frontiers in Oncology

ORIGINAL RESEARCH
published: 28 September 2021
doi: 10.3389/fonc.2021.732644

RNA Sequencing Data for FFPE Tumor Blocks Can Be Used for Robust Estimation of Tumor Mutation Burden in Individual Biosamples

Maxim Sorokin[1,2,3,4*], *Alexander Gorelyshev*[3,4], *Victor Efimov*[3], *Evgenia Zotova*[2], *Marianna Zolotovskaia*[3], *Elizaveta Rabushko*[2], *Denis Kuzmin*[3], *Alexander Seryakov*[5], *Dmitry Kamashev*[6], *Xinmin Li*[7], *Elena Poddubskaya*[8], *Maria Suntsova*[8] *and Anton Buzdin*[1,3,4,6,8*]

[1] Biostatistics and Bioinformatics Subgroup, European Organization for Research and Treatment of Cancer (EORTC), Brussels, Belgium, [2] The Laboratory of Clinical and Genomic Bioinformatics, I.M. Sechenov First Moscow State Medical University, Moscow, Russia, [3] Laboratory for Translational Genomic Bioinformatics, Moscow Institute of Physics and Technology, Dolgoprudny, Russia, [4] OmicsWay Corp., Walnut, CA, United States, [5] Medical Holding SM-Clinic, Moscow, Russia, [6] Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of Sciences, Moscow, Russia, [7] Department of Pathology & Laboratory Medicine, University of California Los Angeles (UCLA) Technology Center for Genomics & Bioinformatics, Los Angeles, CA, United States, [8] World-Class Research Center "Digital Biodesign and Personalized Healthcare", Sechenov First Moscow State Medical University, Moscow, Russia

OPEN ACCESS

Edited by:
Xiao Zhu,
Guangdong Medical University, China

Reviewed by:
Magali Espinosa,
Instituto Nacional de Medicina Genómica (INMEGEN), Mexico
Yanchun Zhang,
Boston Children's Hospital and Harvard Medical School, United States

***Correspondence:**
Maxim Sorokin
sorokin@oncobox.com
Anton Buzdin
buzdin@oncobox.com

Specialty section:
This article was submitted to Cancer Genetics, a section of the journal Frontiers in Oncology

Received: 29 June 2021
Accepted: 06 September 2021
Published: 28 September 2021

Citation:
Sorokin M, Gorelyshev A, Efimov V, Zotova E, Zolotovskaia M, Rabushko E, Kuzmin D, Seryakov A, Kamashev D, Li X, Poddubskaya E, Suntsova M and Buzdin A (2021) RNA Sequencing Data for FFPE Tumor Blocks Can Be Used for Robust Estimation of Tumor Mutation Burden in Individual Biosamples. Front. Oncol. 11:732644. doi: 10.3389/fonc.2021.732644

Tumor mutation burden (TMB) is a well-known efficacy predictor for checkpoint inhibitor immunotherapies. Currently, TMB assessment relies on DNA sequencing data. Gene expression profiling by RNA sequencing (RNAseq) is another type of analysis that can inform clinical decision-making and including TMB estimation may strongly benefit this approach, especially for the formalin-fixed, paraffin-embedded (FFPE) tissue samples. Here, we for the first time compared TMB levels deduced from whole exome sequencing (WES) and RNAseq profiles of the same FFPE biosamples in single-sample mode. We took TCGA project data with mean sequencing depth 23 million gene-mapped reads (MGMRs) and found 0.46 (Pearson)–0.59 (Spearman) correlation with standard mutation calling pipelines. This was converted into low (<10) and high (>10) TMB per megabase classifier with area under the curve (AUC) 0.757, and application of machine learning increased AUC till 0.854. We then compared 73 experimental pairs of WES and RNAseq profiles with lower (mean 11 MGMRs) and higher (mean 68 MGMRs) RNA sequencing depths. For higher depth, we observed ~1 AUC for the high/low TMB classifier and 0.85 (Pearson)–0.95 (Spearman) correlation with standard mutation calling pipelines. For the lower depth, the AUC was below the high-quality threshold of 0.7. Thus, we conclude that using RNA sequencing of tumor materials from FFPE blocks with enough coverage can afford for high-quality discrimination of tumors with high and low TMB levels in a single-sample mode.

Keywords: TMB (tumor mutation burden), RNAseq, FFPE (formalin-fixed paraffin-embedded), machine learning, oncology

INTRODUCTION

Tumor mutation burden (TMB) per million base pairs is a well-known efficacy predictor for checkpoint inhibitor immunotherapy (1). TMB can be calculated in several ways (2). For example, in commercial FDA-approved FoundationOne CDx test for unpaired single tumor samples, TMB is defined as the number of somatic mutations per million base pairs (megabase) of the protein-coding sequence analyzed—including both substitutions and indels, but irrespective of the functional consequences of the variants (3). Highly mutated tumors are more likely to produce tumor neoantigens and become more "visible" to the immune system; thus, TMB is a good proxy for the tumor neoantigen load (4).

To date, TMB assessment is commercially available in the form of clinical and research use-only diagnostic tests (5). In June 2020, TMB was approved for the use of immune checkpoint inhibitor pembrolizumab in the treatment of patients with advanced or metastatic solid cancers, and FoundationOne CDx assay was approved as a companion diagnostic test.

The two major approaches for evaluating TMB are based on using whole exome sequencing (WES) and next-generation sequencing (NGS) panels. WES-TMB was demonstrated first to be associated with tumor responses on immune checkpoint inhibitors and thus proposed as a predictive biomarker (6–9). These early WES-TMB estimates were considering only non-synonymous somatic mutations. Overall, TMB levels were classified as "high" or "low." However, the cutoff values varied from ≥7.4 in esophageal and gastric cancer till ≥23.1 in non-small cell lung cancer (NSCLC) for the number of mutations per megabase DNA and from ≥158 mutations in advanced NSCLC till ≥248 mutations in advanced small cell lung cancer for the whole tumor exome non-synonymous mutation estimates (10).

To address many of the WES-TMB limitations, targeted sequencing panels with exonic sequences of especially frequently mutated genes were developed to estimate TMB (11, 12). Unlike in WES-TMB, NGS panels count both non-synonymous and synonymous mutations as well as indels, which can increase assay sensitivity (3). This approach showed that sufficiently large NGS panels can accurately recapitulate WES-TMB, and demonstrated good agreement between panels-derived and WES-derived TMB values (13, 14).

There are two NGS panels commercially available to date that have been approved by regulatory bodies: FoundationOne CDx assay approved by the FDA as a companion diagnosis for the assessment of TMB and the MSK-IMPACT panel.

Overall, many variable factors can influence TMB estimation and output, including tumor type (15), biosample type (FFPE materials artificially have more mutations than fresh frozen tissue), and sequencing parameters (NGS panel content, size, and sequencing depth; bioinformatic pipeline; and reporting cutoff) (16).

At the moment, DNA analysis is the only standard for TMB assessment and it is largely unclear whether TMB derived from RNA sequencing (RNAseq) corresponds to DNAseq data. In 2020, Jang and coauthors attempted to calculate TMB from single-cell RNA sequencing data (17). However, the authors did not provide any technical rationale for their approach and did not validate it by DNA mutation analysis, thus leaving the adequacy of the results communicated uncertain. In February 2021, DiGuardo and colleagues demonstrated a correlation between RNAseq- and DNAseq-derived TMB using formalin-fixed, paraffin-embedded (FFPE) tumor tissue blocks (18). However, this was more a proof-of-concept study done for only eight individual samples and the methodological limitations of this approach were not explored. Furthermore, this was done for the matched pairs of tumors and adjacent normal tissues, while matching healthy samples are not frequently available in the routine clinical practice. For example, the FoundationOne CDx test utilizes single tumor-only biosamples to return TMB (2).

FFPE cancer tissue biosamples are known to yield highly fragmented nucleic acid preparations that could be hardly applicable to the tasks requiring high RNA integrity like analysis of differential splicing (19). Nucleic acids extracted from FFPE also often contain artifact alterations caused by formalin fixation, thus having an increased rate of C>T substitutions, when compared with nucleic acids from fresh tissues. Thus, sequencing profiles derived from FFPE should be processed differently and the results obtained from fresh tissues may be poorly compatible. However, FFPE materials can be used to properly estimate TMB by the DNA screens (20) and for the clinical-grade estimation of the gene expression levels by analyzing RNA reads (21, 22). From certain points of view, this is the preferred type of biomaterial because of its high availability and stability, as FFPE blocks can be stored at room temperature for years prior to nucleic acid extraction for sequencing purposes (23).

Thus, in this study, we investigated whether FFPE-isolated RNA can be used for TMB estimates without the analysis of adjacent or blood-derived norms. To this end, we used the paired RNAseq–WES profiles for FFPE materials available from The Cancer Genome Atlas (TCGA) project database, in both paired and single-sample modes, and then validated the results using 73 experimental RNAseq–WES profiles obtained for FFPE cancer specimens.

For TMB deduced using TCGA RNAseq data, where mean sequencing depth was ~23 million gene-mapped reads (MGMRs), we obtained 0.46 (Pearson)–0.59 (Spearman) correlation with the standard mutation calling pipelines. This was converted in the classifier of low (<10) and high (>10) TMB per megabase with area under the curve (AUC) 0.757, and the application of machine learning (ML) increased AUC till 0.854. We then compared 73 experimental pairs of WES and RNAseq profiles with lower (mean 11 MGMRs) and higher (mean 68 MGMRs) sequencing depths. We observed 0.85 (Pearson)–0.95 (Spearman) correlation of TMB with standard mutation calling pipelines for higher RNA sequencing depth samples, and ~1 AUC for the high/low TMB classifier. However, for the lower depth, the AUC was below the high-quality threshold of 0.7 even in case of applying ML. Thus, we conclude that using RNA sequencing of tumor materials from FFPE blocks with enough coverage can afford for high-quality discrimination of high and low TMB tumors in a single-sample mode.

MATERIALS AND METHODS

Reference Public Dataset

The available set of matching tumor WES/FFPE RNAseq and normal (blood or adjacent non-cancerous tissue) WES FASTQ files corresponding to the same tumors, a total of 53 samples, was downloaded from The Cancer Genome Atlas (TCGA) international project repository (https://portal.gdc.cancer.gov/), and only FFPE samples of primary tumors were selected (**Table S1**).

Experimental Tissue Samples

All experimental data were obtained for pathologist-verified FFPE tumor tissue blocks with tumor cell content greater than 50%. The sample annotation contained information about the sex, age, and cancer type of the patient (**Table S2**). In all cases, written informed consents to participate in this study were acquired from the patients or from their legal representatives. The consent procedure and the design of the study were approved by the ethical committees of the Karelia Republic Oncological Hospital, Petrozavodsk, Russia, and Vitamed Oncological Clinical Center, Moscow, Russia.

