nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-022-30753-2 OPEN

Check for updates

Estimating tumor mutational burden from RNA-sequencing without a matched-normal sample

Rotem Katzir[1,2], Noam Rudberg[2] & Keren Yizhak[2]✉

Detection of somatic mutations using patients sequencing data has many clinical applications, including the identification of cancer driver genes, detection of mutational signatures, and estimation of tumor mutational burden (TMB). We have previously developed a tool for detection of somatic mutations using tumor RNA and a matched-normal DNA. Here, we further extend it to detect somatic mutations from RNA sequencing data without a matched-normal sample. This is accomplished via a machine-learning approach that classifies mutations as either somatic or germline based on various features. When applied to RNA-sequencing of >450 melanoma samples high precision and recall are achieved, and both mutational signatures and driver genes are correctly identified. Finally, we show that RNA-based TMB is significantly associated with patient survival, showing similar or higher significance level as compared to DNA-based TMB. Our pipeline can be utilized in many future applications, analyzing novel and existing datasets where only RNA is available.

[1] Center for Bioinformatics and Computational Biology, Department of Computer Science and the University of Maryland Institute of Advanced Computer Studies (UMIACS), University of Maryland, College Park, MD 20742, USA. [2] Department of Cell Biology and Cancer Science, Rappaport Faculty of Medicine, Technion-Israel Institute of Technology, Haifa 31096, Israel. ✉email: kyizhak@technion.ac.il

Somatic point mutations accumulate in the DNA of all dividing cells, both normal and neoplastic, and are the most common mechanism for altering gene function[1–4]. Their detection in tumor samples is of high clinical value; first, when accumulated in specific genes termed "drivers", they may lead to the development of cancer. Identifying these mutations is therefore crucial for matching existing targeted therapies and for developing novel ones[5–8]. In addition, somatic mutations are used to determine intra-tumor heterogeneity which is a major mechanism of therapeutic resistance[9], and for identifying mutational signatures that have proven as clinically useful biomarkers[10,11]. More recently, the set of somatic mutations in a tumor has been used to estimate the tumor mutational burden (TMB), an emerging proxy for neoantigen load. TMB is usually defined as the number of non-silent mutations found in a tumor DNA, and was found to be an independent marker of patient response to immune checkpoint inhibitor therapy (ICI), and for predicting patient survival, both in treated and treatment-naive patients[12–18].

Traditionally, detection of somatic point mutations is done using whole exome or genome sequencing of tumor and matched-normal DNA[19–23]. The latter is required for distinguishing between somatic mutations found exclusively in the tumor sample, and germline variants shared by all cells of an individual. Recently, several studies have developed a 'tumor-only' pipeline that uses tumor DNA sequencing to detect somatic mutations without a matched-normal sample, at the cost of lower precision and recall levels[24–26]. An additional extension to these pipelines includes the detection of somatic mutations from RNA sequencing and a matched-normal DNA sample. In a recent publication, we have introduced such a pipeline termed RNA-MuTect, and showed that most of the mutations detected only in the RNA are filtered out by our pipeline, achieving an overall high precision. In addition, high sensitivity for mutations with sufficient detection power was observed, enabling the detection of most driver genes and mutational signatures[27].

In this study, we take our RNA-based approach one step further and develop a pipeline for detecting somatic point mutations from RNA sequencing without a matched-normal sample, named RNA-MuTect-WMN (WMN; without-matched-normal). This is accomplished via a machine learning framework which utilizes a few dozens of features to classify single nucleotide variants as either somatic or germline. Our pipeline is trained and tested on the TCGA melanoma, lung and colon dataset where it achieves high precision and recall. High performance is also achieved in two additional cancer types, lung adenocarcinoma, and colon cancer. In addition, it enables a reliable identification of both driver genes and mutational signatures across the different cancer types. When applied to estimate the TMB from RNA samples alone, we find that its performance is either equivalent or superior to TMB estimated based on DNA with a matched-normal sample. The ability to estimate the TMB using a tumor RNA alone further emphasizes the potential clinical utility of our pipeline.

Results

Identifying somatic mutations from RNA-seq data without a matched-normal sample. To develop a pipeline for detection of somatic point mutations from RNA-seq without a matched-normal sample, we leveraged RNA-seq and matched-normal DNA of 462 melanoma samples from The Cancer Genome Atlas (TCGA)[28]. To obtain the ground truth of somatic and germline variants in these samples, we ran RNA-MuTect[27]; in short, RNA-MuTect works by first running MuTect[19] on tumor RNA and matched-normal DNA, to identify the set of tumor somatic mutations and the set of potential germline variants (Methods). Since this set includes multiple noisy sites unique to RNA, a series of filtering steps is then applied to yield the final set of true somatic mutations (Fig. 1a). To examine the accuracy of RNA-MuTect on this set of samples, we compared the list of somatic mutations to that obtained using tumor and matched normal DNA. As originally reported[27], focusing on the RNA mutations with sufficient detection power in the DNA, 90% were indeed found in the DNA, with a median of only 3 detected mutations per sample found in the RNA alone (Methods).

To classify point mutations as either somatic or germline, we collected a set of genomic features for each variant (Methods). This list includes the number of reference and alternate reads, variant classification type and MuTect likelihood score. In addition, we collected data on germline variants from dbSNP[29], gnomAD[30], 1000 genomes[31] and the Exome Sequencing project[32]. Finally, we utilized both DNA and RNA panel of normal (PoN) which are based on ~8000 TCGA and ~6500 Genotype-Tissue Expression (GTEx) normal samples (Methods)[33]. These PoNs encode the distribution of alternate read counts across the entire sets of normal samples[34].

To test how well our features separate between somatic and germline variants, we performed a two-sided Wilcoxon rank sum test for each feature, and found that all features show a significant difference between these two types of variants (FDR corrected p-values $<= 0.0111$, Supplementary Fig. 1). However, when searching across a range of thresholds in each feature, we found that the Precision-Recall Area Under the Curve (PR-AUC) is very low (<0.08, Supplementary Fig. 2), as well as the F1-score (<0.16, Supplementary Fig. 3). This finding is a result of the substantial overlap between features' values in these two variant types, demonstrating the need for a more complex model.

To this end, we developed a machine learning framework named RNA-MuTect-WMN that gets as input a list of variants with their associated features, and classifies them as either somatic or germline. Specifically, our data is first randomly split into training ($n = 100$) and test sets ($n = 362$). In the training set, each sample contains a list of single nucleotide variants with their genomic features (Methods), and a somatic\germline label based on the RNA-MuTect pipeline, as described above (Fig. 1a). Next, a set of random forest classifiers is trained[35] in a 5-fold cross-validation manner, such that in each iteration 80 samples are used for training and 20 samples are used for validation. We then aggregated our models' predictions over all folds and computed the precision and recall for each sample in the validation sets. The median precision and recall values obtained were 0.82 and 0.83, respectively (mean precision and recall of 0.78 and 0.83, respectively, median F1-score = 0.82, mean F1-score = 0.79, Fig. 1b). To examine our model performance we used the test set of 362 samples and applied the following three steps: (1) we ran MuTect with tumor RNA-seq and *without* a matched-normal sample. In this step, both somatic and germline variants are marked as true somatic mutations, and a subset of sites are removed based on MuTect filtering scheme; (2) we then applied the 5 models built in the training step, and classified each variant as either somatic or germline based on the majority vote; (3) finally, to remove any remaining RNA-specific noise, we applied the RNA-MuTect filtering steps and achieved the final set of predicted somatic mutations. We have decided to run RNA-MuTect filtering steps on the narrowed list achieved after step 2 instead of upfront at step 1, due to a couple of time-consuming steps implemented in that pipeline (realignment and PoN filtering steps[27]). Those could have significantly slowed down the process. The final set of somatic and germline variants was then used to estimate the pipeline's performance, showing a sample-level median precision and recall of 0.85 and 0.83, respectively (mean precision and recall of 0.8 and 0.83,

