nature portfolio

Peer Review File

Predicting immunotherapy response in advanced bladder cancer through a meta-analysis of six independent cohorts

Corresponding Author: Professor M. Mar Albà

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
Boll et al. constructs large meta-dataset of advanced bladder tumors that have been clinically annotated for response to ICI. While they perform comprehensive analysis on the dataset, none of their finds are particularly novel. Instead, they present a series features which have been shown previously, in those same cohorts, to be associated with ICI response, namely TMB, APOBEC activity and immune features. They then go on generate a predictive model, however it is unclear if this adds to or outperforms models from other studies.

While the authors should be applauded work to compile this dataset, there are still significant concerns which need to be address before publication (see below), in addition to my worry that the over representation of IMvigor210 is driving the results.

Specific Comments:

Based on the supplemental data, there seems to be a bias towards specific datasets for features (e.g. TMB is highest in UC-GENOME, HdM-BLCA-1 has increased purity for WES). Are the mutational frequencies, immune features, and clinical variables similar across datasets?

For the mutational signature analysis, are SBS2 and SBS13 similarity correlated to response. Do any of the datasets have survival? Is there an association between prediction of response and prognosis?

How does the authors model compare again other published models of ICI response and what is the minimal set of features needed to retain significant predictive power.

The cohort is ~50% IMvigor210, how is this influencing which features are selected in the model. Would you get similar results if you build the model based on all except IMvigor and then predicted on IMvigor?

In Figure 6E, I agree it is important to remove datasets to see if the model is overly influenced by any give cohort, however the inset legend lists numbers of samples that do not fit with the pattern of excluding one cohort at a time. Could you please clarify the numbers and the calculation.

Minor points:

Dataset color for IMvigor and UC-GENOME are too close together in Figure 1 and hard to tell apart.

Some of the datasets seem to be misattributed.

Reviewer #2

(Remarks to the Author)
Boll et al have performed a large scale analysis of 6 BLCA cohorts treated with anti-PD1/PDL1 therapies, including an in

depth analysis of specific subtypes. For cohorts with available data, they process and harmonize the data to allow comparison. They investigate previously identified BLCA specific biomarkers of response to immunotherapy and find that immune infiltration, TMB were positively associated, while TGF-β/stroma/EMT and CCND1 were negatively associated. Novel biomarkers of immunotherapy response include nonstop mutations, lncRNAs, and ARHGEF12 mutations. Finally, they develop a random forest machine learning model to predict immunotherapy response given these significant biomarkers, which achieves better performance than TMB alone. The methodology is clear and reproducible and we particulary appreciated the individual dataset removal analysis to help to demonstrate that one cohort is not driving the result. Another strength is analysing the different immune signatures in the context of TCGA subtype and immune infiltrate high and low.
As acknowledged in the manuscript, TMB, immune infiltration, and neoantigen abundance biomarkers are already associated with response to immunotherapy in solid tumours including bladder cancer (PMC9462669, https://doi.org/10.1093/annonc/mdx371.003) supporting the findings here. There are some novel insights which will be of significant interest to those in the field of cancer biology, drug development and immunology. However, presented here is a discovery data set. There is risk of over-fitting and results should be considered hypothesis generating until a true validation cohort can be generated. However given the size of the validation cohort required and future trials that will need to be recruited to, this will be a future endeavor. Nether the less, this is important data that should be published and we would support publication in Nat comms. We have no major comments and the following minor comments.
Minor Comments:
• Line 136, since you are not actually comparing the effect sizes between the two. I would refrain from making the statement that there is no difference. Violin plots are nice for visualization but it would also be good to have odds ratios or coefficients (e.g. from logistic regression) to be able to quantify the effect size
• Figure 2F is interesting but I don't think it adds much value to compare mutation types per se – it could be more informative to compare mutations predicted to be high impact (e.g. FS, nonstop, LOF predicted, etc) vs those predicted to be more benign. It also might be worth trying to add in indel and nonsense neoantigens – I believe pvacseq can handle missense, inframe indels, and frameshifts
• It would be good to give the results of a statistical test for the CDKN2A/CDKN2B and CCND1 vs response analysis (line 177-185)
• Figure 3D it is unclear which paper belongs to which signature, especially the CD8 signatures. Have post-hoc multiple testing been applied?
• Figure 3F and 3G are swapped according to the text. Figure 3 legend F and G are incorrect
• The lncRNA combined expression is significantly associated with response but only one cohort independently shows significant association (Supp Figure 3) – I would recommend revising lines 218-219
• Does line 230 mean to reference Figure 3G and supplementary figure 6? Are the relationships between TMB and the immune signatures referenced on lines 231-237 significant? Supplementary figure 4 is interesting, cleaning up the figure would improve readability. Of the variables within the 5 groups, are these significantly correlated with each other? Figure 3G has some comparisons but the relationship between APOBEC and TMB is not analyzed, for example
• Sup. Figure 4 is very difficult to read as legends are overlapping the graph. Be good to amend
• Couldn't the HLA-I expression (Figure 4D) and similarly APM (Figure 4G) be influenced by the higher immune infiltration? The line 276 mis-references Figures 4F-G
• Line 279 – does 'infiltrated tumors' mean basal squamous and luminal infiltrated? Is there any additional expression filter/biomarker to assess if the immune infiltration is exhausted or not? This would also improve the analysis in Figure 5F/line 310. Do the results of this analysis also hold for the respective subtypes?
• Without performing a statistical test, it is difficult to agree that the effects observed visually are "more marked" in the immune infiltrated groups (line 282-283)
• Typo line 345 'aera'
• It would also be good to have a model with TMB 10 mb as this is the FDA threshold. Analysis of correlation between model variables and removal of essentially duplicate ones could reduce noise and improve model performance. A De Long's test to assess if there is a significant difference between the TMB only model and the authors' model would add confidence to the analysis.
• Why does TMB have a 0 feature importance score in Supp Figure 8?
• There are trials that are selecting treatment for BLCA patients on the basis of TCGA/transcriptional subtype. The results presented here might suggest caution for such an approach – eg. In the GUSTO trial, patients with luminal papillary or luminal subtypes are deemed likely non-responsive to chemotherapy and immunotherapy and go straight to radical cystectomy.