RNAseq: Library Preparation and Sequencing

RNA sequencing was performed according to the previous protocol used to generate ANTE collection of healthy tissue RNAseq profiles (23) and several cancer expression collections (22, 24–28). To isolate RNA preps, 10-µM-thick paraffin slices were trimmed from each FFPE tissue block with a microtome. RNA was extracted from FFPE slices using Qiagen RNeasy FFPE kit following the protocol of the manufacturer. RNA 6000 Nano or Qubit RNA Assay kits were used to measure RNA concentration. RNA integrity number (RIN) was measured using Agilent 2100 Bioanalyzer. For depletion of ribosomal RNA and library construction, KAPA RNA Hyper with rRNA erase kit (HMR only) was used. Different adaptors were used for multiplexing samples in a single sequencing run. Library concentrations were measured using the Qubit dsDNA HS Assay kit (Life Technologies) and Agilent TapeStation (Agilent). RNA sequencing was performed at the Department of Pathology and Laboratory Medicine, University of California Los Angeles, using Illumina HiSeq 3000 equipment for single-end sequencing, 50 bp read length, achieving a median of ~217 million raw reads or ~68 million mapped reads per sample. Illumina SAV was used for data quality checks. De-multiplexing was performed with Illumina Bcl2fastq2 Conversion Software v2.17.

Sequencing data were deposited in NCBI Sequencing Read Archive (SRA) under accession ID PRJNA733593.

Whole Exome Sequencing: Library Preparation and Sequencing

DNA WES was performed according to (29). DNA was extracted from the FFPE tissue using the AnaPrep FFPE DNA extraction kit and whole exome DNA was captured from total genomic DNA using the SeqCap EZ System from NimbleGen according to the instructions of the manufacturer. Briefly, genomic DNA was sheared, size selected to roughly 200–250 base pairs, and the ends were repaired and ligated to specific adapters and multiplexing indexes. Fragments were then incubated with SeqCap biotinylated DNA baits followed by the LM-PCR, and the RNA–DNA hybrids were purified using streptavidin-coated magnetic beads. The RNA baits were then digested to release the targeted DNA fragments, followed by a brief amplification of 15 or less PCR cycles. Sequencing was performed on Illumina HiSeq 3000 for a pair read 150 run. Data quality check was done on Illumina SAV. Demultiplexing was performed with Illumina Bcl2fastq2 v 2.17 program.

Processing of RNA Sequencing Data

For RNAseq data, a GATK mutation calling pipeline was used (**Figure 1**) (30). Reads were aligned to the human genome assembly GRCh38 with STAR v2.6.1d software in two-pass mode (31). The following parameters were set to non-default values: sjdbOverhang 100, twopass1readsN 10000000, and twopassMode Basic.

Exon coordinates were taken from Ensembl annotation version 89. Samtools v1.3.1 package was used for BAM file indexing (32). All reads were assigned to a single read group, and read group information editing and duplicate marking were performed with AddOrReplaceReadGroups and MarkDuplicates software, respectively (http://broadinstitute.github.io/picard). GATK v3.8.0 SplitNCigarReads module was used to split reads that aligned to exon junctions. Base quality score recalibration was performed with GATK v4.beta.1 BaseRecalibrator and ApplyBQSR modules. For mutation calling, GATK4 Mutect2 software (33) was used in tumor-only mode with dbSNP version 146 variant database (34) and 1000G gold standard indel database (35). No panel of norms was used for the experimental settings.

Variants were called only in exonic regions of human chromosomes 1–22, X, and Y (mapped according to GENCODE (36), and the PCR indel model parameter was set to "HOSTILE". Variants were filtered with GATK4 FilterMutectCalls (all variants are kept in VCF, only the "FILTER" field is edited). Predicted functional effects of variants identified were annotated using ANNOVAR software (37). For variant annotation, we used the version of ExAC database, which does not contain TCGA samples (38). All tri- or more allelic sites were excluded from further analyses: such mutations were not annotated using ANNOVAR and were not included in TMB calculation.

Processing of WES Data

For WES data analysis, a GATK somatic mutation calling pipeline was used (**Figure 1**). Reads were aligned to the human

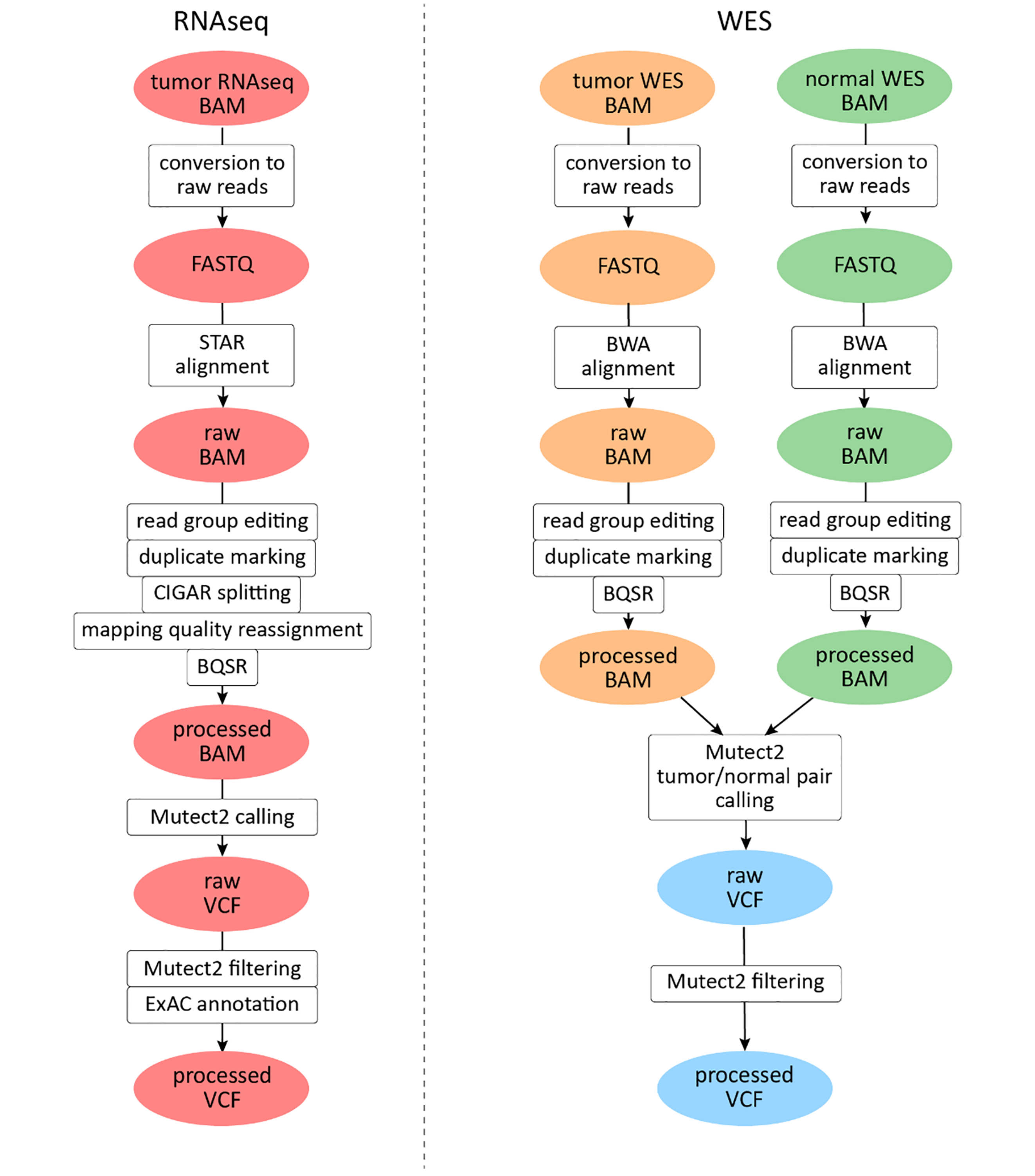

FIGURE 1 | Mutation calling pipeline for The Cancer Genome Atlas (TCGA) data analysis. File instances are shown in ovals, pipeline steps in rectangles. RNAseq files are highlighted in red, tumor whole exome sequencing (WES) files in orange, and normal WES files in green. Data derived from both tumor and normal WES data are shown in light blue.

genome version 38 with BWA mem v0.7.17 software (39). The following parameters were set to non-default values: –k 15, –r 2. The rest of the pre-processing steps were identical to the RNAseq pipeline described above, except for reads splitting and mapping quality editing steps which were skipped.

For mutation calling, GATK4 Mutect2 (33) software was used simultaneously for tumor and matched normal samples, supplied with the same dbSNP and indel databases, regions, and PCR model. Subsequent post-processing steps included filtering GATK4 FilterMutectCalls and annotation with ANNOVAR. All tri- or more allelic sites were excluded from further analyses: such mutations were not annotated using ANNOVAR and were not included in TMB calculation. For parallel computational task management, GNU parallel software was used (40).

Supervised Machine Learning

For filtering with supervised learning, we selected 31 variant features such as reference allele depth, median base quality, or the number of events in the haplotype. From them, 23 features were taken directly from Mutect2 output VCF files and one was obtained from ANNOVAR annotation of VCF with non-TCGA ExAC database. The other seven features were engineered using data from Mutect2 output VCF: four Boolean features of variant being *i*) an insertion, *ii*) a deletion, *iii*) a C>T (G>A) transition, and *iv*) C>A (G>T) transversion were constructed based on values in REF and ALT fields and three integer features: *v*) total depth and *vi–vii*) REF and ALT lengths were constructed based on values in REF, ALT fields, and FORMAT AD field.

Model hyperparameters were selected during a series of randomized grid searches. Parameters adjusted were as follows: learning_rate, n_estimators, min_child_weight, gamma, subsample, colsample_bytree, max_depth, reg_alpha, and reg_lambda; among them, the first two were the most impactful. Cross-validation was five-fold, and receiver operating characteristic (ROC) AUC was used as the metric for hyperparameter selection. The Python code used for ML and visualization is available at https://gitlab.com/oncobox/tmb_rnaseq.

RESULTS

To initially explore the possibility of estimating TMB from FFPE RNAseq data, we used a set with matched WES/RNAseq tumor profiles and, at the same time, with the normal control WES data available from the TCGA project database. Totally, data could be obtained for only 53 tumor cases (paired RNAseq and WES data for tumor samples, and WES data for the matched controls) because the absolute majority of the paired TCGA data were generated for fresh-frozen tissue samples.

We then assessed how different RNAseq data filtering options, including ML, can alter the correlation between TMB-RNAseq and TMB-WES. To this end, we randomly assigned 53 available TCGA samples to the training (n = 27) and validation (n = 26) subsets.

