communications
biology

ARTICLE

https://doi.org/10.1038/s42003-022-03098-1 OPEN

Data-driven design of targeted gene panels for estimating immunotherapy biomarkers

Jacob R. Bradley[1]✉ & Timothy I. Cannings[1]✉

Tumour mutation burden and other exome-wide biomarkers are used to determine which patients will benefit from immunotherapy. However, the cost of whole exome sequencing limits the widespread use of such biomarkers. Here, we introduce a data-driven framework for the design of targeted gene panels for estimating a broad class of biomarkers including tumour mutation burden and tumour indel burden. Our first goal is to develop a generative model for the profile of mutation across the exome, which allows for gene- and variant type-dependent mutation rates. Based on this model, we then propose a procedure for constructing biomarker estimators. Our approach allows the practitioner to select a targeted gene panel of prespecified size and construct an estimator that only depends on the selected genes. Alternatively, our method may be applied to make predictions based on an existing gene panel, or to augment a gene panel to a given size. We demonstrate the excellent performance of our proposal using data from three non small-cell lung cancer studies, as well as data from six other cancer types.

[1] School of Mathematics, University of Edinburgh, Edinburgh, UK. ✉email: j.r.j.bradley@sms.ed.ac.uk; timothy.cannings@ed.ac.uk

It has been understood for a long time that cancer, a disease occurring in many distinct tissues of the body and giving rise to a wide range of presentations, is initiated and driven by the accumulation of mutations in a subset of a person's cells[1]. Since the discovery of immune checkpoint blockade (ICB)[2–4], there has been an explosion of interest in cancer therapies targeting immune response and ICB therapy is now widely used in clinical practice[5]. ICB therapy works by targeting natural mechanisms (or *checkpoints*) that disengage the immune system, for example, the proteins cytotoxic T-lymphocyte-associated protein 4 (CTLA-4) and programmed death ligand 1 (PD-L1)[6]. Inhibition of these checkpoints can promote a more aggressive anti-tumour immune response[7], and in some patients, this leads to long-term remission[8]. However, ICB therapy is not always effective[9] and may have adverse side-effects, so determining which patients will benefit in advance of treatment is vital.

Exome-wide prognostic biomarkers for immunotherapy are now well-established—in particular, tumour mutation burden (TMB) is used to predict response to immunotherapy[10,11]. TMB is defined as the total number of non-synonymous mutations occurring throughout the tumour exome, and can be thought of as a proxy for how easily a tumour cell can be recognised as foreign by immune cells[12]. However, the cost of measuring TMB using whole exome sequencing (WES)[13] currently prohibits its widespread use as standard-of-care. Sequencing costs, both financial and in terms of the time taken for results to be returned, are especially problematic in situations where high-depth sequencing is required, such as when utilising blood-based circulating tumour DNA (ctDNA) from liquid biopsy samples[14]. The same issues are encountered when measuring more recently proposed biomarkers such as tumour indel burden (TIB)[15,16], which counts the number of frameshift insertion and deletion mutations. There is, therefore, demand for cost-effective approaches to estimate these biomarkers[17,18].

In this paper, we propose a data-driven method for biomarker estimation, based on a generative model of how mutations arise in the tumour exome. More precisely, we model mutation counts as independent Poisson variables, where the mean number of mutations depends on the gene of origin and variant type, as well as the background mutation rate (BMR) of the tumour. Due to the ultrahigh-dimensional nature of sequencing data and the fact that in many genes mutations arise purely according to the BMR, we use a regularisation penalty when estimating the parameters of the model. In addition, this identifies a subset of genes that are mutated above or below the background rate. Our model facilitates the construction of an estimator of TMB, based on a weighted linear combination of the number of mutations in each gene. The vector of weights is chosen to be sparse (i.e. have many entries equal to zero), so that our estimator of TMB may be calculated using only the mutation counts of a subset of genes. In particular, this allows for accurate estimation of TMB from a targeted gene panel, where the panel size (and therefore the cost) may be determined by the user. We also provide an R package `ICBioMark`[19] which implements the methodology and reproduces the experimental results in the paper.

We demonstrate the excellent practical performance of our framework using a non-small cell lung cancer (NSCLC) dataset[20], and include a comparison with existing state-of-the-art approaches for estimating TMB. We further validate these results by testing the performance on data from two more NSCLC studies[21,22]. Moreover, since our model allows variant type-dependent mutation rates, it can be adapted easily to predict other biomarkers, such as TIB. Our method may also be used in combination with an existing targeted gene panel. In particular, we can estimate a biomarker directly from the panel, or first augment the panel and then construct an estimator. Finally, in order to further investigate the utility of our proposal across a range of mutation profiles, we use it to select targeted gene panels and estimate TMB in six other cancer types.

Due to its emergence as a biomarker for immunotherapy in recent years, a variety of groups have considered methods for estimating TMB. A simple and common way to estimate TMB is via the proportion of mutated codons in a targeted region. Budczies et al.[23]. investigate how the accuracy of predictions made in this way are affected by the size of the targeted region, where mutations are assumed to occur at a uniform rate throughout the genome. More recently Yao et al.[24] modelled mutations as following a negative binomial distribution while allowing for gene-dependent rates, which are inferred by comparing non-synonymous and synonymous mutation counts. In contrast, our method does not require data including synonymous mutations. Where they are included, we do not assume that synonymous mutations occur at a uniform rate throughout the genome, giving us the flexibility to account for location-specific effects on synonymous mutation rates such as chromatin configuration[25] and transcription-dependent repair mechanisms[26]. Linear regression models have been used for both panel selection[27] and for biomarker prediction[28]. A review of some of the issues arising when dealing with targeted panel-based predictions of TMB biomarkers is given by Wu et al.[29]. Finally, we are unaware of any methods for estimating TIB from targeted gene panels.

Results and discussion

In this section, we demonstrate in detail the practical performance of our proposal using the dataset from Campbell et al.[20]. Our main focus is the prediction of TMB, and we show that our method outperforms state-of-the-art approaches. We also analyse the suitability of our generative model, include a panel augmentation case study with the TST-170 gene panel, and consider the task of predicting the recently proposed biomarker TIB. Finally, in this section, we test our method's generality and robustness by applying it to data from two further NSCLC datasets and then six further cancer types.

Data and terminology. Our methodology can be applied to any annotated mutation dataset obtained by WES. To demonstrate our proposal we make use of the NSCLC dataset produced by Campbell et al.[20], which contains data from 1144 patient-derived tumours. For each sample in this dataset, we have the genomic locations and variant types of all mutations identified. At the time of the study, the patients had a variety of prognoses and smoking histories, were aged between 39 and 90, 41% were female and 59% were male; see Fig. 1a, b. In Fig. 1c we see that mutations counts are distributed over a very wide range, as is the case in many cancer types[30]. For simplicity, we only consider seven non-synonymous variant types: missense mutations (which are the most abundant), nonsense mutations, frameshift insertions/deletions, splice site mutations, in-frame insertions/deletions, nonstop mutations and translation start site mutations. We present the frequencies of these mutation types in Fig. 1d. Frameshift insertion/deletion (also known as indel) mutations are of particular interest when predicting TIB, but contribute only a small proportion (<4%) of non-synonymous mutations.

It is useful at this point to introduce the notation used throughout the paper. The set G denotes the collection of genes that make up the exome. For a gene $g \in G$, let ℓ_g be the length of g in nucleotide bases, defined by maximum coding sequence as collected from the *Ensembl* database[31]. A gene panel is a subset $P \subseteq G$, and we write $\ell_P := \sum_{g \in P} \ell_g$ for its total length. We let S denote the set of variant types in our data (e.g. in the dataset mentioned above, S contains the seven possible non-synonymous variants). Now, for $i = 0, 1, \ldots, n$, let M_{igs} denote the count of

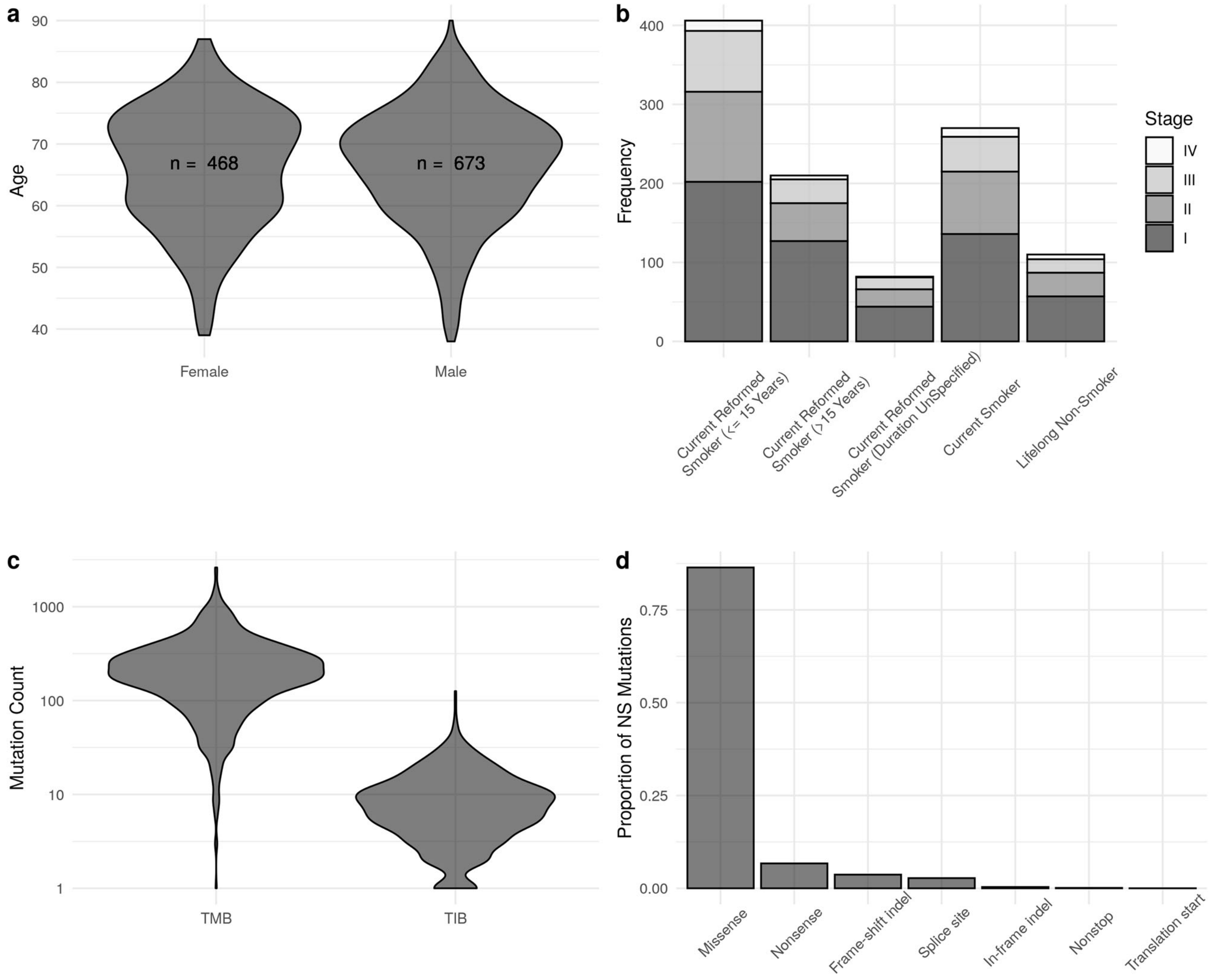

Fig. 1 Population data for the clinical cohort in Campbell et al. NSCLC dataset[20]. a Violin plots of age for patients, stratified by sex. **b** Stacked bar chart of patients' smoking histories, shaded according to cancer stage diagnosis. **c** Violin plot of the distribution of TMB and TIB across training samples. **d** The relative frequency of different non-synonymous mutation types.

