboratories using all 20 samples and by a reference laboratory of the panel provider. In addition, all samples were analyzed using an "ACADEMIC" assay in three hospital laboratories and by applying the F1 assay. The study cohort consisted of 20 samples from patients with LUAD (n = 10), HNSCC (n = 7), and COAD (n = 3). For all tumors, wesTMB using fresh-frozen tumor tissue samples and paired blood samples was available. ACADEMIC, custom-designed academic panel; COAD, colon adenocarcinoma; F1, FoundationOne assay; FFPE, formalin-fixed and paraffin-embedded; HNSCC, head and neck squamous cell carcinoma; LUAD, lung adenocarcinoma; NEOplus, NEOplus RUO assay; OTML, Oncomine Tumor Mutational Load Assay; psTMB, TMB assessed by gene panel sequencing; QIAseq, QIAseq TMB panel; QuIP, Quality in Pathology; TMB, tumor mutational burden; TSO500, TruSight Oncology 500 panel; WES, whole-exome sequencing; wesTMB, TMB assessed by WES.

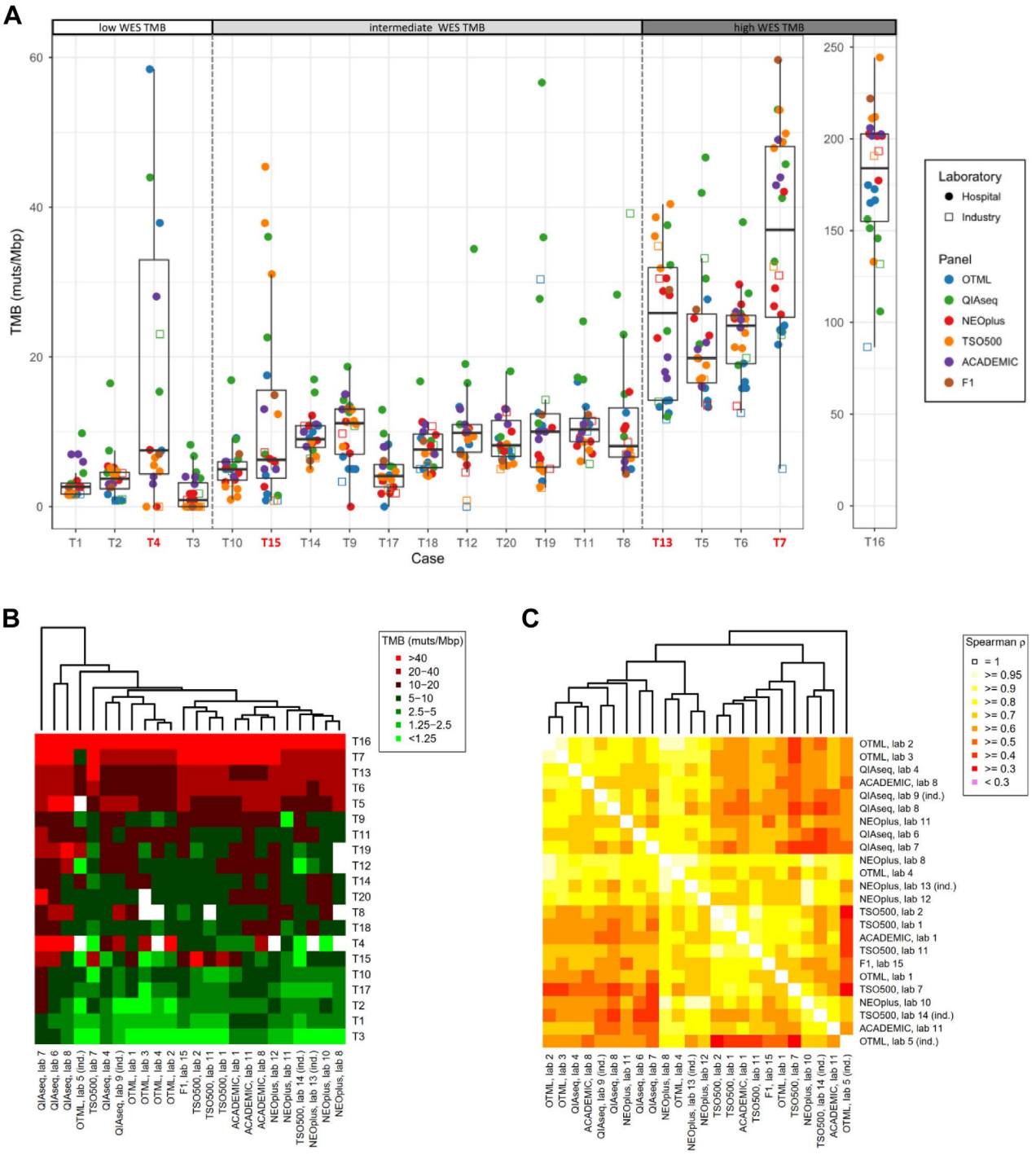

Figure 2. (*A*) Overview of the generated psTMB estimates with tumors ordered by increasing wesTMB levels. Applying a three-tier classification system, four tumors (T1–T4) were classified as TMB low (<100 missense mutations), 11 tumors were classified as TMB intermediate (100–300 missense mutations), and five tumors (T5–T7, T13, and T16) were classified as TMB high (≥300 missense mutations). Four samples stood out by high interquartile ranges and are marked by red IDs. Preanalytic quality parameters were unfavorable for three of these samples (T15: low tumor cellularity; T4 and T13: degraded DNA). (*B*) Heatmap of psTMB levels. Red color indicates psTMB level greater than 10 muts/Mbp. Green color indicates psTMB level less than 10 muts/Mbp. White color indicates insufficient DNA quality. (*C*) Spearman's correlations between psTMB and wesTMB levels in the study cohort. ACADEMIC, custom-designed academic panel; F1, FoundationOne assay; muts/Mbp, mutations per megabase pair; NEOplus, NEOplus RUO assay; OTML, Oncomine Tumor Mutational Load Assay; psTMB, TMB assessed by gene panel sequencing; QIAseq, QIAseq TMB panel; QuIP, Quality in Pathology; TMB, tumor mutational burden; TSO500, TruSight Oncology 500 panel; WES, whole-exome sequencing; wesTMB, TMB assessed by WES.

the case for the two other tumor samples (T7 and T13), in which interpanel variance was an important confounder (Supplementary Figure 2).

