

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

Jacob Bradley

University of Edinburgh <https://orcid.org/0000-0003-1616-4969>

Timothy Cannings (✉ timothy.cannings@ed.ac.uk)

University of Edinburgh <https://orcid.org/0000-0002-2111-4168>

Article

Keywords: cancer, gene panel design, targeted sequencing, tumour indel burden, tumour mutation burden.

Posted Date: March 17th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-223265/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Version of Record: A version of this preprint was published at Communications Biology on February 23rd, 2022. See the published version at <https://doi.org/10.1038/s42003-022-03098-1>.

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

Jacob R. Bradley and Timothy I. Cannings
School of Mathematics, University of Edinburgh

Abstract

We introduce a novel data-driven framework for the design of targeted gene panels for estimating exome-wide biomarkers in cancer immunotherapy. 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 new procedure for estimating biomarkers such as Tumour Mutation Burden and Tumour Indel Burden. Our approach allows the practitioner to select a targeted gene panel of a prespecified size, and then construct an estimator that only depends on the selected genes. Alternatively, the practitioner may apply our method 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 an annotated mutation dataset from 1144 Non-Small Cell Lung Cancer patients.

Keywords: cancer, gene panel design, targeted sequencing, tumour indel burden, tumour mutation burden.

1 Introduction

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 (Boveri, 2008). Since the discovery of Immune Checkpoint Blockade (ICB)¹ (Ishida et al., 1992; Leach et al., 1996), there has been an explosion of interest in cancer therapies targeting immune response and ICB therapy is now widely used in clinical practice (Robert, 2020). 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) (Buchbinder and Desai, 2016). Inhibition of these checkpoints can promote a more aggressive anti-tumour immune response (Pardoll, 2012), and in some patients this leads to long-term remission (Gettinger et al., 2019). However, ICB therapy is not always effective (Nowicki et al., 2018) 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

¹For their work on ICB, James Allison and Tasuku Honjo received the 2018 Nobel Prize for Physiology/Medicine (Ledford et al., 2018).

34 (Zhu et al., 2019; Cao et al., 2019). TMB is defined as the total number of non-synonymous
35 mutations occurring throughout the tumour exome, and can be thought of as a proxy for
36 how easily a tumour cell can be recognised as foreign by immune cells (Chan et al., 2019).
37 However, the cost of measuring TMB using Whole Exome Sequencing (WES) (Sboner et al.,
38 2011) currently prohibits its widespread use as standard-of-care. Sequencing costs, both
39 financial and in terms of the time taken for results to be returned, are especially problematic
40 in situations where high-depth sequencing is required, such as when utilising blood-based
41 Circulating Tumour DNA (ctDNA) from liquid biopsy samples (Gandara et al., 2018). The
42 same issues are encountered when measuring more recently proposed biomarkers such as
43 Tumour Indel Burden (TIB) (Wu et al., 2019b; Turajlic et al., 2017), which counts the
44 number of frameshift insertion and deletion mutations. There is, therefore, demand for cost-
45 effective approaches to estimate these biomarkers (Fancello et al., 2019; Golkaram et al.,
46 2020).

47 In this paper we propose a novel, data-driven method for biomarker estimation, based on
48 a generative model of how mutations arise in the tumour exome. More precisely, we model
49 mutation counts as independent Poisson variables, where the mean number of mutations
50 depends on the gene of origin and variant type, as well as the Background Mutation Rate
51 (BMR) of the tumour. Due to the ultrahigh-dimensional nature of sequencing data and the
52 fact that in many genes mutations arise purely according to the BMR, we use a regularisation
53 penalty when estimating the parameters of the model. In addition, this identifies a subset
54 of genes that are mutated above or below the background rate. Our model facilitates the
55 construction of a new estimator of TMB, based on a weighted linear combination of the
56 number of mutations in each gene. The vector of weights is chosen to be sparse (i.e. have
57 many entries equal to zero), so that our estimator of TMB may be calculated using only
58 the mutation counts in a subset of genes. In particular, this allows for accurate estimation
59 of TMB from a targeted gene panel, where the panel size (and therefore the cost) may
60 be determined by the user. We demonstrate the excellent practical performance of our
61 framework using a Non-Small Cell Lung Cancer (NSCLC) dataset (Chalmers et al., 2017),
62 and include a comparison with existing state-of-the-art approaches for estimating TMB.
63 Moreover, since our model allows variant type-dependent mutation rates, it can be adapted
64 easily to predict other biomarkers, such as TIB. Finally, our method may also be used in
65 combination with an existing targeted gene panel. In particular, we can estimate a biomarker
66 directly from the panel, or first augment the panel and then construct an estimator.

67 Due to its emergence as a biomarker for immunotherapy in recent years, a variety of
68 groups have considered methods for estimating TMB. A simple and common way to esti-
69 mate TMB is via the proportion of mutated codons in a targeted region. Budczies et al.
70 (2019) investigate how the accuracy of predictions made in this way are affected by the size
71 of the targeted region, where mutations are assumed to occur at uniform rate throughout the
72 genome. More recently Yao et al. (2020) modelled mutations as following a negative binomial
73 distribution while allowing for gene-dependent rates, which are inferred by comparing non-
74 synonymous and synonymous mutation counts. In contrast, our method does not require
75 data including synonymous mutations. Where they are included, we do not assume that
76 synonymous mutations occur at a uniform rate throughout the genome, giving us the flexi-
77 bility to account for location-specific effects on synonymous mutation rate such as chromatin
78 configuration (Makova and Hardison, 2015) and transcription-dependent repair mechanisms
79 (Fong et al., 2013). Linear regression models have been used for both panel selection (Lyu

80 et al., 2018) and for biomarker prediction (Guo et al., 2020). A review of some of the issues
 81 arising when dealing with targeted panel-based predictions of TMB biomarkers is given by
 82 Wu et al. (2019a). Finally, we are unaware of any methods for estimating TIB from targeted
 83 gene panels.

84 The remainder of the paper is as follows. In Section 2, we introduce our data sources,
 85 and provide a detailed description of our methodological proposal. Experimental results
 86 are given in Section 3 and we conclude in Section 4. Finally, we also provide an R package
 87 ICBioMark (Bradley and Cannings, 2021) which implements the methodology and reproduces
 88 the experimental results in the paper.

89 2 Methodology

90 2.1 Data and terminology

91 Our methodology can be applied to any annotated mutation dataset obtained by WES. To
 92 demonstrate our proposal we make use of the NSCLC dataset produced by Campbell et al.
 93 (2016), which contains data from 1144 patient-derived tumours. For each sample in this
 94 dataset we have the genomic locations and variant types of all mutations identified. At the
 95 time of the study, the patients had a variety of prognoses and smoking histories, were aged
 96 between 39 and 90, 41% were female and 59% were male; see Figure 1. In Figure 2A we
 97 see that mutations counts are distributed over a very wide range, as is the case in many
 98 cancer types (Chalmers et al., 2017). For simplicity, we only consider seven nonsynonymous
 99 variant types: missense mutations (which are the most abundant), nonsense mutations,
 100 frameshift insertions/deletions, splice site mutations, in-frame insertions/deletions, nonstop
 101 mutations and translation start site mutations. We present the frequencies of these mutation
 102 types in Figure 2B. Frameshift insertion/deletion (also known as indel) mutations are of
 103 particular interest when predicting TIB, but contribute only a small proportion ($< 4\%$) of
 104 nonsynonymous mutations.

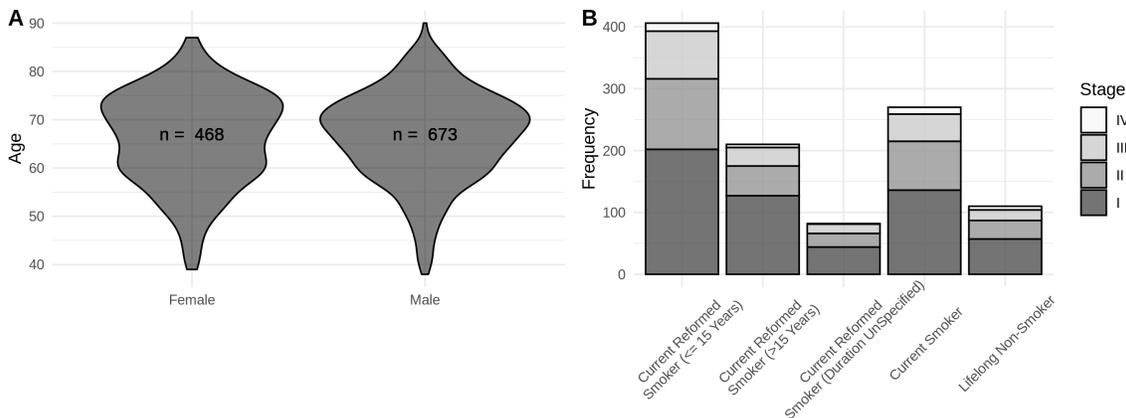


Figure 1: Demographic data for the clinical cohort in Campbell et al. (2016). **A**: Violin plots of age for patients, stratified by sex. **B**: Stacked bar chart of patients’ smoking histories, shaded according to cancer stage diagnosis.