mutations in gene $g \in G$ of type $s \in S$ in the ith sample. Here the index $i = 0$ is used to refer to an unseen test sample for which we would like to make a prediction, while the indices $i = 1, \ldots, n$ enumerate the samples in our training dataset. In order to define the exome-wide biomarker of particular interest, we specify a subset of mutation types $\bar{S} \subseteq S$, and let

$$T_{i\bar{S}} := \sum_{g \in G} \sum_{s \in \bar{S}} M_{igs}, \tag{1}$$

for $i = 0, \ldots, n$. For example, including all non-synonymous mutation types in $\bar{S}$ specifies $T_{i\bar{S}}$ as the TMB of sample i, whereas letting $\bar{S}$ contain only indel mutations gives TIB.

Our main goal is to predict $T_{0\bar{S}}$ based on $\{M_{0gs}: g \in P, s \in S\}$, where the panel $P \subseteq G$ has length ℓ_P satisfying some upper bound. When it is clear from context that we are referring to the test sample and a specific choice of biomarker (i.e. $\bar{S}$ is fixed), we will simply write T in place of $T_{0\bar{S}}$.

Since we are only looking to produce estimators for TMB and TIB, we group mutations into two categories—indel mutations and all other non-synonymous mutations—so that $|S| = 2$. This simplifies the presentation of our results and reduces the computational cost of fitting the generative model. In order to assess the performance of each of the methods in this section, we randomly split the dataset into training, validation and test sets, which contain $n_{\text{train}} = n = 800$, $n_{\text{val}} = 171$ and $n_{\text{test}} = 173$ samples, respectively. Mutations are observed in $|G| = 17{,}358$ genes. Our training set comprises samples with an average TMB of 252 and TIB of 9.25.

Generative model fit. The first step in our analysis is to fit our generative model using only the training dataset. In particular, we obtain estimates of the model parameters using equation (4), where the tuning parameter κ_1 is determined using tenfold cross-validation as described in the Methodology section. Diagnostics and model validation statistics are presented in Supplementary Fig. 1 and Supplementary Table 1. The best choice of κ_1 produces estimates of λ and η with 44.4 and 77.8% sparsity respectively, i.e. that proportion of their components are estimated to be exactly zero. We plot $\hat{\lambda}$ and $\hat{\eta}$ for this value of κ_1 in Fig. 2a, b. Genes with $\hat{\lambda}_g = 0$ are interpreted to be mutating according to the background mutation rate, and genes with $\hat{\eta}_{\text{g,indel}} = 0$ are interpreted as having no specific selection pressure for or against indel mutations. In Fig. 2a, b we highlight genes with large (in absolute value) parameter estimates, some of which have known biological relevance in oncology; see our Conclusion for further discussion. Finally, note that the average fitted value of μ_i among current smokers is 5.40 (with a standard deviation of 0.76), amongst reformed smokers is 5.26 (0.84), and among lifelong non-smokers

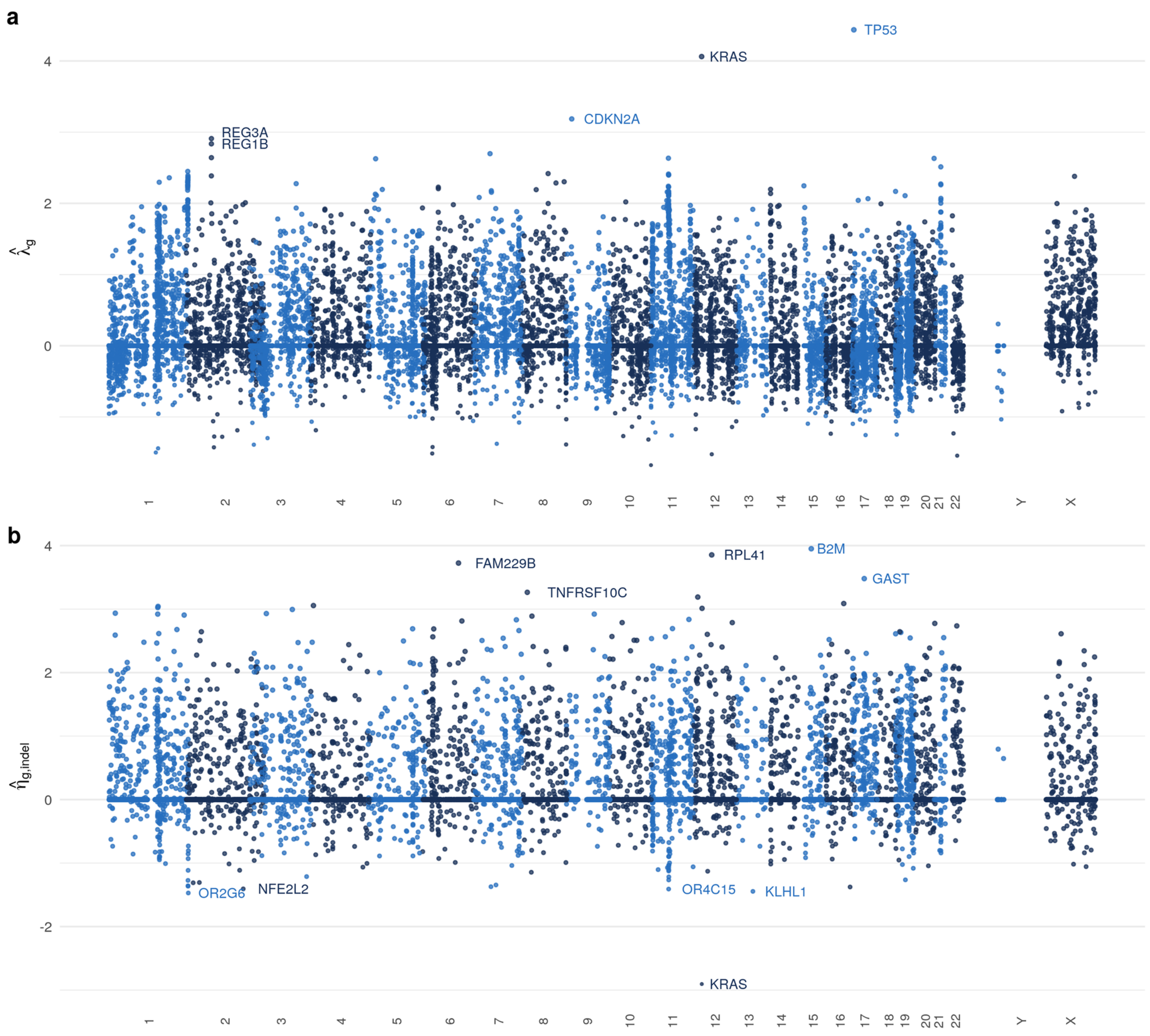

Fig. 2 Manhattan plots of generative model parameters. a Manhattan plot of fitted parameters $\hat{\lambda}_g$ and their associated genes' chromosomal locations. The genes with the five largest positive parameter estimates are labelled. **b** Manhattan plot of fitted parameters $\hat{\eta}_{g,\text{indel}}$ and their associated genes' chromosomal locations. The five largest positive and negative genes are labelled.

is 4.04 (1.12). This suggests that smokers may have higher BMRs, as would be expected.

Predicting tumour mutation burden. We now demonstrate the practical performance of our procedure for estimating TMB. First, it is shown that our method can indeed select gene panels of the size specified by the practitioner and those good predictions can be made even with small panel sizes (i.e. ≤1 Mb). We then compare the performance of our proposal with state-of-the-art estimation procedures based on a number of widely used gene panels.

In order to evaluate the predictive performance of an estimator we calculate the R^2 score on the validation data as follows: given predictions of TMB, $\hat{t}_1, \ldots, \hat{t}_{n_{val}}$, for the observations in the validation set with true TMB values $t_1, \ldots, t_{n_{val}}$. Let $\bar{t} := \frac{1}{n_{val}} \sum_{i=1}^{n_{val}} t_i$, and define

$$R^2 := 1 - \frac{\sum_{i=1}^{n_{val}} (t_i - \hat{t}_i)^2}{\sum_{i=1}^{n_{val}} (t_i - \bar{t})^2}.$$

Other works have aimed to classify tumours into two groups (high TMB, low TMB); see, for example, Buttner et al.[32] and Wu et al.[29]. Here we also report the estimated area under the precision-recall curve (AUPRC) for a classifier based on our estimator. We define the classifier as follows: first, in line with major clinical studies[33,34] the true class membership of a tumour is defined according to whether it has $t^* := 300$ or more exome mutations (~10 Mut/Mb). In the validation set, this gives 47(27.5%) tumours with high TMB and 124(72.5%) with low TMB. Now, for a cutoff $t \geq 0$, we can define a classifier by assigning a tumour to the high TMB class if its estimated TMB value is greater than or equal to t. For such a classifier, we have precision and recall (estimated over the validation set) given by

$$p(t) := \frac{\sum_{i=1}^{n_{val}} \mathbb{1}_{\{\hat{t}_i \geq t, t_i \geq t^*\}}}{\sum_{i=1}^{n_{val}} \mathbb{1}_{\{\hat{t}_i \geq t\}}} \quad \text{and} \quad r(t) := \frac{\sum_{i=1}^{n_{val}} \mathbb{1}_{\{\hat{t}_i \geq t, t_i \geq t^*\}}}{\sum_{i=1}^{n_{val}} \mathbb{1}_{\{t_i \geq t^*\}}},$$

respectively. The precision-recall curve then is $\{(r(t),\ p(t)) \colon t \in [0,\ \infty)\}$. Note that a perfect classifier achieves a AUPRC of 1, whereas a random guess, in this case, would have an average AUPRC of 0.275 (the prevalence of the high TMB class).