In a heatmap, including hierarchical clustering of the psTMB levels, data readouts based on the same sequencing panel often clustered together, indicating independence from the operating laboratory (Fig. 2B). Among most of the sequencing results, moderate to strong pairwise correlations of psTMB measurements were observed: of all pairwise Spearman correlations, 65.9% were greater than or equal to 0.7, 87.7% were greater than or equal to 0.6, and 95.7% were greater than or equal to 0.5 (Fig. 2C). In the study cohort, the strength of Pearson's correlations was dependent on the inclusion or exclusion of a single sample (T16, *POLE*-mutated colorectal carcinoma) that had a very high TMB (>100 muts/Mbp) (Supplementary Figure 3). Hence, the Spearman's correlation was a more suitable approach for the measurement of the psTMB correlations than the Pearson's method.

Bridging From psTMB to wesTMB

Linear regression models were fitted for bridging from psTMB to wesTMB (Fig. 3). To this end, we performed LS but also WLS (see the Materials and Methods section for details) for each of the panels tested in the study. Bridging factors (BFs) for transformation of psTMB to wesTMB were calculated as reciprocals of the regression slopes (Supplementary Table 3). For most of the assays, the BF determined by WLS was very close to the BF determined by LS. However, for the ACADEMIC assay, the WLS BF was slightly lower than the LS BF (17.7 versus 19.8), whereas it was considerably lower for the QIAseq assay (15.8 versus 25.6).

A clinically relevant psTMB cut point of 10 muts/Mbp in NSCLC was established in the CheckMate 568 study using the F1 panel, evaluated in the CheckMate 227 study, and bridged to a wesTMB cut point of 199 mutations using data from the CheckMate 026 study.[18-20] Based on these findings, psTMB cut points corresponding to 199 mutations were calculated for each of the investigated assays (Supplementary Table 3). For most of the psTMB assays, the calculated cut points were consistently in the range of 9.4 to 11.5 muts/Mbp. There were two exceptions, as follows: considerably different cut points were obtained for the OTML assay (LS: 7.8 muts/Mbp, WLS: 7.9 muts/Mbp) and the QIAseq assay (LS: 7.8 muts/Mbp, WLS: 12.6 muts/Mbp).

TMB Classification

Next, we evaluated and compared a two-tier system with a three-tier system for TMB classification (Fig. 4) after a recent indication to improve the misclassification ratio. For the two-tier approach, a dichotomization into "low TMB" and "high TMB" was conducted using a wesTMB cut point of 199 mutations. The three-tier approach classified TMB as "low" (<100 mutations), "intermediate" (100–300 mutations), and "high" (≥300 mutations). Classification with alternative cut points (150 and 250 mutations) is found in Supplementary Figure 4. For each of the panel-sequencing platforms, psTMB values were converted to wesTMB values using the BFs obtained by WLS regression. Altogether (20 samples × 24 experiments), we observed an agreement between psTMB and wesTMB classifications in 74.9% of the cases using the two-tier approach. For the three-tier approach, a "strong misclassification" was defined by a high TMB tumor classified as low TMB or vice versa (difference spanning two tiers), whereas a misclassification by a single tier (e.g., intermediate TMB to low TMB) was termed "weak misclassification." Here, we observed an agreement in 76.7% of cases, compared with a weak and strong misclassification in 21.8% and 1.5% of the cases, respectively. Of note, strong misclassification occurred only for a single tumor sample (T4) that was classified as low TMB by WES but as high TMB in seven psTMB assays and was not analyzable in five psTMB approaches. Assessment of this tumor (LUAD) was priori expected to be challenging owing to highly degraded DNA.

TMB classifications using BFs determined either by WLS or LS regression were similar, as LS regression resulted in 74.3% agreement for two-tier classification and 75.0% agreement, 23.1% weak misclassifications, and 1.9% strong misclassifications for the three-tier classification.

Interlaboratory Comparison of the Identified Variants

In-depth analysis of called variants included in the calculation of TMB identified key factors that influence precise psTMB estimation from the FFPE tissue (Fig. 5). A sequencing approach without an application for PCR duplicate removal, known as deduplication, has a higher probability of erroneous calling of C>T or G>A fixation artifacts and subsequent overestimation of psTMB, especially in highly fragmented, low-quality DNA samples. Methods for deduplication include specialized software solutions and the use of unique molecular identifiers (or molecular barcodes).

False-positive variants in the generated data set were identified by a side-to-side comparison of all variants identified by the different laboratories using the same panel. Variants were classified into nonreproducible variants (detected by a single laboratory), partially reproducible variants (detected by more than one laboratory,