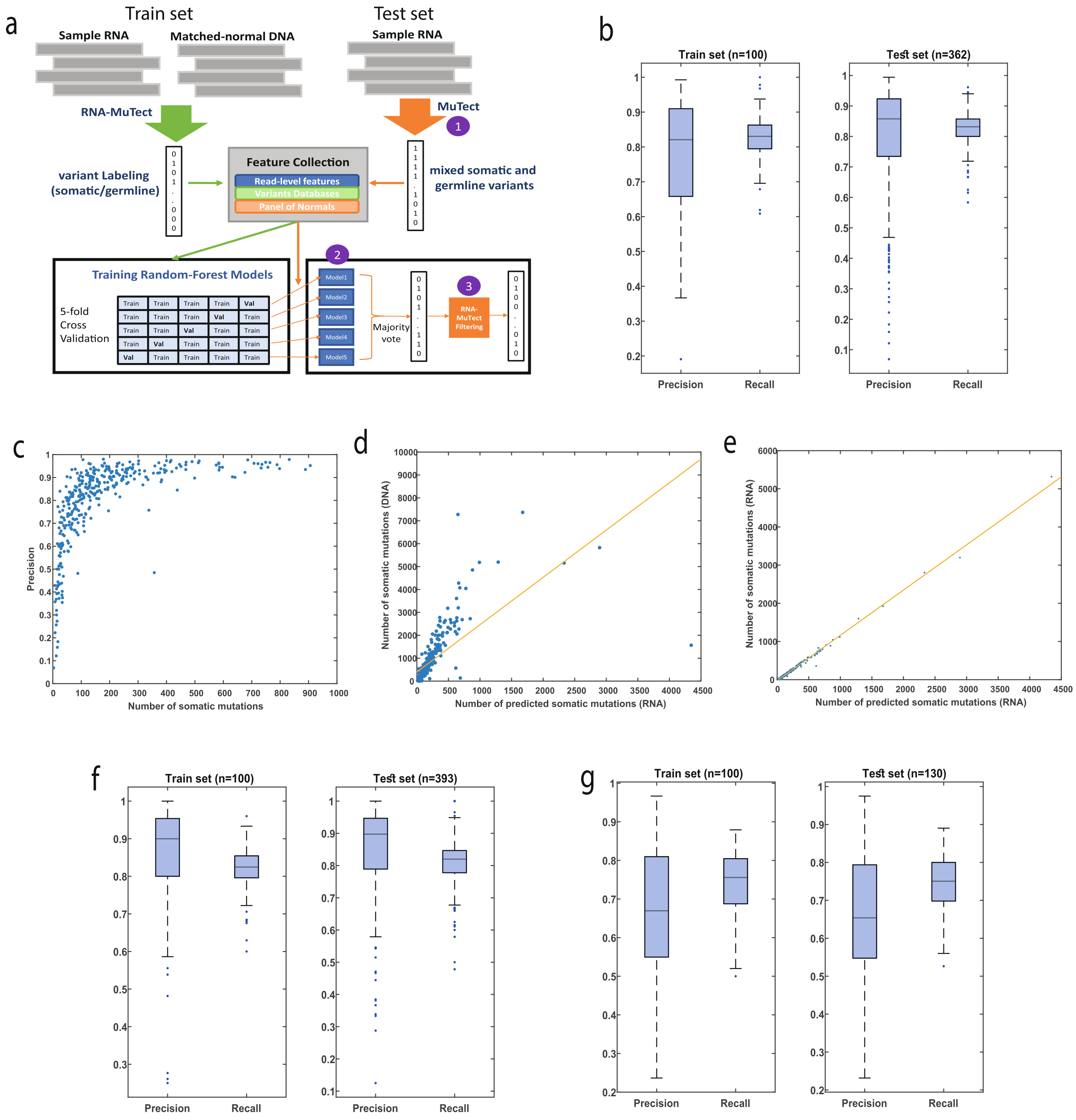

Fig. 1 Summary of pipeline predictions. a An overview of the RNA-MuTect-WMN pipeline: In the training set ($n = 100$, green arrows), RNA-MuTect is applied on tumor RNA and a matched-normal DNA to obtain a list of variants labeled as somatic or germline. A random forest classifier is then trained with the collected set of features for each variant in a 5-fold cross validation manner. In the test set (orange arrows), 3 steps are performed: (1) MuTect is applied with tumor RNA and without a matched-normal sample, to yield a list of mixed somatic and germline variants. (2) The five trained models are then applied to this set of variants and classify them as either somatic or germline in a majority vote manner. (3) Finally, the predicted set of variants is further filtered by the RNA-MuTect filtering steps. **b** Distribution of precision and recall values on validation (left) and test (right) sets computed for each sample. Box plots show median, 25th, and 75th percentiles. The whiskers extend to the most extreme data points not considered outliers, and the outliers are represented as dots. **c** Precision as the function of the number of true somatic mutations per sample. **d** Correlation between the number of predicted somatic mutations and the number of somatic mutations as determined by DNA with a matched-normal DNA sample. **e** Correlation between the number of predicted somatic mutations and the number of somatic mutations as determined by RNA with a matched-normal DNA sample. **f** Distribution of precision and recall values on validation (left) and test (right) sets computed for each sample in the lung dataset. Box plots show median, 25th, and 75th percentiles. The whiskers extend to the most extreme data points not considered outliers, and the outliers are represented as dots. **g** Distribution of precision and recall values on validation (left) and test (right) sets computed for each sample in the colon dataset. Box plots show median, 25th, and 75th percentiles. The whiskers extend to the most extreme data points not considered outliers, and the outliers are represented as dots. Source data are provided as a Source Data file.

respectively, median F1-score = 0.84, mean F1-score = 0.8, Fig. 1b). These precision and recall levels correspond to a median of 21 type I and 17 type II errors.

Further investigating our results, we observed that a few samples achieved a precision value <0.6. We found that all these samples had a relatively low number of mutations and a similar number of type I and II errors, except for a single outlier (Fig. 1c). The median precision on the remaining samples is 0.89. In addition, to circumvent the possibility that the high performance obtained by our model is a result of low purity levels which will in turn result in substantially different allele fractions for somatic and germline variants, we examined the correlation between tumor purity and the obtained precision and recall levels. Encouragingly, we found this correlation to be insignificant (Spearman R = −0.0040, −0.0874, p-value = 0.93, 0.09, for precision and recall, respectively). These results testify that the sensitivity of our model remains high even in samples having lower coverage due to normal contamination. Indeed, we found insignificant or low correlation between sample coverage and precision or recall values (Spearman R = −0.05, 0.26, p-value = 0.34, $2.2*10^{-7}$, for precision and recall, respectively, Supplementary Fig. 4).

To better characterize our model we next examined which features are the most important in distinguishing between somatic and germline variants, using the feature importance score (Methods). We found that a few of the PoN features as well as the gnomAD feature are the most influential in our model (Supplementary Data 1). Finally, we computed the Spearman correlation between the number of predicted somatic mutations and the number of mutations detected by DNA or RNA with a matched-normal DNA sample. In both cases, we found it to be highly significant (R = 0.92, p-value = 4.15^{-151} for DNA and R = 0.98, p-value $<8.7*10^{-286}$ for RNA, Fig. 1d,e, respectively, Supplementary Data 2).

When comparing our models' performance to that achieved by previously published methods, we obtained a median precision of 0.017 and recall of 0.32, respectively, for Jessen et al.[36], and a median precision and recall of 0.067 and 0.43, respectively, for Coudray et al.[37]. This inferior performance is expected given that these methods apply a small number of filtering steps, representing a small subset of features included in our ML model, and mainly rely on dbSNP. As discussed above, using this database and even more comprehensive ones such as gnomAD alone, is insufficient for achieving high precision and recall values (Supplementary Fig. 2).