Now recall that TMB is given by equation (1) with $\bar{S}$ being the set of all non-synonymous mutation types. Thus to estimate TMB we apply our proposed estimator with $\bar{S} = S$, where the model parameters are estimated as described in our generative model section. In Supplementary Fig. 2, we present the R^2 and AUPRC for the first-fit and refitted estimators (see (6) and (8)) as the selected panel size varies from 0Mb to 2 Mb in length. We see that we obtain a more accurate prediction of TMB, both in terms of regression and classification, as the panel size increases, and that good estimation is possible even with very small panels (as low as 0.2 Mb). As expected, the refitted estimator slightly outperforms the first-fit estimator. We show good robustness of these results to permutations of the training set in Supplementary Fig. 3.

We now compare our method with state-of-the-art estimators applied to commonly used gene panels, as well as a panel selected by the proposal of Lyu et al.[27]. The three next-generation sequencing panels that we consider are chosen for their relevance to TMB. These are TST-170[35], Foundation One[36] and MSK-IMPACT[37]. Further, the panel selected by the approach in Lyu et al.[27] consists of the genes that are mutated more than 10% of the time, that are less than 0.015 Mb in length and for which the presence of a mutation in the gene is significantly associated with higher TMB values. For each panel $P \subseteq G$, we use four different methods to predict TMB:

(i) Our refitted estimator applied to the panel P: we estimate TMB using $T(\hat{w}_P)$, where $\hat{w}_P \in \mathrm{argmin}_{w \in W_P}\{f(w)\}$, and W_P is defined in (7).

(ii) Estimation and classification of tumour mutation burden (ecTMB): the procedure proposed by Yao et al.[24].

(iii) A count estimator: TMB is estimated by $\frac{\ell_G}{\ell_P}\sum_{g \in P}\sum_{s \in \bar{S}} M_{0gs}$, i.e. rescaling the mutation burden in the genes of P.

(iv) A linear model: we estimate TMB via ordinary least-squares linear regression of TMB against $\left\{\sum_{s \in S} M_{0gs} : g \in P\right\}$.

The latter three comprise existing methods for estimating TMB available to practitioners. The second (ecTMB), which is based on a negative binomial model, is state-of-the-art. The third is a standard practical procedure for the estimation of TMB from targeted gene panels. The fourth is the approach proposed by Lyu et al.[27]. The refitted estimator applied to panel P is also included here, in order to demonstrate the utility of our approach even with a prespecified panel.

We present the results of these comparisons in Fig. 3. First, for each of the four panels considered here, we see that our refitted estimator applied to the panel outperforms all existing approaches in terms of regression performance and that for smaller panels we are able to improve regression accuracy even further by selecting a panel (perhaps even of smaller size) based on the training data. For instance, in comparison to predictions based on the TST-170 panel, our procedure can achieve higher R^2 with a selected panel of half the size (with 0.2 Mb we obtain an R^2 of 0.78). The best available existing method based on the TST-170 panel, in this case, the linear estimator, has an R^2 of 0.74. Moreover, data-driven selection of panels considerably increases the classification performance for the whole range of panel sizes considered. In particular, even for the smallest panel size shown in Fig. 3 (~0.2 Mb), the classification performance of our method outperforms the best existing methodology applied to the MSK-IMPACT panel, despite being almost a factor of six times smaller. The full proposal of Lyu et al.[27], which involves applying the linear regression estimate to the panel selected as described above, also performs well here.

Finally, in this subsection, we demonstrate the practical performance of our method using the test set, which until this point has been held out. Based on the validation results above, we take the panel of size 0.6 Mb selected by our procedure and use our refitted estimator on that panel to predict TMB for the 173 samples in the test set. For comparison, we also present predictions from ecTMB, the count-based estimator and the linear regression estimator applied to the same panel. In Fig. 4 we see that our procedure performs well; we obtain an R^2 value (on the test data) of 0.85. The other methods have R^2 values of 0.67 (ecTMB), −36 (count) and 0.64 (linear regression). The count-based estimator here gives predictions which are reasonably well

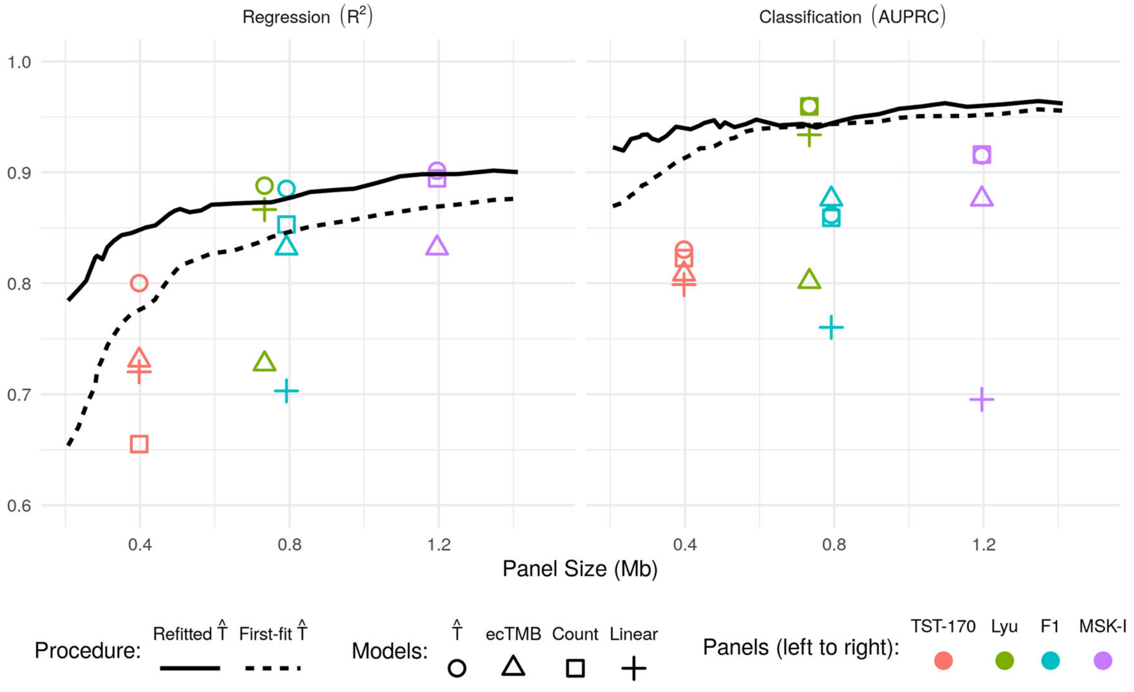

Fig. 3 Comparison with existing estimators. The performance of our TMB estimator in comparison to existing approaches. Left: R^2, Right: AUPRC.

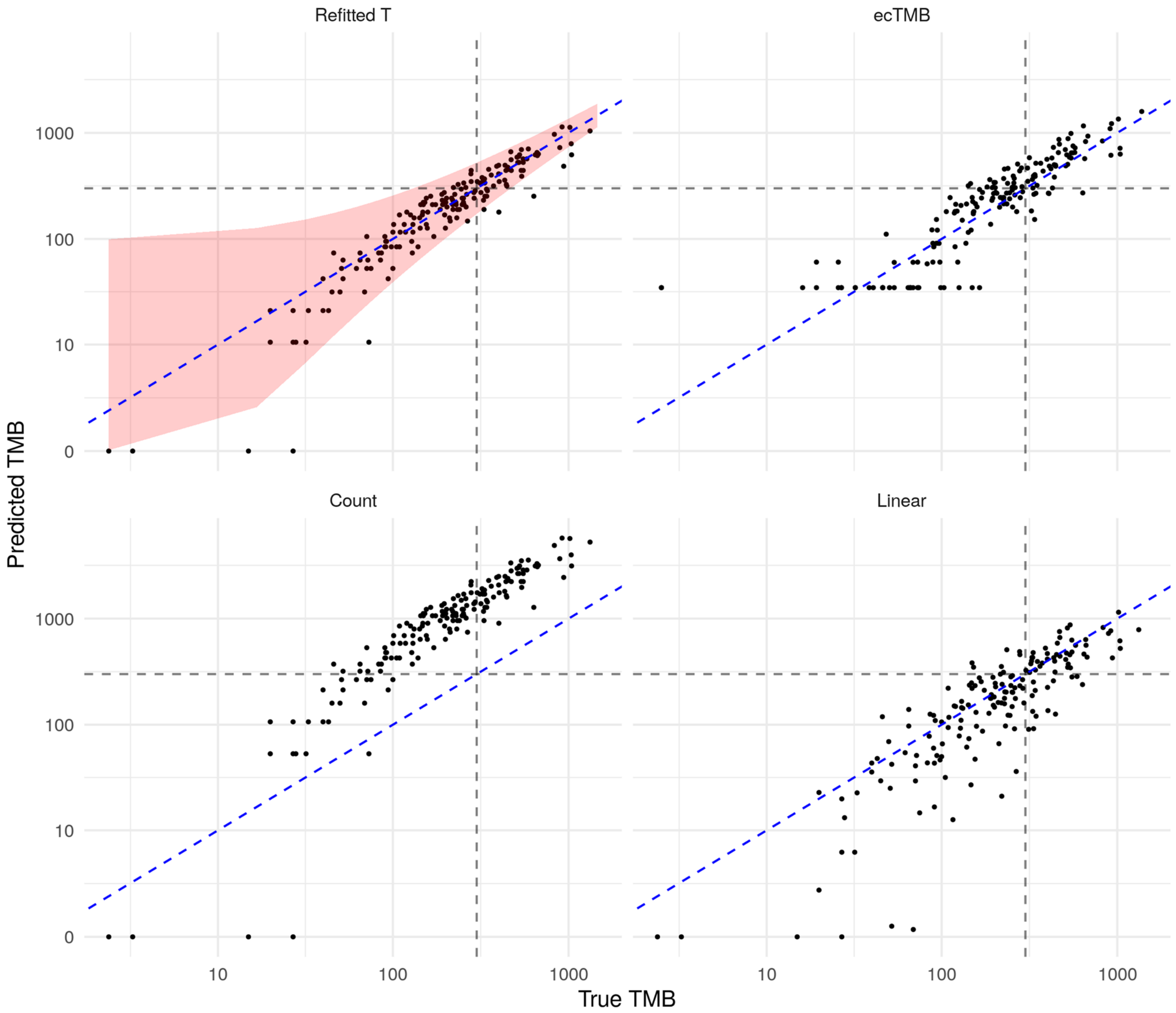

Fig. 4 Prediction of TMB on the test dataset. Dashed blue (diagonal) line represents perfect prediction and the black dashed lines indicate true and predicted TMB values of 300.

correlated to the true values of TMB but are positively biased. This is because our selection procedure tends to favour genes with higher overall mutation rates and thus a count estimator based on the highly mutated genes will overestimate the total number of mutations. We also include a red shaded region comprising all points for which heuristic 90% prediction intervals (as described in our Practical considerations section) include the true TMB value. We find in this case that 93.6% of the observations in the test set fall within this region, giving valid empirical coverage.