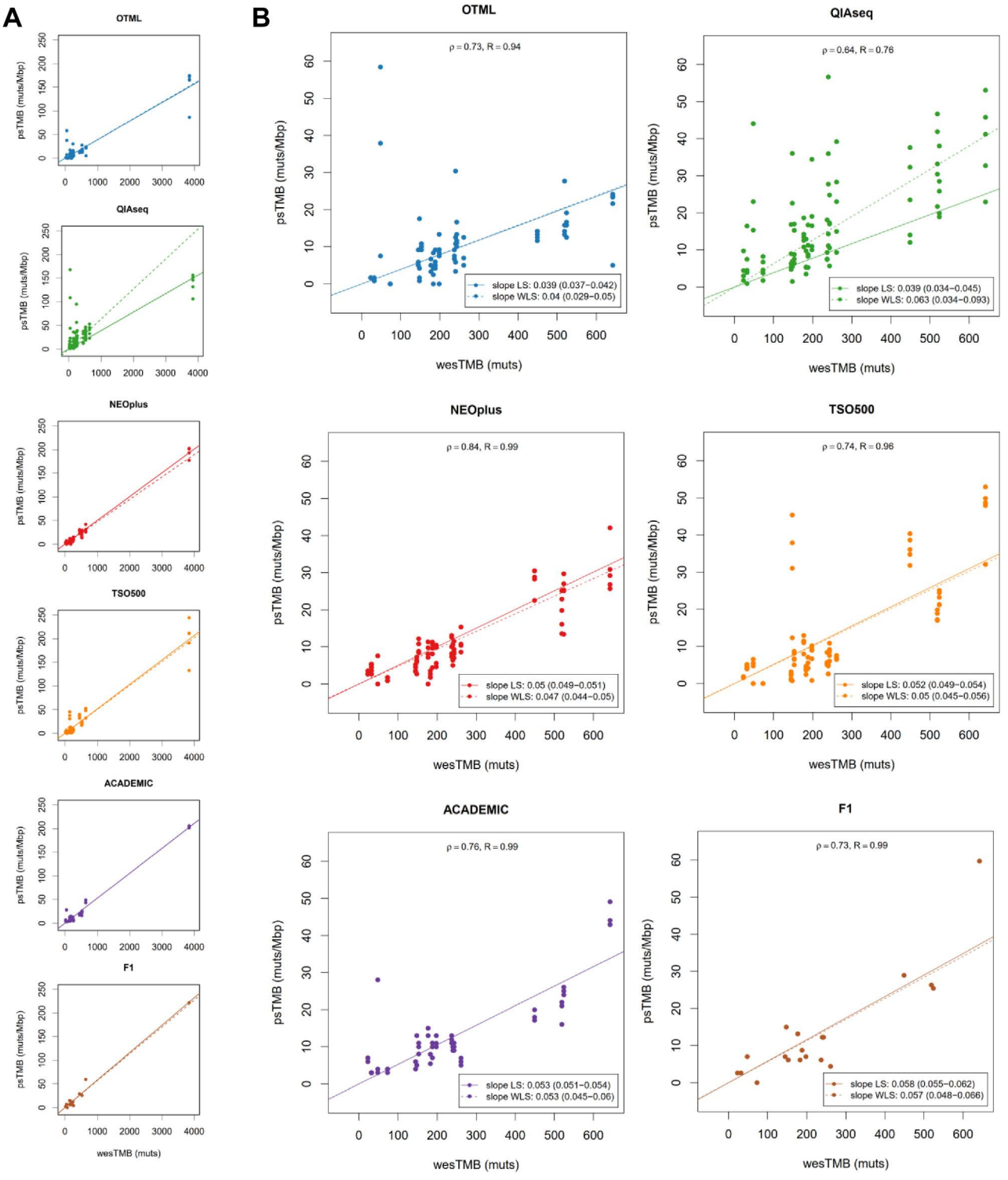

Figure 3. Calibration of TMB measured by psTMB against wesTMB. Linear fits using LS and WLS regression. (A) Overview plots revealing all psTMB and wesTMB measurements. (B) Zoom-ins to the intervals (0, 650) of wesTMB and (0, 65) of psTMB. The intercepts in the linear regression models were set to zero. ACADEMIC, custom-designed academic panel; F1, FoundationOne assay; LS, least squares; muts/Mbp, mutations per megabase pair; NEOplus, NEOplus RUO assay; OTML, Oncomine Tumor Mutational Load Assay; psTMB, TMB assessed by gene panel sequencing; QIAseq, QIAseq TMB panel; TMB, tumor mutational burden; TSO500, TruSight Oncology 500 panel; WES, whole-exome sequencing; wesTMB, TMB assessed by WES; WLS, weighted least squares.