105 It is useful at this point to introduce the notation used throughout the paper. The set
 106 G denotes the collection of genes that make up the exome. For a gene $g \in G$, let ℓ_g be the

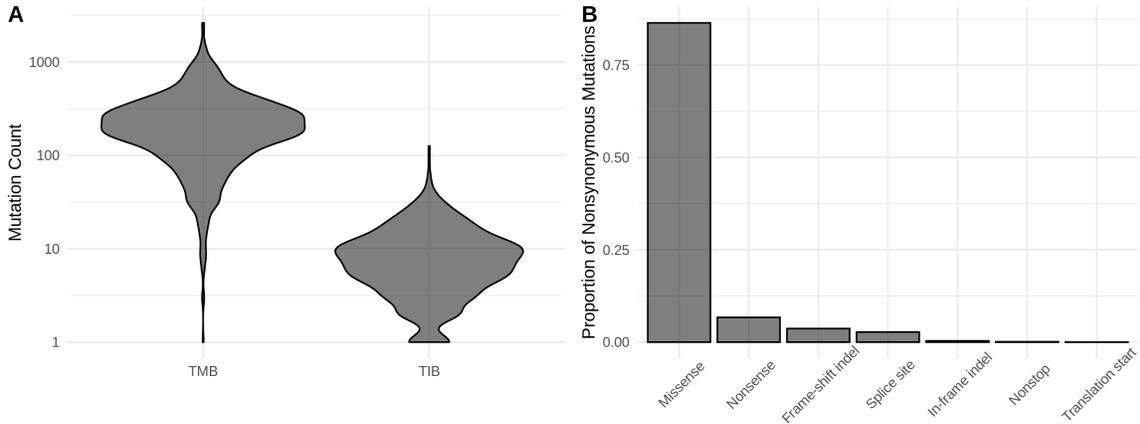


Figure 2: Dataset-wide distribution of mutations. **A**: Violin plot of the distribution of TMB and TIB across training samples. **B**: The relative frequency of different nonsynonymous mutation types.

length of g in nucleotide bases, defined by the maximum coding sequence². 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 $i = 0, 1, \dots, n$, let M_{igs} denote the count of mutations in gene $g \in G$ of type $s \in S$ in the i th 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 $i = 1, \dots, n$ enumerate the samples in our training data set. In order to define the exome-wide biomarker of particular interest, we specify a subset of mutation types $\bar{S} \subseteq S$, and let

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

for $i = 0, \dots, n$. 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 $T_{0\bar{S}}$.

2.2 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.

²The maximum coding sequence is defined as the collection of codons that may be translated for some version of a gene, even if all the codons comprising the maximum coding sequence are never simultaneously translated. Gene coding lengths are extracted from the *Ensembl* database (Yates et al., 2020).

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

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

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

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

134 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 $g \in G$.

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

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

148 Writing $\mu := (\mu_1, \dots, \mu_n)$, $\lambda := (\lambda_g : g \in G)$, $\nu := (\nu_s : s \in S)$ and $\eta := (\eta_{gs} : g \in G, s \in$
 149 $S)$, and given training observations $M_{igs} = m_{igs}$, we let

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

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

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

151 where $\kappa_1 \geq 0$ is a tuning parameter that controls the number of non-zero components in $\hat{\lambda}$
 152 and $\hat{\eta}$, which we choose using cross-validation (see Section 2.5 for more detail).

153 2.3 Proposed estimator

154 We now attend to our main goal of estimating a given exome-wide biomarker for the unseen
 155 test sample. Fix $\bar{S} \subseteq S$ and recall that we write $T = T_{0\bar{S}}$. We wish to construct an estimator
 156 of T that only depends on the mutation counts in a gene panel $P \subset G$, subject to a constraint
 157 on ℓ_P . To that end, we consider estimators of the form³

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

³Note that our estimator may use the 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).

158 for $w \in \mathbb{R}^{|G| \times |S|}$. In the remainder of this subsection we explain how the weights w are
 159 chosen to minimise the expected squared error of $T(w)$ based on the generative model in
 160 Section 2.2.

161 Of course, setting $w_{gs} = 1$ for $g \in G$ and $s \in \bar{S}$ (and $w_{gs} = 0$ otherwise) will give
 162 $T(w) = T$. However, our aim is to make predictions based on a concise gene panel. If,
 163 for a given $g \in G$, we have $w_{gs} = 0$ for all $s \in S$, then $T(w)$ does not depend on the
 164 mutations in g and therefore the gene does not need to be included in the panel. In order to
 165 produce a suitable gene panel (i.e. with many $w_{gs} = 0$), we penalise non-zero components of
 166 w when minimising the expected squared error. We define our final estimator via a refitting
 167 procedure, which improves the predictive performance by reducing the bias, and is also
 168 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

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

169 This depends on the unknown parameters μ_0, λ_g, ν_s and η_{gs} , the latter three of which are
 170 replaced by their estimates given in (4). It is also helpful to then rescale (5) as follows: write
 171 $\hat{\phi}_{0gs} = \ell_g \exp(\hat{\lambda}_g + \hat{\nu}_s + \hat{\eta}_{gs})$, and define

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

172 Then let

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

173 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
 174 the error in (5) (with the true parameters λ, ν, η replaced by the estimates $\hat{\lambda}, \hat{\nu}, \hat{\eta}$), we will
 175 choose w to minimise $f(w)$.

176 Note that f only depends on μ_0 via the $K(\mu_0)$ term, which can be interpreted as a penalty
 177 factor controlling the bias of our estimator. For example, we may insist that the squared
 178 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
 179 to choose the penalty K based on the training data; see Section 2.5.

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

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

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

187 Now, given a cost restriction L , our goal is to minimise $f(w)$ such that $\|w\|_{G,0} \leq L$.
 188 In practice this problem is non-convex and so computationally infeasible. As is common
 189 in high-dimensional optimisation problems, we consider a convex relaxation as follows: let
 190 $\|w\|_{G,1} := \sum_{g \in G} \ell_g \|w_g\|_2$, where $w_g = (w_{gs} : s \in S) \in \mathbb{R}^{|S|}$, for $g \in G$, and $\|\cdot\|_2$ is the
 191 Euclidean norm. Define

$$\hat{w}^{\text{first-fit}} \in \arg \min_w \{f(w) + \kappa_2 \|w\|_{G,1}\}, \quad (6)$$

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

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

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

195 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),
 196 and define

$$\hat{w}^{\text{refit}} \in \arg \min_{w \in W_{\hat{P}}} \{f(w)\}. \quad (8)$$

197 We then estimate T using $\hat{T} := T(\hat{w}^{\text{refit}})$, which only depends on mutations in genes contained
 198 in the selected panel \hat{P} . The performance of our estimator is investigated in Section 3, for
 199 comparison we also include the performance of the first-fit estimator $T(\hat{w}^{\text{first-fit}})$.

200 2.4 Panel augmentation

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

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

$$\hat{w}_{P_0, Q_0}^{\text{first-fit}} \in \arg \min_{w \in W_{G \setminus Q_0}} \left\{ f(w) + \kappa_2 \sum_{g \in G \setminus P_0} \ell_g \|w_g\|_2 \right\}. \quad (9)$$

212 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
 213 in P_0 , while restricting our optimisation to $W_{G \setminus Q_0}$ excludes the genes in Q_0 by definition.
 214 This has the effect of augmenting the predetermined panel P_0 with additional genes selected
 215 to improve predictive performance. We then perform refitting as described above. We
 216 demonstrate this procedure by augmenting the TST-170 gene panel in Section 3.4.

217 **2.5 Practical considerations**

218 In this section, we discuss some practical aspects of our proposal. Our first consideration
 219 concerns the choice of the tuning parameter κ_1 in (4). As is common for the Least Absolute
 220 Shrinkage and Selection Operator (LASSO) estimator in generalised linear regression (see,
 221 for example, [Michoel \(2016\)](#) and [Friedman et al. \(2020\)](#)), we will use 10-fold cross-validation.
 222 To highlight one important aspect of our cross-validation procedure, recall that we consider
 223 the observations M_{igs} as independent across the sample index $i \in \{1, \dots, n\}$, the gene
 224 $g \in G$ and the mutation type $s \in S$. Our approach therefore involves splitting the entire set
 225 $\{(i, g, s) : i = 1, \dots, n, g \in G, s \in S\}$ of size $n|G||S|$ (as opposed to the sample set $\{1, \dots, n\}$)
 226 into 10 folds uniformly at random. We then apply the estimation method in (4) to each of
 227 the 10 folds separately on a grid of values (on the log scale) of κ_1 , and select the value that
 228 results in the smallest average deviance across the folds. The model is then refitted using
 229 all the data for this value of κ_1 .

230 The estimated coefficients in (6) depend on the choice of $K(\mu_0)$ and κ_2 . As mentioned
 231 above, we could set $K(\mu_0) = \infty$ to give an unbiased estimator, however in practice we found
 232 that a finite choice of $K(\mu_0)$ leads to improved predictive performance. Our recommendation
 233 is to use $K(\mu_0) = K(\max_{i=1, \dots, n} \{\hat{\mu}_i\})$, where $\hat{\mu}_i = \log(T_i / \sum_{g,s} \ell_g \exp(\hat{\lambda}_g + \hat{\nu}_s + \hat{\eta}_{gs}))$ is a
 234 pseudo-MLE (in the sense of [Gong and Samaniego \(1981\)](#)) for μ_i , so that the penalisation
 235 is broadly in proportion with the largest values of μ_i in the training dataset. The tuning
 236 parameter κ_2 controls the size of the gene panel selected in (6): given a panel length L , we
 237 set $\kappa_2(L) = \max\{\kappa_2 : \ell_{\hat{p}} \leq L\}$ in order to produce a suitable panel.