A panel-augmentation case study. We may wish to include genes from a given panel, but use our methodology to augment the panel to include additional genes with the goal of obtaining more accurate predictions of TMB (or other biomarkers). In this section we demonstrate how this can be done starting with the TST-170 panel (~0.4 Mb) and augmenting to 0.6 Mb in length, demonstrating impressive gains in predictive performance.

We apply the augmentation method described in the Panel augmentation methods section, with P_0 taken to be the set of TST-170 genes and Q_0 to be empty. The genes added to the panel are determined by the first-fit estimator in equation (9). To evaluate the performance, we then apply the refitted estimator on the test dataset, after selecting the augmented panel of size 0.6 Mb. For comparison, we apply our refitted estimator to the TST-170 panel directly. We also present the results obtained by the other estimators described above, both before and after the panel augmentation, in Table 1. We find that by augmenting the panel we improve predictive performance with our refitted $\hat{T}$ estimator, both in terms of regression and classification. The

Table 1 Predictive performance of models on TST-170 (0.4 Mb) versus augmented TST-170 (0.6 Mb) panels on the test set.

Model	Regression (R^2)		Classification (AUPRC)	
	TST-170	Aug. TST-170	TST-170	Aug. TST-170
Refitted $\hat{T}$	**0.58**	**0.84**	**0.84**	**0.94**
ecTMB	0.37	0.51	0.80	0.88
Count	0.18	0.18	0.83	**0.94**
Linear	0.47	0.74	0.78	0.89

Highest values for each column are in bold.

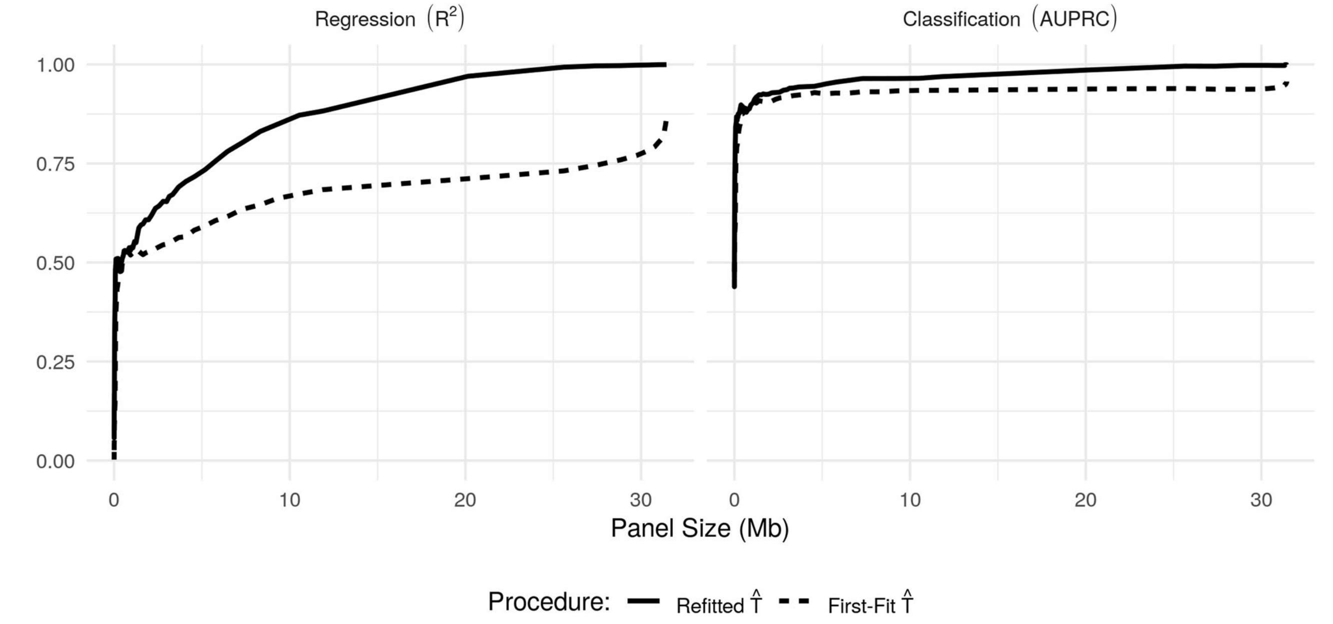

Fig. 5 Estimating TIB on the validation dataset. Performance of our first-fit and refitted estimators of TIB as the selected panel size varies. Left: R^2, Right: AUPRC.

refitted estimator provides better estimates than any other model on the augmented panel by both metrics.

Predicting tumour indel burden. In this section, we demonstrate how our method can be used to estimate TIB. This is more challenging than estimating TMB due to the low abundance of indel mutations relative to other variant types (see Fig. 1d), as well as issues involved in sequencing genomic loci of repetitive nucleotide constitution[38]. Indeed, in contrast to the previous section, we are not aware of any existing methods designed to estimate TIB from targeted gene panels. We, therefore, investigate the performance of our method across a much wider range (0–30 Mb) of panel sizes and find that we are able to accurately predict TIB with larger panels. Our results also demonstrate that accurate classification of TIB status is possible even with small gene panels.

We let S_{indel} be the set of all frameshift insertion and deletion mutations, and apply our method introduced in the methods section with $\bar{S} = S_{\text{indel}}$. As in the previous section, we assess regression and classification performance via R^2 and AUPRC, respectively, where in this case tumours are separated into two classes: high TIB (ten or more indel mutations) and low TIB (otherwise). In the validation dataset, this gives 57(33.3%) tumours in the high TIB class.

The results are presented in Fig. 5. We comment first on the regression performance: as expected, we see that the R^2 values for our first-fit and refitted estimators are much lower than what we achieved in estimating TMB. The refitted approach improves for larger panel sizes, while the first-fit estimator continues to perform relatively poorly. On the other hand, we see that the classification performance is impressive, with AUPRC values of above 0.8 for panels of less than 1 Mb in size.

We now assess the performance on the test set of our refitted estimator of TIB applied to a selected panel of size 0.6 Mb, and we compare with a count-based estimator and linear regression estimator. We do not compare with ecTMB here, since it is designed to estimate TMB as opposed to TIB. The count-based estimator, in this case, scales the total number of non-synonymous mutations across the panel by the ratio of the length of the panel to that of the entire exome, and also by the relative frequency of indel mutations versus all non-synonymous mutations in the training dataset:

$$\frac{\ell_G}{\ell_P}\frac{\sum_{i=1}^{n}\sum_{g\in G}\sum_{s\in S_{\text{indel}}}M_{igs}}{\sum_{i=1}^{n}\sum_{g\in G}\sum_{s\in S}M_{igs}}\sum_{g\in P}\sum_{s\in S}M_{0gs}.$$

In Fig. 6 we present the predictions on the test set of our refitted estimator ($R^2 = 0.35$); the count estimator ($R^2 = -44$); and the linear regression estimator ($R^2 = -0.15$). We also include (shaded in red) the set of points for which 90% prediction intervals contain the true value. In this case, we find that 97.7% of test set points fall within this region.

External testing and classification of response to immunotherapy. The aim of this section is to further test our proposed estimator of TMB by making use of two external NSCLC datasets for which the response to immunotherapy is available: Hellmann et al.[21], which contains 75 samples with an average TMB of 261; and Rizvi et al.[22], which contains 34 samples with an average TMB of 258.

We first use our refitted estimator trained on the same data as in the section above on predicting TMB to predict TMB for the samples in the new datasets using the selected panel of size 0.6 Mb. The predictions are given in Fig. 7a; the corresponding regression performance is $R^2 = 0.70$ across the two datasets, with a joint AUPRC for classifying tumours to high or low TMB classes of 0.91.

These datasets also allow us to assess the practical utility of using our estimated TMB values to predict response to immunotherapy. Of the 75 samples in the Hellman et al.[21] study, 37 were identified as having a *Durable Clinical Benefit* (Class 1) in response to immunotherapy (PD-L1+CTLA-4 blockade), and the remaining 38 were deemed to have *No Benefit* (Class 0). Of the 34 samples in the Rizvi et al.[22] studies, 14 were identified as having a *Durable clinical benefit beyond 6 months* (Class 1) in response to immunotherapy (Pembrolizumab), while the remaining 20 were deemed not to have such benefit (Class 0). Since the treatment and outcome definitions differ between studies, we separate them for analysis of response. We construct two simple classifiers for comparison, the first assigning a sample to Class 1 if