Modeling of WES-RNAseq-TMB Correlation on TCGA Dataset With Matched Normal WES Controls

In this application, WES tumor mutation calling was performed while taking into account the available WES profiles for the healthy control biosamples from the same patients. Our first step was to determine the biggest possible correlation between RNAseq and WES-TMB estimates for the above FFPE TCGA dataset. To that end, we collected variants common to the WES and RNAseq callsets, as identified by genomic coordinate, reference, and alternative allele. We used variants common to both WES and RNAseq only to estimate maximal possible correlation between TMB derived from WES *vs.* TMB derived from RNAseq ("isec"). For further rule- and ML-based filtering of RNAseq variants presented below, all RNAseq variants were used regardless of their presence in WES data. The WES variants were filtered by the condition "FILTER==PASS," while RNAseq variants were left unfiltered. The Pearson correlation coefficient between TMB-RNAseq and TMB-WES estimates calculated for the n = 206 subset was 0.91, $p = 2.1 \times 10^{-10}$ (**Figure 2**, "isec"). No discernable sequencing batch bias was detected during both correlation analysis and primary component analysis: samples from different TCGA sequencing centers were randomly clustered on the PCA (**Figure S1**).

We then built the ROC curve and calculated the AUC metric for it. ROC AUC is widely used to assess the performance of biomarkers in oncology (41–43), and it depends on their sensitivity and specificity (44). It varies between 0.5 and 1, and the robustness criterion of biomarkers is typically AUC greater than 0.7 (45).

We selected three clinically relevant thresholds of TMB per megabase: of TMB ≥6, ≥10, and ≥20 (11, 46). Among these, the cutoff value of 10 is currently especially frequently used in clinical studies, although no formal consensus has been reached yet (10). Seven out of 26 samples (27%) had TMB ≥10 Mut/Mb. This is higher than in previous studies investigating pan-cancer patient cohorts with different solid tumors (47–49). Considering WES-derived TMB as the gold standard among the available TMB data, we obtained the AUC scores for TMB-RNAseq of 0.925, 0.903, and 1 for the abovementioned thresholds, respectively (**Table 1**). Note that these estimates considered only the fraction of genome that was sufficiently covered by both WES and RNAseq reads and left apart other sequences that were covered by the RNAseq or WES reads separately.

We then simulated a situation when there are no available WES data for matched healthy controls. To this end, we assessed the performance of algorithmic RNAseq data filtering, treating it as a baseline. The selected variants used for TMB calculations were *i*) marked as "germline_risk" or "panel_of_normals, germline_risk" by Mutect2, *ii*) had ExAC frequency <0.000033

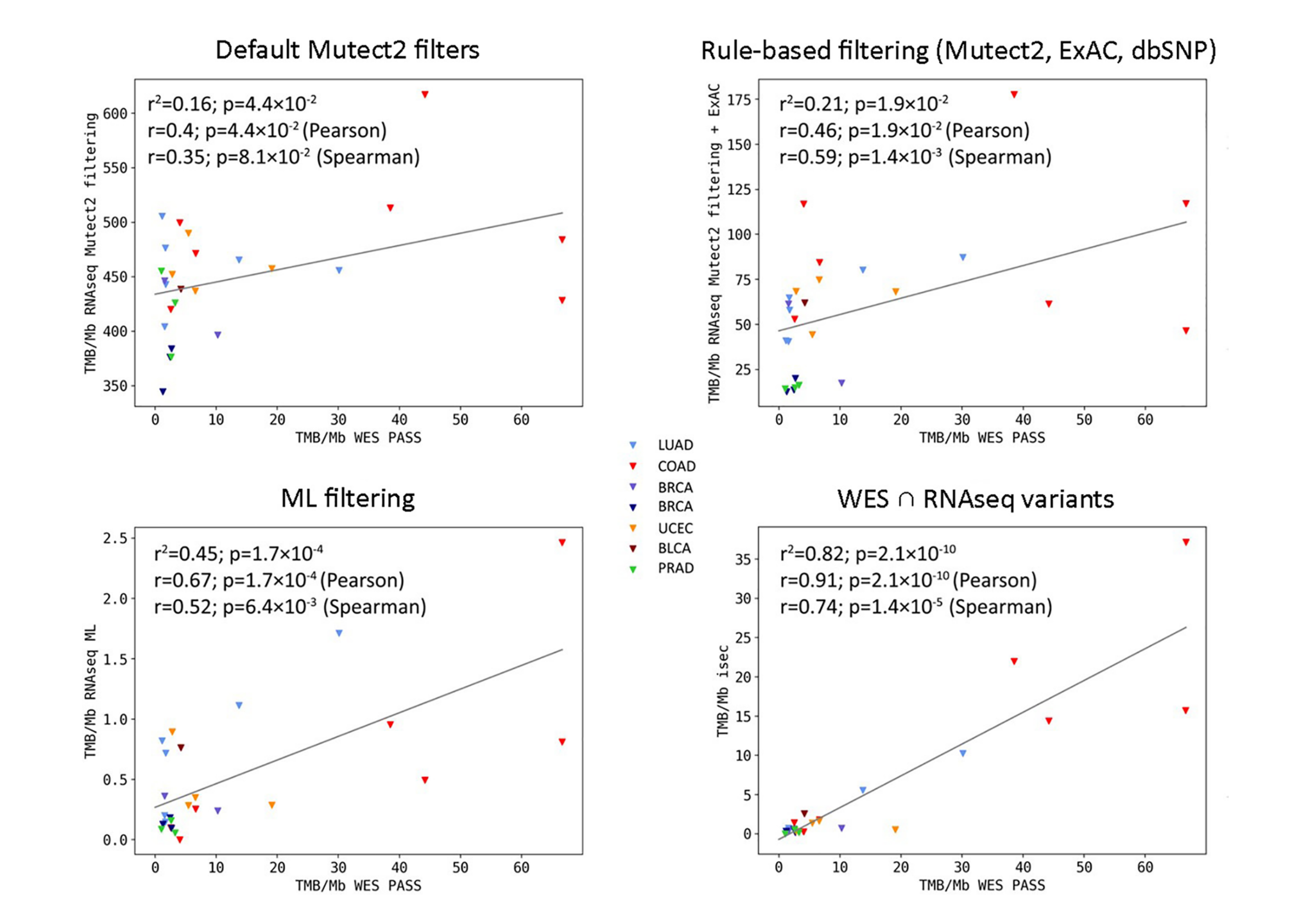

FIGURE 2 | Correlations between RNAseq- and WES-derived tumor mutation burden (TMB) estimates (TCGA FFPE dataset, *n* = 26), with matched normal WES references. Samples are colored according to cancer type. Cancer type abbreviations used: LUAD, lung adenocarcinoma; COAD, colon adenocarcinoma; BRCA, breast invasive carcinoma; UCEC, uterine corpus endometrial carcinoma; BLCA, urothelial bladder carcinoma; PRAD, prostate adenocarcinoma.

TABLE 1 | ROC AUC scores for predicting TMB (≥6, ≥10, and ≥20 thresholds) from FFPE RNAseq data, according to TMB-WES standards calculated with matched normal WES profiles.

Filtering method	TMB ≥ 6	TMB ≥ 10	TMB ≥ 20
TCGA (matched norm was used for WES mutation calling)			
WES ∩ RNAseq variants	0.925	0.903	1
Default Mutect2 filtering	0.694	0.701	0.781
Default Mutect2 filtering + ExAC <0.000033	0.825	0.757	0.81
ML filtering using XGBoost method	0.825	0.854	0.905
TCGA (matched norm was not used for WES mutation calling)			
WES ∩ RNAseq variants	0.953	0.921	1
Default Mutect2 filtering	0.573	0.64	0.726
Default Mutect2 filtering + ExAC <0.000033	0.693	0.763	0.857
ML filtering using XGBoost method	0.7	0.868	0.857
Experimental dataset (high coverage)			
WES ∩ RNAseq variants	0.75	1	NA
Default Mutect2 filtering	0.812	0.857	NA
Default Mutect2 filtering + ExAC <0.000033	1	1	NA
Experimental dataset (low coverage)			
WES ∩ RNAseq variants	0.719	0.748	NA
Default Mutect2 filtering	0.407	0.226	NA
Default Mutect2 filtering + ExAC <0.000033	0.622	0.674	NA
ML filtering using XGBoost method	0.53	0.613	NA

NA, Not applicable.

(3), and *iii*) had no associated dbSNP150 identifier. Mutect2 software filters were chosen here to accommodate for the lack of paired normal samples for RNAseq data. Using this approach, we observed a weak yet statistically significant correlation between the TMB-RNAseq and TMB-WES estimates with Pearson correlation 0.46, $p = 0.019$ (**Figure 2B**), suggesting the need for a more advanced filtering method.

To improve the correlation and decrease the signal-to-noise ratio, we developed a supervised ML binary classifier, based on the XGBoost algorithm (50) (**Figure 3**). The $n = 27$ subset (458,957 variants called) was used here as a training dataset, and the $n = 26$ subset (375,148 variants called) as a test subset. Prior to the analysis, unfiltered RNAseq VCF files were merged within each subset. Each variant in the training subset was labeled as either "signal" or "noise" based on whether a variant with the same genomic coordinate was discovered in the accompanying WES "gold standard" data. Confusion matrices for the ML model predictions on the training and testing subsets are shown in **Table 2**, and feature importance scores (gain), assigned by the ML model, are shown in **Figure S2**. These scores reflect the value of each feature in the construction of the boosted decision trees within the model. The more an attribute is used to