Reviewer #3

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Reviewer #4

(Remarks to the Author)
Boll and coauthors present a meta-analysis of six independent cohorts aiming to identify factors that can predict response to immunotherapy in advanced bladder cancer. The topic is clinically important, especially if these critical factors are easy to test and include in the model. Overall, the paper is robust and clearly written. Several minor suggestions are offered below.

Please define “advanced cancer”. Can it be replaced by “metastatic cancer” in the title and elsewhere for clarity?

Ln. 85. ECOG >1 was significantly associated with worse outcomes, but Suppl Tables list “ECOG_under0”. Please clarify.

Please provide p-values for correlation coefficients everywhere

p.271. Luminal-papillary tumors, not patients.

The study's main limitation is, of course, the lack of independent validation for predicting response. The model could have been developed based on 5 cohorts, using the 6th as the validation set.

Abstract, ln. 20 – “to build highly accurate predictive models”. Please include an AUC value or a range of values. Again, the models are entirely based on training datasets and independent prediction is not yet validated. A statement about “highly accurate predictive models” is too strong at this point.

Figures are very fuzzy in the pdf and details are hard to see.

Supplemental File 2 should be provided with submission as a Source file.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
I would like to thank the authors for their work to address my comments and revise the manuscript. While a majority of my comments were addressed, there still remains a few outstanding issues (See below).

In the UC-GENOME paper (PMID: 36333289) the authors identified an increased presence of stroma-rich subtype within metastatic cohorts. Also, the stroma-rich group had increased immune cell fractions, by using the TCGA subtype scheme as opposed to the Consensus Subtypes (PMID: 31563503) are you losing information.

In Figure 5, you have divided samples into immune-infiltrated and non-infiltrated group by subtype, however immune phenotype and subtype are not a 1 to 1 surrogate. As both the IMvigor(PMID: 29443960) and UC-GENOME papers have shown response to be associated with exclusion vs. inflamed, are the subtype based grouping appropriate?

Unfortunately, Figure 6 labels were unreadable therefore I was unable to evaluate the figures. However, I still have concerns regarding the value of a predictor with an AUC = 0.76.

Reviewer #3

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Reviewer #4

(Remarks to the Author)
The authors have appropriately addressed all my questions and comments. The addition of the new validation dataset was very valuable.

Version 2:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
I would like to thank the authors for their responses, however, they failed to completely address several my concerns and comments in meaningful manner.

1) In the response to comment 1, the authors state that they are not losing any information by using the TCGA classifications, however present no analysis comparing the use of the TCGA subtypes to the use of Consensus subtypes. Instead, they point to a prior study where there was concordance between the subtypes, however it is unclear how the non-concordant samples in that study may factor into the current work.

2) I still have concerns regarding the results in Figure 3-5 are being driven by a singular study. Additionally, in the IMvigor cohort, there is a mix of samples that have platinum therapy and are platinum naïve, does that influence any of the data or drive any results?

3) The authors should discuss the limitations of their study within the discussion.

4) Finally, while the model is interesting, the AUC is still below that of a model which could be clinically used, nor it is clear that the model outperforms any of the prior models/signatures of ICI response

RESPONSE TO REVIEWERS' COMMENTS

Reviewer #1 (Remarks to the Author):

Boll et al. constructs large meta-dataset of advanced bladder tumors that have been clinically annotated for response to ICI. While they perform comprehensive analysis on the dataset, none of their finds are particularly novel. Instead, they present a series features which have been shown previously, in those same cohorts, to be associated with ICI response, namely TMB, APOBEC activity and immune features. They then go on generate a predictive model, however it is unclear if this adds to or outperforms models from other studies.

While the authors should be applauded work to compile this dataset, there are still significant concerns which need to be address before publication (see below), in addition to my worry that the over representation of IMvigor210 is driving the results.

Reply:

We would like to thank the reviewer for the valuable comments, which have clearly helped to improve this manuscript. We have now performed an external validation of the model on the JAVELIN Bladder 100 trial cohort (we only recently gained access to data from this cohort) and the model has proven to be robust, with an AUC value that is slightly higher to that obtained with the internal validation method (0.764 vs 0.767). In addition to identifying biomarkers of response, the compilation of data from different cohorts has allowed us to compare response rate of different bladder cancer subtypes. We have also obtained novel insights into which could be the more relevant biomarkers in each case. We believe this adds to the novelty of the study and could be of value in decision making in the clinics. We address the specific comments below in more detail.