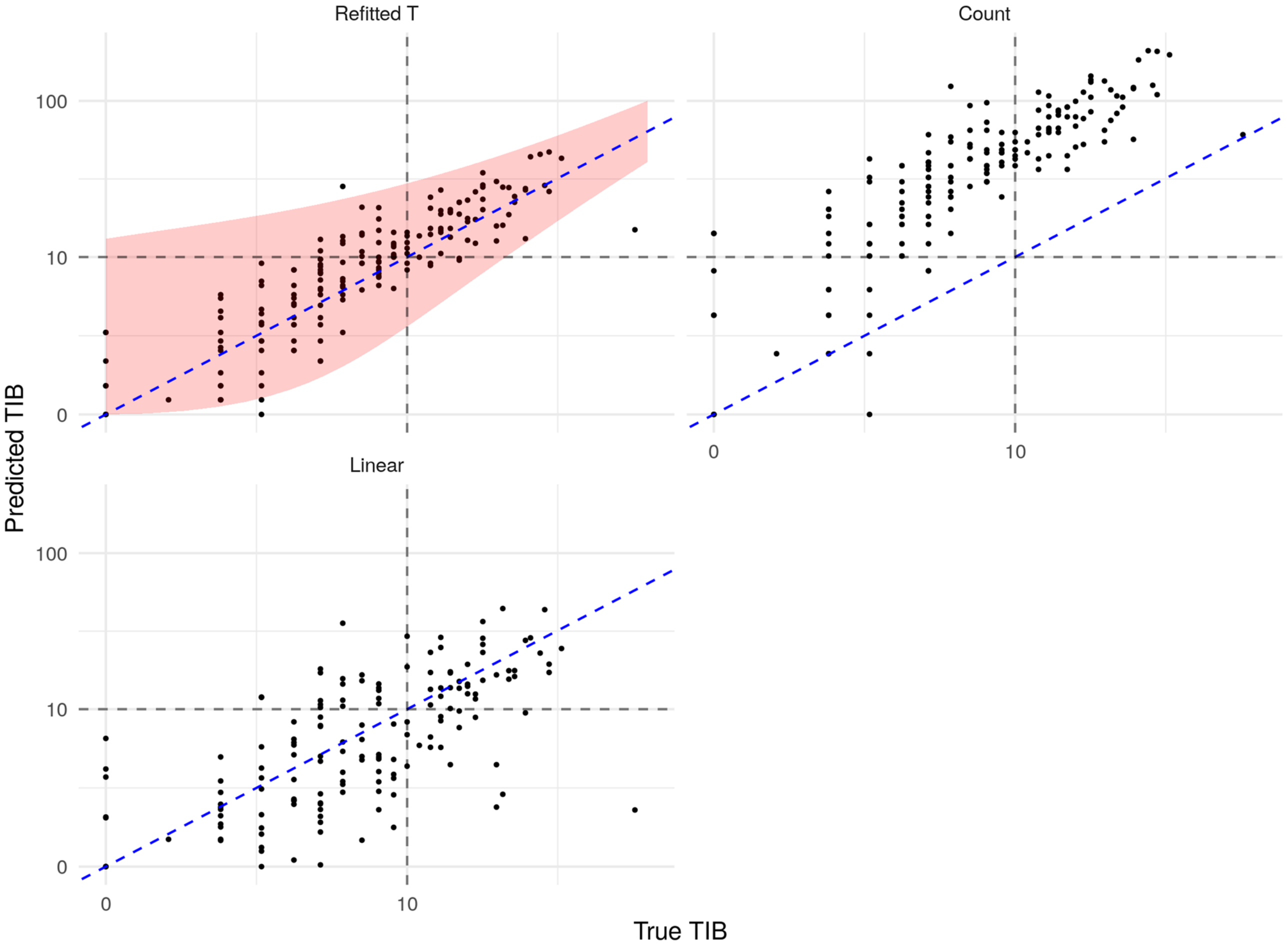

Fig. 6 Estimation of TIB on the test dataset. Dashed blue (diagonal) line represents perfect prediction and the grey dashed lines indicate true and predicted TIB values of 10.

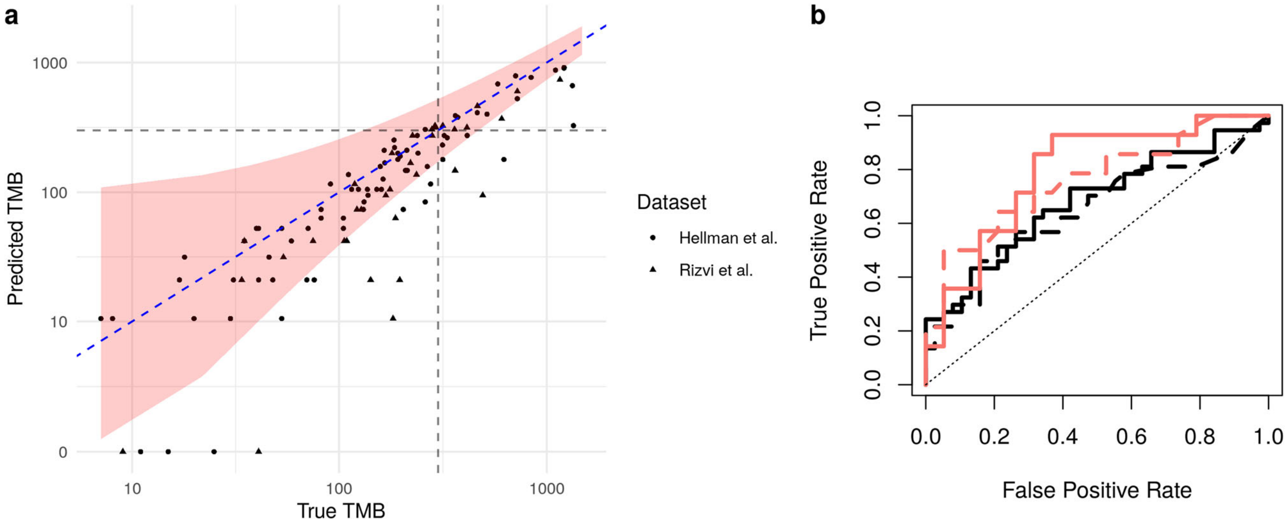

Fig. 7 External test set regression and classification performance. a Performance of our model trained on the Campbell et al.[20] dataset used to predict TMB based on the panel of size 0.6 Mb selected by our method on the external test datasets of Hellman et al.[21] and Rizvi et al.[22]. **b** ROC curves for classifying the response to immunotherapy in the Hellman et al.[21] (black) and Rizvi et al.[22] (red) datasets using the true TMB values (solid) and estimated TMB values (dashed) based on the panel of size 0.6 Mb selected by our method.

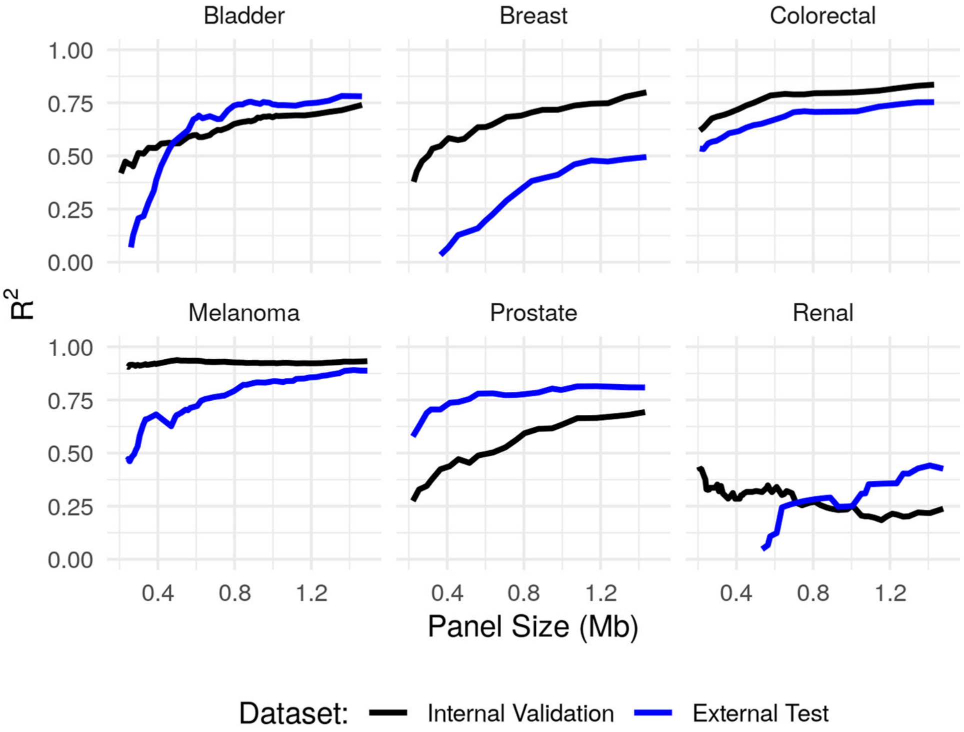

Fig. 8 Predictive performance on six further cancer types. The performance of our refitted TMB estimator in the six further cancer types.

the true TMB value is greater than some threshold *t*, and the second using our estimated value of TMB in the same way. In Fig. 7b, we plot the receiver operating characteristic (ROC) curve (that is the false positive rate against the true positive rate as the classification threshold *t* varies). The area under the ROC curve is 0.68 for the Hellman et al.[21] dataset when using the true TMB value and is 0.64 when using the estimated TMB value. Rizvi et al.[22] have an area under the ROC curve of 0.79 using true TMB values and 0.76 using estimated TMB values. We see that, in both cases, very little is lost in terms of predicting response to immunotherapy when using our estimated value of TMB.

Further testing in other cancer types. The aim of this section is to further demonstrate the performance of our proposed framework in a number of other cancer types. We apply our method for estimating TMB in six more cancer types, namely bladder cancer, breast cancer, colorectal cancer, melanoma, prostate cancer and renal cell cancer. For each cancer, data from two studies are used. Data from the first study is (randomly) split into a training and validation set; the training data is used to construct our estimator for a range of panel sizes, we then evaluate the predictive performance on the validation set (note that in contrast to our analysis above, we do not require a separate test set since the panel size is not selected based on the data). Further, in order to test the robustness of our approach to study effects, for each cancer type, we will also apply our fitted estimator (trained using data from the first study) to predict TMB values for tumours from the second study.

The twelve datasets[39–47] used are detailed in Supplementary Table 2. These datasets have a range of mutation rates, specifically the average TMB values in the training datasets are 247 (bladder cancer), 91 (breast cancer), 339 (colorectal cancer), 568 (melanoma), 63 (prostate cancer) and 77 (renal cell cancer).

In Fig. 8, the black lines plot the R^2 values obtained on the internal validation set from the first study for the six cancer types as the panel size varies from 0.25 Mb to 1.25 Mb. The blue lines show the R^2 values obtained when predicting TMB for tumours in the external test set from the second study. We see that the performance on the internal validation set is very good and broadly in line with the performance we obtained for the NSCLC dataset (with the exception of renal cell cancer). The main factor affecting the performance appears to be the overall mutation rate; our method performs very well in cancer types with large mutation rates (colorectal cancer and melanoma), but less well in the cancers with lower overall mutation rates (prostate and renal cell). The performance on the renal cell dataset is particularly poor due to the combination of low sample size and the low average mutation rate.