Finally, we repeated our analysis in two additional TCGA datasets, lung adenocarcinoma[38] (n = 493) and colon cancer[39] (n = 230). For the lung dataset, we obtained a median precision and recall of 0.89 and 0.82, respectively, on the train set (n = 100; median F1-score = 0.86), and a median precision and recall of 0.9 and 0.82 on the test set (n = 393; median F1-score = 0.85; Fig. 2f). The mean precision and recall achieved were 0.86 and 0.81 on the train set (mean F1-score = 0.83.) and a mean of 0.85 and 0.81 on the test set, respectively (mean F1-score = 0.83). These results correspond to a median of 15 and 7, type I and II errors, respectively, which is similar to what was found in the melanoma dataset.

In colon, we obtained a median precision and recall of 0.67 and 0.74, respectively, on the train set (n = 100; median F1-score = 0.7), and a median precision and recall of 0.65 and 0.75 on the test set (n = 130; median F1-score = 0.7; Fig. 2g). The mean precision and recall achieved were 0.67 and 0.74 on the train set and (mean F1-score = 0.7) and a mean of 0.66 and 0.74 on the test set, respectively (mean F1-score = 0.7). Similarly, this is equivalent to a median of 23 type I and 14 type II errors. Overall, these results testify for the robustness of our model across different cancer types.

Detecting mutational signatures and significantly mutated genes without a matched-normal sample. The overall high performance of RNA-MuTect-WMN enabled us to apply our standard analysis pipelines for identifying mutational signatures and significantly mutated genes. To this end, we applied SignatureAnalyzer[40,41] using the set of predicted somatic mutations, and identified 4 signatures (Fig. 2a): UV signature (SBS7, cosine similarity = 0.95) which is common in melanoma[42,43], signature 5 (SBS5, cosine similarity = 0.87) which is common in various cancer types, including melanoma, and a signature enriched with C > A mutations that was previously found in ultraviolet light associated melanomas (SBS38, cosine similarity = 0.78). Importantly, the same three signatures were identified in the DNA (Supplementary Fig. 5). In addition, a signature enriched with T > G mutations was detected. This signature was not detected in the DNA but was detected in the RNA when somatic mutations were identified with a matched-normal DNA sample (Supplementary Fig. 5). Indeed, we found that out of 552 mutations that are associated with this signature, 489 were detected only in the RNA. While it is hard to conclude whether this signature is a true RNA signature or a result of RNA-specific noise, it is important to note that its detection is not specific to our pipeline in which a matched-normal sample is not used. Performing the same analysis on the lung and colon datasets, we have found that in both cases the same set of mutational signatures is identified in both the DNA, RNA with a matched-normal sample and our predicted set of somatic mutations (Supplementary Figs. 7 and 9).

Next, we identified significantly mutated genes by applying MutSig2CV[44] on the set of predicted somatic mutations. Out of 24 identified genes, 22 were found to be significantly mutated also when the matched-normal sample is taken into account (Fig. 2b), and only 2 were missed by our pipeline (Supplementary Data 3; p-values are computed by MutSig using the Fisher's method). Importantly, 13 out of the 24 genes were also identified as significantly mutated based on a DNA analysis, a rate that is similar to our previous report (Supplementary Fig. 6)[27]. This difference is a result of the higher rate of non-silent to silent mutations detected in these genes in the RNA, and is not due to germline contamination.

Finally, we examined our pipeline's performance in identifying a set of 55 known melanoma somatic driver genes found in the COSMIC database[45] (Supplementary Data 4). We found that for the 43 genes in this group that carried at least one true somatic mutation in our dataset, our pipeline achieves an even higher precision and recall levels, with median values of 1 and 0.95, respectively, further demonstrating its high value. Finally, applying MutSig2CV to the lung and colon datasets similar results were obtained: 19 out of 24 genes in the lung dataset and 10 out of 12 genes in the colon dataset were found to be significantly mutated in the RNA using our pipeline, as compared to when a matched-normal sample is used (Supplementary Figs. 8 and 10).

TMB predicted by RNA-MuTect-WMN is associated with patient survival. The development of ICI therapy such as anti-PD1 and anti-CTLA4 has revolutionized cancer therapy and resulted in long-lasting tumor responses in patients with a variety of cancers[46]. As a result, these drugs have been FDA-approved for many cancer types, including melanoma, non-small cell lung cancer, Urothelial carcinoma, Head and Neck squamous cell carcinoma, and more[47]. Recently, an accelerated approval for anti-PD1 for the treatment of adult and pediatric with tumor mutational burden-high (TMB-H, ≥10 mut/Mb) has been granted, making it a critical metric in the clinical decision process.