Specific Comments:

1. Based on the supplemental data, there seems to be a bias towards specific datasets for features (e.g. TMB is highest in UC-GENOME, HdM-BLCA-1 has increased purity for WES). Are the mutational frequencies, immune features, and clinical variables similar across datasets?1. Based on the supplemental data, there seems to be a bias towards specific datasets for features (e.g. TMB is highest in UC-GENOME, HdM-BLCA-1 has increased purity for WES). Are the mutational frequencies, immune features, and clinical variables similar across datasets?

Reply:

In general, the different variables investigated are very similar across cohorts. We have added a table with descriptive statistics for all variables in the different datasets (Suppl. file 2). TMB is a particular case because for UC-GENOME and UNC-108 it was estimated from the available DNA mutation panels, which resulted in higher values than when TMB was inferred from WES data (for the rest of datasets). But all TMB values were normalized to z-scores in each of the datasets prior to any analyses, and this homogenized the values. We have added boxplots of the raw TMB and the zscore TMB to supplementary figure 14. While the difference in the original TMB values between datasets is significant (Kruskal-Wallis p-value < 0.001), no difference between datasets can be found after normalization (Kruskal-Wallis p-value = 0.18).

Regarding the purity estimates, which were used to calculate mutation clonality for those datasets with WES data, the differences between datasets are not statistically significant (Kruskal-Wallis p-value = 0.92).

1. For the mutational signature analysis, are SBS2 and SBS13 similarity correlated to response. Do any of the datasets have survival? Is there an association between prediction of response and prognosis?

   Reply:

   To understand the impact of the two APOBEC signatures SBS2 and SBS13 on response, we have run a logistic regression using these variables. While both signatures are significantly related to response, signature SBS2 has a stronger effect (B = 5.40, p-value = 0.0048) than SBS13 (B = 1.91, p-value = 0.0107).

   We have included this information in the Results: “The two APOBEC signatures SBS2 and SBS13 were significantly associated with the response, but the first one had a stronger effect (linear regression values B = 5.40 and p-value = 0.0048) than the second one (B = 1.91 and p-value = 0.0107).” (lines 143-147)

2. How does the authors model compare again other published models of ICI response and what is the minimal set of features needed to retain significant predictive power.

   Reply:

   By testing the complete model on a portion of the data we obtained an average AUC of 0.767. This is quite comparable to previous estimates for the model built combining data from several cancer types by Litchfield *et al.* (2021, PMID: 33508232). Depending on the dataset they tested the model against these authors obtained AUC values between 0.66 and 0.86. The elastic net logistic regression model by Damrauer *et al.* (2022, PMID: 36333289), built on the bladder cancer IMvigor210 dataset, showed an AUC of 0.84 when tested on the validation portion of the same dataset, 0.82 when tested on UNC-108 and 0.65 when tested on UC-GENOME.

While the manuscript was being reviewed we gained access to the data from the JAVELIN Bladder 100 trial (Powles et al., PMID: 3294563), which includes mutation and gene expression data from 123 advanced urothelial cancer patients treated with avelumab (anti-PD-L1). When we tested our model against this external dataset we obtained an AUC of 0.764, which is slightly higher than the value of 0.747 obtained with internal validation (using 1000 different seeds in the train/test phase for the same RNA + TMB model)(see updated Figure 6). This indicates that the model is robust and applicable to other bladder cancer datasets unrelated to the training set.

Regarding the minimal set of features needed to retain significant predictive power we developed models with less features, but all of them resulted in lower AUC values. It is worth saying that a model that approaches the accuracy of the complete model is the RNA + TMB model (Figure 6, AUC 0.747 *versus* 0.761). Using this model can be useful when no WES data is available and we only have TMB estimates.

We have added the following paragraph to the Discussion: “Using part of the data for validation we obtained an AUC estimate of 0.761 for the model that used the complete set of variables (0.793 for the immune infiltrated subtypes) and 0.747 for the RNA + TMB model (gene expression variables and TMB). This represented an improvement over using TMB alone (0.678). When we tested the RNA + TMB model against the JAVELIN Bladder 100 trial we obtained an AUC of 0.764. These predictive values are in the range of those obtained by a previously obtained model that combined data from different types of cancer (Litchfield et al., 2021); the AUC values were between 0.66 and 0.86 depending on the test dataset. The elastic net logistic regression model by Damrauer et al. (2022), built on the bladder cancer IMvigor210 dataset, showed an AUC of 0.84 when tested on the validation portion of the same dataset, 0.82 when tested on UNC-108 and 0.65 when tested on UC-GENOME.” (lines 519-528)

3. The cohort is ~50% IMvigor210, how is this influencing which features are selected in the model. Would you get similar results if you build the model based on all except IMvigor and then predicted on Imvigor?