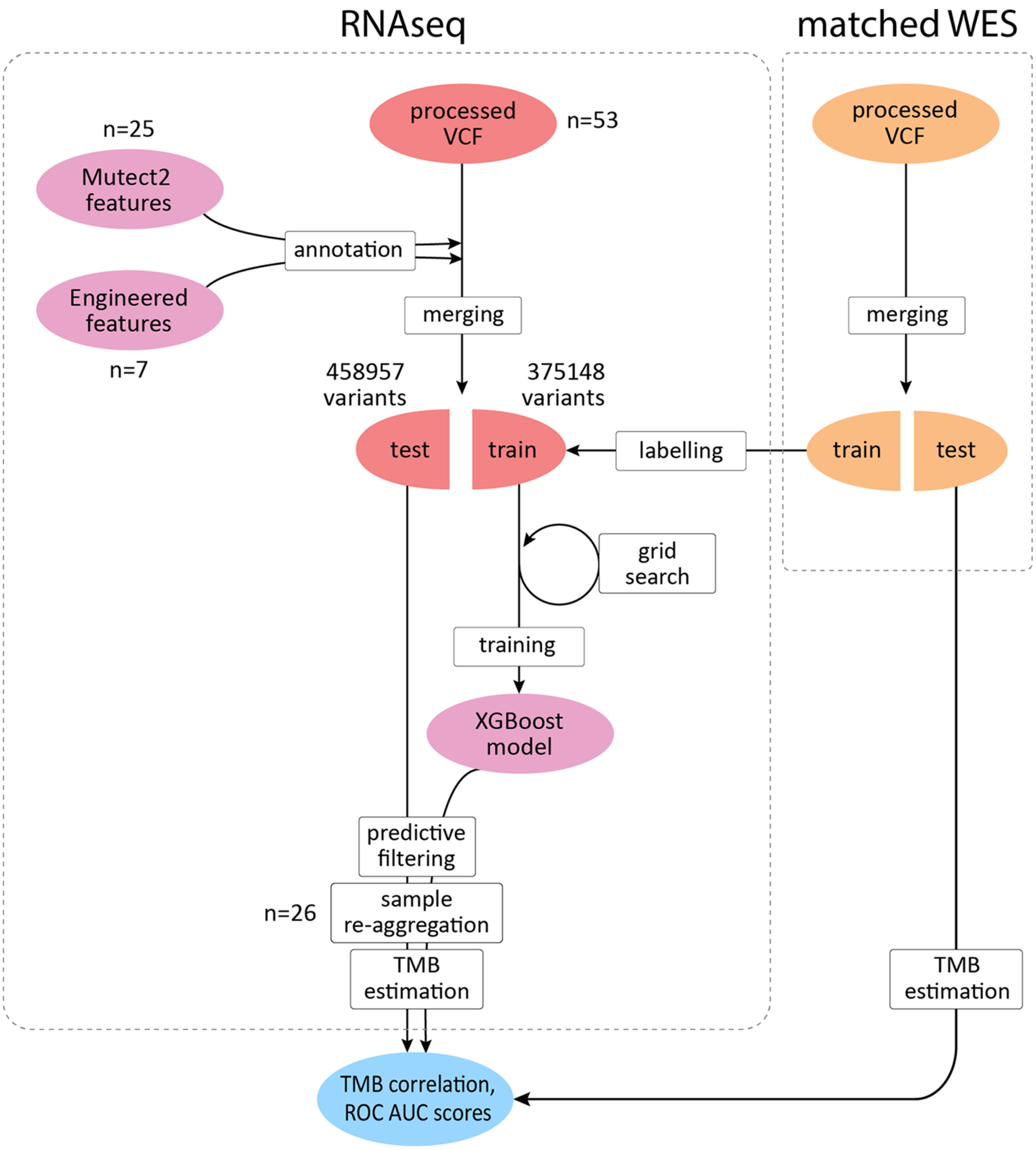

FIGURE 3 | XGBoost binary classifier development workflow. TCGA FFPE samples with matched WES and RNAseq data were reanalyzed to produce callsets in the VCF format. Thirty-two features were introduced to the model in total. RNAseq files along with matching WES files were randomly assigned to two subgroups with variants merged respectively to obtain two sets of variants for each data source. RNAseq variants from the training subgroup were labeled by cross-referencing with the WES callset. Variants matched in WES callset by genomic coordinate were labeled as "signal" and the rest as "noise." After the model was trained to distinguish between the two classes and validated, variants in the testing subset were reaggregated per sample. Filtering out variants predicted as "noise," testing per-sample callsets were used to calculate TMB estimates and compared against the respective WES-derived estimates to obtain correlation coefficients and ROC AUC scores.

TABLE 2 | Confusion matrices for the XGBoost binary classifier predictions (TCGA FFPE dataset, according to WES-TMB standards calculated with matched normal WES profiles).

	No WES variant	WES variant	
A. Training dataset	457,276 (TN)	1,050 (FN)	Predicted noise
	60 (FP)	571 (TP)	Predicted signal
B. Testing dataset	373,731 (TN)	1,218 (FN)	Predicted noise
	101 (FP)	98 (TP)	Predicted signal

make key decisions for the trees, the higher is its relative importance.

The top 5 features that obtained the highest scores by the ML model were as follows: *i*) ExAC_nontcga_ALL (allele frequency as observed in the complete ExAC database, excluding only participants from the TCGA project), *ii*) SA_MAP_AF_2 (the maximum likelihood estimate of the allele fraction given no artifacts on either strand), *iii*) INS (whether or not the variant is an insertion), *iv*) TLOD (log-odds that the variant is present in the tumor sample relative to the expected noise), and *v*) SA_MAP_AF_1 (the maximum likelihood estimate of the allele fraction given an artifact on the reverse strand). SA_MAP_AF_1/2 parameters are used for TLOD calculation, which, in turn, together with ExAC allele frequency, were used for rule-based filtering. Thus, ML outperforms rule-based filtering possibly due to more optimal thresholds (or their combinations) for these parameters. Whether the variant is an insertion or not was not included into the rule-based filtering; however, according to the model used, this might be an important parameter for selecting true variants. The latter may be due to possible bias of insertions between true and false variants. Other features considered by the ML model built are shown in **Table S3**.

When applying data filtering with the ML model obtained, we observed on the test subset a significantly better Pearson correlation between TMB-RNAseq and TMB-WES of 0.67, $p = 1.7 \times 10^{-4}$ (**Figure 2C**). Importantly, these results were obtained despite significant sample-to-sample variability in precision and universally low recall (**Figure S3**). The ROC AUC values obtained were also relatively high: 0.825, 0.854, and 0.905 for TMB thresholds of ≥6, ≥10, and ≥20 mutations per megabase, respectively (**Figure 4** and **Table 1**).

Modeling of WES-RNAseq-TMB Correlation on TCGA Dataset Without Matched Normal WES Controls

We then tested whether RNAseq-TMB data are congruent with the WES-TMB data when no matched healthy tissue TBM profile is available. To this end, we took the same $n = 26$ and $n = 27$ subsets of 53 TCGA profiles. The difference was that the matched healthy WES profiles were not used when calculating the "gold standard" WES-TMB values for tumor samples. The data processing workflow was modified accordingly. First, the following specific criteria were added to WES variant filtering: *i*) annotated ExAC ALL frequency <0.000033 and *ii*) no associated dbSNP150 identifier. These filtering rules were chosen to help discern between germline and somatic variants in the absence of paired WES norms.

We observed that algorithmic filtering of RNAseq data resulted in a modest, yet significant correlation between RNAseq-TMB and WES-TMB data: Pearson correlation 0.48, $p = 0.016$ (**Figure 5**). The ROC AUC values obtained were 0.693, 0.763, and 0.857 for TMB thresholds of ≥6, ≥10, and ≥20, respectively (**Figure 6** and **Table 1**).

In turn, using XGBoost ML filtering allowed to increase this correlation up to 0.72 ($p = 5.9 \times 10^{-5}$), while the theoretically deduced maximum possible correlation was 0.94 ($p = 2.6 \times 10^{-12}$, **Figure 5**). In the case of ML filtering, the ROC AUC values increased up to 0.7 and 0.868 for TMB thresholds of ≥6 and ≥10, respectively, while ROC AUC for the threshold of ≥20 did not change (0.857). ML-based prediction for ≥10 TMB threshold was very close to the possible maximum (AUC 0.87 *vs.* 0.92).

Thus, we conclude that using RNAseq-based TMB assessment in one-sample mode can be robust for all three TMB thresholds of ≥6, ≥10, and ≥20, as evidenced by the TCGA quality-paired WES/RNAseq dataset.

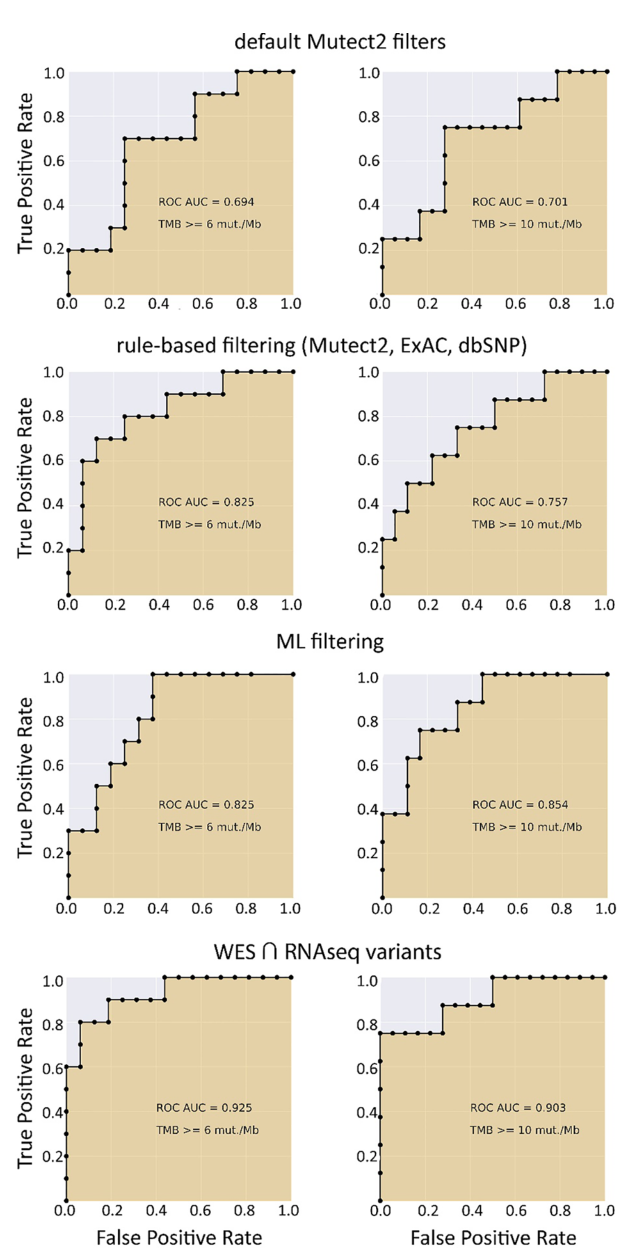

FIGURE 4 | ROC AUC scores for predicting WES-TMB per megabase with RNAseq-TMB per megabase (TCGA FFPE dataset, $n = 26$).

Experimental Evaluation of RNAseq-TMB in Comparison With WES-TMB

The easiest diagnostic solution would ideally comprise analysis of just one tumor sample. We found on the previous step with the TCGA model dataset that RNAseq-TMB estimates can afford for robust discrimination between the high and low TMB groups even without matched healthy controls. We