238 We now comment briefly on some computational aspects of our method. The generative
 239 model fit in (4) can be solved via coordinate descent (see, for example, [Friedman et al., 2010](#)),
 240 which has a computational complexity of $O(N|G|^2|S|^2)$ per iteration. We fit the model 10
 241 times, one for each fold in our cross-validation procedure. This is the most computationally
 242 demanding part of our proposal – in our experiments below, it takes approximately an hour
 243 to solve on a standard laptop – but it only needs to be carried out once for a given dataset.
 244 The convex optimisation problem in (6) can be solved by any method designed for the group
 245 LASSO; see, for example, [Yang and Zou \(2015\)](#). In our experiments in Section 3, we use the
 246 `gglasso` R package ([Yang et al., 2020](#)), which takes around 10 minutes to reproduce the plot
 247 in Figure 6. Note also that the solutions to (6) and (8) are unique; see, for example, [Roth
 248 and Fischer \(2008, Theorem 1\)](#). The last step of our proposal, namely making predictions
 249 for new test observations based on a selected panel, carries negligible computational cost.

250 Finally we describe a heuristic procedure for producing prediction intervals around our
 251 point estimates. In particular, for a given confidence level $\alpha \in (0, 1)$, we aim to find an
 252 interval $[\hat{T}_L, \hat{T}_U]$ such that $\mathbb{P}(\hat{T}_L \leq T \leq \hat{T}_U) \geq 1 - \alpha$. To that end, let $t_\alpha := \mathbb{E}\{(\hat{T} - T)^2\}/\alpha$,
 253 then by Markov’s inequality we have that $\mathbb{P}(|\hat{T} - T|^2 \geq t_\alpha) \leq \alpha$. It follows that $[\hat{T} - t_\alpha^{1/2}, \hat{T} +$
 254 $t_\alpha^{1/2}]$ is a $(1 - \alpha)$ -prediction interval for T . Of course, the mean squared error $\mathbb{E}\{(\hat{T} - T)^2\}$
 255 defined in (5) depends on the parameters λ, η, ν and μ_0 , which are unknown. Our approach is
 256 to utilise the estimates $\hat{\lambda}, \hat{\eta}, \hat{\nu}$ (see (4)) and replace μ_0 with $\log(\hat{T} / \sum_{g,s} \ell_g \exp(\hat{\lambda}_g + \hat{\nu}_s + \hat{\eta}_{gs}))$.
 257 While this is not an exact $(1 - \alpha)$ -prediction interval for T , we will see in our experimental
 258 results in Sections 3.2 and 3.3 that in practice this approach provides intervals with valid
 259 empirical coverage.

3 Experimental results

In this section we demonstrate the practical performance of our proposal using the dataset from Campbell et al. (2016), which we introduced in Section 2.1. Our main focus is the prediction of TMB, and we show that our approach outperforms the state-of-the-art approaches. We also analyse the suitability of our generative model, consider the task of predicting the recently proposed biomarker TIB, and include a panel augmentation case study with the Foundation One gene panel.

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 $|G| = 17358$ genes. Our training set comprises samples with an average TMB of 252 and TIB of 9.25.

3.1 Generative model fit and validation

The first step in our analysis is to fit the model proposed in Section 2.2 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 10-fold cross-validation as described in Section 2.5. The results are presented in Figure 3. 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 Figures 4 and 5. Genes with $\hat{\lambda}_g = 0$ are interpreted to be mutating according to the background mutation rate, and genes with $\hat{\eta}_{g,\text{indel}} = 0$ are interpreted as having no specific selection pressure for or against indel mutations. In Figures 4 and 5 we highlight genes with large (in absolute value) parameter estimates, some of which have known biological relevance in oncology; see Section 4 for further discussion.

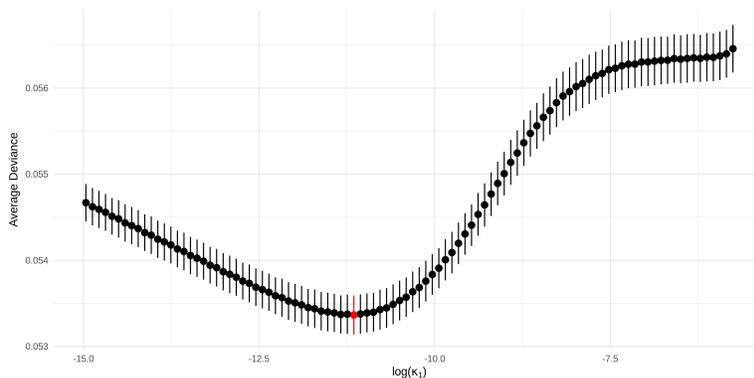


Figure 3: The average deviance (with one standard deviation) across the 10 folds in our cross-validation procedure plotted against $\log(\kappa_1)$. The minimum average deviance is highlighted red.

We now validate our model in (3) by comparing with the following alternatives:

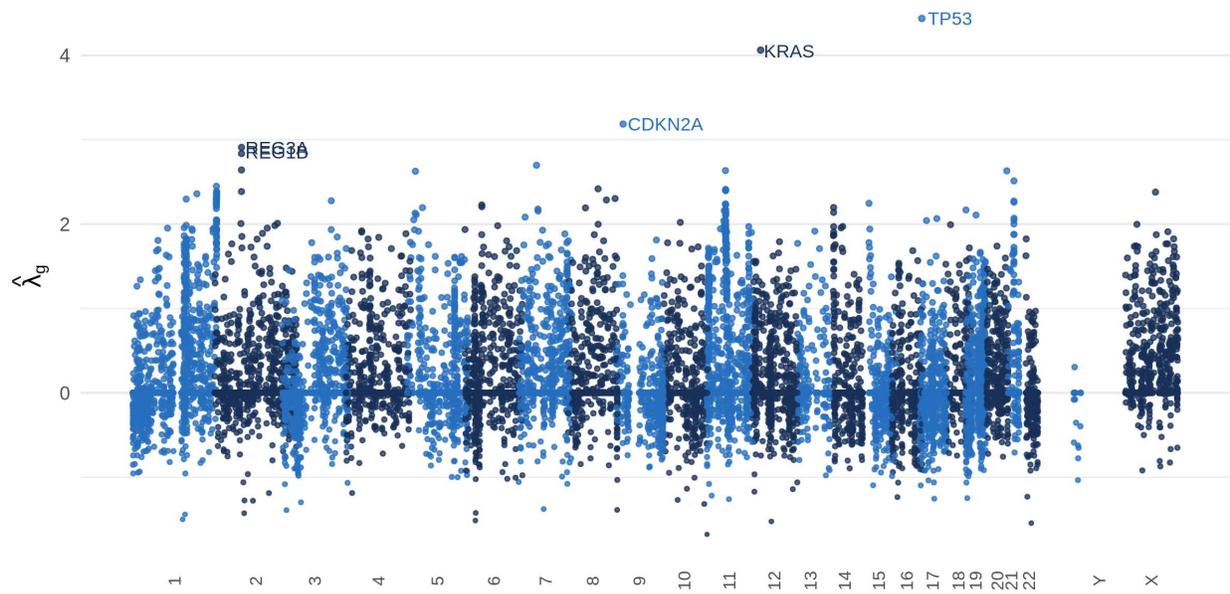


Figure 4: Manhattan plot of fitted parameters $\hat{\lambda}_g$ and their associated genes' chromosomal locations. The genes with the five largest positive parameter estimates are labelled.

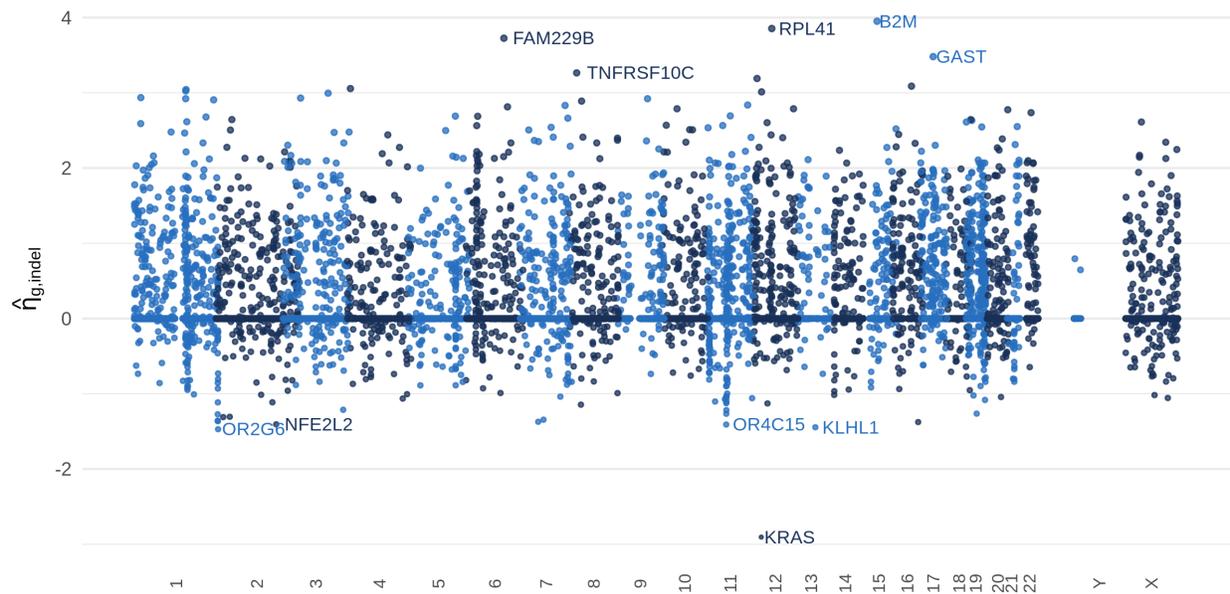


Figure 5: Manhattan plot of fitted parameters $\hat{\eta}_{g,\text{indel}}$ and their associated genes' chromosomal locations. The five largest positive and negative genes are labelled.