Reply:

IMvigor 210 is the largest dataset but the distribution of the different variables is similar across datasets. RNA + TMB models in which we removed one dataset at a time, and which had a sample size between 178 (when removing IMvigor 210) and 327 (when removing HdM-BLCA-1) showed similar ROC curves (AUC 0.731-0.741, see Figure 6F), indicating that loss of information from IMvigor210 does not compromise the accuracy of the predictive models.

Given the availability of data from the JAVELIN Bladder 100 trial cohort (n=123), we have focused on this external cohort to validate the model. In this dataset our model achieves an AUC of 0.764 (new Figure 6C), indicating that there has been no or little overfitting in the train/test approach. The new findings have been added to the manuscript: “As a validation cohort, we tested the model on data from the JAVELIN Bladder 100 trial (Powles et al., 2021), which includes advanced urothelial cancer patients treated with avelumab (anti-PD-L1). While no raw data was available, we were provided with processed data. With this data, we tested the model TMB + RNA using data from 123 patients from the JAVELIN Bladder 100 trial (27 responders, 96 non-responders). The model achieved an AUC of 0.764 in the validation run, surpassing the averaged AUC of 0.747 obtained from 1000 seeds in the train/test phase (Figure 6C). This result underscores the robustness of our model, highlighting its potential in predicting outcomes for patients treated with immune checkpoint inhibitors.” (lines 385-392)

4. In Figure 6E, I agree it is important to remove datasets to see if the model is overly influenced by any give cohort, however the inset legend lists numbers of samples that do not fit with the pattern of excluding one cohort at a time. Could you please clarify the numbers and the calculation.

   Reply:

   We apologize for the mistake in the sample numbers. The figure (now Figure 6F) has been updated to contain the correct sample sizes.

   Minor points:

5. Dataset color for IMvigor and UC-GENOME are too close together in Figure 1 and hard to tell apart.

   We have corrected this.

6. Some of the datasets seem to be misattributed.

   We appreciate the attention to detail of the reviewer. We have updated the color for IMvigor and corrected the typos in the dataset names of Figure 1.

Reviewer #2 (Remarks to the Author):

Boll et al have performed a large scale analysis of 6 BLCA cohorts treated with anti-PD1/PDL1 therapies, including an in depth analysis of specific subtypes. For cohorts with available data, they process and harmonize the data to allow comparison. They investigate previously identified BLCA specific biomarkers of response to immunotherapy and find that immune infiltration, TMB were positively associated, while TGF-β/stroma/EMT and CCND1 were negatively associated. Novel biomarkers of immunotherapy response include nonstop mutations, lncRNAs, and ARHGEF12 mutations. Finally, they develop a random forest machine learning model to predict immunotherapy response given these significant biomarkers, which achieves better performance than TMB alone. The methodology is clear and reproducible and we particulary appreciated the individual dataset removal analysis to help to demonstrate that one cohort is not driving the result. Another strength is analysing the different immune signatures in the context of TCGA subtype and immune infiltrate high and low.

As acknowledged in the manuscript, TMB, immune infiltration, and neoantigen abundance biomarkers are already associated with response to immunotherapy in solid tumours including bladder cancer (PMC9462669, https://doi.org/10.1093/annonc/mdx371.003) supporting the findings here. There are some novel insights which will be of significant interest to those in the field of cancer biology, drug development and immunology. However, presented here is a discovery data set. There is risk of over-fitting and results should be considered hypothesis generating until a true validation cohort can be generated. However given the size of the validation cohort required and future trials that will need to be recruited to, this will be a future endeavor. Nether the less, this is important data that should be published and we would support publication in Nat comms. We have no major comments and the following minor comments.

Reply:

We would like to thank the reviewer for the useful and constructive comments. We have added the reference PMC9462669 to the manuscript. We would also like to add that we have now added the results of an external validation cohort (JAVELIN Bladder 100 trial, n=123).

Minor Comments:

1. Line 136, since you are not actually comparing the effect sizes between the two. I would refrain from making the statement that there is no difference. Violin plots are nice for visualization but it would also be good to have odds ratios or coefficients (e.g. from logistic regression) to be able to quantify the effect size.

Reply:

We agree with the reviewer that we should refrain from making the statement that there is no difference. Actually, we have now checked the effect size and subclonal mutations have a larger effect than clonal ones. The odds ratio obtained from a logistic regression is higher for subclonal TMB (OR = 1.55, 95%CI [1.30; 1.92]) than for clonal TMB (OR = 1.24; 95%CI [1.10; 1.42]). We have included this information in the results: "While both subclonal and clonal mutations were significantly associated with the response (Figure 2C), the first ones had a larger effect size. The odds ratio obtained from a logistic regression was higher for subclonal TMB (OR = 1.55, 95%CI [1.30; 1.92]) than for clonal TMB (OR = 1.24; 95%CI [1.10; 1.42])." (lines 137-140)

2. Figure 2F is interesting but I don't think it adds much value to compare mutation types per se – it could be more informative to compare mutations predicted to be high impact (e.g. FS, nonstop, LOF predicted, etc) vs those predicted to be more benign. It also might be worth trying to add in indel and nonsense neoantigens – I believe pvacseq can handle missense, inframe indels, and frameshifts.