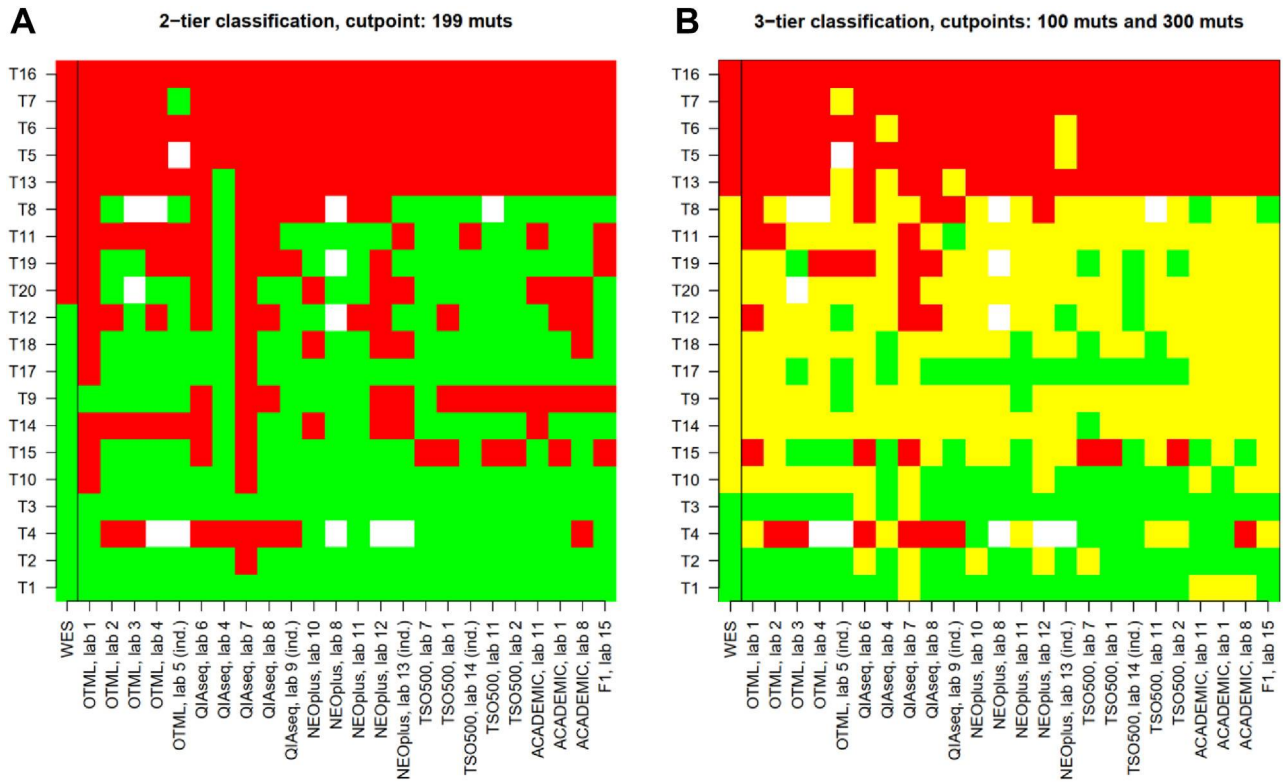

Figure 4. TMB classification by panel sequencing compared with TMB classification by WES. Measurements of psTMB were converted to wesTMB using the bridging factors in Supplementary Table 3. (A) Two-tier classification using the cut point of 199 mutations. Misclassifications: 25.1%. (B) Three-tier classification using the cut points of 100 and 300 mutations. Red indicates high TMB, yellow indicates intermediate TMB, and green indicates low TMB. Strong misclassifications (=misclassifications mixing TMB high and TMB low cases): 1.5%. Weak misclassifications (=misclassifications mixing intermediate TMB cases with TMB high or TMB low cases): 21.8%. ACADEMIC, custom-designed academic panel; F1, FoundationOne assay; NEOplus, NEOplus RUO assay; OTML, Oncomine Tumor Mutational Load Assay; psTMB, TMB assessed by gene panel sequencing; QIAseq, QIAseq TMB panel; TMB, tumor mutational burden; TSO500, TruSight Oncology 500 panel; WES, whole-exome sequencing; wesTMB, TMB assessed by WES.

but not by all laboratories), and fully reproducible variants (detected by all laboratories). Variant allele frequencies (VAFs) were considerably lower for variants with low degrees of interlaboratory reproducibility, and many of the nonreproducible variants had VAFs close to the VAF cut point (Fig. 5A). Thus, low-frequency variants close to the VAF cut point contributed considerably to psTMB variation. To minimize the rate of false-positive calls, specific thresholds for VAF were used for each panel according to the assay provider: VAFs greater than or equal to 10% was applied for the OTML and NEOplus panels and VAFs greater than or equal to 5% for the remaining panels. The number of nonreproducible variants was considerably higher for the OTML assay (3497 variants), which did not include deduplication, compared with the other assays (QIAseq: 1055; NEOplus: 94; TSO500: 70; ACADEMIC: 691). In addition, as illustrated in Figure 5B, the ratio of C>T or G>A transitions was considerably higher for nonreproducible variants (red pie charts) detected by the OTML panel (86%) compared with the other panels (22%–42%), and compared with the ratio of C>T or G>A of variants that were detected by all laboratories (gray pie charts). These data identify paraffin fixation artifacts and resulting C>T or G>A transitions as important parameters contributing to false-positive variant detection for assays that do not employ deduplication.

False-negative calls (defined here as mutations called by all but one laboratory) can be connected to insufficient depth of coverage at the respective positions. Because the pipelines for capture-based fragment libraries typically include deduplication and unique molecular identifier filtering, the depth of coverage directly correlates with the amount of DNA input, as found representatively for the TSO500 panel in Figure 5C. Here, the median exon coverage that could be analyzed was significantly higher ($p < 0.01$) in laboratory 1 using 80 ng as DNA input compared with 40 ng that was used for the other TSO500 approaches (laboratories 2, 7, 11, and Illumina). Furthermore, the amount of DNA input had a strong impact on the average size of the covered sequencing region (Fig. 5C, middle). Although the