- 288 (i) *Saturated model*: the model in (2), where each observation has an associated free
 289 parameter (i.e. $\phi_{igs} > 0$ is unrestricted);
- 290 (ii) *No sample-specific effects*: the model in (3), with $\mu_i = 0$ for all $i \in \{1, \dots, n\}$;
- 291 (iii) *No gene-specific effects*: the model in (3), with $\lambda_g = \eta_{gs} = 0$ for all $g \in G$ and $s \in S$;
- 292 (iv) *No gene/mutation type interactions*: the model in (3), with $\eta_{gs} = 0$ for all $g \in G$ and
 293 $s \in S$.

294 In Table 1 we present the residual deviance and the residual degrees of freedom between
 295 our model and each of the models above. We see that our model is preferred over the
 296 saturated model, and all three submodels of (3).

Table 1: Model comparisons on the basis of residual deviance statistics.

Comparison Model	Residual Deviance (dev)	Residual Degrees of Freedom (df)	dev/df	p -value
(i)	$1.43E + 06$	$2.74E + 07$	$5.22E - 02$	1.00
(ii)	$1.42E + 05$	$8.00E + 02$	$1.77E + 02$	0.00
(iii)	$1.10E + 05$	$1.33E + 04$	$8.24E + 00$	0.00
(iv)	$1.70E + 04$	$1.82E + 03$	$9.33E + 00$	0.00

297 3.2 Predicting tumour mutation burden

298 We now demonstrate the excellent practical performance of our procedure for estimating
 299 TMB. First it is shown that our method can indeed select gene panels of size specified by
 300 the practitioner and that good predictions can be made even with small panel sizes (i.e. \leq
 301 1Mb). We then compare the performance of our proposal with state-of-the-art estimation
 302 procedures based on a number of widely used gene panels.

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

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

306 Other existing works have aimed to classify tumours into two groups (high TMB, low
 307 TMB); see, for example, Büttner et al. (2019) and Wu et al. (2019a). Here we also report
 308 the estimated area under the precision-recall curve (AUPRC) for a classifier based on our
 309 estimator. We define the classifier as follows: first, in line with major clinical studies (e.g.
 310 Hellmann et al., 2018; Ramalingam et al., 2018) the true class membership of a tumour
 311 is defined according to whether it has $t^* := 300$ or more exome mutations (approximately
 312 10 Mut/Mb). In the validation set, this gives 47 (27.5%) tumours with high TMB and
 313 124 (72.5%) with low TMB. Now, for a cutoff $t \geq 0$, we can define a classifier by assigning a
 314 tumour to the high TMB class if its estimated TMB value is greater than or equal to t . For
 315 such a classifier, we have precision and recall (estimated over the validation set) given by

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

316 respectively. The precision-recall curve then is $\{(r(t), p(t)) : t \in [0, \infty)\}$. Note that a perfect
 317 classifier achieves a AUPRC of 1, whereas a random guess in this case would have an average
 318 AUPRC of 0.308 (the prevalence of the high TMB class).

319 Now recall that TMB is given by equation (1) with \bar{S} being the set of all non-synonymous
 320 mutation types. Thus to estimate TMB we apply our procedure in Section 2.3 with $\bar{S} = S$,
 321 where the model parameters are estimated as described in Section 3.1. In Figure 6, we
 322 present the R^2 and AUPRC for the first-fit and refitted estimators (see (6) and (8)) as the
 323 selected panel size varies from 0Mb to 2Mb in length. We see that we obtain a more accurate
 324 prediction of TMB, both in terms of regression and classification, as the panel size increases,
 325 and that good estimation is possible even with very small panels (as low as 0.2Mb). Finally,
 326 as expected, the refitted estimator slightly outperforms the first-fit estimator.

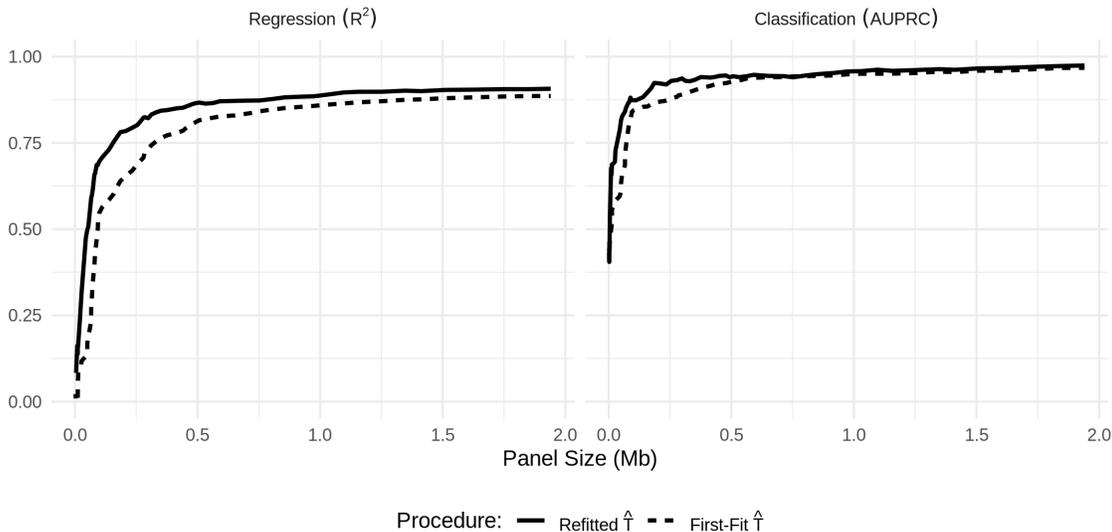


Figure 6: Performance of our first-fit and refitted estimators of TMB as the selected panel size varies. **Left:** R^2 , **Right:** AUPRC.

327 We now compare our method with state-of-the-art estimators applied to commonly used
 328 gene panels. The three next-generation sequencing panels that we consider are chosen for
 329 their relevance to TMB. These are TST-170 (Heydt et al., 2018), Foundation One (Frampton
 330 et al., 2013) and MSK-IMPACT (Cheng et al., 2015). For each panel $P \subseteq G$, we use four
 331 different methods to predict TMB:

- 332 (i) Our refitted estimator applied to the panel P : we estimate TMB using $T(\hat{w}_P)$, where
 333 $\hat{w}_P \in \arg \min_{w \in W_P} \{f(w)\}$, and W_P is defined in (7).
- 334 (ii) Estimation and Classification of Tumour Mutation Burden (ecTMB): the procedure
 335 proposed by Yao et al. (2020).
- 336 (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 muta-
 337 tion burden in the genes of P .
- 338 (iv) A linear model: we estimate TMB via ordinary least-squares linear regression of TMB
 339 against $\{\sum_{s \in \bar{S}} M_{0gs} : g \in P\}$.

340 The latter three comprise existing methods for estimating TMB available to practitioners.
 341 The second (ecTMB), which is based on a negative binomial model, is the state-of-the-art.

342 The third and fourth are standard practical procedures for the estimation of TMB from
 343 targeted gene panels. The refitted estimator applied to the panel P is also included here, in
 344 order to demonstrate the utility of our approach even with a prespecified panel.

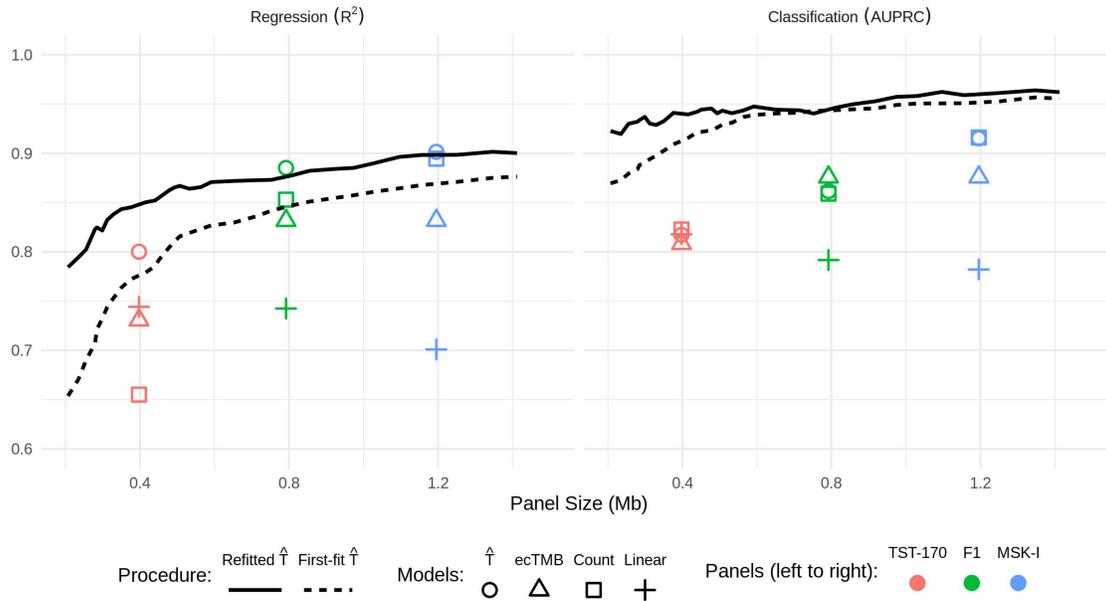


Figure 7: The performance of our TMB estimator in comparison to existing approaches. **Left:** R^2 , **Right:** AUPRC.

345 We present results of these comparisons in Figure 7. First, for each of the three pan-
 346 els considered here, we see that our refitted estimator applied to the panel outperforms all
 347 existing approaches in terms of regression performance, and that for smaller panels we are
 348 able to improve regression accuracy even further by selecting a panel based on the training
 349 data. For instance, in comparison to predictions based on the TST-170 panel, our procedure
 350 with a selected panel of the same size (0.4Mb) achieves an R^2 of 0.85. The best available
 351 existing method based on the TST-170 panel, in this case the linear estimator, has an R^2 of
 352 0.74. Moreover, data-driven selection of panels considerably increases the classification per-
 353 formance for the whole range of panel sizes considered. In particular, even for the smallest
 354 panel size shown in Figure 7 (~ 0.2 Mb), the classification performance of our method out-
 355 performs the best existing methodology applied to the MSK-IMPACT panel, despite being
 356 almost a factor of six times smaller.