Reply:

We have used the classification of the ENSEMBL annotation tool vep to separate the non-synonymous mutations into two groups: high and moderate impact. Responders have significantly more high but also more moderate impact mutations (Wilcoxon test p-value< 0.001 in both cases). Next, we constructed peptides resulting from frameshift INDELs and the protein elongations of nonstop mutations. We did not find any significant differences in the putative number of neoantigens derived from these mutations between responders and non-responders.

These results have been added to the manuscript text: "Missense mutations were significantly associated with response (Figure 2F, Suppl. Table 3), independently of the predicted impact of the mutation on protein function (Suppl. Figure 1). We found that non-stop mutations (single nucleotide variants that cause a stop codon loss) were significantly associated with response, but this was not the case for frameshift or in-frame insertions and deletions (INDELs). We also predicted neoantigens derived from missense mutations using the NetMHCpan software [23]. As previously described for different cancer types [6,18,24], we found that the number of predicted neoantigens derived from missense mutations was significantly associated with response. Finally, we constructed peptides resulting from frameshift INDELs and the protein extensions generated by nonstop mutations. For these mutations we found no difference in the number of putative neoantigens between responders and non-responders (Suppl. Figure 1)." (lines 176-186)

3. It would be good to give the results of a statistical test for the CDKN2A/CDKN2B and CCND1 vs response analysis (line 177-185).

   Reply:

   We used the Chi-square test to compare the number of amplifications and deletions between the two response groups. For CCND1 75 non-responders and 43 responders were found to have an amplification of the gene (Chi-square test p-value=0.65). For *CDKN2A*/*CDKN2B* we report 30 non-responders and 16 responders to have a copy number deletion of the gene (Chi-square test p-value=0.709). We have added this information in the manuscript text “While we found CDKN2A/CDKN2B to be the most common deletion in our data (16 responders and 30 non-responders), and every second patient to have an amplification in the CCND1 gene region (51.30%, 75 non-responders and 43 responders), none of the copy number alterations were found to be significantly associated to ICI response (Chi-square test p-value > 0.05).” (lines 191-195)

4. Figure 3D it is unclear which paper belongs to which signature, especially the CD8 signatures. Have post-hoc multiple testing been applied?

   Reply:

   A complete list of gene signatures with the corresponding publications are provided in the Methods section. Additionally, a full list of the genes included in each signature can be found in Suppl. file 2.

   We have added the following information to the legend of Figure 3: ·All comparisons in panel A and D were significant after multiple testing adjustment, except for the genes TGFB1 and CCND1, as well as the gene signatures T cells inflamed and Stroma/EMT”

   We provide additional details in Methods: “We used the Bonferroni correction post-hoc to assess the significance of the effect of expression of single genes or gene signatures. As 15 gene signatures were tested, the adjusted alpha for gene signatures was 0.00033. For genes, the alpha for adjusted p-values was 0.00029.” (lines 796-799)

5. Figure 3F and 3G are swapped according to the text. Figure 3 legend F and G are incorrect.

   Reply:

The letters in Figure 3 have been updated accordingly.

6. The lncRNA combined expression is significantly associated with response but only one cohort independently shows significant association (Supp Figure 3) – I would recommend revising lines 218-219.

Reply:

We agree with the reviewer that this should be clarified. We have added text to the corresponding section to provide more information: “When the analysis was performed separately for each cohort, only IMvigor210 achieved statistical significance (Suppl. Figure 4).” (lines 231-233)

7. Does line 230 mean to reference Figure 3G and supplementary figure 6? Are the relationships between TMB and the immune signatures referenced on lines 231-237 significant? Supplementary figure 4 is interesting, cleaning up the figure would improve readability. Of the variables within the 5 groups, are these significantly correlated with each other? Figure 3G has some comparisons but the relationship between APOBEC and TMB is not analyzed, for example.

Reply:

The text refers to figure 3G that was previously mislabeled and is now corrected (see comment 5). The correlations between TMB and the mentioned immune signatures are significant, we have added the p-values in the text. We have reformatted the Suppl. Figure 4 (which is now Suppl. Figure 5) so that the results of the PCA can be clearly seen. The variables within the 5 groups show different degrees of correlation, we now provide a table with all correlation coefficients and associated p-values in Suppl. data file 2. The comparison of mutation-based variables such as APOBEC and TMB can be found in Figure 2E.

8. Sup. Figure 4 is very difficult to read as legends are overlapping the graph. Be good to amend.

Reply:

We agree with the reviewer and have improved the clarity of the figure (see previous question).

9. Couldn't the HLA-I expression (Figure 4D) and similarly APM (Figure 4G) be influenced by the higher immune infiltration? The line 276 mis-references Figures 4F-G

Reply:

We found high correlation coefficients between the HLA-I and APM gene signatures and other signatures of immune infiltration. This was unanticipated, as expression of HLA class I molecules and proteins of the APM reflect antigen processing and presentation by the tumor cell. But any interpretation of why this happens remains speculative (we cannot establish causality) so we have refrained from drawing further conclusions.

10. Line 279 – does 'infiltrated tumors' mean basal squamous and luminal infiltrated? Is there any additional expression filter/biomarker to assess if the immune infiltration is exhausted or not? This would also improve the analysis in Figure 5F/line 310. Do the results of this analysis also hold for the respective subtypes?