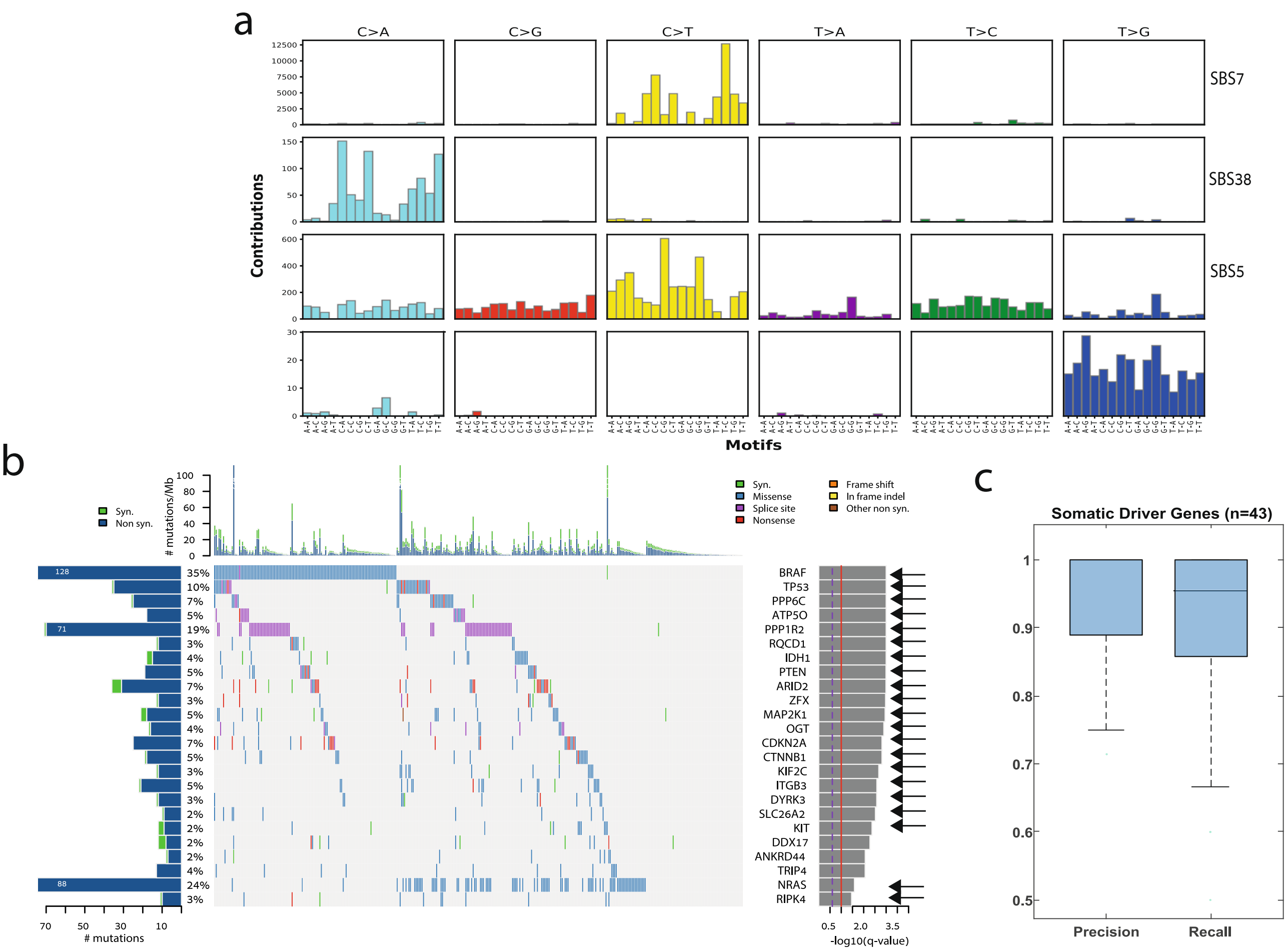

Fig. 2 Identifying mutational signatures and driver genes. a Mutational signatures identified by SignatureAnalyzer[41] on the basis of predicted somatic mutations; **b** Co-mutation plot based on predicted somatic mutations in our test set. Overall frequencies, allele fractions, and significance levels of candidate cancer genes ($Q < 0.05$) identified by MutSig2CV[44] are shown. Genes marked with an arrow were also identified as significantly mutated based on the set of somatic mutations detected using RNA and a matched-normal DNA. Source data are provided as a Source Data file. **c** Precision and recall on the set of know melanoma drivers. Box plots show median, 25th, and 75th percentiles. The whiskers extend to the most extreme data points not considered outliers, and the outliers are represented as dots.

Indeed, the TMB which is traditionally estimated via DNA sequencing, has been found to be associated with patient survival to different extents, depending on cancer type[48] as well as prior and current treatment[49–51].

Considering that only expressed mutations can serve as neoantigens, we here used the set of predicted somatic mutations from RNA sequencing alone, to estimate the TMB, defined as the number of non-silent somatic mutation in each sample (Methods). We then divided the patients into two groups with high- and low-TMB levels, using the median TMB as the cutoff value. We found that patients with high-TMB had a mild but significant increase in survival time as compared to those with low-TMB (log-rank p-value = 0.02, Fig. 3a). Of note, performing the same analysis using the set of somatic mutations detected based on tumor and matched-normal DNA, similar results are obtained (logrank p-value = 0.02, Fig. 3b), further demonstrating the utility of our pipeline.

Performing a multivariate Cox proportional hazards regression analysis with patient age, tumor stage and our RNA-based TMB estimates as the covariates, we found that TMB is the prognostic factor most associated with increased survival (HR = 0.59, 95% CI = 0.36–0.96, p-value <0.03, Fig. 3c).

The extent of association between TMB and patient survival vary widely between the different datasets according to cancer type and prior therapy. A recent publication by Valero et al. showed that among patients that were not treated with ICI, a very high TMB at the top percentiles is associated with poor survival[52]. Given that most of the patients in the TCGA cohort were not treated with ICI, we set to examine this observation in our data as well. Indeed, when we divide the patients into three groups with very high, high, and low TMB levels, using the top 10^{th} percentile for the very high group, and median for the remaining samples, we find that those with the highest TMB values have a poor survival (logrank p-value = 0.04 between high and very high TMB), and those with a median high TMB have an improved survival as compared to those with low TMB (logrank p-value = $6.6*10^{-4}$, Fig. 3d). Importantly, these results remain robust and even become more significant for thresholds between 40 and 60 percentiles (logrank p-value = $5*10^{-3}$–$5*10^{-7}$, Supplementary Data 5). Performing the same analysis based on DNA revealed the same trends, though with a lower significance levels (logrank p-value = 0.03, 0.002, respectively, Fig. 3e, Supplementary Data 5). Repeating the Cox regression analysis while removing the top 10^{th} percentile, the association of TMB with survival became more significant (HR = 0.31, 95% CI = 0.17-0.58, p-value <$2*10^{-4}$, Fig. 3c).

Overall, these results demonstrate that estimating TMB based on RNA alone is feasible and of a high predictive power.

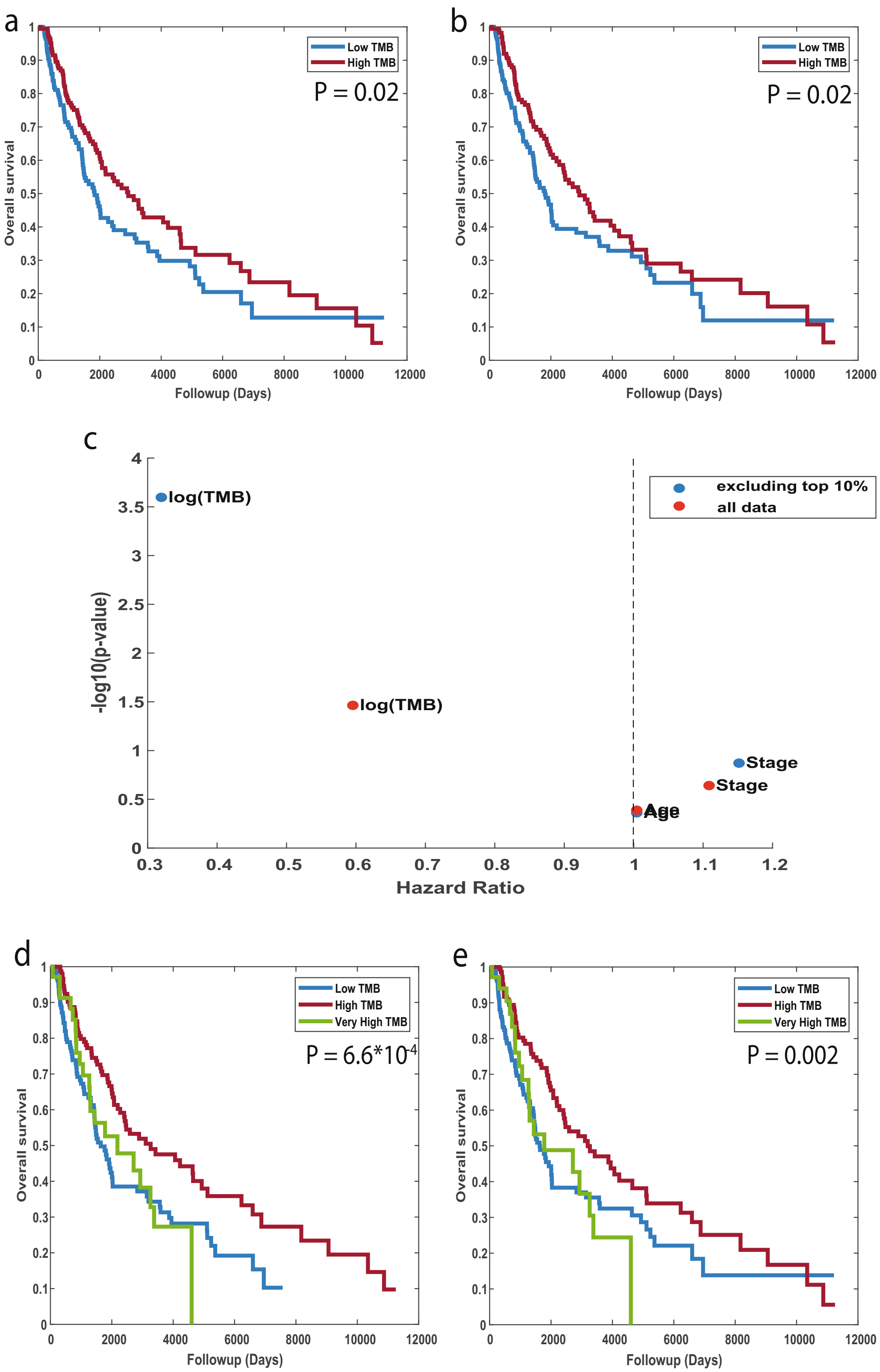

Fig. 3 Association between Tumor Mutational Burden (TMB) and patient survival. Kaplan–Meier survival curves for patients with high vs. low TMB estimated on the basis of predicted somatic mutations from RNA alone (**a**); or on the basis of DNA with a matched-normal sample (**b**). The median TMB value is used to define the 'low TMB' and 'high TMB' subgroups. P-values are computed using a log-rank test. **c** Hazard Ratio vs. $-\log_{10}(p\text{-value})$, obtained by a multivariate Cox proportional hazards regression analysis. Red dots represent the values obtained when all samples are used and blue dots represent the values obtained after excluding the top 10% of samples (very high TMB). **d**, **e** Kaplan–Meier survival curves as in **a** and **b**, respectively, with patients divided into three groups with very-high vs. high vs. low TMB. *p*-values are computed using logrank test. Subgroups were split by using the top 10th percentile for the very high group, and median for the remaining samples. Source data are provided as a Source Data file.