357 Finally in this section we demonstrate the practical performance of our method using
 358 the test set, which until this point has been held out. Based on the validation results above,
 359 we take the panel of size 0.6Mb selected by our procedure and use our refitted estimator
 360 on that panel to predict TMB for the 173 samples in the test set. For comparison, we
 361 also present predictions from ecTMB, the count-based estimator and the linear regression
 362 estimator applied to the same panel. In Figure 8 we see that our procedure performs well;
 363 we obtain an R^2 value (on the test data) of 0.85. The other methods have R^2 values of 0.67
 364 (ecTMB), -36 (count) and 0.64 (linear regression). The count-based estimator here gives
 365 predictions which are reasonably well correlated to the true values of TMB but are positively
 366 biased. This is as expected, since our selection procedure tends to favour genes with higher
 367 overall mutation rates. We also include a red shaded region comprising all points for which

368 heuristic 90% prediction intervals (as described in Section 2.5) include the true TMB value.
 369 We find in this case that 93.6% of the observations in the test set fall within this region,
 370 giving valid empirical coverage.

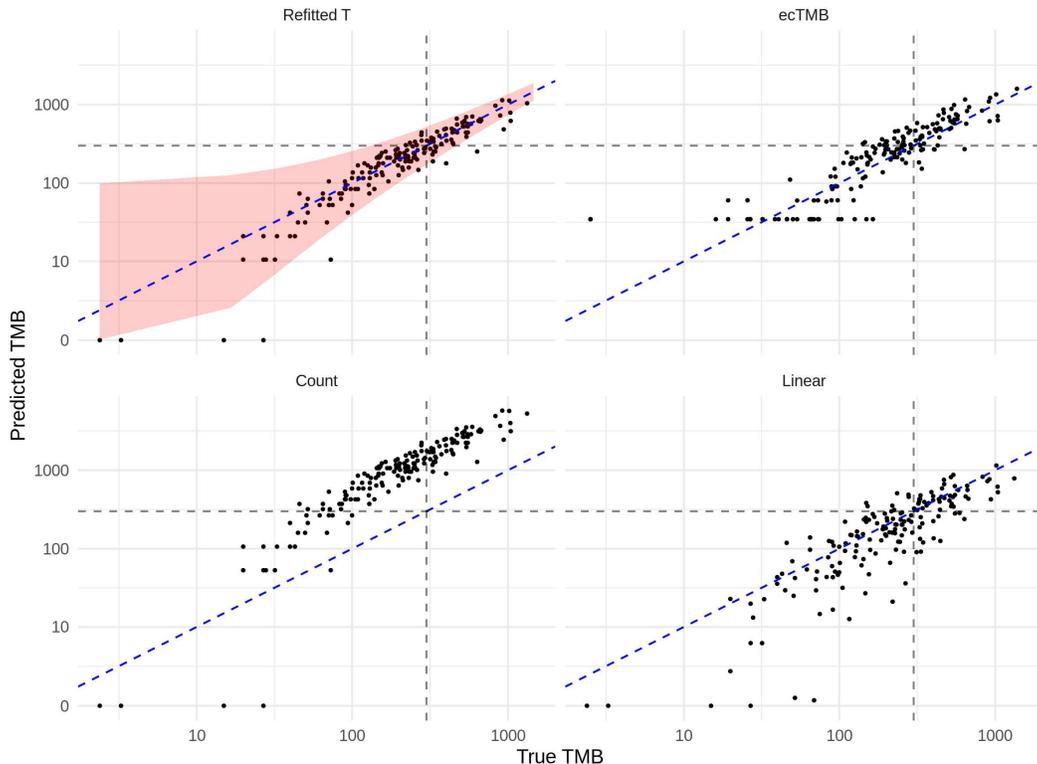


Figure 8: Prediction of TMB on the test dataset. Dashed blue (diagonal) line represents perfect prediction, and the black dashed lines indicate true and predicted TMB values of 300.

371 3.3 Predicting tumour indel burden

372 In this section we demonstrate how our method can be used to estimate TIB. This is more
 373 challenging than estimating TMB due to the low abundance of indel mutations relative to
 374 other variant types (see Figure 2), as well as issues involved in sequencing genomic loci
 375 of repetitive nucleotide constitution (Narzisi and Schatz, 2015). Indeed, in contrast to the
 376 previous section, we are not aware of any existing methods designed to estimate TIB from
 377 targeted gene panels. We therefore investigate the performance of our method across a much
 378 wider range (0-30Mb) of panel sizes, and find that we are able to accurately predict TIB
 379 with larger panels. Our results also demonstrate that accurate classification of TIB status
 380 is possible even with small gene panels.

381 We let S_{indel} be the set of all frameshift insertion and deletion mutations, and apply our
 382 method introduced in Section 2.3 with $\tilde{S} = S_{\text{indel}}$. As in the previous section, we assess
 383 regression and classification performance via R^2 and AUPRC, respectively, where in this
 384 case tumours are separated into two classes: high TIB (10 or more indel mutations) and low
 385 TIB (otherwise). In the validation dataset, this gives 57 (33.3%) tumours in the high TIB
 386 class.

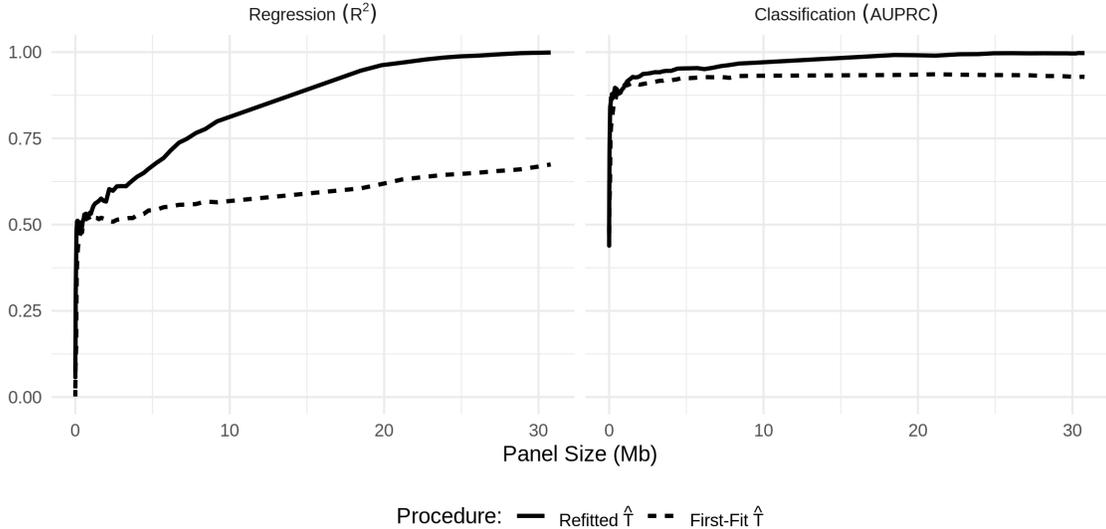


Figure 9: Performance of our first-fit and refitted estimators of TIB as the selected panel size varies. **Left:** R^2 , **Right:** AUPRC.

387 The results are presented in Figure 9. We comment first on the regression performance:
 388 as expected, we see that the R^2 values for our first-fit and refitted estimators are much lower
 389 than what we achieved in estimating TMB. The refitted approach improves for larger panel
 390 sizes, while the first-fit estimator continues to perform relatively poorly. On the other hand,
 391 we see that the classification performance is impressive, with AUPRC values of above 0.8
 392 for panels of less than 1Mb in size.

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

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

400 In Figure 10 we present the predictions on the test set of our refitted estimator ($R^2 = 0.35$);
 401 the count estimator ($R^2 = -0.44$); and the linear regression estimator ($R^2 = -0.15$). We also
 402 include (shaded in red) the set of points for which 90% prediction intervals contain the true
 403 value. In this case we find that 97.7% of test set points fall within this region.

404 3.4 A panel-augmentation case study

405 As discussed in Section 2.4, we may wish to include genes from a given panel, but use our
 406 methodology to augment the panel to include additional genes with goal of obtaining more
 407 accurate predictions of TMB (or other biomarkers). In this section we demonstrate how this
 408 can be done starting with the TST-170 panel (~ 0.4 Mb) and augmenting to 0.6Mb in length,
 409 demonstrating impressive gains in predictive performance.