Reply:

Infiltrated tumors indeed means basal squamous and luminal infiltrated. We have added more information to the text to make this more clear: "We next analyzed the different variables in the immune-infiltrated (luminal infiltrated and basal-squamous) and non-immune infiltrated (luminal-papillary, luminal and neuronal) subtypes separately. Although immune infiltrated tumors had significantly lower TMB than non-immune infiltrated ones (Figure 5A).." (lines 317-320)

Although several biomarkers, such as PD-1, have been related to T cell exhaustion, we do not think that we can provide information on this matter in our study. All patients in the study have been treated with anti-PD-(L)1 and we do not have adequate controls for T cell exhaustion/non-exhaustion states.

We have repeated the analysis presented in Figure 5F for luminal infiltrated and basal-squamous separately. When separated by subtypes, the results do not achieve statistical significance, probably due to smaller sample size, however we can observe the same trend as when combined (see new Suppl. Figure 9). We have added the following text: "When we performed the analysis per subtype the same trend was observed, especially for luminal infiltrated, although the results did not achieve statistical significance (Suppl. Figure 9)." (lines 336-338)

11. Without performing a statistical test, it is difficult to agree that the effects observed visually are “more marked” in the immune infiltrated groups (line 282-283).

Reply:

Because the shape of the distributions is quite different across groups and variables, we have compared the median values of responders and non-responders in each case. We find that, except for TGF-β, the difference between medians of responders vs non-responders is always higher in the immune infiltrated subtypes than in the non-immune infiltrated subtypes. We have included this information in the manuscript as follows: “The effect of immune activation and immunosuppression signatures on the response was significant in both groups, but in general more marked in the case of the immune-infiltrated subtypes (Figure 5C and 5D, Suppl. Figures 6 and 7). For the IFN-γ signature, the difference between the median of the response group and the median of the non-response group was 0.727 for immune-infiltrated and 0.229 for non-immune-infiltrated. We observed the same trend for the CIBERSORT score (0.364 and 0.112, respectively), M1 macrophages (0.099 and 0.037, respectively) and CD8 T cells (0.07 and 0.026, respectively).” (lines 321-327)

12. Typo line 345 ‘aera’

The word has been corrected.

13. It would also be good to have a model with TMB 10 mb as this is the FDA threshold. Analysis of correlation between model variables and removal of essentially duplicate ones could reduce noise and improve model performance. A De Long’s test to assess if there is a significant difference between the TMB only model and the authors’ model would add confidence to the analysis.

Reply:

We have generated a ROC curve using TMB 10 mutations/Mb as a threshold. In this model, patients with more than 10 mutations/Mb are predicted to be responders and the rest non-responders. The AUC value is 0.61. We have added this information to the manuscript text: “A random forest model considering only z-score TMB had an AUC of 0.678. A threshold model in which tumor samples with more than 10 mutations/Mb were predicted to be from responders and those with less mutations from non-responders was associated with an AUC of 0.61 (Suppl. Figure 10).”

We have not performed the De Long test because this test is for comparing two individual ROC curves, and our procedure involves building 1000 models with the

70/30 train/test method and taking the average values as an estimate. Taken together, the results of our study support than variables other than TMB are informative on the response and can increase the predictive power of the models.

14. Why does TMB have a 0 feature importance score in Supp Figure 8?

Reply:

It seems there was an error when exporting the plot. We have updated the figure (now it is Suppl. Figure 10).

15. There are trials that are selecting treatment for BLCA patients on the basis of TCGA/transcriptional subtype. The results presented here might suggest caution for such an approach – eg. In the GUSTO trial, patients with luminal papillary or luminal subtypes are deemed likely non-responsive to chemotherapy and immunotherapy and go straight to radical cystectomy.

Reply:

We thank the reviewer for bringing this interesting clinical trial to our attention. We have added information on this issue to the discussion: “Surprisingly, we found that the two subtypes with the highest immune infiltration - basal squamous and luminal infiltrated - did not show an overall better response than the other subtypes. This is likely to be relevant for trials that select treatment on the basis of cancer subtype [55].” (lines 485-488)

Reviewer #3 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Reviewer #4 (Remarks to the Author):

Boll and coauthors present a meta-analysis of six independent cohorts aiming to identify factors that can predict response to immunotherapy in advanced bladder cancer. The topic is clinically important, especially if these critical factors are easy to test and include in the model. Overall, the paper is robust and clearly written. Several minor suggestions are offered below.

1. Please define “advanced cancer”. Can it be replaced by “metastatic cancer” in the title and elsewhere for clarity?

   Reply:

   Advanced cancer includes locally advanced surgically incurable and metastatic cancer and it is more appropriate to describe the complete set of patients in this study. We have clarified it in the manuscript: “We extracted clinical and omics data for a total of 707 advanced BLCA patients treated with anti-PD-L1/PD-1 (Figure 1). The patients were diagnosed with metastatic or locally advanced BLCA.” (lines 73-75). For consistency, we have also replaced metastatic by advanced in the abstract.

2. Ln. 85. ECOG >1 was significantly associated with worse outcomes, but Suppl Tables list “ECOG_under0”. Please clarify.