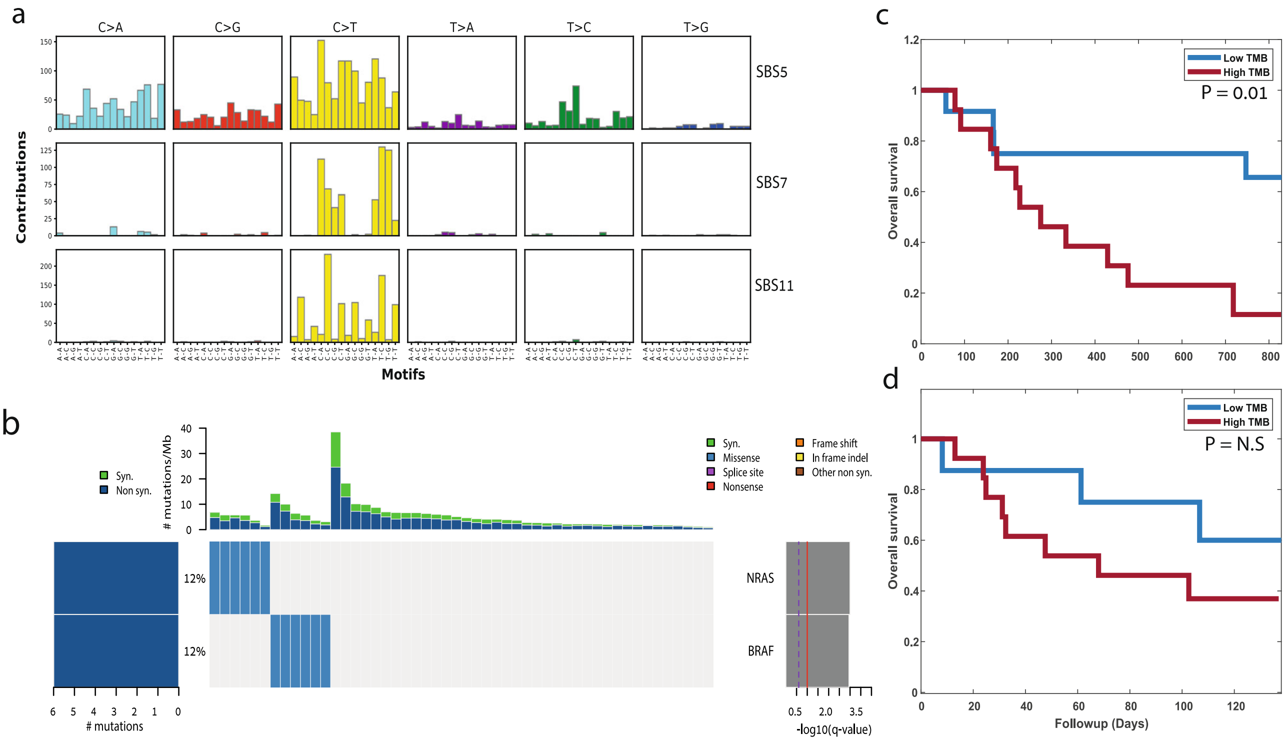

Fig. 4 Pipeline application to an independent dataset. **a** Mutational signatures identified based on the set of predicted somatic mutations using the RNA-seq data of 50 pre-therapy biopsies from the Riaz et al.[15] dataset. **b** Co-mutation plot based on predicted set of somatic mutations. Overall frequencies, allele fractions, and significance levels of candidate cancer genes identified by MutSig2CV[44] are shown. **c**–**d** Kaplan–Meier survival curves for patients that have previously progressed on ipilimumab ($n = 25$). Tumor Mutational Burden (TMB) is estimated based on predicted somatic mutations from RNA (**c**) or based on the list of somatic mutations detected by the authors using tumor and matched-normal DNA (**d**). *p*-values are computed using logrank test. The patients are split to two groups of low and high TMB using the median TMB value as the cutoff. Source data are provided as a Source Data file.

An improved RNA-based TMB estimation in patients treated with ICI. We next examined the prediction power of our model on an additional set of melanoma patients that were treated with nivolumab (anti-PD1), some were treatment-naive and some had previously progressed on ipilimumab (anti-CTLA4)[15]. Raw RNA-sequencing data from 50 pre-therapy biopsies were obtained and aligned to the reference genome (Methods). Then, the 5 models obtained by RNA-MuTect-WMN were applied to identify and classify the set of somatic mutations in each sample. To validate our calls we first applied SignatureAnalyzer and identified the set of mutational signatures that are active in these samples. Encouragingly, we found the UV signature (SBS7), along with the TMZ signature (SBS11) and SBS5 that were also found by the authors based on DNA (cosine similarity = 0.86, 0.95, and 0.78, respectively, Fig. 4a). In addition, when applying MutSig2CV to identify significantly mutated genes, both NRAS and BRAF, known melanoma drivers, were found to be significantly mutated (Fig. 4b).

Finally, we estimated the TMB based on the set of predicted somatic mutation. Interestingly, when considering the set of treatment-naive patients for which both DNA and RNA sequencing is available, no significant association between TMB and patient survival is found, based on neither DNA nor RNA. However, when considering the set of patients that were previously progressed on ipilimumab, a significant association between high TMB and poor survival is found (logrank *p*-value = 0.01, Fig. 4c). This is in similar to the trend reported by the authors using DNA (Fig. 4d), which was insignificant. Overall, in this independent dataset as well we find that estimating the TMB from tumor RNA alone is feasible and results with similar trends to those obtained with tumor and matched-normal DNA.

Discussion

In this study we introduce RNA-MuTect-WMN, a computational method that identifies somatic mutations from RNA-seq data without a matched-normal sample. Our pipeline is based on the RNA-MuTect method[27] which is designed to detect somatic mutations from tumor RNA-seq and a matched-normal DNA. To extend it to a 'tumor-only' mode we developed an ML model that distinguishes between somatic and germline variants using various features, including both mutation-specific ones and those derived from large panels and databases of normal individuals. Our model was trained and tested on the TCGA melanoma dataset, achieving high precision and recall levels. Importantly, we find that when using the set of predicted somatic mutations which are derived from RNA-seq samples alone, all mutational signature and >90% of the driver genes are correctly identified, as compared to RNA-MuTect where a matched-normal sample is used.

When calling somatic mutations directly from RNA-seq we are clearly limited to the set of mutations that are sufficiently expressed. While this reduces sensitivity for certain downstream analyses, it may increase it in others. Specifically, we hypothesized that estimating the tumor mutational burden from RNA rather than from DNA would improve prediction power, as only expressed mutations can become neoantigens and elicit an immune response. Indeed, we first show that estimating TMB from tumor RNA-seq without a matched-normal sample is feasible, and that the exact same trends as those found using tumor DNA with a matched-normal sample are observed. Moreover, the prediction power of RNA-based TMB is either equivalent or higher than that estimated by DNA. As previously shown, we find that in melanoma patients that were not treated with ICI, very high TMB is associated with poor survival[52], while median high is associated with improved survival as compared to patients with low TMB. In addition, in treated patients that were previously progressed on anti-PD1, we find that high TMB is significantly associated with poor survival compared to low TMB. These results are in concordance with the original findings[15].