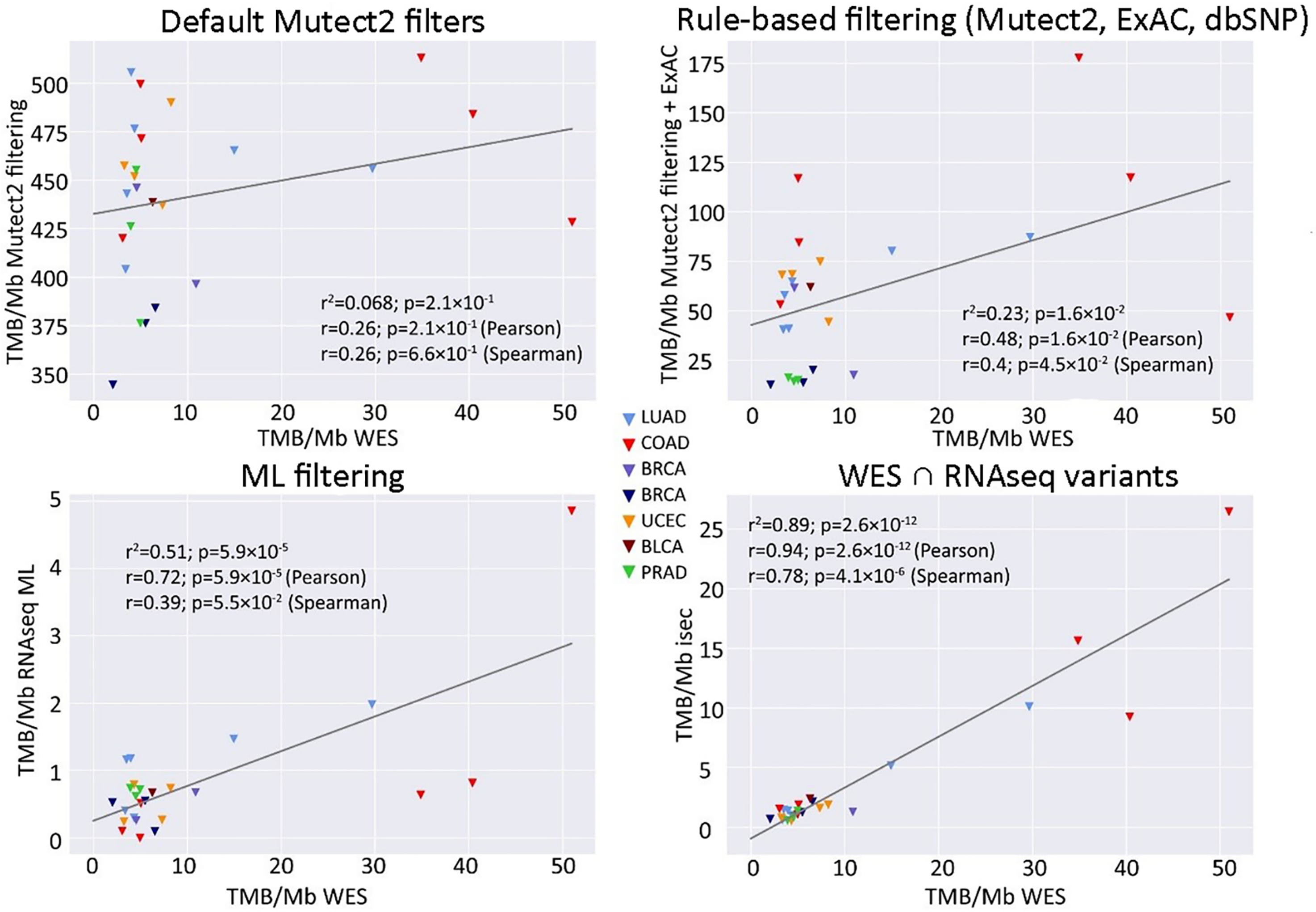

FIGURE 5 | Correlations between RNAseq- and WES-derived TMB estimates (TCGA FFPE dataset, *n* = 26), without matched normal WES references. Samples are colored according to cancer type. Cancer type abbreviations used: LUAD, lung adenocarcinoma; COAD, colon adenocarcinoma; BRCA, breast invasive carcinoma; UCEC, uterine corpus endometrial carcinoma; BLCA, urothelial bladder carcinoma; PRAD, prostate adenocarcinoma.

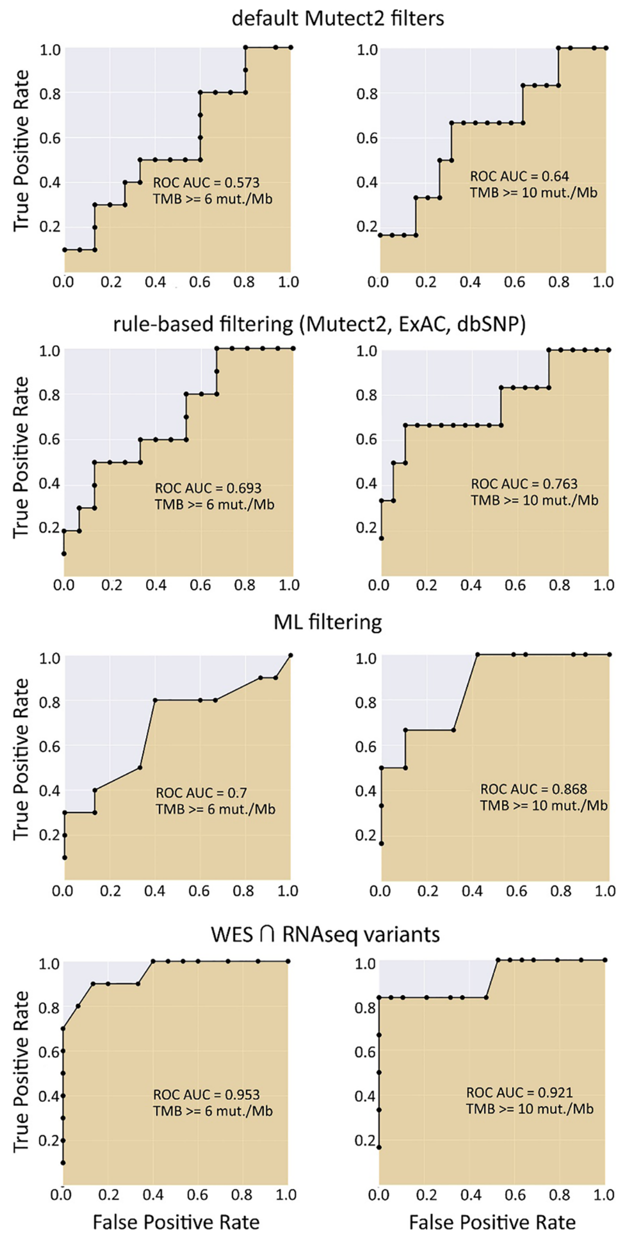

FIGURE 6 | ROC AUC scores for predicting WES-TMB per megabase with RNAseq-TMB per megabase (TCGA FFPE dataset, *n* = 26, no matched norm was used for mutation calling in WES).

hypothesized that RNAseq-TMB estimates may depend on the RNA sequencing coverage. To test this hypothesis, we did experimental paired RNAseq and WES sequencing for 73 FFPE solid tumor tissue biosamples from different cancer patients (**Table S3**). Among them, for 65 samples, we aimed to obtain RNAseq profiles with ~2.5 times lower coverage than in the model TCGA dataset, and for 8 samples, with ~2.5 times higher coverage (**Figure 7**).

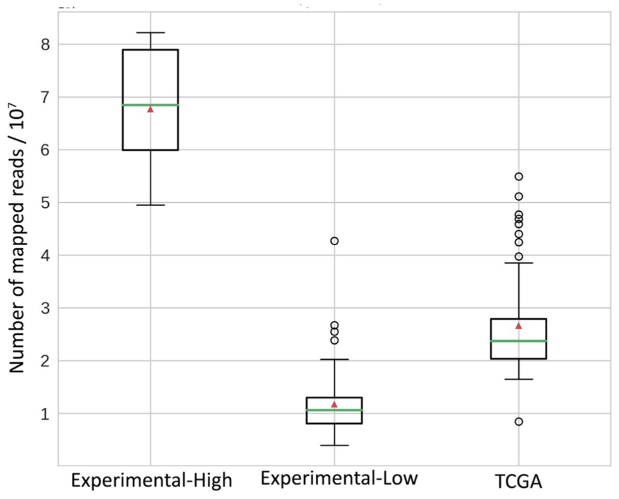

FIGURE 7 | Distribution of mapped reads in RNAseq BAM files. All reads were approximately 50 bases long. Green horizontal lines indicate the median, red triangles indicate the mean.

TMB Evaluation for Experimental RNAseq Data With “Low” Coverage

We then explored the situation when RNAseq read coverage was ~2.5 times lower than that in the model TCGA dataset. In the XGBoost model, we used 34 samples as the training subset and 31 samples as the validation subset. Among the variants common to the WES and RNAseq callsets, we observed a statistically significant Pearson correlation of 0.4, $p = 0.023$. However, when applying the same rule-based filtering as with the TCGA dataset, Pearson correlation dropped to being non-significant: −0.048, $p = 0.79$ (**Figure 8**). The application of the above ML model could not significantly improve RNAseq-TMB performance and resulted in a Pearson correlation as low as 0.16 ($p = 0.38$, **Figure 8**). In all the cases, the ROC AUC scores calculated for this dataset for the thresholds of TMB >6; 10 (no significant fraction of experimental samples with TMB >20 was available, and this threshold was not characterized) were below the AUC quality threshold of 0.7 (**Figure 9** and **Table 1**).

TMB Evaluation for Experimental RNAseq Data With “High” Coverage

In the opposite case (when RNAseq reads coverage significantly exceeded the TCGA dataset), among the variants common to the WES and RNAseq callsets, we observed a statistically significant Pearson correlation of 0.93, $p = 7.1 * 10^{-4}$ (**Figure 6**). We also detected a strong correlation between the experimental RNAseq-TMB and WES-TMB estimates: Pearson correlation 0.86, $p = 0.0056$ (**Figure 10**) with AUC ~1 for TMB per megabase thresholds of >6; 10 (**Figure 11** and **Table 1**); no significant fraction of experimental samples with TMB >20 was available, and this threshold was not characterized.