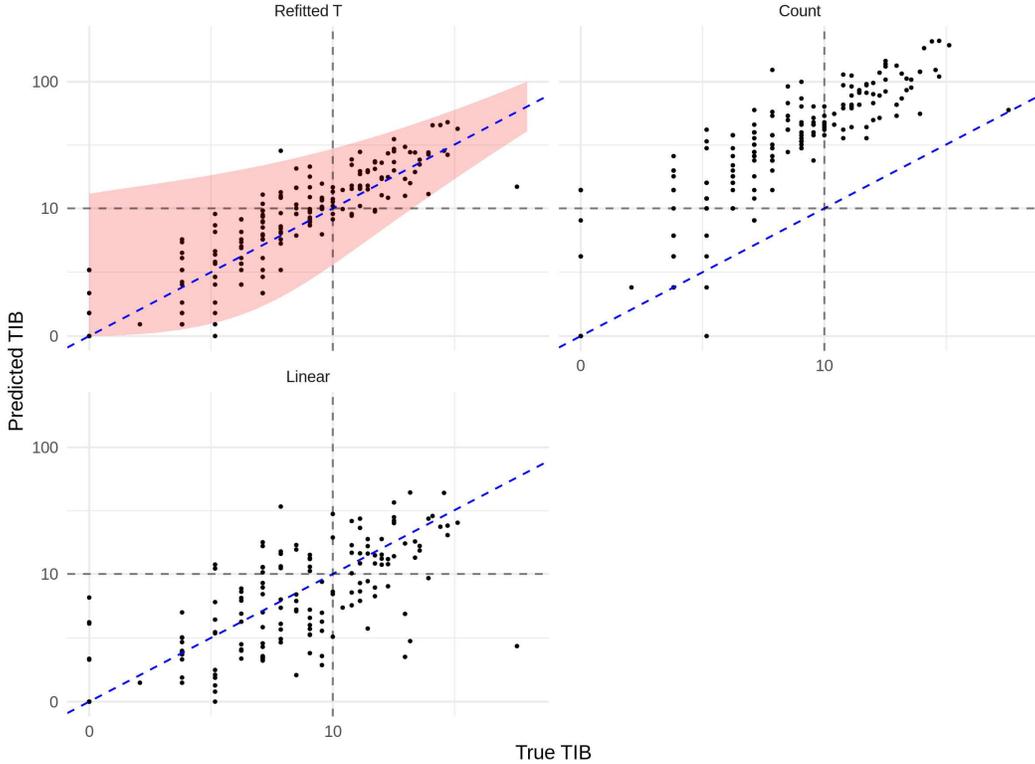


Figure 10: Estimation of TIB on the test dataset. Dashed blue (diagonal) line represents perfect prediction, and the grey dashed lines indicate true and predicted TIB values of 10.

410 We apply the augmentation method described in Section 2.4, with P_0 taken to be the
 411 set of TST-170 genes and Q_0 to be empty. The genes added to the panel are determined
 412 by the first-fit estimator in equation (9). To evaluate the performance, we then apply the
 413 refitted estimator on the test dataset, after selecting the augmented panel of size 0.6Mb. For
 414 comparison, we apply our refitted estimator to the TST-170 panel directly. We also present
 415 the results obtained by the other estimators described above, both before and after the panel
 416 augmentation, in Table 2. We find that by augmenting the panel we improve predictive
 417 performance with our refitted \hat{T} estimator, both in terms of regression and classification.
 418 The refitted estimator provides better estimates than any other model on the augmented
 419 panel by both metrics.

Table 2: Predictive performance of models on TST-170 (0.4Mb) versus augmented TST-170 (0.6Mb) 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	0.83	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

4 Conclusions

We have introduced a new 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 dataset from [Campbell et al. \(2016\)](#), 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. Our framework can be applied to any tumour dataset containing annotated mutations, and we provide an R package ([Bradley and Cannings, 2021](#)) which implements the methodology.

Our work also has the scope to help understand mutational processes. For example, the parameters of our fitted model in Section 3.1 have interesting interpretations: of the five genes highlighted in Figure 4 as having the highest mutation rates relative to the BMR, three (*TP53*, *KRAS*, *CDKN2A*) are known tumour suppressors ([Olivier et al., 2010](#); [Jančík et al., 2010](#); [Foulkes et al., 1997](#)). Furthermore, indel mutations in *KRAS* are known to be deleterious for tumour cells ([Lee et al., 2018](#)) – in our work the *KRAS* gene has a large negative indel-specific parameter (see Figure 5). Our methodology identifies a number of other genes with large parameter estimates.

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 data sources (e.g. on different cancer types and tissues of origin).

Data availability

All data used in this manuscript is publicly available. The NSCLC dataset of [Campbell et al. \(2016\)](#) and the *Ensembl* gene length dataset are available as part of our R package ICBioMark ([Bradley and Cannings, 2021](#)) – see below for more detail. The BED files for the gene panels used in Section 3.2 can be downloaded from https://github.com/cobrbra/TargetedPanelEstimation_Paper.

Code availability

All figures and tables in this manuscript may be reproduced using the code available at https://github.com/cobrbra/TargetedPanelEstimation_Paper. We also provide an open access R package ICBioMark ([Bradley and Cannings, 2021](#)), which is available on CRAN <https://cran.r-project.org>. Alternatively, the package may be accessed and downloaded at <https://github.com/cobrbra/ICBioMark>.

Acknowledgements

We gratefully acknowledge funding provided by Cambridge Cancer Genomics (CCG) through their PhD Scholarship at the University of Edinburgh. We also benefited from discussions

456 with several individuals, including Adnan Akbar, Philip Beer, Harry Clifford, Aleksandra
457 Jartseva, Morton, Kevin Myant, William Orchard, Nirmesh Patel and Charlotte Paterson.

458 **References**

459 T. Boveri. Concerning the Origin of Malignant Tumours. Translated and annotated by
460 Henry Harris. *Journal of Cell Science*, 121(Supplement 1):1–84, Jan. 2008. ISSN 0021-
461 9533, 1477-9137. doi: 10.1242/jcs.025742. Publisher: The Company of Biologists Ltd
462 Section: Article.

463 J. R. Bradley and T. I. Cannings. ICBioMark: Data-Driven Design of Targeted Gene Pan-
464 els for Estimating Immunotherapy Biomarkers (R Package), Jan. 2021. URL <https://github.com/cobrbra/ICBioMark>.
465

466 E. I. Buchbinder and A. Desai. CTLA-4 and PD-1 Pathways: Similarities, Differences, and
467 Implications of Their Inhibition. *American Journal of Clinical Oncology*, 39(1):98–106,
468 Feb. 2016. ISSN 1537-453X. doi: 10.1097/COC.000000000000239.

469 J. Budczies, M. Allgäuer, and K. Litchfield. Optimizing panel-based tumor mutational
470 burden (TMB) measurement. *Annals of Oncology: Official Journal of the European Society
471 for Medical Oncology*, 30(9):1496–1506, 2019. ISSN 1569-8041. doi: 10.1093/annonc/
472 mdz205.

473 R. Büttner, J. W. Longshore, and F. López-Ríos. Implementing TMB measurement in
474 clinical practice: considerations on assay requirements. *ESMO Open*, 4(1):e000442, Jan.
475 2019. ISSN 2059-7029. doi: 10.1136/esmoopen-2018-000442.

476 J. D. Campbell, A. Alexandrov, and J. Kim. Distinct patterns of somatic genome alterations
477 in lung adenocarcinomas and squamous cell carcinomas. *Nature Genetics*, 48(6):607–616,
478 2016. ISSN 1546-1718. doi: 10.1038/ng.3564.

479 D. Cao, H. Xu, and X. Xu. High tumor mutation burden predicts better efficacy of
480 immunotherapy: a pooled analysis of 103078 cancer patients. *Oncoimmunology*, 8(9):
481 e1629258, 2019. ISSN 2162-4011. doi: 10.1080/2162402X.2019.1629258.

482 Z. R. Chalmers, C. F. Connelly, and D. Fabrizio. Analysis of 100,000 human cancer genomes
483 reveals the landscape of tumor mutational burden. *Genome Medicine*, 9, Apr. 2017. ISSN
484 1756-994X. doi: 10.1186/s13073-017-0424-2.

485 T. A. Chan, M. Yarchoan, and E. Jaffee. Development of tumor mutation burden as an
486 immunotherapy biomarker: utility for the oncology clinic. *Annals of Oncology*, 30(1):
487 44–56, Jan. 2019. ISSN 0923-7534, 1569-8041. doi: 10.1093/annonc/mdy495.

488 D. T. Cheng, T. N. Mitchell, and A. Zehir. Memorial Sloan Kettering-Integrated Mutation
489 Profiling of Actionable Cancer Targets (MSK-IMPACT). *The Journal of Molecular Diag-
490 nostics : JMD*, 17(3):251–264, May 2015. ISSN 1525-1578. doi: 10.1016/j.jmoldx.2014.
491 12.006.