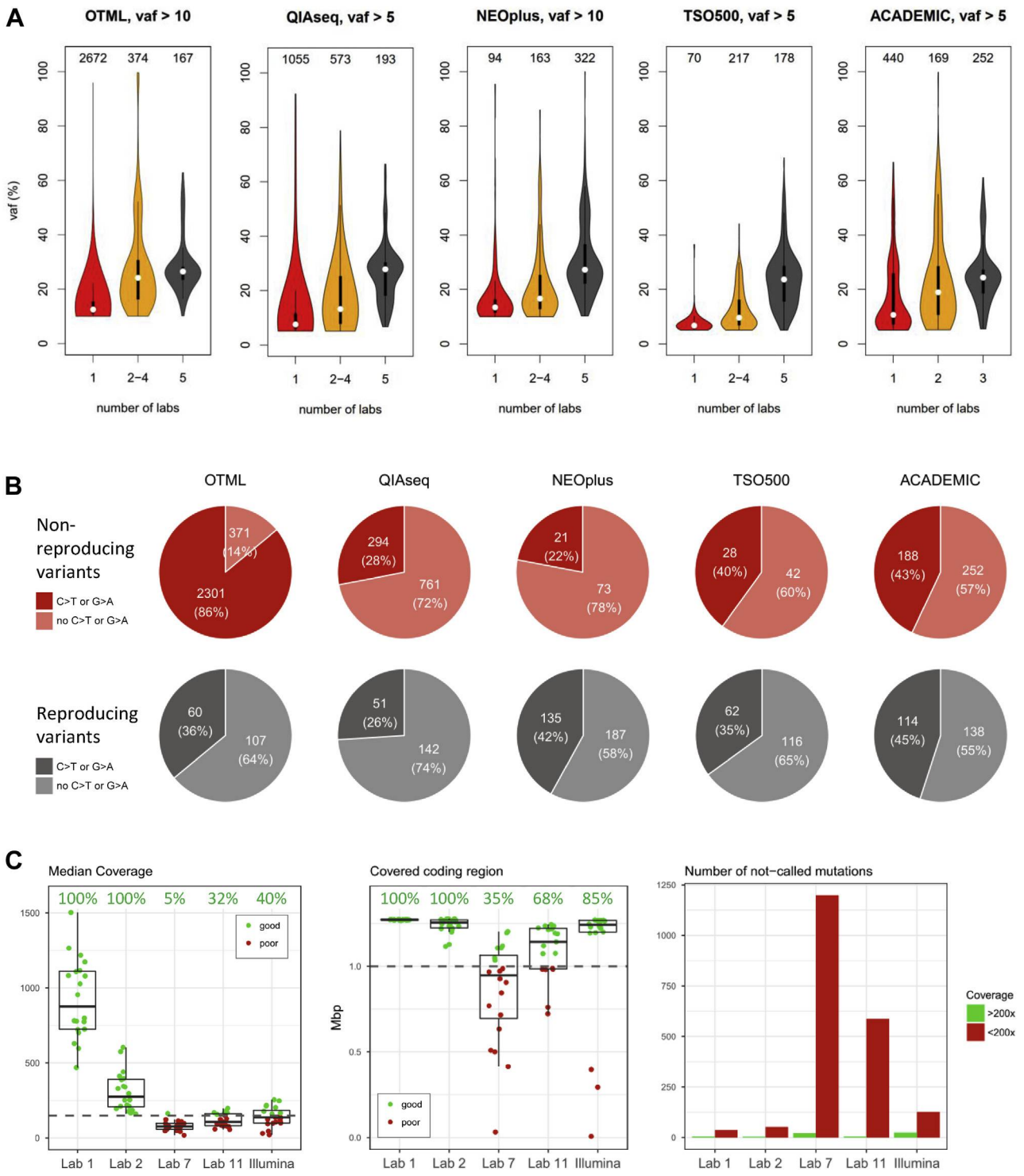

Figure 5. Interlaboratory reproducibility of the detected mutations (pooled analysis of 20 samples). (*A*) Distribution of VAFs in dependence of the number of laboratories that detected the mutation. (*B*) Mutation type (C>T, G>A, or other) of the mutations detected by a single laboratory. (*C*) Impact on DNA input is representatively revealed for the TSO500 panel. DNA input: 80 ng (laboratory 1), 40 ng (labs 2, 7, 11, and Illumina). Left: Median exon coverage for each sample; the number on top gives the percentage of cases with a median exon coverage of more than 150 times. Middle: covered coding region size for each sample. The number on top gives the percentage of cases with a covered coding region of more than 1.0 Mbp. Right: coverage of mutations not-called by a single laboratory (false negatives). ACADEMIC, custom-designed academic panel; NEOplus, NEOplus RUO assay; OTML, Oncomine Tumor Mutational Load Assay; QIAseq, QIAseq TMB panel; TSO500, TruSight Oncology 500 panel; VAF, variant allele frequency.

maximum covered coding region size of 1.28 Mbp was reached for all samples using 80 ng (laboratory 1), lower DNA input resulted in significantly ($p < 0.01$) lower covered coding region sizes, which were larger than 1.0 Mbp in 35% to 100% of the analyzed samples. To enhance specificity, only mutations with minimum coverage of 100 times were included in the psTMB calculation. Therefore, and connected to the lower coverage, we observed a higher rate of false-negative variants in analyses using 40 ng DNA (laboratories 2, 7, 11, and Illumina) compared with 80 ng (Fig. 5C, right). Similar findings were seen for 100-ng versus 200-ng DNA input using the NEOplus assay (data not shown).

Germline Mutation Filtering

Germline mutation filtering is an important step in the calculation of psTMB because only the tumor's somatic mutations are relevant for recognition by the immune system. In the absence of sequencing of paired normal tissue or blood samples in most diagnostic scenarios, germline mutation filtering needs to be performed in silico. For all assays evaluated in the current TMB harmonization study, the bioinformatic pipelines include a step of negative filtering for entries in single-nucleotide polymorphism (SNP) databases, such as gnomAD, ExAC, and dbSNP. In addition, some of the pipelines include further steps, for example, filtering by algorithms specifically designed to distinguish somatic versus germline mutations such as somatic-germline zygosity or filtering with respect to the mutations detected by panel sequencing of reference cohorts of normal samples (e.g., NEOplus and ACADEMIC panel).[21] We evaluated the performance of filtering using SNP databases for the LUAD samples (n = 10) (Supplementary Figure 5). Variants detected by WES in matched blood samples were used as a reference. The sensitivity for classifying mutations as somatic was 87%, 90%, and 79%, with corresponding positive predictive values of 90%, 90%, and 91% when using gnomAD, ExAC, and dbSNP for filtering (pooled analysis of the 10 tumor samples). Filtering out only common SNPs (minor allele frequency > 0.001 in gnomAD) increased sensitivity to 98% but decreased positive predictive value to 81%.

Although germline mutation filtering using gnomAD and ExAC performed well, rs-filtering (dbSNP) seemed to be too stringent. Restriction of filtering to common SNPs considerably decreased the number of false negatives but increased the number of false positives. Additional filters that are implemented in the panel-specific bioinformatic pipelines, such as somatic-germline zygosity algorithm or the TSO500 "proxy filter" (Supplementary Figure 6), can further improve germline mutation filtering.

Discussion

Tumor versus matched blood WES was used in many initial clinical immuno-oncology studies and may be considered a reference standard for TMB assessment. However, clinical implementation of WES-based TMB assessment may be impractical, considering the financial costs and the limited availability of appropriately preserved samples or quality DNA, and matched normal samples for germline sequencing. Gene panel assays offer a number of economical and practical advantages for clinical assessment of patient samples, including increased sequencing depth, in silico germline subtraction (negating the requirement for matched samples), and concurrent evaluation of actionable mutations.

The QuIP study provides a thorough analysis of real-world performances of six select TMB panels. Using real-world diagnostic FFPE samples, which included different types of challenging cases with poor DNA quality, heavy fixation artifacts, or low tumor cellularity, our results reveal that, in principle, all assays tested in this study were able to estimate TMB values and could be applied in a diagnostic setting.