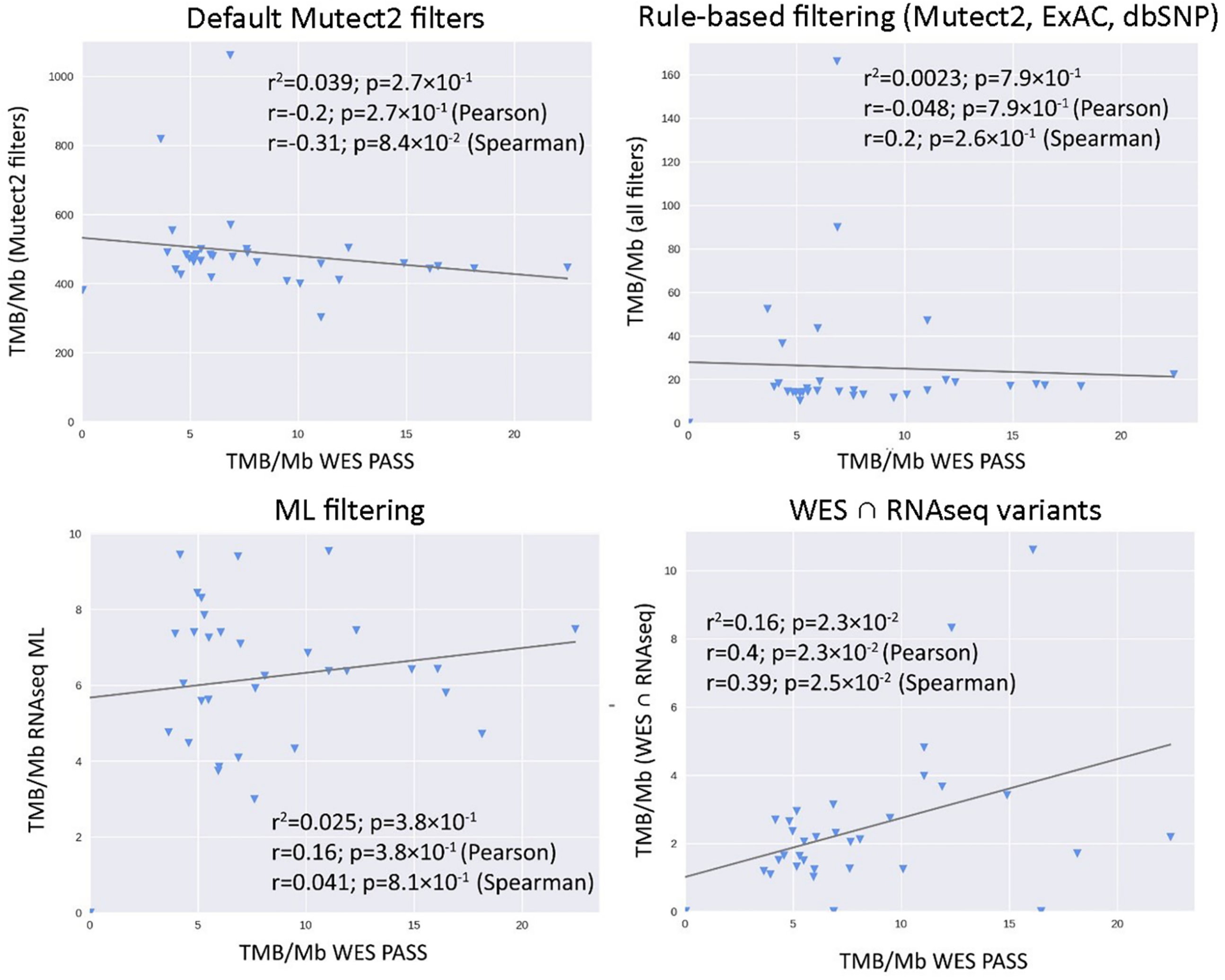

FIGURE 8 | Correlations between RNAseq- and WES-derived tumor mutation burden estimates in experimental low coverage FFPE dataset, *n* = 31.

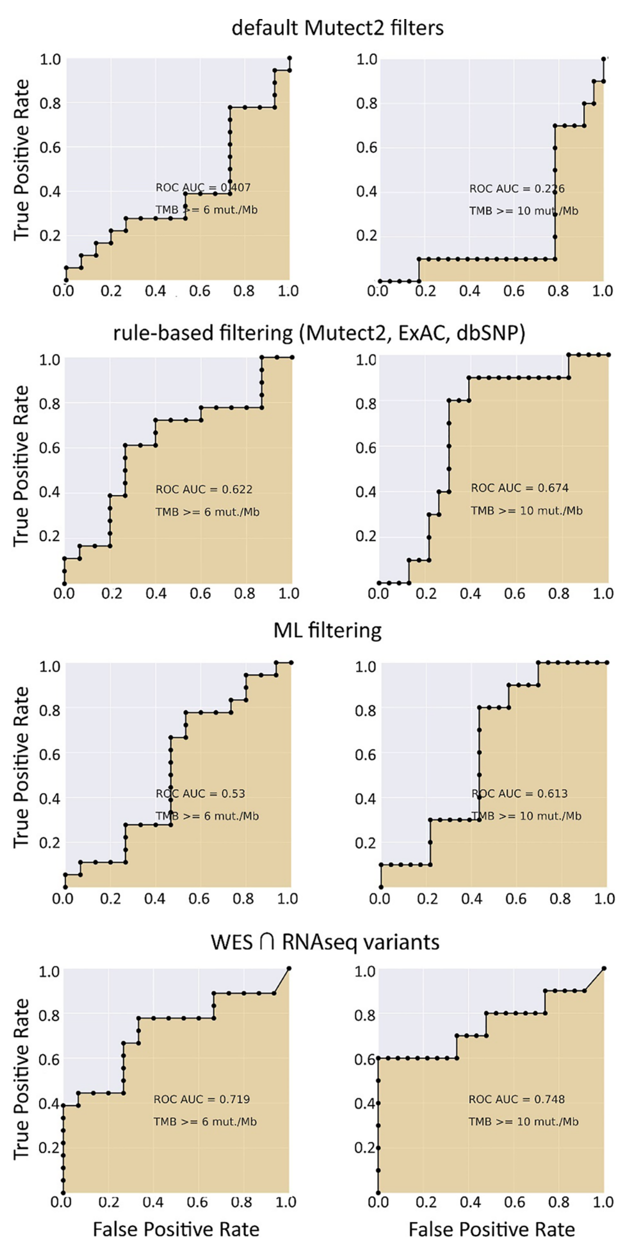

FIGURE 9 | ROC AUC scores for predicting WES-TMB per megabase with RNAseq-TMB per megabase (experimental low coverage transcriptomes, *n* = 31).

Taken together, these results suggest that for the RNAseq datasets with relatively high coverage (~2.5 times higher than in TCGA), the RNAseq-TMB classifier is so strong, that no further ML-assisted improvement is needed to increase its performance.

DISCUSSION

Our findings suggest that RNA sequencing data for FFPE tumor tissue samples can be used to robustly assess TMB levels even in the single-sample mode. As quantified by AUC metric for the high/low TMB classification models, the performance of RNAseq-TMB estimates clearly depends on the sequencing depth. When average RNAseq depth was ~68 MGMRs, the obtained RNAseq-TMB was very well correlated with the "gold standard" WES-TMB, and the performance of the high–low binary TMB classifier was very high (AUC close to 1; **Figure 11**). In the case of mean sequencing depth of ~23 MGMRs as for the model TCGA dataset, using ML algorithms may be needed to improve the classifier robustness, thus giving the classifier AUC of ~0.8–0.9 (**Figure 4**). Finally, when the RNAseq depth is low (as ~11 MGMRs as in the low-coverage experimental dataset), the WES-TMB and RNAseq-TMB correlations are poor, and no good-quality classifier can be built even using the ML approach that used to be successful for the 23-MGMR dataset (**Figure 9**).

Our data also suggest that no healthy tissue control is needed for the FFPE-derived RNAseq data assessment to robustly estimate TMB level when the sequencing depth is sufficient (**Figure 6**).

Furthermore, WES-TMB data obtained for tumor biosamples with healthy controls correlated well with the data for the same biosamples without healthy controls (**Figure S4**). This is in line with the broad clinical practice of using targeted NGS TMB panels like FoundationOne CDx assay that does not require a healthy norm to estimate TMB (2).

Gene expression profiling by RNAseq is the alternative type of high-throughput genetic analysis that can inform clinical decision-making (51, 52). It was recently published that RNAseq data can serve as the alternative to immunohistochemical tests for several major cancer markers like HER2, ESR1, PGR, and PD-L1 (22). It can reliably estimate concentrations of cancer drug targets (53), which is also true for the emerging non-protein molecular target ganglioside GD2 (27). In addition, RNAseq data obtained for FFPE biosamples may be used to identify clinically actionable or new fusion oncogenes (26) and to generate gene signatures that can establish statuses of important tumor biomarkers like microsatellite instability (25, 54, 55) and oncogenic mutations (56) or that can predict individual sensitivity of a tumor to targeted (28, 57, 58) and non-targeted (59) therapies. Thus, adding a new option of TMB level assessment may strongly benefit this approach.

As RNAseq focuses exclusively on transcribed allele sequences, TMB calculated from such data, or another possible alternative—frequency of neoantigens, might theoretically even surpass the predictive power of WES, WGS, or target panel DNAseq with respect to the efficiency of checkpoint inhibitor immunotherapy.

Here, we provide the first experimental assessment of TMB quality for FFPE-derived RNAseq data obtained with different coverage in tumor sample-only mode. Although the results obtained are quite encouraging, the practical implementation of this technology and more detailed clinical guidelines should be a matter of further investigations with greater patient cohorts and (possibly) more specifically selected cancer types. In addition, sufficient sequencing depth threshold should be established in future studies comparing the correlation

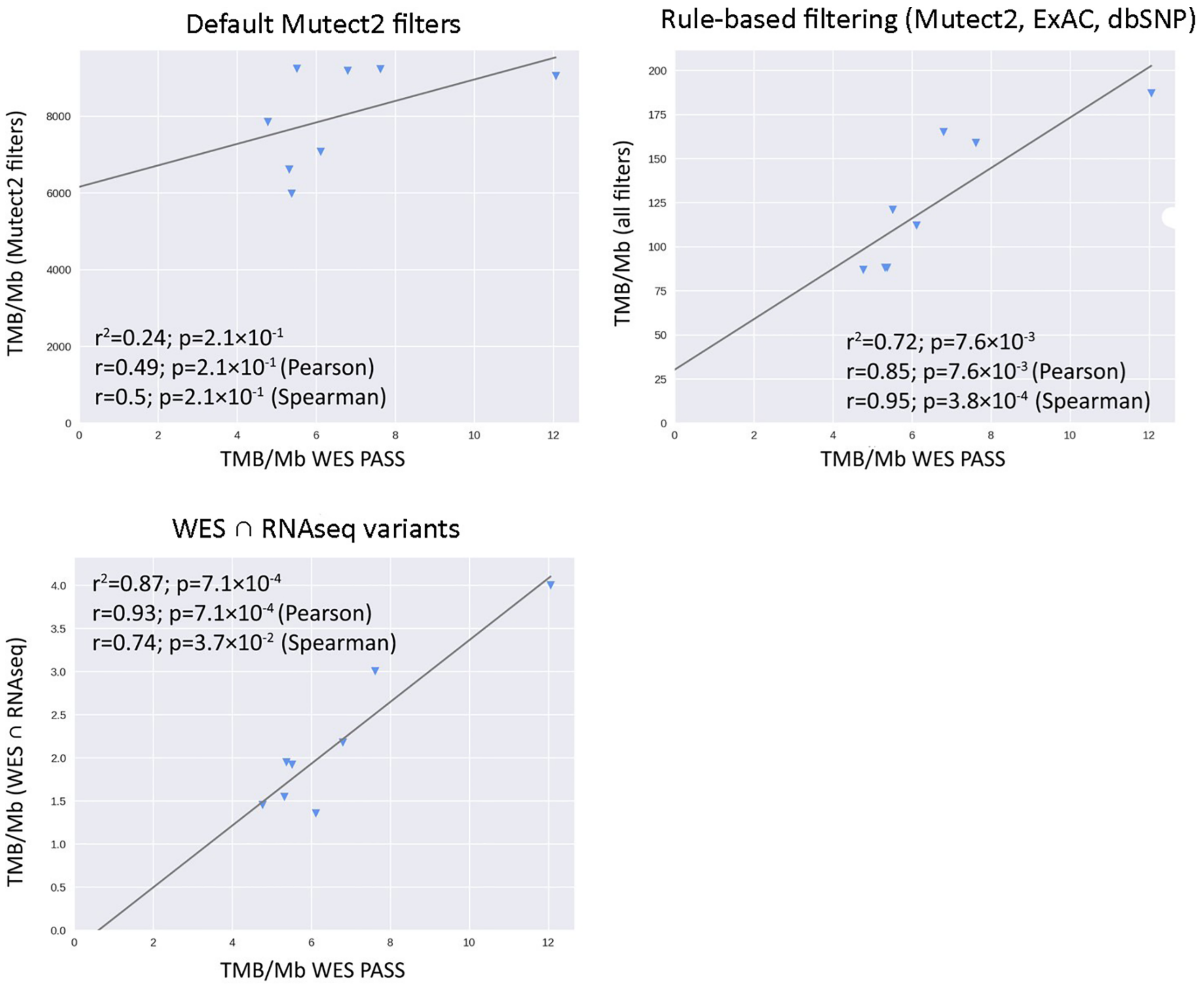

FIGURE 10 | Correlations between RNAseq- and WES-derived TMB estimates in experimental high coverage RNAseq dataset, n = 8.