The results on the external test datasets are more mixed; there is a drop-off in performance in comparison with the internal validation results for breast cancer and melanoma, but apparent improvement for prostate cancer. This highlights that study effects, such as differences in patient demographics and clinical profiles, as well as variations in sequencing technologies need to be considered carefully. In practice, one should ensure that the patients in the training data used to fit the model have similar characteristics to the intended test cohort.

Conclusions

We have introduced a data-driven framework for designing targeted gene panels which allows for cost-effective estimation of exome-wide biomarkers. Using the non-small cell lung cancer datasets from Campbell et al.[20], Hellman et al.[21] and Rizvi et al.[22], we have demonstrated the excellent predictive performance of our proposal for estimating tumour mutation burden and tumour indel burden, and shown that it outperforms the state-of-the-art procedures. We further tested the applicability and robustness of our method, by applying it to datasets on several other cancer types. Our framework can be applied to any tumour dataset containing annotated mutations, and we provide an R package[19], which implements the methodology.

The main use of TMB is often to help identify patients that are more likely to respond to immunotherapy. While TMB is a good single predictor of response[10,11], it is of course desirable to improve the predictive performance by including other factors. For instance, these may include cancer type (and subtype), specific mutational signatures, aneuploidy and tumor histology, as well as other variables, such as gender, age and exogenous factors. Indeed, Litchfield et al.[48] show that, by including markers of T-cell infiltration and other factors, a multivariate predictor of response to immunotherapy significantly improves the classification performance in comparison to using TMB alone. Nevertheless, one would certainly like to include TMB (or a closely related measure) as a factor in any classifier of response.

Our work also has the scope to help understand mutational processes. For example, the parameters of our fitted model have interesting interpretations: of the five genes highlighted in Fig. 2a as having the highest mutation rates relative to the BMR, two (*TP53, CDKN2A*) are known tumour suppressors[49,50] and *KRAS* is an oncogene[51]. Furthermore, indel mutations in *KRAS* are known to be deleterious for tumour cells[52]—in our work, the *KRAS* gene has a large negative indel-specific parameter (see Fig. 2b). Our methodology identifies a number of other genes with large parameter estimates. Of course, any such associations need to be carefully investigated in follow-up studies.

Finally, we believe there are many ways in which our general framework can be extended. For example, it may be adapted to incorporate alternate data types (e.g. transcriptomics); we may seek to predict other features (e.g. outcomes such as survival); or we may wish to extend the method to incorporate multiple potentially incomplete data sources.

Methods

Generative model. We now describe the main statistical model that underpins our methodology. In order to account for selective pressures and other factors within the tumour, we allow the rate at which mutations occur to depend on the gene and type of mutation. Our model also includes a sample-dependent parameter to account for the differing levels of mutagenic exposure of tumours, which may occur due to exogenous (e.g. UV light, cigarette smoke) or endogenous (e.g. inflammatory, free radical) factors.

We model the mutation counts M_{igs} as independent Poisson random variables with mutation rates $\phi_{igs} > 0$. More precisely, for $i = 0, 1, \ldots, n$, $g \in G$ and $s \in S$, we have

$$M_{igs} \sim \text{Poisson}(\phi_{igs}), \tag{2}$$

where M_{igs} and $M_{i'g's'}$ are independent for $(i, g, s) \neq (i', g', s')$. Further, to model the dependence of the mutation rate on the sample, gene and mutation type, we use a log link function and let

$$\log(\phi_{igs}) = \mu_i + \log(\ell_g) + \lambda_g + \nu_s + \eta_{gs}, \tag{3}$$

for $\mu_i, \lambda_g, \nu_s, \eta_{gs} \in \mathbb{R}$, where for identifiability we set $\eta_{gs_1} = 0$, for some $s_1 \in S$ and all $g \in G$.

The terms in our model can be interpreted as follows. First, the parameter μ_i corresponds to the BMR of the ith sample. The offset $\log(\ell_g)$ accounts for a mutation rate that is proportional to the length of a gene, so that a non-zero value of λ_g corresponds to an increased or decreased mutation rate relative to the BMR. The parameters ν_s and η_{gs} account for differences in frequency between mutation types for each gene.

The model in (2) and (3) (discounting the unseen test sample $i = 0$) has $n + |S| + |G||S|$ free parameters and we have $n|G||S|$ independent observations in the training dataset. In principle, we could attempt to fit our model directly using maximum likelihood estimation. However, we wish to exploit the fact that most genes do not play an active role in the development of a tumour, and will be mutated approximately according to the BMR. This corresponds to the parameters λ_g and η_{gs} being zero for many $g \in G$. We, therefore, include an ℓ_1-penalisation term applied to the parameters λ_g and η_{gs} when fitting our model. We do not penalise the parameters ν_s or μ_i since we expect that different mutation types occur at different rates and that the BMR is different in each sample.

Writing $\mu := (\mu_1, \ldots, \mu_n)$, $\lambda := (\lambda_g\colon g \in G)$, $\nu := (\nu_s\colon s \in S)$ and $\eta := (\eta_{gs}\colon g \in G, s \in S)$, and given training observations $M_{igs} = m_{igs}$, we let

$$\mathcal{L}(\mu, \lambda, \nu, \eta) = \sum_{i=1}^{n} \sum_{g \in G} \sum_{s \in S} \Big(\phi_{igs} - m_{igs} \log \phi_{igs} \Big)$$

be the negative log-likelihood of the model specified by (2) and (3). We then define

$$(\hat{\mu}, \hat{\lambda}, \hat{\nu}, \hat{\eta}) = \underset{\mu,\lambda,\nu,\eta}{\operatorname{argmin}} \left\{ \mathcal{L}(\mu, \lambda, \nu, \eta) + \kappa_1 \left(\sum_{g \in G} |\lambda_g| + \sum_{g \in G} \sum_{s \in S} |\eta_{gs}| \right) \right\}, \tag{4}$$

where $\kappa_1 \geq 0$ is a tuning parameter that controls the number of non-zero components in $\hat{\lambda}$ and $\hat{\eta}$, which we choose using cross-validation.

Proposed estimator. We now attend to our main goal of estimating a given exome-wide biomarker for the unseen test sample. Fix $\bar{S} \subseteq S$ and recall that we write $T = T_{0\bar{S}}$. We wish to construct an estimator of T that only depends on the mutation counts in a gene panel $P \subset G$, subject to a constraint on ℓ_P. To that end, we consider estimators of the form

$$T(w) := \sum_{g \in G} \sum_{s \in S} w_{gs} M_{0gs},$$

for $w \in \mathbb{R}^{|G| \times |S|}$. Note that our estimator may use the full set S of variant types, rather than just those in $\bar{S}$. In other words, our estimator may utilise information from every mutation type, not just those that directly constitute the biomarker of interest. This is important when estimating mutation types in $\bar{S}$ that are relatively scarce (e.g. for TIB). In the remainder of this subsection, we explain how the weights w are chosen to minimise the expected squared error of $T(w)$ based on the generative model described in the previous section.

Of course, setting $w_{gs} = 1$ for $g \in G$ and $s \in \bar{S}$ (and $w_{gs} = 0$ otherwise) will give $T(w) = T$. However, our aim is to make predictions based on a concise gene panel. If, for a given $g \in G$, we have $w_{gs} = 0$ for all $s \in S$, then $T(w)$ does not depend on the mutations in g and therefore the gene does not need to be included in the panel. In order to produce a suitable gene panel (i.e. with many $w_{gs} = 0$), we penalise non-zero components of w when minimising the expected squared error. We define our final estimator via a refitting procedure, which improves the predictive performance by reducing the bias, and is also helpful when applying our procedure to panels with predetermined genes.

To construct our estimator, note that under our model in (2) we have $\mathbb{E}M_{0gs} = \text{Var}(M_{0gs}) = \phi_{0gs}$, and it follows that the expected squared error of $T(w)$ is

$$\begin{aligned} \mathbb{E}\left[\{T(w) - T\}^2\right] &= \text{Var}(T(w)) + \text{Var}(T) - 2\text{Cov}(T(w), T) + [\mathbb{E}\{T(w) - T\}]^2 \\ &= \sum_{g \in G} \sum_{s \in \bar{S}} (1 - w_{gs})^2 \phi_{0gs} + \sum_{g \in G} \sum_{s \in S \setminus \bar{S}} w_{gs}^2 \phi_{0gs} \\ &\quad + \left(\sum_{g \in G} \sum_{s \in S} w_{gs} \phi_{0gs} - \sum_{g \in G} \sum_{s \in \bar{S}} \phi_{0gs} \right)^2. \end{aligned} \tag{5}$$

This depends on the unknown parameters μ_0, λ_g, ν_s and η_{gs}, the latter three of which are replaced by their estimates given in (4). It is also helpful to then rescale (5) as follows: write $\hat{\phi}_{0gs} = \ell_g \exp(\hat{\lambda}_g + \hat{\nu}_s + \hat{\eta}_{gs})$, and define

$$p_{gs} := \frac{\hat{\phi}_{0gs}}{\sum_{g' \in G} \sum_{s' \in \bar{S}} \hat{\phi}_{0g's'}} = \frac{\ell_g \exp(\hat{\lambda}_g + \hat{\nu}_s + \hat{\eta}_{gs})}{\sum_{g' \in G} \sum_{s' \in \bar{S}} \ell_{g'} \exp(\hat{\lambda}_{g'} + \hat{\nu}_{s'} + \hat{\eta}_{g's'})}.$$

Then let

$$f(w) := \sum_{g \in G} \sum_{s \in \bar{S}} p_{gs} (1 - w_{gs})^2 + \sum_{g \in G} \sum_{s \in S \setminus \bar{S}} p_{gs} w_{gs}^2 + K(\mu_0) \left(1 - \sum_{g \in G} \sum_{s \in S} p_{gs} w_{gs} \right)^2,$$

where $K(\mu_0) = \exp(\mu_0) \sum_{g \in G} \sum_{s \in \bar{S}} \ell_g \exp(\hat{\lambda}_g + \hat{\nu}_s + \hat{\eta}_{gs})$. Since f is a rescaled version of the error in (5) (with the true parameters λ, ν, η replaced by the estimates $\hat{\lambda}, \hat{\nu}, \hat{\eta}$), we will choose w to minimise $f(w)$.

Note that f only depends on μ_0 via the $K(\mu_0)$ term, which can be interpreted as a penalty factor controlling the bias of our estimator. For example, we may insist that the squared bias term $(1 - \sum_{g \in G} \sum_{s \in S} p_{gs} w_{gs})^2$ is zero by setting $K(\mu_0) = \infty$. In practice, we propose to choose the penalty K based on the training data.