- 492 L. Fancello, S. Gandini, and P. G. Pelicci. Tumor mutational burden quantification from
493 targeted gene panels: major advancements and challenges. *Journal for ImmunoTherapy*
494 *of Cancer*, 7(1):183, July 2019. ISSN 2051-1426. doi: 10.1186/s40425-019-0647-4.
- 495 Y. W. Fong, C. Cattoglio, and R. Tjian. The intertwined roles of transcription and repair
496 proteins. *Molecular Cell*, 52(3):291–302, Nov. 2013. ISSN 1097-4164. doi: 10.1016/j.
497 molcel.2013.10.018.
- 498 W. D. Foulkes, T. Y. Flanders, and P. M. Pollock. The CDKN2A (p16) gene and human
499 cancer. *Molecular Medicine*, 3(1):5–20, Jan. 1997. ISSN 1076-1551.
- 500 G. M. Frampton, A. Fichtenholtz, and G. A. Otto. Development and validation of a clin-
501 ical cancer genomic profiling test based on massively parallel DNA sequencing. *Nature*
502 *Biotechnology*, 31(11):1023–1031, Nov. 2013. ISSN 1546-1696. doi: 10.1038/nbt.2696.
- 503 J. Friedman, T. Hastie, and R. Tibshirani. glmnet: Lasso and Elastic-Net Regularized
504 Generalized Linear Models, June 2020. URL [https://CRAN.R-project.org/package=
505 glmnet](https://CRAN.R-project.org/package=glmnet).
- 506 J. H. Friedman, T. Hastie, and R. Tibshirani. Regularization Paths for Generalized Linear
507 Models via Coordinate Descent. *Journal of Statistical Software*, 33(1):1–22, Feb. 2010.
508 ISSN 1548-7660. doi: 10.18637/jss.v033.i01.
- 509 D. R. Gandara, S. M. Paul, and M. Kowanz. Blood-based tumor mutational bur-
510 den as a predictor of clinical benefit in non-small-cell lung cancer patients treated
511 with atezolizumab. *Nature Medicine*, 24(9):1441–1448, 2018. ISSN 1546-170X. doi:
512 10.1038/s41591-018-0134-3.
- 513 S. Gettinger, H. Borghaei, and J. Brahmer. 5-Year Outcomes From the Randomized, Phase
514 3 Trials CheckMate 017/057: Nivolumab vs Docetaxel in Previously Treated NSCLC.
515 *Journal of Thoracic Oncology*, 14(10):S244–S245, Oct. 2019. ISSN 1556-0864. doi: 10.
516 1016/j.jtho.2019.08.486.
- 517 M. Golkaram, C. Zhao, and K. Kruglyak. The interplay between cancer type, panel size
518 and tumor mutational burden threshold in patient selection for cancer immunotherapy.
519 *PLOS Computational Biology*, 16(11):e1008332, Nov. 2020. ISSN 1553-7358. doi: 10.1371/
520 journal.pcbi.1008332.
- 521 G. Gong and F. J. Samaniego. Pseudo Maximum Likelihood Estimation: Theory and Ap-
522 plications. *The Annals of Statistics*, 9(4):861–869, 1981. ISSN 0090-5364.
- 523 W. Guo, Y. Fu, and L. Jin. An Exon Signature to Estimate the Tumor Mutational Burden
524 of Right-sided Colon Cancer Patients. *Journal of Cancer*, 11(4):883–892, 2020. ISSN
525 1837-9664. doi: 10.7150/jca.34363.
- 526 M. D. Hellmann, T.-E. Ciuleanu, and A. Pluzanski. Nivolumab plus Ipilimumab in Lung
527 Cancer with a High Tumor Mutational Burden. *New England Journal of Medicine*, 378
528 (22):2093–2104, May 2018. ISSN 0028-4793. doi: 10.1056/NEJMoa1801946.

- 529 C. Heydt, R. Pappesch, and K. Stecker. Evaluation of the TruSight Tumor 170 (TST170)
530 assay and its value in clinical research. *Annals of Oncology*, 29:vi7–vi8, Sept. 2018. ISSN
531 0923-7534, 1569-8041. doi: 10.1093/annonc/mdy318.003.
- 532 Y. Ishida, Y. Agata, and K. Shibahara. Induced expression of PD-1, a novel member of the
533 immunoglobulin gene superfamily, upon programmed cell death. *The EMBO journal*, 11
534 (11):3887–3895, Nov. 1992. ISSN 0261-4189.
- 535 S. Jančík, J. Drábek, and D. Radzioch. Clinical Relevance of KRAS in Human Cancers.
536 *Journal of Biomedicine and Biotechnology*, 2010, 2010. ISSN 1110-7243. doi: 10.1155/
537 2010/150960.
- 538 D. R. Leach, M. F. Krummel, and J. P. Allison. Enhancement of antitumor immunity by
539 CTLA-4 blockade. *Science (New York, N.Y.)*, 271(5256):1734–1736, Mar. 1996. ISSN
540 0036-8075. doi: 10.1126/science.271.5256.1734.
- 541 H. Ledford, H. Else, and M. Warren. Cancer immunologists scoop medicine Nobel prize.
542 *Nature*, 562(7725):20–21, Oct. 2018. doi: 10.1038/d41586-018-06751-0. Number: 7725
543 Publisher: Nature Publishing Group.
- 544 W. Lee, J. H. Lee, and S. Jun. Selective targeting of KRAS oncogenic alleles by
545 CRISPR/Cas9 inhibits proliferation of cancer cells. *Scientific Reports*, 8(1):11879, Aug.
546 2018. ISSN 2045-2322. doi: 10.1038/s41598-018-30205-2. Number: 1 Publisher: Nature
547 Publishing Group.
- 548 G.-Y. Lyu, Y.-H. Yeh, and Y.-C. Yeh. Mutation load estimation model as a predictor of the
549 response to cancer immunotherapy. *npj Genomic Medicine*, 3(1):1–9, Apr. 2018. ISSN
550 2056-7944. doi: 10.1038/s41525-018-0051-x. Number: 1 Publisher: Nature Publishing
551 Group.
- 552 K. D. Makova and R. C. Hardison. The effects of chromatin organization on variation in
553 mutation rates in the genome. *Nature Reviews Genetics*, 16(4):213–223, Apr. 2015. ISSN
554 1471-0064. doi: 10.1038/nrg3890. Number: 4 Publisher: Nature Publishing Group.
- 555 T. Michoel. Natural coordinate descent algorithm for L1-penalised regression in generalised
556 linear models. *Computational Statistics & Data Analysis*, 97:60–70, May 2016. ISSN
557 0167-9473. doi: 10.1016/j.csda.2015.11.009.
- 558 G. Narzisi and M. C. Schatz. The Challenge of Small-Scale Repeats for Indel Discovery.
559 *Frontiers in Bioengineering and Biotechnology*, 3, Jan. 2015. ISSN 2296-4185. doi: 10.
560 3389/fbioe.2015.00008.
- 561 T. S. Nowicki, S. Hu-Lieskovan, and A. Ribas. Mechanisms of Resistance to PD-1 and
562 PD-L1 blockade. *Cancer journal (Sudbury, Mass.)*, 24(1):47–53, 2018. ISSN 1528-
563 9117. doi: 10.1097/PPO.0000000000000303. URL [https://www.ncbi.nlm.nih.gov/pmc/
564 articles/PMC5785093/](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5785093/).
- 565 M. Olivier, M. Hollstein, and P. Hainaut. TP53 Mutations in Human Cancers: Origins,
566 Consequences, and Clinical Use. *Cold Spring Harbor Perspectives in Biology*, 2(1), Jan.
567 2010. ISSN 1943-0264. doi: 10.1101/cshperspect.a001008.

- 568 D. M. Pardoll. The blockade of immune checkpoints in cancer immunotherapy. *Nature*
569 *reviews. Cancer*, 12(4):252–264, Mar. 2012. ISSN 1474-175X. doi: 10.1038/nrc3239.
- 570 S. S. Ramalingam, M. D. Hellmann, and M. M. Awad. Tumor mutational burden (TMB)
571 as a biomarker for clinical benefit from dual immune checkpoint blockade with nivolumab
572 (nivo) + ipilimumab (ipi) in first-line (1L) non-small cell lung cancer (NSCLC). *Cancer*
573 *Research*, 78(13 Supplement):CT078–CT078, July 2018. ISSN 0008-5472, 1538-7445. doi:
574 10.1158/1538-7445.AM2018-CT078.
- 575 C. Robert. A decade of immune-checkpoint inhibitors in cancer therapy. *Nature Communi-*
576 *cations*, 11(1):3801, July 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-17670-y.
- 577 V. Roth and B. Fischer. The Group-Lasso for generalized linear models: uniqueness of
578 solutions and efficient algorithms. In *Proceedings of the 25th international conference on*
579 *Machine learning, ICML '08*, pages 848–855, New York, NY, USA, July 2008. Association
580 for Computing Machinery. ISBN 9781605582054. doi: 10.1145/1390156.1390263.
- 581 A. Sboner, X. J. Mu, and D. Greenbaum. The real cost of sequencing: higher than you think!
582 *Genome Biology*, 12(8):125, Aug. 2011. ISSN 1474-760X. doi: 10.1186/gb-2011-12-8-125.
- 583 S. Turajlic, K. Litchfield, and H. Xu. Insertion-and-deletion-derived tumour-specific neoanti-
584 gens and the immunogenic phenotype: a pan-cancer analysis. *The Lancet. Oncology*, 18
585 (8):1009–1021, 2017. ISSN 1474-5488. doi: 10.1016/S1470-2045(17)30516-8.
- 586 H.-X. Wu, Z.-X. Wang, and Q. Zhao. Designing gene panels for tumor mutational burden
587 estimation: the need to shift from 'correlation' to 'accuracy'. *Journal for Immunotherapy*
588 *of Cancer*, 7(1):206, 2019a. ISSN 2051-1426. doi: 10.1186/s40425-019-0681-2.
- 589 H.-X. Wu, Z.-X. Wang, and Q. Zhao. Tumor mutational and indel burden: a systematic
590 pan-cancer evaluation as prognostic biomarkers. *Annals of Translational Medicine*, 7(22):
591 640, Nov. 2019b. ISSN 2305-5847. doi: 10.21037/31486. Number: 22.
- 592 Y. Yang and H. Zou. A fast unified algorithm for solving group-lasso penalize learning
593 problems. *Statistics and Computing*, 25(6):1129–1141, Nov. 2015. ISSN 1573-1375. doi:
594 10.1007/s11222-014-9498-5.
- 595 Y. Yang, H. Zou, and S. Bhatnagar. gglasso: Group Lasso Penalized Learning Using a Unified
596 BMD Algorithm, Mar. 2020. URL <https://CRAN.R-project.org/package=gglasso>.
- 597 L. Yao, Y. Fu, and M. Mohiyuddin. ecTMB: a robust method to estimate and classify
598 tumor mutational burden. *Scientific Reports*, 10(1):1–10, Mar. 2020. ISSN 2045-2322.
599 doi: 10.1038/s41598-020-61575-1. Number: 1 Publisher: Nature Publishing Group.
- 600 A. D. Yates, P. Achuthan, and W. Akanni. Ensembl 2020. *Nucleic Acids Research*, 48(D1):
601 D682–D688, Jan. 2020. ISSN 0305-1048. doi: 10.1093/nar/gkz966.
- 602 J. Zhu, T. Zhang, and J. Li. Association Between Tumor Mutation Burden (TMB) and
603 Outcomes of Cancer Patients Treated With PD-1/PD-L1 Inhibitions: A Meta-Analysis.
604 *Frontiers in Pharmacology*, 10, June 2019. ISSN 1663-9812. doi: 10.3389/fphar.2019.
605 00673.