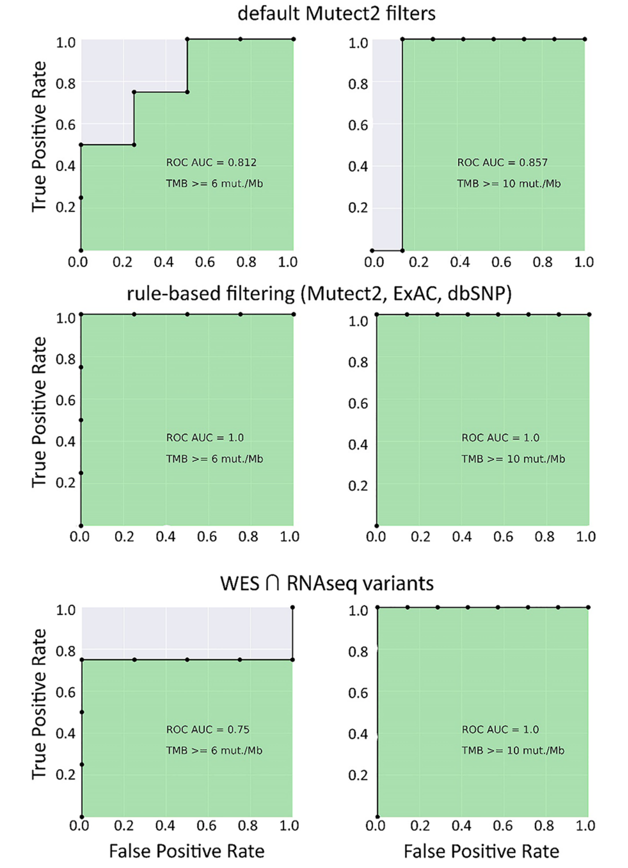

FIGURE 11 | ROC AUC scores for predicting WES-TMB per megabase with RNAseq-TMB per megabase (experimental high coverage transcriptomes, *n* = 8).

between DNA- and RNAseq-derived TMB in groups of samples with different mean coverage.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The original sequencing data were deposited to NCBI SRA with accession number PRJNA663280 allowing a free access.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Karelia Republic Oncological Hospital, Petrozavodsk, Russia, and Vitamed Oncological Clinical Center, Moscow, Russia. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

AB, DKu, and MSo contributed to conception and design of the study. VE, EZ, and AG organized the database and performed the ML experiments. ER, MSu, XL, and DKa organized and performed molecular analyses. MZ, VE, and AG performed the statistical analysis. MSo, EP, and AS provided the biomaterial and clinical history of the patients. AG and AB wrote the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This study was supported by the Russian Science Fund grant 20-75-10071.

ACKNOWLEDGMENTS

We acknowledge the Amazon and Microsoft Azure support with cloud-based computations and OmicsWay Corp. for providing access to Oncobox software and cloud computing infrastructure.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2021.732644/full#supplementary-material

Supplementary Figure 1 | Principal component analysis (TCGA FFPE dataset). Color shows IDs of TCGA sequencing center that generated the corresponding RNAseq profile.

Supplementary Figure 2 | XGBoost binary classifier feature importance scores (gain). 31 variant features were selected to train the model. 23 features are variant attributes, assigned by Mutect2. Exac_nontcga_ALL was obtained from ANNOVAR output. 7 features were engineered: Boolean features of the variant being an insertion (INS), a deletion (DEL), a transition (CT_GA) or a transversion (CA_GT)—and integer features: total depth (TOTAL_DP), as well as REF and ALT lengths (len_REF and len_ALT, respectively). See text and **Table S1** for a more detailed explanation.

Supplementary Figure 3 | Precision and recall metrics for the XGBoost binary classifier predictions on the TCGA FFPE test subset, based on WES-TMB data calculated with matched normal WES profiles.

Supplementary Figure 4 | TMB calculated for TCGA WXS samples called with matched normal data (x-axis) *vs* TMB calculated for TCGA WXS samples called without matched normal data (y-axis).

Supplementary Table 1 | TCGA FFPE biosamples used for training and validation of XGBoost model for filtering RNAseq variants.

Supplementary Table 2 | Clinical characteristics and coverage information for experimental samples.

Supplementary Table 3 | Features used in training the XGBoost classifier for filtering RNAseq variant calls.

REFERENCES

1. Zhu J, Zhang T, Li J, Lin J, Liang W, Huang W, et al. Association Between Tumor Mutation Burden (TMB) and Outcomes of Cancer Patients Treated With PD-1/PD-L1 Inhibitions: A Meta-Analysis. *Front Pharmacol* (2019) 10:673. doi: 10.3389/fphar.2019.00673
2. Wang Y, Tong Z, Zhang W, Buzdin AA, Mu X, Yan Q, et al. FDA-Approved and Emerging Next Generation Predictive Biomarkers for Immune Checkpoint Inhibitors in Cancer Patients. *Front Oncol* (2021) 11:683419. doi: 10.3389/FONC.2021.683419
3. Chalmers ZR, Connelly CF, Fabrizio D, Gay L, Ali SM, Ennis R, et al. Analysis of 100,000 Human Cancer Genomes Reveals the Landscape of Tumor Mutational Burden. *Genome Med* (2017) 9:34. doi: 10.1186/s13073-017-0424-2
4. Fancello L, Gandini S, Pelicci PG, Mazzarella L. Tumor Mutational Burden Quantification From Targeted Gene Panels: Major Advancements and Challenges. *J Immunother Cancer* (2019) 7:183. doi: 10.1186/s40425-019-0647-4
5. Büttner R, Longshore JW, López-Ríos F, Merkelbach-Bruse S, Normanno N, Rouleau E, et al. Implementing TMB Measurement in Clinical Practice: Considerations on Assay Requirements. *ESMO Open* (2019) 4:e000442. doi: 10.1136/esmoopen-2018-000442
6. Snyder A, Makarov V, Merghoub T, Yuan J, Zaretsky JM, Desrichard A, et al. Genetic Basis for Clinical Response to CTLA-4 Blockade in Melanoma. *N Engl J Med* (2014) 371:2189–99. doi: 10.1056/nejmoa1406498
7. Rizvi NA, Hellmann MD, Snyder A, Kvistborg P, Makarov V, Havel JJ, et al. Mutational Landscape Determines Sensitivity to PD-1 Blockade in non-Small Cell Lung Cancer. *Sci (80)* (2015) 348:124–8. doi: 10.1126/science.aaa1348
8. Riaz N, Havel JJ, Makarov V, Desrichard A, Urba WJ, Sims JS, et al. Tumor and Microenvironment Evolution During Immunotherapy With Nivolumab. *Cell* (2017) 171:934–49.e15. doi: 10.1016/j.cell.2017.09.028
9. Eroglu Z, Zaretsky JM, Hu-Lieskovan S, Kim DW, Algazi A, Johnson DB, et al. High Response Rate to PD-1 Blockade in Desmoplastic Melanomas. *Nature* (2018) 553:347–50. doi: 10.1038/nature25187
10. Klempner SJ, Fabrizio D, Bane S, Reinhart M, Peoples T, Ali SM, et al. Tumor Mutational Burden as a Predictive Biomarker for Response to Immune Checkpoint Inhibitors: A Review of Current Evidence. *Oncologist* (2020) 25: e147–59. doi: 10.1634/theoncologist.2019-0244
11. Goodman AM, Kato S, Bazhenova L, Patel SP, Frampton GM, Miller V, et al. Tumor Mutational Burden as an Independent Predictor of Response to Immunotherapy in Diverse Cancers. *Mol Cancer Ther* (2017) 16:2598–608. doi: 10.1158/1535-7163.MCT-17-0386
12. Rizvi H, Sanchez-Vega F, La K, Chatila W, Jonsson P, Halpenny D, et al. Molecular Determinants of Response to Anti-Programmed Cell Death (PD)-1 and Anti-Programmed Death-Ligand 1 (PD-L1) Blockade in Patients With non-Small-Cell Lung Cancer Profiled With Targeted Next-Generation Sequencing. *J Clin Oncol* (2018) 36:633–41. doi: 10.1200/JCO.2017.75.3384
13. Carbone DP, Reck M, Paz-Ares L, Creelan B, Horn L, Steins M, et al. First-Line Nivolumab in Stage IV or Recurrent Non–Small-Cell Lung Cancer. *N Engl J Med* (2017) 376:2415–26. doi: 10.1056/nejmoa1613493
14. Buchhalter I, Rempel E, Endris V, Allgäuer M, Neumann O, Volckmar AL, et al. Size Matters: Dissecting Key Parameters for Panel-Based Tumor Mutational Burden Analysis. *Int J Cancer* (2019) 144:848–58. doi: 10.1002/ijc.31878
15. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV, et al. Signatures of Mutational Processes in Human Cancer. *Nature* (2013) 500:415–21. doi: 10.1038/nature12477
16. Stenzinger A, Allen JD, Maas J, Stewart MD, Merino DM, Wempe MM, et al. Tumor Mutational Burden Standardization Initiatives: Recommendations for Consistent Tumor Mutational Burden Assessment in Clinical Samples to Guide Immunotherapy Treatment Decisions. *Genes Chromosom Cancer* (2019) 58:578–88. doi: 10.1002/gcc.22733
17. Jang BS, Han W, Kim IA. Tumor Mutation Burden, Immune Checkpoint Crosstalk and Radiosensitivity in Single-Cell RNA Sequencing Data of Breast Cancer. *Radiother Oncol* (2020) 142:202–9. doi: 10.1016/j.radonc.2019.11.003
18. DiGuardo MA, Davila JI, Jackson RA, Nair AA, Fadra N, Minn KT, et al. RNA-Seq Reveals Differences in Expressed Tumor Mutation Burden in Colorectal and Endometrial Cancers With and Without Defective DNA-Mismatch Repair. *J Mol Diagnostics* (2021) 23:555–64. doi: 10.1016/j.jmoldx.2021.01.008
19. Schmeller J, Wessolly M, Mairinger E, Borchert S, Hager T, Mairinger T, et al. Setting Out the Frame Conditions for Feasible Use of FFPE Derived RNA. *Pathol Res Pract* (2019) 215:381–6. doi: 10.1016/j.prp.2018.12.027
20. Chaudhary R, Quagliata L, Martin JP, Alborelli I, Cyanam D, Mittal V, et al. A Scalable Solution for Tumor Mutational Burden From Formalin-Fixed, Paraffin-Embedded Samples Using the Oncomine Tumor Mutation Load Assay. *Transl Lung Cancer Res* (2018) 7:616–30. doi: 10.21037/tlcr.2018.08.01
21. Buzdin A, Sorokin M, Garazha A, Glusker A, Aleshin A, Poddubskaya E, et al. RNA Sequencing for Research and Diagnostics in Clinical Oncology. *Semin Cancer Biol* (2019) 60:311–23. doi: 10.1016/j.semcancer.2019.07.010
22. Sorokin M, Ignatev K, Poddubskaya E, Vladimirova U, Gaifullin N, Lantsov D, et al. RNA Sequencing in Comparison to Immunohistochemistry for Measuring Cancer Biomarkers in Breast Cancer and Lung Cancer Specimens. *Biomedicines* (2020) 8:114. doi: 10.3390/BIOMEDICINES8050114
23. Suntsova M, Gaifullin N, Allina D, Reshetun A, Li X, Mendeleeva L, et al. Atlas of RNA Sequencing Profiles for Normal Human Tissues. *Sci Data* (2019) 6:36. doi: 10.1038/s41597-019-0043-4
24. Borisov N, Sergeeva A, Suntsova M, Raevskiy M, Gaifullin N, Mendeleeva L, et al. Machine Learning Applicability for Classification of PAD/VCD Chemotherapy Response Using 53 Multiple Myeloma RNA Sequencing Profiles. *Front Oncol* (2021) 11:1124. doi: 10.3389/fonc.2021.652063
25. Adamyan L, Aznaurova Y, Stepanian A, Nikitin D, Garazha A, Suntsova M, et al. Gene Expression Signature of Endometrial Samples From Women With and Without Endometriosis. *J Minim Invasive Gynecol* (2021). doi: 10.1016/j.jmig.2021.03.011
26. Vladimirova U, Rumiantsev P, Zolotovskaia M, Albert E, Abrosimov A, Slashchuk K, et al. DNA Repair Pathway Activation Features in Follicular and Papillary Thyroid Tumors, Interrogated Using 95 Experimental RNA Sequencing Profiles. *Heliyon* (2021) 7:e06408. doi: 10.1016/j.heliyon.2021.e06408
27. Sorokin M, Kholodenko I, Kalinovsky D, Shamanskaya T, Doronin I, Konovalov D, et al. RNA Sequencing-Based Identification of Ganglioside GD2-Positive Cancer Phenotype. *Biomedicines* (2020) 8:142. doi: 10.3390/BIOMEDICINES8060142
28. Sorokin M, Poddubskaya E, Baranova M, Glusker A, Kogoniya L, Markarova E, et al. RNA Sequencing Profiles and Diagnostic Signatures Linked With Response to Ramucirumab in Gastric Cancer. *Cold Spring Harb Mol Case Stud* (2020) 6:a004945. doi: 10.1101/mcs.a004945
29. Moisseev A, Albert E, Lubarsky D, Schroeder D, Clark J. Transcriptomic and Genomic Testing to Guide Individualized Treatment in Chemoresistant Gastric Cancer Case. *Biomedicines* (2020) 8:67. doi: 10.3390/biomedicines8030067
30. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, et al. From FastQ Data to High Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Curr Protoc Bioinforma* (2013) 43:11.10.1–33. doi: 10.1002/0471250953.bi1110s43
31. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: Ultrafast Universal RNA-Seq Aligner. *Bioinformatics* (2013) 29:15–21. doi: 10.1093/bioinformatics/bts635
32. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map Format and SAMtools. *Bioinformatics* (2009) 25:2078–9. doi: 10.1093/bioinformatics/btp352
33. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, et al. Sensitive Detection of Somatic Point Mutations in Impure and Heterogeneous Cancer Samples. *Nat Biotechnol* (2013) 31:213–9. doi: 10.1038/nbt.2514
34. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: The NCBI Database of Genetic Variation. *Nucleic Acids Res* (2001) 29:308–11. doi: 10.1093/nar/29.1.308
35. Mills RE, Luttig CT, Larkins CE, Beauchamp A, Tsui C, Pittard WS, et al. An Initial Map of Insertion and Deletion (INDEL) Variation in the Human Genome. *Genome Res* (2006) 16:1182–90. doi: 10.1101/gr.4565806
36. Frankish A, Diekhans M, Ferreira AM, Johnson R, Jungreis I, Loveland J, et al. GENCODE Reference Annotation for the Human and Mouse Genomes. *Nucleic Acids Res* (2019) 47:D766–73. doi: 10.1093/nar/gky955
37. Wang K, Li M, Hakonarson H. ANNOVAR: Functional Annotation of Genetic Variants From High-Throughput Sequencing Data. *Nucleic Acids Res* (2010) 38:e164–4. doi: 10.1093/nar/gkq603

38. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of Protein-Coding Genetic Variation in 60,706 Humans. *Nature* (2016) 536:285–91. doi: 10.1038/nature19057
39. Li H, Durbin R. Fast and Accurate Long-Read Alignment With Burrows–Wheeler Transform. *Bioinformatics* (2010) 26:589–95. doi: 10.1093/bioinformatics/btp698
40. Tange O. *Gnu Parallel-the Command-Line Power Tool* (2011). Available at: usenix.org.
41. Kamashev D, Sorokin M, Kochergina I, Drobyshev A, Vladimirova U, Zolotovskaia M, et al. Human Blood Serum can Donor-Specifically Antagonize Effects of EGFR-Targeted Drugs on Squamous Carcinoma Cell Growth. *Heliyon* (2011) 7:E06394. doi: 10.1016/j.heliyon.2021.e06394
42. Liu T, Cheng G, Kang X, Xi Y, Zhu Y, Wang K, et al. Noninvasively Evaluating the Grading and IDH1 Mutation Status of Diffuse Gliomas by Three-Dimensional Pseudo-Continuous Arterial Spin Labeling and Diffusion-Weighted Imaging. *Neuroradiology* (2018) 60:693–702. doi: 10.1007/s00234-018-2021-5
43. Tanioka M, Fan C, Parker JS, Hoadley KA, Hu Z, Li Y, et al. Integrated Analysis of RNA and DNA From the Phase III Trial CALGB 40601 Identifies Predictors of Response to Trastuzumab-Based Neoadjuvant Chemotherapy in HER2-Positive Breast Cancer. *Clin Cancer Res* (2018) 24:5292–304. doi: 10.1158/1078-0432.CCR-17-3431
44. Borisov N, Sorokin M, Tkachev V, Garazha A, Buzdin A. Cancer Gene Expression Profiles Associated With Clinical Outcomes to Chemotherapy Treatments. *BMC Med Genomics* (2020) 13:111. doi: 10.1186/s12920-020-00759-0
45. Boyd JC. Mathematical Tools for Demonstrating the Clinical Usefulness of Biochemical Markers. *Scand J Clin Lab Invest Suppl* (1997) 227:46–63. doi: 10.1080/00365519709168308
46. Schumacher TN, Schreiber RD. Neoantigens in Cancer Immunotherapy. *Sci (80)* (2015) 348:69–74. doi: 10.1126/science.aaa4971
47. Shao C, Li G, Huang L, Pruitt S, Castellanos E, Frampton G, et al. Prevalence of High Tumor Mutational Burden and Association With Survival in Patients With Less Common Solid Tumors. *JAMA Netw Open* (2020) 3:2025109. doi: 10.1001/jamanetworkopen.2020.25109
48. Marabelle A, Fakih M, Lopez J, Shah M, Shapira-Frommer R, Nakagawa K, et al. Association of Tumour Mutational Burden With Outcomes in Patients With Advanced Solid Tumours Treated With Pembrolizumab: Prospective Biomarker Analysis of the Multicohort, Open-Label, Phase 2 KEYNOTE-158 Study. *Lancet Oncol* (2020) 21:1353–65. doi: 10.1016/S1470-2045(20)30445-9
49. Chan TA, Yarchoan M, Jaffee E, Swanton C, Quezada SA, Stenzinger A, et al. Development of Tumor Mutation Burden as an Immunotherapy Biomarker: Utility for the Oncology Clinic. *Ann Oncol* (2019) 30:44–56. doi: 10.1093/annonc/mdy495
50. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: *Proc 22nd ACM SIGKDD Int Conf Knowl Discov Data Min - KDD '16*. Association for Computing Machinery (2016). p. 785–94. doi: 10.1145/2939672.2939785
51. Buzdin A, Sorokin M, Garazha A, Glusker A, Aleshin A, Poddubskaya E, et al. RNA Sequencing for Research and Diagnostics in Clinical Oncology. *Semin Cancer Biol* (2020) 60:311–23. doi: 10.1016/j.semcancer.2019.07.010
52. Buzdin A, Skvortsova II, Li X, Wang Y. Editorial: Next Generation Sequencing Based Diagnostic Approaches in Clinical Oncology. *Front Oncol* (2021) 10:3276. doi: 10.3389/fonc.2020.635555
53. Tkachev V, Sorokin M, Garazha A, Borisov N, Buzdin A. "Oncobox Method for Scoring Efficiencies of Anticancer Drugs Based on Gene Expression Data". In: Astakhova K, Bukhari SA, editors. *Methods Mol Biol*. New York, NY: Springer US (2020). p. 235–55. doi: 10.1007/978-1-0716-0138-9_17
54. Fu Y, Qi L, Guo W, Jin L, Song K, You T, et al. A Qualitative Transcriptional Signature for Predicting Microsatellite Instability Status of Right-Sided Colon Cancer. *BMC Genomics* (2019) 20:769. doi: 10.1186/s12864-019-6129-8
55. Wang H, Wang X, Xu L, Zhang J, Cao H. Analysis of the Transcriptomic Features of Microsatellite Instability Subtype Colon Cancer. *BMC Cancer* (2019) 19:605. doi: 10.1186/s12885-019-5802-2
56. Petrov I, Suntsova M, Ilnitskaya E, Roumiantsev S, Sorokin M, Garazha A, et al. Gene Expression and Molecular Pathway Activation Signatures of MYCN-Amplified Neuroblastomas. *Oncotarget* (2017) 8:83768–80. doi: 10.18632/oncotarget.19662
57. Zhu Q, Izumchenko E, Aliper AM, Makarev E, Paz K, Buzdin AA, et al. Pathway Activation Strength Is a Novel Independent Prognostic Biomarker for Cetuximab Sensitivity in Colorectal Cancer Patients. *Hum Genome Var* (2015) 2:15009. doi: 10.1038/hgv.2015.9
58. Rodon J, Soria JC, Berger R, Miller WH, Rubin E, Kugel A, et al. Genomic and Transcriptomic Profiling Expands Precision Cancer Medicine: The WINTHER Trial. *Nat Med* (2019) 25:751–8. doi: 10.1038/s41591-019-0424-4
59. Wheler JJ, Parker BA, Lee JJ, Atkins JT, Janku F, Tsimberidou AM, et al. Unique Molecular Signatures as a Hallmark of Patients With Metastatic Breast Cancer: Implications for Current Treatment Paradigms. *Oncotarget* (2014) 5:2349–54. doi: 10.18632/oncotarget.1946

Conflict of Interest: Authors MSo, AG and AB were employed by OmicsWay Corp.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.