At this point $f(w)$ is minimised by choosing w to be such that $w_{gs} = 1$ for all $g \in G, s \in \bar{S}$, and $w_{gs} = 0$ otherwise. As mentioned above, in order to form a concise panel while optimising predictive performance, we impose a constraint on the cost of sequencing the genes used in the estimation. More precisely, for a given w, an appropriate cost is

$$\| w \|_{G,0} := \sum_{g \in G} \ell_g \mathbb{1}\{w_{gs} \neq 0 \text{ for some } s \in S\}.$$

This choice acknowledges that the cost of a panel is roughly proportional to the length of the region of genomic space sequenced, and that once a gene has been sequenced for one mutation type there is no need to sequence again for other mutation types.

Now, given a cost restriction L, our goal is to minimise $f(w)$ such that $\|w\|_{G,0} \leq L$. In practice, this problem is non-convex and so computationally infeasible. As is common in high-dimensional optimisation problems, we consider a convex relaxation as follows: let $\|w\|_{G,1} := \sum_{g \in G} \ell_g \|w_g\|_2$, where $w_g = (w_{gs} : s \in S) \in \mathbb{R}^{|S|}$, for $g \in G$, and $\| \cdot \|_2$ is the Euclidean norm. Define

$$\hat{w}^{\text{first-fit}} \in \underset{w}{\operatorname{argmin}} \Big\{ f(w) + \kappa_2 \| w \|_{G,1} \Big\}, \tag{6}$$

where $\kappa_2 \geq 0$ is chosen to determine the size of the panel selected.

The final form of our estimator is obtained by a refitting procedure. First, for $P \subseteq G$, let

$$W_P := \{w \in \mathbb{R}^{|G| \times |S|} : w_g = (0, \ldots, 0) \text{ for } g \in G \setminus P\}. \tag{7}$$

Let $\hat{P} := \{g \in G : \| \hat{w}_g^{\text{first-fit}} \|_2 > 0\}$ be the panel selected by the first-fit estimator in (6), and define

$$\hat{w}^{\text{refit}} \in \underset{w \in W_{\hat{P}}}{\operatorname{argmin}}\{f(w)\}. \tag{8}$$

We then estimate T using $\hat{T} := T(\hat{w}^{\text{refit}})$, which only depends on mutations in genes contained in the selected panel $\hat{P}$.

Panel augmentation. In practice, when designing gene panels a variety of factors contribute to the choice of genes included. For example, a gene may be included due to its relevance to immune response or its known association with a particular cancer type. If this is the case, measurements for these genes will be made regardless of their utility for predicting exome-wide biomarkers. When implementing our methodology, therefore, there is no additional cost to incorporate observations from these genes into our prediction if they will be helpful. Conversely, researchers may wish to exclude genes from a panel, or at least from actively contributing to the estimation of a biomarker, for instance, due to technical difficulties in sequencing a particular gene.

We can accommodate these restrictions by altering the structure of our regularisation penalty in (6). Suppose we are given (disjoint sets of genes) P_0, $Q_0 \subseteq G$ to be included and excluded from our panel, respectively. In this case, we replace $\hat{w}^{\text{first-fit}}$ in (6) with

$$\hat{w}_{P_0,Q_0}^{\text{first-fit}} \in \underset{w \in W_{G \setminus Q_0}}{\operatorname{argmin}}\left\{ f(w) + \kappa_2 \sum_{g \in G \setminus P_0} l_g \| w_g \|_2 \right\}. \tag{9}$$

Excluding the elements of P_0 from the penalty term means that $\hat{w}_{P_0,Q_0}^{\text{first-fit}} \neq 0$ for the genes in P_0, while restricting our optimisation to $W_{G \setminus Q_0}$ excludes the genes in Q_0 by definition. This has the effect of augmenting the predetermined panel P_0 with additional genes selected to improve predictive performance. We then perform refitting as described above. We demonstrate this procedure by augmenting the TST-170 gene panel.

Practical considerations. In this section, we discuss some practical aspects of our proposal. Our first consideration concerns the choice of the tuning parameter κ_1 in (4). As is common for the Least Absolute Shrinkage and Selection Operator (LASSO) estimator in generalised linear regression (see, for example, Michoel[53] and Friedman et al.[54]), we will use tenfold cross-validation. To highlight one important aspect of our cross-validation procedure, recall that we consider the observations M_{igs} as independent across the sample index $i \in \{1, \ldots, n\}$, the gene $g \in G$ and the mutation type $s \in S$. Our approach, therefore, involves splitting the entire set $\{(i, g, s): i = 1, \ldots, n, g \in G, s \in S\}$ of size $n|G||S|$ (as opposed to the sample set $\{1, \ldots, n\}$) into tenfolds uniformly at random. We then apply the estimation method in (4) to each of the tenfolds separately on a grid of values (on the log scale) of κ_1, and select the value that results in the smallest average deviance across the folds. The model is then refitted using all the data for this value of κ_1.

The estimated coefficients in (6) depend on the choice of $K(\mu_0)$ and κ_2. As mentioned above, we could set $K(\mu_0) = \infty$ to give an unbiased estimator, however, in practice, we found that a finite choice of $K(\mu_0)$ leads to improved predictive performance. Our recommendation is to use $K(\mu_0) = K(\max_{i=1,\ldots,n}\{\hat{\mu}_i\})$, where $\hat{\mu}_i = \log(T_i / \sum_{g,s} \ell_g \exp(\hat{\lambda}_g + \hat{\nu}_s + \hat{\eta}_{gs}))$ is a pseudo-MLE (in the sense of Gong and Samaniego[55]) for μ_i, so that the penalisation is broadly in proportion with the largest values of μ_i in the training dataset. The tuning parameter κ_2 controls the size of the gene panel selected in (6): given a panel length L, we set $\kappa_2(L) = \max\{\kappa_2 : \ell_{\hat{P}} \leq L\}$ in order to produce a suitable panel.

We now comment briefly on some computational aspects of our method. The generative model fit in (4) can be solved via coordinate descent (see, for example, Friedman et al.[56]), which has a computational complexity of $O(N|G|^2|S|^2)$ per iteration. We fit the model ten times, one for each fold in our cross-validation procedure. This is the most computationally demanding part of our proposal—in our experiments below, it takes approximately an hour to solve on a laptop—but it only needs to be carried out once for a given dataset. The convex optimisation problem in (6) can be solved by any method designed for the group LASSO; see, for example, Yang and Zou[57]. In our experiments, we use the `gglasso` R package[58], which takes around 10 min to reproduce the plot in Supplementary Fig. 2. Note also that the solutions to (6) and (8) are unique; see, for example, Theorem 1 of Roth and Fischer[59]. The last step of our proposal, namely making predictions for new test observations based on a selected panel, carries a negligible computational cost.

Finally, we describe a heuristic procedure for producing prediction intervals around our point estimates. In particular, for a given confidence level $\alpha \in (0, 1)$, we aim to find an interval $[\hat{T}_{\mathrm{L}}, \hat{T}_{\mathrm{U}}]$ such that $\mathbb{P}(\hat{T}_{\mathrm{L}} \leq T \leq \hat{T}_{\mathrm{U}}) \geq 1 - \alpha$. To that end, let $t_\alpha := \mathbb{E}\{(\hat{T} - T)^2\}/\alpha$, then by Markov's inequality we have that $\mathbb{P}(|\hat{T} - T|^2 \geq t_\alpha) \leq \alpha$. It follows that $[\hat{T} - t_\alpha^{1/2}, \hat{T} + t_\alpha^{1/2}]$ is a $(1 - \alpha)$-prediction interval for T. Of course, the mean squared error $\mathbb{E}\{(\hat{T} - T)^2\}$ defined in (5) depends on the parameters λ, η, ν and μ_0, which are unknown. Our approach is to utilise the estimates $\hat{\lambda}, \hat{\eta}, \hat{\nu}$ (see (4)) and replace μ_0 with $\log(\hat{T} / \sum_{g,s} \ell_g \exp(\hat{\lambda}_g + \hat{\nu}_s + \hat{\eta}_{gs}))$. While this is not an exact $(1 - \alpha)$-prediction interval for T, we see in our experimental results that in practice this approach provides intervals with valid empirical coverage.

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

Data availability

All data used in this manuscript are publicly available. The NSCLC dataset of Campbell et al.[20] and the *Ensembl* gene length dataset are available as part of our `R` package `ICBioMark`[19] - see below for more detail. The BED files for the gene panels can be downloaded from https://github.com/cobrbra/TargetedPanelEstimation_Paper[60], while data citations for the six further cancer types are given in Supplementary Table 2.

Code availability

All figures and tables in this manuscript may be reproduced using the code available at the DOI-minted repository https://github.com/cobrbra/TargetedPanelEstimation_Paper[60]. We also provide an open-access `R` package `ICBioMark`[19], which is available on CRAN https://cran.r-project.org. Alternatively, the package may be accessed and downloaded at https://github.com/cobrbra/ICBioMark.

Received: 15 November 2021; Accepted: 31 January 2022;
Published online: 23 February 2022