In this study we applied our pipeline to three different cancer types. However, the RNA-MuTect-WMN approach is generic and can be easily applied to any cancer type, given a sufficient number of samples with RNA-seq of the tumor, along with tumor and matched normal DNA for validation. Clearly, melanoma is a highly mutated cancer with a sufficient number of somatic mutations that can be used for model training, and where the fraction of germline contamination predicted by our model is negligible. Nevertheless, our approach was able to achieve high performance also in a less mutated cancer type such as colon adenocarcinoma, with a consistent small number of type I and II errors. In addition, we here used a relatively small number of samples in our train set to demonstrate the applicability of our method to datasets with a smaller number of samples. However, performing a 10-fold cross-validation on 200 melanoma samples increased the model's performance, with a median precision and recall of 0.89 and 0.87, respectively, and a median F1-score of 0.87 (mean precision and recall of 0.83 and 0.85, respectively, and mean F1-score of 0.83). To overcome potential limitations in datasets where a smaller number of somatic mutations is available for training, and where the fraction of germline contamination can become substantial, one can perform down-sampling of the germline group, or combine multiple datasets together. Moreover, as germline databases continue to grow, we expect that the performance achieved by our approach can be further improved. Overall, as in any ML approach, the number of variants available for training is crucial. We therefore suggest to first build the models on datasets of whole exome or genome sequencing. Those models can then be used on datasets with a smaller number of available variants such as in targeted sequencing. Finally, it should be noted that since germline databases currently available to the community represent mainly individuals of European Ancestry, it is expected that our model will work best on data derived from these lineages.

Overall, we believe that the motivation for using RNA-MuTect-WMN is three-fold: first, for future studies, it can reduce the sequencing cost of a matched-normal sample for the purpose of computing TMB. It can also be used for more established tasks such as identifying driver genes, mutational signatures, tumor heterogeneity, and others, though with reduced sensitivity. This reduction is a result of the low to lack of expression in various genomic regions, which will affect the statistical power required for detecting changes in these regions. Second, it enables the analysis of RNA-seq data in retrospective studies where RNA was originally sequenced for expression-based analyses, and no matched-normal sample is available. Third, it enables a combined analysis where both genetic and phenotypic data can be inferred from the exact same sample. This is especially crucial in cancer where different regions of a tumor from which DNA and RNA are extracted may be significantly different due to tumor heterogeneity. Overall, these applications can significantly increase the number of samples analyzed and thus aid biomarker and drug target discovery.

Methods

DNA Mutation calling pipeline. TCGA DNA BAM files were aligned to the NCBI Human Reference Genome Build GRCh37 (hg19). Sample contamination by DNA originating from a different individual was assessed using ContEst[53]. Somatic single nucleotide variations (sSNVs) were then detected using MuTect[19]. Following this standard procedure, we filtered sSNVs by: (1) removing potential DNA oxidation artifacts[54]; (2) realigning identified sSNVs with NovoAlign V2 (www.novocraft.com) and performing an additional iteration of MuTect with the newly aligned BAM files; and (3) removing technology- and site-specific artifacts using a panel of ~8000 TCGA normal samples (PoN filtering, as in[34]). Specifically, each genomic position in the PoN is binned into one of eight bins using its allele fraction as follows:

total counts <8 (insufficient coverage)
total counts > = 8 (and no alternate reads above subsequent thresholds)
alt count > = 1 and alt fraction > = 0.1%
alt count > = 2 and alt fraction > = 0.3%
alt count > = 3 and alt fraction > = 1%
alt count > = 3 and alt fraction > = 3%
alt count > = 3 and alt fraction >20%
alt count > = 10 and alt fraction > = 20%

Then, a log likelihood score is computed using bins 3–8, as follows:

For a given position, the vector of bin counts is denoted as $\vec{h}$. For each variant call, its allele fraction is represented as a beta distribution parameterized by its alternate and reference read counts (to account for numerical uncertainty when converting read counts to allele fraction):

$$f \sim beta(n_{alt} + 1, n_{ref} + 1) \tag{1}$$

The beta distribution's PDF is then sliced according to the alt fraction bins encoded by the PoN, i.e.,

$$\vec{f} = \left[\int_{0}^{0.1\%} df\, p(f), \int_{0.1\%}^{0.3\%} df\, p(f) \ldots , \int_{20\%}^{100\%} df\, p(f) \right] \tag{2}$$

Finally, a score for this position is computed by weighting each element of $\vec{f}$ by its corresponding histogram bin counts: $S = \vec{f} \cdot \vec{h}$. Mutations with a score of $\log10(S) \geq -2.5$ are filtered out.

Finally, sSNVs were annotated using Oncotator[55]. All steps were run in Terra (https://terra.bio/).

RNA mutation calling pipeline (RNA-MuTect). RNA FASTQ files were downloaded from the Genomic Data Commons database and aligned to the NCBI Human Reference Genome Build GRCh37 (hg19) using STAR[56]. The RNA-MuTect pipeline was applied as previously described[27], and includes the following steps: (1) Applying MuTect to STAR-aligned RNA-seq BAMs with the ALLOW_N_CIGAR_READS flag, considering only mutations supported by ≥3 reads; (2) Removing technology- and site-specific artifacts using a panel of ~8000 TCGA normal samples, as described above; (3) A realignment filter for RNA-seq data where all aligned reads that span a candidate variant position from both the tumor (case) and normal (control) samples are realigned using HISAT2[57]. Then, an additional iteration of MuTect with the newly aligned BAM files is performed; only mutations that are kept by MuTect using both the STAR-aligned and HISAT-aligned BAMs are kept for the next step; (4) Applying an RNA-seq PoN built based on a panel of ~6500 GTEx samples. The RNA-based PoN accounts for various RNA potential and recurrent artifacts, such as errors caused by reverse transcription of RNA to cDNA, RNA modifications, and more; (5) Removing sites that were found in the ExAC[58] database with minor allele frequency of more than 5%; (6) Removing mutations mapped to non-coding regions such as introns, intergenic regions, and non-coding RNAs; (7) Removing RNA editing sites listed in the DARNED[59] and RADAR[60] databases; (8) In cases where multiple reads that support a variant were aligned to the exact same positions, we kept only a single read; (9) Removing mutations that may be caused by sequencing leakage errors, which are identified as alternate bases that have at least 3 bases that match the alternate base in a ± 3 base window around the variant base; (10) Removing sites falling into pseudogenes or IgG genes, which have noisy alignments.

This set of filtering steps is the one used in step 3 of the RNA-MuTect-WMN pipeline. The RNA-MuTect pipeline was run in Terra (https://terra.bio/).

Power analysis. Given a mutation with an alternate allele count of x and a reference allele count of y in the RNA, we computed the power to detect it given a coverage N in the DNA. This was done by applying a beta-binomial model for observing at least k reads: (3) $P(k, |, x, y, N) = \binom{N}{k} \frac{B(k+x+1, N-k+y+1)}{B(x+1, y+1)}$ where B is the Beta function. To determine the minimal number of reads k, we first computed the error rate at the variant site, r, using the matched normal sample by taking the maximal allele fraction of the three possible alternate alleles and applying the Laplace correction with $\alpha = 1$. We then identified k as the number of alternate reads that have a probability <1% to be generated by the noise. Eventually, powered mutations were considered as those with power > 0.95 and alternate read count > = 4.