The effect of panel size and coverage on the accuracy of psTMB assessment has been previously studied using in silico simulations of gene panels derived from WES data.[17,22] The gene panels used in the laboratory-developed tests covered at least 1 Mbp of the coding sequence, which was found to be essential for valid panel-based TMB assessment.[22] However, even with these large panels, variability of the TMB score can be expected because psTMB measurement has a probabilistic nature: the overall TMB is extrapolated by investigating only a fraction (about 1:30) of the exome. Simulating five commercial panels in WES data, only 17% to 28% of additional error occurred on top of the unavoidable probabilistic error, demonstrating that sufficient panel size is more critical than the particular localization of the panel in the exome.[17]

There is a multitude of other wet-laboratory parameters, ranging from biological factors (e.g., tumor heterogeneity) and preanalytics (e.g., DNA quality) to sequencing (e.g., coverage) and bioinformatics parameters (e.g., germline subtraction) that can influence TMB scores.[23-26] Hence, as expected, absolute TMB values slightly varied. This scenario is not unknown to pathology in general and immune oncology response prediction in particular: just as for TMB, the established PD-L1 IHC assay for NSCLC quantifies a continuous variable in tumor cells ranging from 0% to 100% PD-L1 expressing cells, and several parameters, such as tumor heterogeneity and fixation, are known to influence PD-L1 scores.[4,27] Just as with PD-L1, for clinical purposes, TMB as a continuous measure must be categorized. In our

approach, we stratified samples into one of three groups, which categorized the continuum of TMB ranging from 0 to greater than 200 muts/Mbp: low, intermediate, and high TMB, according to a concept proposed by us recently.[17] In contrast to a two-tier system with one defined cutoff, this concept allows for a definition of a certain "intermediate" gray zone of TMB measurements in an area around the currently proposed clinical cut point. Using cut points of 100 mutations (corresponding to approximately 5 muts/Mbp) and 300 mutations (corresponding to approximately 15 muts/Mbp), strong misclassifications occurred only for a single tumor sample (T4), a case that was particularly challenging because of poor DNA quality, which would justify to decline analysis in a clinical setting. Misclassification of other highly degraded samples (T12 and T19) or critical cases with a low tumor-cell content (T15), high-level microsatellite instability status (T13 and T15), or a loss-of-function mutation in *POLE* (T16) was prevented using the three-tier system instead of the two-tier system.

We observed an overall low influence of the specific laboratory performing the analysis; data generated by the industrial partners for their specific panel were in the range of the respective TMB scores determined by the hospital laboratories. Moreover, we found that most panels had moderate to strong correlations with TMB measured by WES ($\rho = 0.64–0.84$).

Our study also found that germline subtraction using bioinformatic pipelines can be used to identify likely somatic variants in the probabilistic setting of psTMB measurement. Nevertheless, as revealed by us recently, incorrect filtering can influence TMB scores in individual cases, and future studies are warranted to further investigate the influence of in silico versus blood-based subtraction of germline events.[25,28] As current germline variant databases are biased toward, for example, white populations, ethnicity-related aspects require careful analysis in this context.

We identified assay-independent and assay-specific parameters (Fig. 6) that will require careful control when psTMB is implemented in routine diagnostics. Of these, the effects of tumor-cell content, DNA input, and coverage are most critical to prevent the miss of mutations which would result in too low psTMB scores. Another important aspect are deamination artifacts (C>T transitions) created by formalin fixation, which can be diagnostically challenging when left uncontrolled. In this regard, DNA amplification during panel sequencing