   Reply:

   The variable should be “ECOG ≥ 1” as it separates the patients into ECOG scores of 0 and ECOG over 0. We have corrected it throughout the manuscript and the supplementary data.

3. Please provide p-values for correlation coefficients everywhere.

   Reply:

   We have added the p-values for all correlation coefficients (Suppl. file 2).

4. p.271. Luminal-papillary tumors, not patients.

   Reply:

The line has been corrected accordingly.

5. The study's main limitation is, of course, the lack of independent validation for predicting response. The model could have been developed based on 5 cohorts, using the 6th as the validation set.

   Reply:

   While the manuscript was being reviewed we gained access to the data from the JAVELIN Bladder 100 trial (Powles et al., PMID: 3294563), which includes mutation and gene expression data from 123 advanced urothelial cancer patients treated with avelumab (anti-PD-L1). When we tested our model against this external dataset we obtained an AUC of 0.764. This indicates that the model is robust and applicable to other bladder cancer datasets unrelated to the training set.

   We have added the following information to the results section: “As a validation cohort, we tested the model on data from the JAVELIN Bladder 100 trial 45, which includes advanced urothelial cancer patients treated with avelumab (anti-PD-L1). While no raw data was available, we were provided with processed data. With this data, we tested the model TMB + RNA using data from 123 patients from the JAVELIN Bladder 100 trial (27 responders, 96 non-responders). The model achieved an AUC of 0.764 in the validation run, surpassing the averaged AUC of 0.747 obtained from 1000 seeds in the train/test phase (Figure 6C). This result underscores the robustness of our model, highlighting its potential in reliably predicting outcomes for patients treated with immune checkpoint inhibitors.” (lines 385-392)

6. Abstract, ln. 20 – “to build highly accurate predictive models”. Please include an AUC value or a range of values. Again, the models are entirely based on training datasets and independent prediction is not yet validated. A statement about “highly accurate predictive models” is too strong at this point.

   Reply:

   As mentioned in point 5 we have now validated the model with an independent external cohort. The AUC value is actually slightly better than that obtained using internal validation, supporting that the model is highly accurate.

7. Figures are very fuzzy in the pdf and details are hard to see.

   Reply:

All figures are now available as separate vector-based svg format, which should ensure high resolution.

8. Supplemental File 2 should be provided with submission as a Source file.

Reply:

The supplementary data file 2 is now being provided as a source file.

RESPONSE TO REVIEWERS' COMMENTS

Reviewer #1 (Remarks to the Author):

I would like to thank the authors for their work to address my comments and revise the manuscript. While a majority of my comments were addressed, there still remains a few outstanding issues (See below).

1. In the UC-GENOME paper (PMID: 36333289) the authors identified an increased presence of stroma-rich subtype within metastatic cohorts. Also, the stroma-rich group had increased immune cell fractions, by using the TCGA subtype scheme as opposed to the Consensus Subtypes (PMID: 31563503), are you losing information?

The stroma-rich subtype of the Consensus classification shows a good correspondence to the TCGA subgroup Luminal Infiltrated according to the publication by Kamoun et al. on the Consensus groups (PMID: 31563503). Therefore we are not losing information by using the TCGA subtypes. We have added this information to the manuscript: "The luminal infiltrated subtype shows low tumor purity and muscle-related signatures, and corresponds to the stroma-rich subtype in other classifiers[11]" (Introduction, lines 52 and 53).

2. In Figure 5, you have divided samples into immune-infiltrated and non-infiltrated group by subtype, however immune phenotype and subtype are not a 1 to 1 surrogate. As both the IMvigor (PMID: 29443960) and UC-GENOME papers have shown response to be associated with exclusion vs. inflamed, are the subtype based grouping appropriate?

Both the basal squamous and the luminal infiltrated subtypes are characterized by high levels of immune biomarkers when compared to the other subtypes (PMID: 28988769); this is why we labelled them as immune infiltrated. But we performed a large number of analyses that were independent of the subtype, and which allowed us to study the effect of differently components of the immune system in the response to ICI. These analyses indicate that there is a positive correlation between immune activating biomarkers, such as CD8, and response (Figure 3). TGF-beta, on the contrary, was associated with lack of response (Figure 5F), which is consistent with the results of the IMvigor study.

Unfortunately, Figure 6 labels were unreadable therefore I was unable to evaluate the figures. However, I still have concerns regarding the value of a predictor with an AUC = 0.76.

We have uploaded a new Figure 6 in PDF format, which should cause no problems. In this Figure one can see the results of applying the model to an external cohort (JAVELIN Bladder 100 trial cohort). We have obtained a very similar AUC value than when testing the model on the validation part of the same dataset, reinforcing the robustness of our approach. We would also like to emphasize that the AUC value of 0.76 corresponds to the complete dataset. For some subtypes we achieved higher accuracy, for example for basal

squamous the AUC was 0.82 (Suppl. Table 5). The large number of tumor samples analyzed here allows us to examine different groups of patients separately.

Reviewer #3 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Reviewer #4 (Remarks to the Author):

The authors have appropriately addressed all my questions and comments. The addition of the new validation dataset was very valuable.