The RNA-MuTect-WMN pipeline

Data preprocessing. For the training set where a matched-normal sample is used we first labeled as somatic the set of variants passing our entire calling pipeline, as described in the 'RNA Mutation calling pipeline' section. Germline variants were determined based on MuTect annotation, 'normal_lod', 'germline_risk' or 'alt_allele_in_normal'. The analysis is focused only on chromosome 1–22, X, and Y.

Feature collection. the following features were used in our pipeline:

1. T_ref_count - number of tumor reads supporting the reference allele
2. T_alt_count - number of tumor reads supporting the alternate allele
3. T_lod_fstar – The LOD score computed by MuTect
4. Tumor_f – tumor allele fraction

5–12. For each of the germline variants database (dbSNP, gnomAD, 1000Genome, ESP) two vectors were created: (a) A Boolean indicating whether the variant is present (1) or not (0) in each database; (b) variant allele fraction (AF), when available, and a mean AF value over all variants in the database when this data is missing.

13. Variant_classification - if the variant classification as defined by Oncotator[55] was either IGR, Intron, RNA, lincRNA this feature was set to be (1) and (0) otherwise. While introns and intergenomic regions should not appear in RNA sequencing, such transcripts can sometimes arise due to the “fuzzy” transcription of known genes that extends beyond the annotated boundaries[61]. Since there are 4 orders of magnitude more germline mutations than somatic ones, there’s a higher chance to find them in these regions. We included this feature following an analysis showing that almost no somatic mutations were annotated with this variant classification.

14–31 DNA and RNA Panel of Normals –each genomic position in each PoN is binned into one of eight bins using its allele fraction as follows:

total counts <8 (insufficient coverage)
total counts > = 8 (and no alternate reads above subsequent thresholds)
alt count > = 1 and alt fraction > = 0.1%
alt count > = 2 and alt fraction > = 0.3%
alt count > = 3 and alt fraction > = 1%
alt count > = 3 and alt fraction > = 3%
alt count > = 3 and alt fraction >20%
alt count > = 10 and alt fraction > = 20%

The 9th feature for each PoN is then the log-likelihood score computed as described above under the “DNA mutation calling pipeline” section.

Model training. 100 samples were randomly selected and defined as the training set. These samples were then divided to 5 pairs of training and validation sets with 80 and 20 samples in each group, respectively. A random forest classifier was applied on each of the training sets, using the somatic and germline labels, with 50 trees and number of features that equals the square root of the total number of features. Each resulting model was then tested on the corresponding validation set, and the precision and recall were calculated per sample to evaluate the model’s performance. Using a different number of trees ranging from 30 to 100 resulted in an identical performance.

Model testing. To test the models generated in the training step we first applied MuTect on our test set composed of the remaining samples, using tumor RNA-seq and without the matched normal sample. As a result, we obtained a list of variants containing both somatic, germline, and RNA-specific noise. For each of these variants, we collected the set of features described above and applied the 5 trained models. Each variant was then assigned a somatic or germline label based on a majority vote of the 5 models. Finally, the predicted group of somatic mutations was further filtered using RNA-MuTect filtering steps, as described in[27] and above.

Mutational Signature analysis. To identify mutational signatures, we used the SignatureAnalyzer tool: https://github.com/broadinstitute/getzlab-SignatureAnalyzer[41]. A cosine similarity score was used as a measure of closeness to known signatures. This score ranges between zero and one, where similarity of one represents identical signatures and similarity of zero represents completely different mutational signatures. The similarity was measured against the latest version (V3.2) of SBS signatures in COSMIC.

MutSig2CV for RNA-seq data. To apply MutSig2CV[44] for RNA-seq data we utilized an RNA-based gene-level coverage model that reflected which bases were typically sufficiently covered in each gene using GTEx RNA-seq data, as previously done[27]. Specifically, this model contains information about the sequencing coverage achieved for each gene and sample. Within each gene-sample bin, the coverage is broken down further according to the category (e.g., A: T basepairs, C: G basepairs), and also according to the zone (silent/non-silent/noncoding)[5].

We considered genes as significantly mutated if they had an FDR-corrected Q value <0.05.

Riaz data analysis. Raw RNA sequencing data for 50 available pre-treatment samples was aligned to the NCBI Human Reference Genome Build GRCh37 (hg19) using STAR[56]. As described in the main text (Fig. 1a), we then ran the three steps of our pipeline: (1) running MuTect with tumor RNA alone; (2) applying the 5 models trained on TCGA data to get an initial list of predicted somatic mutations; (3) applying RNA-MuTect filtering steps. The final list of somatic variants was used for identifying significantly mutated genes, mutational signatures and for estimating the TMB.

Tumor mutational burden analysis. To compute tumor mutation burden (TMB), we counted the number of non-silent somatic SNVs per sample. For the survival analysis, the absolute TMB value was used for determining the median TMB and the top 10 percentile values. For the continuous Cox regression models we used $\log_{10}$(TMB) together with patient age and tumor stage. A logrank test was used to estimate the significance level of the survival analysis.

Feature importance. To determine feature importance we used the built-in feature importance scores of scikit-learn, also known as GINI importance (or mean decreased impurity). We obtained the feature importance scores for each of the 5 trained models, and computed the final importance score for each feature based on the average score across all 5 models.

Estimating precision/recall level for single features. For each feature, we computed the difference between germline and somatic variants using a two-sided Wilcoxon rank-sum test.

To calculate the best precision and recall that can be achieved, we computed the F1-score across a range of thresholds and report the maximal one. The precision-recall AUC was computed in a standard way across a range of thresholds.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Access to TCGA raw sequencing data (DNA and RNA) was obtained via dbGap authorization from accession number phs000178. The DNA PoN is available in Google Cloud upon dbGap authorization (gs://firecloud-tcga-dcc-closed-access/reference/PoNs/final_summed_tokens.hist.bin)

The Riaz[15] bulk RNA dataset used in this study is available under BioProject accession number PRJNA356761 and in the NCBI Gene Expression Omnibus (GEO) GSE91061.

The RNA PoN is available upon dbGap GTEx approval (phs000424) under ‘Available Phenotype and Genotype Files/Genotype Files/phg000830.v1.GTEx_WES.panel-of-normals.c1.GRU.tar’.

In addition, NCBI Human Reference Genome Build GRCh37 (hg19) was used [https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.13/]. Source data are provided with this paper.

Code availability

The code for running the ML pipeline is available as Supplementary Code, as pseudo code in the Supplementary Material as Supplementary Note 1, and in GitHub: https://github.com/yizhak-lab-ccg/RNA_MUTECt_WMN. All features are collected using the software and databases described above.

Received: 22 July 2021; Accepted: 12 May 2022;
Published online: 02 June 2022

References

1. Vijg, J. Somatic mutations, genome mosaicism, cancer, and aging. *Curr. Opin. Genet. Dev.* **26**, 141–149 (2014).
2. Iñigo, M. & Campbell, P. J. Somatic mutation in cancer and normal cells. *Science* **349**, 1483–1489 (2015).
3. Lodato, M. A. et al. Somatic mutation in single human neurons tracks developmental and transcriptional history. *Science* **350**, 94–98 (2015).
4. Blokzijl, F. et al. Tissue-specific mutation accumulation in human adult stem cells during life. *Nature* **538**, 260–264 (2016).
5. Lawrence, M. S. et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).
6. Davoli, T. et al. Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. *Cell* **155**, 948–962 (2013).
7. Garraway, L. A. & Lander, E. S. Lessons from the cancer genome. *Cell* **153**, 17–37 (2013).
8. Vogelstein, B. et al. Cancer genome landscapes. *Science* **339**, 1546–1558 (2013).
9. Dentro, S. C. et al. Characterizing genetic intra-tumor heterogeneity across 2658 human cancer genomes. *Cell* **184**, 2239–2254.e39 (2021).
10. Ma, J., Setton, J., Lee, N. Y., Riaz, N. & Powell, S. N. The therapeutic significance of mutational signatures from DNA repair deficiency in cancer. *Nat. Commun.* **9**, 3292 (2018).
11. Vanderstichele, A., Busschaert, P., Olbrecht, S., Lambrechts, D. & Vergote, I. Genomic signatures as predictive biomarkers of homologous recombination deficiency in ovarian cancer. *Eur. J. Cancer* **86**, 5–14 (2017).