References

1. Boveri, T. Concerning the origin of malignant tumours. Translated and annotated by Henry Harris. *J. Cell Sci.* **121**, 1–84 (2008).
2. Ishida, Y., Agata, Y., Shibahara, K. & Honjo, T. Induced expression of PD-1, a novel member of the immunoglobulin gene superfamily, upon programmed cell death. *EMBO J.* **11**, 3887–3895 (1992).
3. Leach, D. R., Krummel, M. F. & Allison, J. P. Enhancement of antitumor immunity by CTLA-4 blockade. *Science* **271**, 1734–1736 (1996).
4. Ledford, H., Else, H. & Warren, M. Cancer immunologists scoop medicine Nobel prize. *Nature* **562**, 20–21 (2018).
5. Robert, C. A decade of immune-checkpoint inhibitors in cancer therapy. *Nat. Commun.* **11**, 3801 (2020).
6. Buchbinder, E. I. & Desai, A. CTLA-4 and PD-1 pathways: similarities, differences, and implications of their inhibition. *Am. J. Clin. Oncol.* **39**, 98–106 (2016).
7. Pardoll, D. M. The blockade of immune checkpoints in cancer immunotherapy. *Nat. Rev. Cancer* **12**, 252–264 (2012).
8. Borghaei, H. et al. Five-year outcomes from the randomized, phase III trials checkMate 017 and 057: nivolumab versus docetaxel in previously treated non-small-cell lung cancer. *J. Clin. Oncol.* **39**, 723–733 (2021).
9. Nowicki, T. S., Hu-Lieskovan, S. & Ribas, A. Mechanisms of resistance to PD-1 and PD-L1 blockade. *Cancer J.* **24**, 47–53 (2018).
10. Zhu, J. et al. Association between tumor mutation burden (TMB) and outcomes of cancer patients treated with PD-1/PD-L1 inhibitions: a meta-analysis. *Front. Pharmacol.* **10**, 673 (2019).
11. Cao, D., Xu, H., Xu, X., Guo, T. & Ge, W. High tumor mutation burden predicts better efficacy of immunotherapy: a pooled analysis of 103078 cancer patients. *Oncoimmunology* **8**, e1629258 (2019).
12. Chan, T. A. et al. Development of tumor mutation burden as an immunotherapy biomarker: utility for the oncology clinic. *Ann. Oncol.* **30**, 44–56 (2019).
13. Sboner, A., Mu, X. J., Greenbaum, D., Auerbach, R. K. & Gerstein, M. B. The real cost of sequencing: higher than you think! *Genome Biol.* **12**, 125 (2011).
14. Gandara, D. R. et al. Blood-based tumor mutational burden as a predictor of clinical benefit in non-small-cell lung cancer patients treated with atezolizumab. *Nat. Med.* **24**, 1441–1448 (2018).
15. Wu, H.-X. et al. Tumor mutational and indel burden: a systematic pan-cancer evaluation as prognostic biomarkers. *Ann. Transl. Med.* **7**, 640 (2019).
16. Turajlic, S. et al. Insertion-and-deletion-derived tumour-specific neoantigens and the immunogenic phenotype: a pan-cancer analysis. *Lancet Oncol.* **18**, 1009–1021 (2017).
17. Fancello, L., Gandini, S., Pelicci, P. G. & Mazzarella, L. Tumor mutational burden quantification from targeted gene panels: major advancements and challenges. *J. Immunother. Cancer* **7**, 183 (2019).
18. Golkaram, M., Zhao, C., Kruglyak, K., Zhang, S. & Bilke, S. The interplay between cancer type, panel size and tumor mutational burden threshold in patient selection for cancer immunotherapy. *PLOS Comput. Biol.* **16**, e1008332 (2020).

19. Bradley, J. R. & Cannings, T. I. ICBioMark: data-driven design of targeted gene panels for estimating immunotherapy biomarkers https://CRAN.R-project.org/package=ICBioMark (2021).
20. Campbell, J. D. et al. Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. *Nat. Genet.* **48**, 607–616 (2016).
21. Hellmann, M. D. et al. Genomic features of response to combination immunotherapy in patients with advanced non-small-cell lung cancer. *Cancer Cell* **33**, 843–852.e4 (2018).
22. Rizvi, N. A. et al. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science* **348**, 124–128 (2015).
23. Budczies, J. et al. Optimizing panel-based tumor mutational burden (TMB) measurement. *Ann. Oncol.* **30**, 1496–1506 (2019).
24. Yao, L., Fu, Y., Mohiyuddin, M. & Lam, H. Y. K. ecTMB: a robust method to estimate and classify tumor mutational burden. *Sci. Rep.* **10**, 4983 (2020).
25. Makova, K. D. & Hardison, R. C. The effects of chromatin organization on variation in mutation rates in the genome. *Nat. Rev. Genet.* **16**, 213–223 (2015).
26. Fong, Y. W., Cattoglio, C. & Tjian, R. The intertwined roles of transcription and repair proteins. *Mol. Cell* **52**, 291–302 (2013).
27. Lyu, G.-Y., Yeh, Y.-H., Yeh, Y.-C. & Wang, Y.-C. Mutation load estimation model as a predictor of the response to cancer immunotherapy. *NPJ Genom. Med.* **3**, 12 (2018).
28. Guo, W. et al. An exon signature to estimate the tumor mutational burden of right-sided colon cancer patients. *J. Cancer* **11**, 883–892 (2020).
29. Wu, H.-X., Wang, Z.-X., Zhao, Q., Wang, F. & Xu, R.-H. Designing gene panels for tumor mutational burden estimation: the need to shift from ’correlation’ to ’accuracy’. *J. Immunother. Cancer* **7**, 206 (2019).
30. Chalmers, Z. R. et al. Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. *Genome Med.* **9**, 34 (2017).
31. Yates, A. D. et al. Ensembl 2020. *Nucleic Acids Res.* **48**, D682–D688 (2020).
32. Büttner, R. et al. Implementing TMB measurement in clinical practice: considerations on assay requirements. *ESMO Open* **4**, e000442 (2019).
33. Hellmann, M. D. et al. Nivolumab plus Ipilimumab in lung cancer with a high tumor mutational burden. *N. Engl. J. Med.* **378**, 2093–2104 (2018).
34. Ramalingam, S. S. et al. Tumor mutational burden (TMB) as a biomarker for clinical benefit from dual immune checkpoint blockade with nivolumab (nivo) + ipilimumab (ipi) in first-line (1L) non-small cell lung cancer (NSCLC): identification of TMB cutoff from CheckMate 568. *Cancer Res.* **78**, CT078–CT078 (2018).
35. Heydt, C. et al. Evaluation of the TruSight tumor 170 (TST170) assay and its value in clinical research. *Ann. Oncol.* **29**, vi7–vi8 (2018).
36. Frampton, G. M. et al. Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. *Nat. Biotechnol.* **31**, 1023–1031 (2013).
37. Cheng, D. T. et al. Memorial Sloan Kettering-integrated mutation profiling of actionable cancer targets (MSK-IMPACT): a hybridization capture-based next-generation sequencing clinical assay for solid tumor molecular oncology. *J. Mol. Diag.* **17**, 251–264 (2015).
38. Narzisi, G. & Schatz, M. C. The challenge of small-scale repeats for indel discovery. *Front. Bioeng. Biotechnol.* **3**, 8 (2015).
39. Guo, G. et al. Whole-genome and whole-exome sequencing of bladder cancer identifies frequent alterations in genes involved in sister chromatid cohesion and segregation. *Nat. Genet.* **45**, 1459–1463 (2013).
40. Kan, Z. et al. Multi-omics profiling of younger Asian breast cancers reveals distinctive molecular signatures. *Nat. Commun.* **9**, 1725 (2018).
41. Giannakis, M. et al. Genomic correlates of immune-cell infiltrates in colorectal carcinoma. *Cell Rep.* **15**, 857–865 (2016).
42. Seshagiri, S. et al. Recurrent R-spondin fusions in colon cancer. *Nature* **488**, 660–664 (2012).
43. Cancer Genome Atlas Network. Genomic classification of cutaneous melanoma. *Cell* **161**, 1681–1696 (2015).
44. Krauthammer, M. et al. Exome sequencing identifies recurrent somatic RAC1 mutations in melanoma. *Nat. Genet.* **44**, 1006–1014 (2012).
45. Armenia, J. et al. The long tail of oncogenic drivers in prostate cancer. *Nat. Genet.* **50**, 645–651 (2018).
46. Kumar, A. et al. Substantial interindividual and limited intraindividual genomic diversity among tumors from men with metastatic prostate cancer. *Nat. Med.* **22**, 369–378 (2016).
47. Guo, G. et al. Frequent mutations of genes encoding ubiquitin-mediated proteolysis pathway components in clear cell renal cell carcinoma. *Nat. Genet.* **44**, 17–19 (2011).
48. Litchfield, K. et al. Meta-analysis of tumor- and T cell-intrinsic mechanisms of sensitization to checkpoint inhibition. *Cell* **184**, 596–614.e14 (2021).
49. Olivier, M., Hollstein, M. & Hainaut, P. TP53 mutations in human cancers: origins, consequences, and clinical use. *Cold Spring Harb. Perspect. Biol.* **2**, a001008 (2010).
50. Foulkes, W. D., Flanders, T. Y., Pollock, P. M. & Hayward, N. K. The CDKN2A (p16) gene and human cancer. *Mol. Med.* **3**, 5–20 (1997).
51. Jancík, S., Drábek, J., Radzioch, D. & Hajdúch, M. Clinical relevance of KRAS in human cancers. *J. Biomed. Biotechnol.* **2010**, 150960 (2010).
52. Lee, W., Lee, J. H., Jun, S., Lee, J. H. & Bang, D. Selective targeting of KRAS oncogenic alleles by CRISPR/Cas9 inhibits proliferation of cancer cells. *Sci. Rep.* **8**, 11879 (2018).
53. Michoel, T. Natural coordinate descent algorithm for L1-penalised regression in generalised linear models. *Comput. Stat. Data Anal.* **97**, 60–70 (2016).
54. Friedman, J. et al. glmnet: Lasso and elastic-net regularized generalized linear models. https://CRAN.R-project.org/package=glmnet (2021).
55. Gong, G. & Samaniego, F. J. Pseudo maximum likelihood estimation: theory and applications. *Ann. Stat.* **9**, 861–869 (1981).
56. Friedman, J. H., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–22 (2010).
57. Yang, Y. & Zou, H. A fast unified algorithm for solving group-lasso penalize learning problems. *Stat. Comput.* **25**, 1129–1141 (2015).
58. Yang, Y., Zou, H. & Bhatnagar, S. gglasso: group lasso penalized learning using a unified BMD algorithm https://CRAN.R-project.org/package=gglasso (2020).
59. Roth, V. & Fischer, B. The group-lasso for generalized linear models: uniqueness of solutions and efficient algorithms. In *Proc. 25th international conference on Machine learning, ICML '08, 848-855* (Association for Computing Machinery, 2008).
60. cobrbra. https://doi.org/10.5281/zenodo.5796134 (2021).

Acknowledgements

We are grateful for the constructive feedback from the anonymous reviewers, which helped to improve the paper. We gratefully acknowledge funding provided by Cambridge Cancer Genomics (CCG) through their Ph.D. Scholarship at the University of Edinburgh. We also benefited from discussions with several individuals, including Adnan Akbar, Philip Beer, Harry Clifford, Linden Disney-Hogg, Aleksandra Jartseva, Morton, Kevin Myant, William Orchard, Nirmesh Patel, Charlotte Paterson and Hannah Thompson.

Author contributions

JRB was the main contributor, working under the supervision of TIC.

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/s42003-022-03098-1.

Correspondence and requests for materials should be addressed to Jacob R. Bradley or Timothy I. Cannings.

Peer review information *Communications Biology* thanks the anonymous reviewers for their contribution to the peer review of this work. Primary Handling Editors: Brooke LaFlamme and Christina Karlsson-Rosenthal. Peer reviewer reports are available.

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.