RESPONSE TO REVIEWERS' COMMENTS

Reviewer #1 (Remarks to the Author):

I would like to thank the authors for their responses, however, they failed to completely address several my concerns and comments in meaningful manner.

1. In the response to comment 1, the authors state that they are not losing any information by using the TCGA classifications, however present no analysis comparing the use of the TCGA subtypes to the use of Consensus subtypes. Instead, they point to a prior study where there was concordance between the subtypes, however it is unclear how the non-concordant samples in that study may factor into the current work.

All patients for which we had RNA-Seq data could be classified in one of the TCGA subtypes (N=420). A previous comparison by Kamoun et al. (PMID: 31563503) showed that the stroma-rich subtype of the Consensus classification shows a good correspondence to the TCGA subgroup Luminal Infiltrated. We have added this information to the manuscript: "The luminal infiltrated subtype shows low tumor purity and muscle-related signatures, and corresponds to the stroma-rich subtype in other classifiers" (Kamoun et al., 2020)(Introduction, lines 52 and 53).

2. I still have concerns regarding the results in Figure 3-5 are being driven by a singular study. Additionally, in the IMvigor cohort, there is a mix of samples that have platinum therapy and are platinum naïve, does that influence any of the data or drive any results?

IMvigor 210 is the largest dataset but the distribution of the different variables is similar across datasets (e.g. for the lncRNA signature, Sup Figure 4). We also provide a table with descriptive statistics for all variables of the models in the different datasets (supplementary data file, sheet "Statistics_pre_dataset_and_resp"). RNA + TMB models in which we removed one dataset at a time, and which had a sample size between 178 (when removing IMvigor 210) and 327 (when removing HdM-BLCA-1) showed similar AUC curves (0.731-0.741, see Figure 6F), indicating that loss of information from IMvigor210 does not compromise the accuracy of the predictive models. The robustness of our results was also validated in an external cohort. In this dataset our model achieves an AUC of 0.764 (Figure 6C), indicating that there has been no or little overfitting in the train/test approach.

The patient data analysed comes from different clinical trials where patients sometimes received chemo- or radiation therapy previous or in combination with immunotherapy. While platinum therapy can impact the immune system and tumor microenvironment, we did not find that platinum treatment was related to immunotherapy response in the IMvigor210 cohort (p-value=0.156), nor in the UC-GENOME cohort (p-value=1) or SNY-2017 cohort (p-value=0.311). For 138 patients, there was no treatment history available.

3. The authors should discuss the limitations of their study within the discussion.

The limitations of this study are discussed in the second last paragraph of the discussion.

4. Finally, while the model is interesting, the AUC is still below that of a model which could be clinically used, nor it is clear that the model outperforms any of the prior models/signatures of ICI response

We agree with the reviewer that the proposed model has no direct clinical application, however we believe that the presented results add to the understanding of the interplay between the tumor, the immune microenvironment and response to immunotherapy.
By testing the complete model on a portion of the data we obtained an average AUC of 0.76. This is quite comparable to previous estimates for the model built combining data from several cancer types by Litchfield et al. (2021, PMID: 33508232). Depending on the dataset they tested the model against, these authors obtained AUC values between 0.66 and 0.86. The elastic net logistic regression model by Damrauer et al. (2022, PMID: 36333289), built on the bladder cancer IMvigor210 dataset, showed an AUC of 0.84 when tested on the validation portion of the same dataset, 0.82 when tested on UNC-108 and 0.65 when tested on UC-GENOME.

When we tested our model against an external dataset we obtained an AUC of 0.764, which is slightly higher than the value of 0.747 obtained from 1000 different seeds in the train/test phase for the same model (RNA + TMB)(see updated Figure 6). This indicates that the model is robust and applicable to other bladder cancer datasets unrelated to the training set. Additionally, our model obtained very similar AUC values on different subsets of the data (AUCs between 0.731-0.740 in leave-one-out, figure 6F), reinforcing the robustness of our approach. We would also like to emphasize that the AUC value of 0.76 corresponds to the complete dataset. For some subtypes we achieved higher accuracy, for example for basal squamous the AUC was 0.82 (Suppl. Table 5).

Reviewer #3 (Remarks to the Author):

1. I think the authors adequately addressed Reviewer 1's additional comments 1 and 2. For the last point, the authors could also add that the model used will not necessarily be used clinically, but to show that these features collectively perform better than TMB alone. A De Long's test could demonstrate statistically significant improvement.

As we have already discussed in the first round of revisions, a DeLong test is not applicable in this case. This is because we are presenting the average value for 1000 seeds applying a 70/30 train/test method rather than a single model run. Furthermore, the test is used to compare predicted probabilities between two models while we are looking at a binary outcome (responder/non-responder). Additionally, there is evidence that DeLong test should not be applied in a nested model such as ours(PMID: 22415937).

As an alternative to evaluate change in AUC, we have calculated confidence intervals (CI) with R/CRAN package cvAUC (PMID: 26279737).

> *TMB-only: AUC 0.702 CI 95% [0.699, 0.705]*
> *TMB+RNA: AUC 0.744 CI 95% [0.741, 0.747]*

These results show non-overlapping CI, supporting that the TMB+RNA model performs better than the TMB-only model.