12. Snyder, A. et al. Genetic basis for clinical response to CTLA-4 blockade in melanoma. *N. Engl. J. Med.* **371**, 2189–2199 (2014).
13. Van Allen, E. M. et al. Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. *Science* **350**, 207 LP–207211 (2015).
14. Hugo, W. et al. Genomic and transcriptomic features of response to anti-PD-1 therapy in metastatic melanoma. *Cell* **165**, 35–44 (2016).
15. Riaz, N. et al. Tumor and microenvironment evolution during immunotherapy with Nivolumab. *Cell* **171**, 934–949.e16 (2017).
16. Rizvi, N. A. et al. Mutational landscape determines sensitivity to PD-1 blockade in non–small cell lung cancer. *Science* **348**, 124 LP–124128 (2015).
17. Miao, D. et al. Genomic correlates of response to immune checkpoint blockade in microsatellite-stable solid tumors. *Nat. Genet.* **50**, 1271–1281 (2018).
18. Łuksza, M. et al. A neoantigen fitness model predicts tumour response to checkpoint blockade immunotherapy. *Nature* **551**, 517–520 (2017).
19. Cibulskis, K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotech.* **31**, 213–219 (2013).
20. Koboldt, D. C. et al. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**, 568–576 (2012).
21. Kim, S. et al. Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods* **15**, 591–594 (2018).
22. McKenna, A. et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
23. Lai, Z. et al. VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res.* **44**, e108–e108 (2016).
24. Hiltemann, S., Jenster, G., Trapman, J., van der Spek, P. & Stubbs, A. Discriminating somatic and germline mutations in tumor DNA samples without matching normals. *Genome Res.* **25**, 1382–1390 (2015).
25. Teer, J. K. et al. Evaluating somatic tumor mutation detection without matched normal samples. *Hum. Genom.* **11**, 22 (2017).
26. Sun, J. X. 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.* **14**, e1005965 (2018).
27. Yizhak, K. et al. RNA sequence analysis reveals macroscopic somatic clonal expansion across normal tissues. *Science* **364**, eaaw0726 (2019).
28. Akbani, R. et al. Genomic classification of cutaneous melanoma. *Cell* **161**, 1681–1696 (2015).
29. Sherry, S. T., Ward, M. & Sirotkin, K. dbSNP—database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome Res.* **9**, 677–679 (1999).
30. Karczewski, K. J., Francioli, L. C. & Tiao, G. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443.
31. Authon, A. et al. The 1000 Genomes Project Consortium. *Nature* **526**, 68–74.
32. Server, E. V. NHLBI GO Exome Sequencing Project (ESP) https://www.mendeley.com/catalogue/2e7deb8d-b0d8-3893-9fa6-809f16dfae0b/?utm_source=desktop&utm_medium=1.19.4&utm_campaign=open_catalog&userDocumentId=%7B28fa4f6c-a9f1-4b05-99e1-bd6eae5eaca6%7D.
33. Lonsdale, J. et al. The genotype-tissue expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
34. Ellrott, K. et al. Scalable open science approach for mutation calling of tumor exomes using multiple genomic pipelines. *Cell Syst.* **6**, 271–281.e7 (2018).
35. Ho, T. K. The random subspace method for constructing decision forests. *IEEE Trans. Pattern Anal. Mach. Intell.* **20**, 832–844 (1998).
36. Jessen, E., Liu, Y., Davila, J., Kocher, J.-P. & Wang, C. Determining mutational burden and signature using RNA-seq from tumor-only samples. *BMC Med. Genom.* **14**, 65 (2021).
37. Coudray, A., Battenhouse, A. M., Bucher, P. & Iyer, V. R. Detection and benchmarking of somatic mutations in cancer genomes using RNA-seq data. *PeerJ* **6**, e5362 (2018).
38. Collisson, E. A. et al. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543–550 (2014).
39. Muzny, D. M. et al. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330–337 (2012).
40. Alexandrov, L. B. et al. The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
41. Kim, J. et al. Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. *Nat. Genet* **48**, 600–606 (2016).
42. Martincorena, I. et al. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**, 880–886 (2015).
43. Saini, N. et al. The impact of environmental and endogenous damage on somatic mutation load in human skin fibroblasts. *PLoS Genet.* **12**, 1–25 (2016).
44. Lawrence, M. S. et al. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495–501 (2014).
45. Bamford, S. et al. The COSMIC (Catalogue of somatic mutations in cancer) database and website. *Br. J. Cancer* **91**, 355–358 (2004).
46. Pardoll, D. M. The blockade of immune checkpoints in cancer immunotherapy. *Nat. Rev. Cancer* **12**, 252 (2012).
47. Wei, S. C., Duffy, C. R. & Allison, J. P. Fundamental mechanisms of immune checkpoint blockade therapy. *Cancer Discov.* **8**, 1069 LP–1061086 (2018).
48. McGrail, D. J. et al. High tumor mutation burden fails to predict immune checkpoint blockade response across all cancer types. *Ann. Oncol.* **32**, 661–672 (2021).
49. Samstein, R. M. et al. Tumor mutational load predicts survival after immunotherapy across multiple cancer types. *Nat. Genet.* **51**, 202–206 (2019).
50. Klempner, S. J. et al. Tumor mutational burden as a predictive biomarker for response to immune checkpoint inhibitors: a review of current evidence. *Oncologist* **25**, e147–e159 (2020).
51. Litchfield, K. et al. Meta-analysis of tumor- and T cell-intrinsic mechanisms of sensitization to checkpoint inhibition. *Cell* **184**, 596–614.e14 (2021).
52. Valero, C. et al. The association between tumor mutational burden and prognosis is dependent on treatment context. *Nat. Genet.* **53**, 11–15 (2021).
53. Cibulskis, K. et al. ContEst: estimating cross-contamination of human samples in next-generation sequencing data. *Bioinformatics* **27**, 2601–2602 (2011).
54. Costello, M. et al. Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucl. Acids Res.* **41**, e67–e67 (2013).
55. Ramos, A. H. et al. Oncotator: cancer variant annotation tool. *Hum. Mutat.* **36**, E2423–E2429 (2015).
56. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
57. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
58. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
59. Kiran, A. & Baranov, P. V. DARNED: a Database of RNa editing in humans. *Bioinformatics* **26**, 1772–1776 (2010).
60. Ramaswami, G. & Li, J. B. RADAR: a rigorously annotated database of A-to-I RNA editing. *Nucleic Acids Res.* **42**, D109–D113 (2014).
61. Agostini, F., Zagalak, J., Attig, J., Ule, J. & Luscombe, N. M. Intergenic RNA mainly derives from nascent transcripts of known genes. *Genome Biol.* **22**, 136 (2021).

Acknowledgements

We would like to thank Dr. Moshe Sade-Feldman, Dr. Paz Polak, and Dr. Raphy Zarecki for helpful comments on the manuscript. This work was supported by the Dr. Sherman-Saifer Career Advancement Chair and the Alon fellowship.

Author contributions

R.K. and K.Y. conceived the idea. R.K. and K.Y. designed the study. R.K. and N.R. performed the analysis. R.K. and K.Y. wrote the manuscript.

Competing interests

The authors declare no competing interests

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-022-30753-2.

Correspondence and requests for materials should be addressed to Keren Yizhak.

Peer review information *Nature Communications* thanks Carla Robles-Espinoza and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.