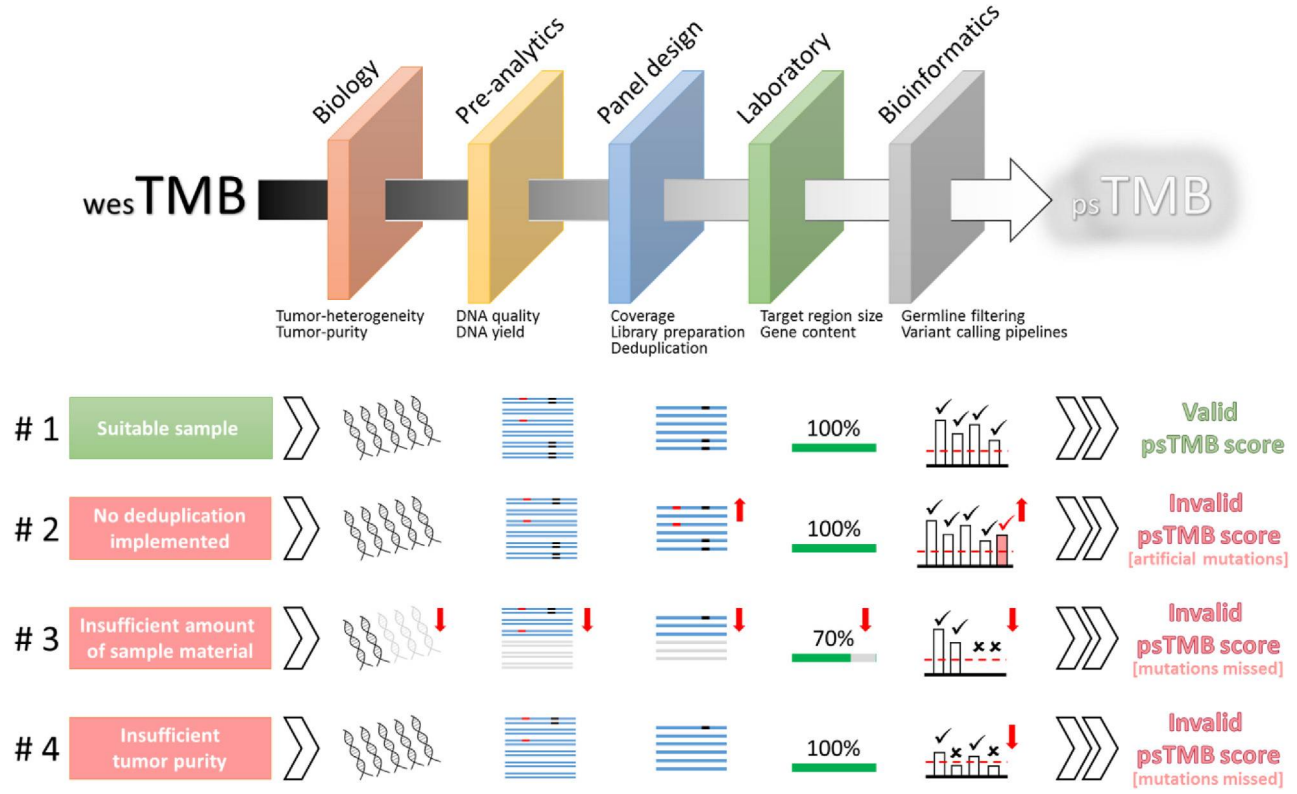

Figure 6. Schematic representation of assay-independent and assay-specific parameters influencing the accuracy of psTMB scores. Lower lane: four representative samples revealing the effect of deduplication strategies (#2), insufficient sample material (#3) or low tumor purity (#4) on DNA input, coverage, covered coding sequence, variant calling, and the resulting psTMB scores. Red arrow pointing down indicates false-negative effect. Red arrow pointing up indicates false-positive effect. psTMB, TMB assessed by gene panel sequencing; wesTMB, TMB assessed by WES.

can be critical as overamplification of artifacts or additional errors during replication can occur, leading to false-positive mutation calls and subsequently to overestimated psTMB scores. This issue can be compensated by setting an appropriate limit of detection (LOD) for the allelic frequency and especially by applying in silico or technical (molecular barcodes) approaches, or both, for deduplication (removal of PCR duplicates). In the present data set, a LOD of 5% in combination with deduplication yielded reliable mutation calling, and eventually TMB values. Hybrid-capture–based target enrichment was favorable in this context. In panels without deduplication, deamination artifacts may be controlled by increasing the LOD to, for example, 10%, rendering cases with low tumor cellularity challenging owing to the impaired sensitivity. Recent reports indicate that the application of uracil-DNA glycosylase, an enzyme selectively digesting uracil-containing nucleic acid, can reduce deamination artifacts, when assessing TMB in FFPE samples using assays without a deduplication approach.[29-31] However, the effect of this approach was not tested in this study.

We also calculated BFs to convert psTMB to wesTMB for the assessed panels. Although future studies exploring larger sample sets will likely improve this analysis, we believe that the data revealed here provide a strong and sound basis that will facilitate the comparison of TMB values obtained by different panels.

A limitation of our study is the limited number of cases and the use of three different cancer types. The latter selection was influenced by (1) a case mix that reflects the continuum of TMB, (2) avoiding result bias owing to a single cancer type, (3) tissue availability for the entire study and all partners, and (4) availability of corresponding WES data. Because the predictive power of TMB is currently being tested in many immuno-oncology trials across various cancer types, and as our study is primarily aimed at investigating the ability of panels to measure TMB, we believe that these points do not interfere with our results and conclusions.

In summary, the QuIP study provides real-world evidence that all panels tested in this study can be used to estimate TMB by panel sequencing from FFPE samples in a routine diagnostic setting. However, this study has identified several critical parameters, including sample fixation, DNA input, sequencing depth, genome coverage, and VAF cut points, that may confound psTMB estimation and require careful control to achieve successful and reliable psTMB analysis. Beyond TMB, in conjunction with efforts by the Friends of Cancer Research, this study provides a blueprint and framework for the systematic analysis of complex or composite predictive biomarkers, which will likely play an increasing role in guiding oncological therapy.

Acknowledgments

The authors thank all teams of the participating institutions for their contributions and the QuIP team for excellent administration. The authors also thank the entire team of the Center for Molecular Pathology (CMP) at the Institute of Pathology Heidelberg (IPH) for expert technical assistance and fruitful discussions. The support of the Heidelberg Center for Personalized Oncology (HIPO) program and Genomics and Proteomics Core Facility (GPCF) (both DKFZ) is gratefully acknowledged. Editorial assistance was provided by Stuart Rulten, PhD, and Jay Rathi, MA, of Spark Medica Inc., funded by Bristol-Myers Squibb. The authors thank David Fabrizio for fruitful discussions related to this project. The authors thank Sandra Siesing for expert editorial assistance and overall handling of the manuscript. The study was partly sponsored by Bristol-Myers Squibb, Illumina, Merck & Co, Inc, Foundation Medicine, Inc, Neo New Oncology, QIAGEN, F. Hoffmann-La Roche, AG, and Thermo Fisher Scientific.

MD, RB, SMB, FH, WW, AJ, JM, PS, MH, TK, HM, DK, JB, VE, and AS conceived and designed the study; DH, SF, PS, HG, AS, and VE provided samples; AS, VE, JB, SMB, DK, WD, NP, US, MH, SH, JA, MZ, LT, ER, MS, HG, SF, JA, DH, GB, CW, MT, ME, HM, TK, RB, OS, AJ, FH, WW, and MD sequenced and analyzed cases; JB, ER, VE, DK, and AS conducted statistical analyses; AS, DK, VE, JB, PS, and MD contributed to the preparation of the manuscript draft; all authors contributed to the writing of the manuscript. All authors read and approved the manuscript.

References

1. Havel JJ, Chowell D, Chan TA. The evolving landscape of biomarkers for checkpoint inhibitor immunotherapy. *Nat Rev Cancer*. 2019;19:133–150.
2. Jenkins RW, Barbie DA, Flaherty KT. Mechanisms of resistance to immune checkpoint inhibitors. *Br J Cancer*. 2018;118:9–16.
3. Brahmer JR, Lacchetti C, Schneider BJ, et al. Management of immune-related adverse events in patients treated with immune checkpoint inhibitor therapy: American Society of Clinical Oncology Clinical Practice Guideline. *J Clin Oncol*. 2018;36:1714–1768.
4. Buttner R, Gosney JR, Skov BG, et al. Programmed death-ligand 1 immunohistochemistry testing: a review of analytical assays and clinical implementation in non-small-cell lung cancer. *J Clin Oncol*. 2017;35:3867–3876.
5. Le DT, Uram JN, Wang H, et al. PD-1 blockade in tumors with mismatch-repair deficiency. *N Engl J Med*. 2015;372:2509–2520.

6. Campbell BB, Light N, Fabrizio D, et al. Comprehensive analysis of hypermutation in human cancer. *Cell*. 2017;171:1042–1056.e10.
7. Carbone DP, Reck M, Paz-Ares L, et al. First-line nivolumab in stage IV or recurrent non-small-cell lung cancer. *N Engl J Med*. 2017;376:2415–2426.
8. Rizvi NA, Hellmann MD, Snyder A, et al. Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science*. 2015;348:124–128.
9. Snyder A, Makarov V, Merghoub T, et al. Genetic basis for clinical response to CTLA-4 blockade in melanoma. *N Engl J Med*. 2014;371:2189–2199.
10. Van Allen EM, Miao D, Schilling B, et al. Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. *Science*. 2015;350:207–211.
11. Chalmers ZR, Connelly CF, Fabrizio D, et al. Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. *Genome Med*. 2017;9:34.
12. Gandara DR, Paul SM, Kowanetz M, et al. Blood-based tumor mutational burden as a predictor of clinical benefit in non-small-cell lung cancer patients treated with atezolizumab. *Nat Med*. 2018;24:1441–1448.
13. Zehir A, Benayed R, Shah RH, et al. Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. *Nat Med*. 2017;23:703–713.
14. Merino DM, McShane LM, Fabrizio D, et al. Establishing guidelines to harmonize tumor mutational burden (TMB): in silico assessment of variation in TMB quantification across diagnostic platforms: phase I of the Friends of Cancer Research TMB harmonization project. *J Immunother Cancer*. 2020. In press.
15. Merino DM, McShane L, Butler M, et al. TMB standardization by alignment to reference standards: phase II of the Friends of Cancer Research TMB Harmonization Project [abstract]. *J Clin Oncol*. 2019;37(suppl 15):2624.
16. R Core Team. R: a language and environment for statistical computing. http://www.r-project.org. 2018. Accessed March 11, 2020.
17. Budczies J, Allgauer M, Litchfield K, et al. Optimizing panel-based tumor mutational burden (TMB) measurement. *Ann Oncol*. 2019;30:1496–1506.
18. Hellmann MD, Ciuleanu TE, Pluzanski A, et al. Nivolumab plus ipilimumab in lung cancer with a high tumor mutational burden. *N Engl J Med*. 2018;378:2093–2104.
19. Hellmann MD, Paz-Ares L, Bernabe Caro R, et al. Nivolumab plus ipilimumab in advanced non-small-cell lung cancer. *N Engl J Med*. 2019;381:2020–2031.
20. Ready N, Hellmann MD, Awad MM, et al. First-line nivolumab plus ipilimumab in advanced non-small-cell lung cancer (CheckMate 568): outcomes by programmed death ligand 1 and tumor mutational burden as biomarkers. *J Clin Oncol*. 2019;37:922–1000.
21. Sun JX, He Y, Sanford E, et al. A computational approach to distinguish somatic vs. germline origin of genomic alterations from deep sequencing of cancer specimens without a matched normal. *PLoS Comput Biol*. 2018;14, e1005965.
22. Buchhalter I, Rempel E, Endris V, et al. Size matters: dissecting key parameters for panel-based tumor mutational burden analysis. *Int J Cancer*. 2019;144:848–858.
23. Ascierto PA, Bifulco C, Palmieri G, Peters S, Sidiropoulos N. Preanalytic variables and tissue stewardship for reliable next-generation sequencing (NGS) clinical analysis. *J Mol Diagn*. 2019;21:756–767.
24. Buttner R, Longshore JW, López-Rios F, et al. Implementing TMB measurement in clinical practice: considerations on assay requirements. *ESMO Open*. 2019;4: e000442.
25. Kazdal D, Endris V, Allgäuer M, et al. Spatial and temporal heterogeneity of panel-based tumor mutational burden in pulmonary adenocarcinoma: separating biology from technical artifacts. *J Thorac Oncol*. 2019;14:1935–1947.
26. Stenzinger A, Allen JD, Maas J, et al. Tumor mutational burden standardization initiatives: recommendations for consistent tumor mutational burden assessment in clinical samples to guide immunotherapy treatment decisions. *Genes Chromosomes Cancer*. 2019;58:578–588.
27. Kerr KM, Tsao MS, Nicholson AG, et al. Programmed death-ligand 1 immunohistochemistry in lung cancer: in what state is this art? *J Thorac Oncol*. 2015;10:985–989.
28. Chang H, Sasson A, Srinivasan S, et al. Bioinformatic methods and bridging of assay results for reliable tumor mutational burden assessment in non-small cell lung cancer. *Mol Diagn Ther*. 2019;23:507–520.
29. Alborelli I, Leonards K, Rothschild SI, et al. Tumor mutational burden assessed by targeted NGS predicts clinical benefit from immune checkpoint inhibitors in non-small cell lung cancer. *J Pathol*. 2020;250:19–29.
30. Do H, Molania R, Mitchell PL, Vaiskunaite R, Murdoch JD, Dobrovic A. Reducing artifactual *EGFR* T790M mutations in DNA from formalin-fixed paraffin-embedded tissue by use of thymine-DNA glycosylase. *Clin Chem*. 2017;63: 1506–1514.
31. Heeke S, Benzaquen J, Long-Mira E, et al. In-house implementation of tumor mutational burden testing to predict durable clinical benefit in non-small cell lung cancer and melanoma patients. *Cancers (Basel)*. 2019;11:1271.