Figures

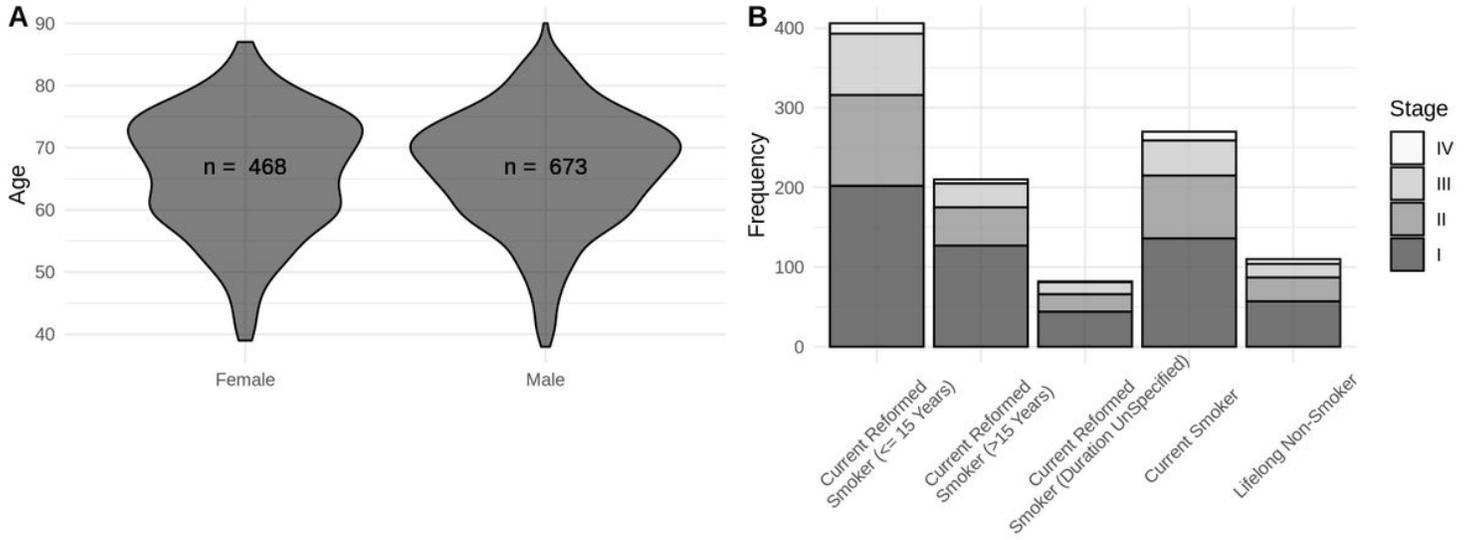


Figure 1

Demographic data for the clinical cohort in Campbell et al. (2016). A: Violin plots of age for patients, stratied by sex. B: Stacked bar chart of patients' smoking histories, shaded according to cancer stage diagnosis.

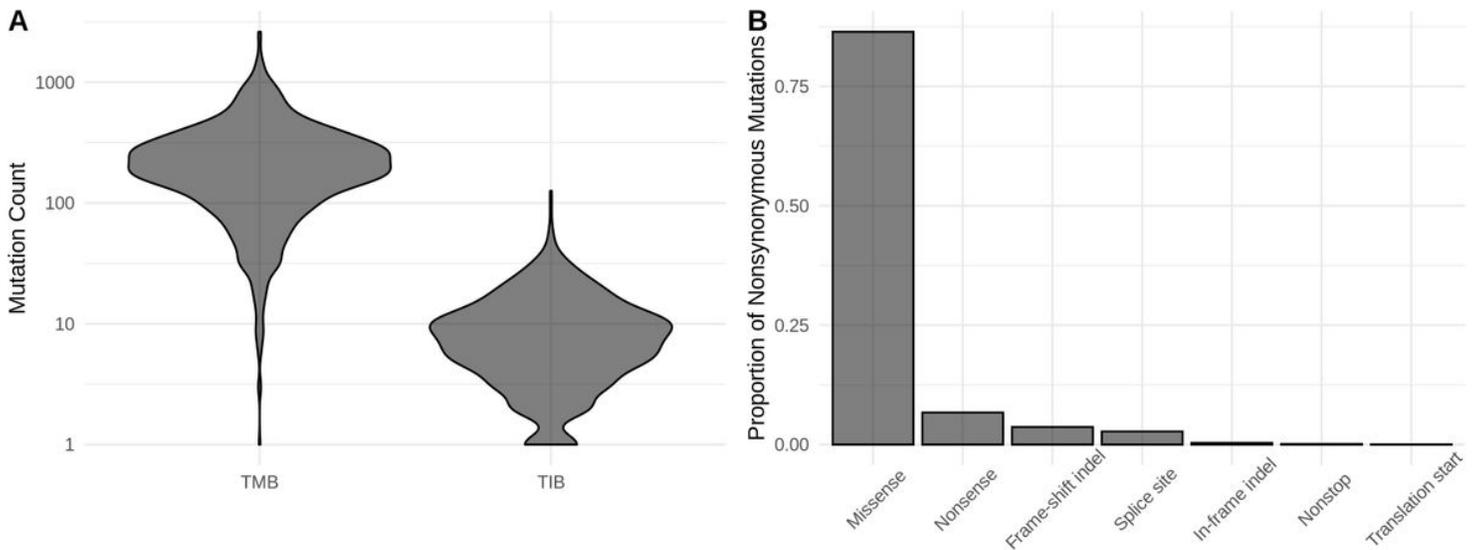


Figure 2

Dataset-wide distribution of mutations. A: Violin plot of the distribution of TMB and TIB across training samples. B: The relative frequency of different nonsynonymous mutation types.

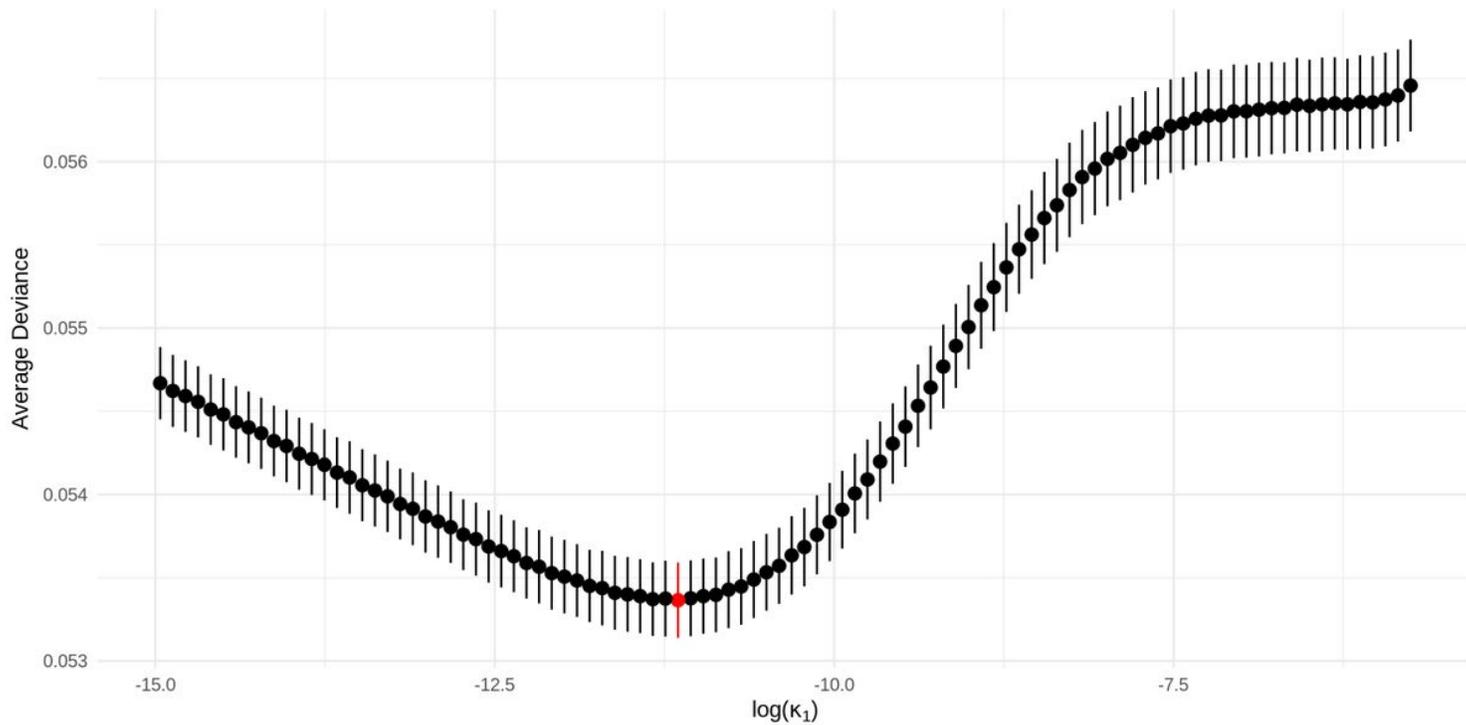


Figure 3

The average deviance (with one standard deviation) across the 10 folds in our crossvalidation procedure plotted against $\log(\kappa_1)$. The minimum average deviance is highlighted red.

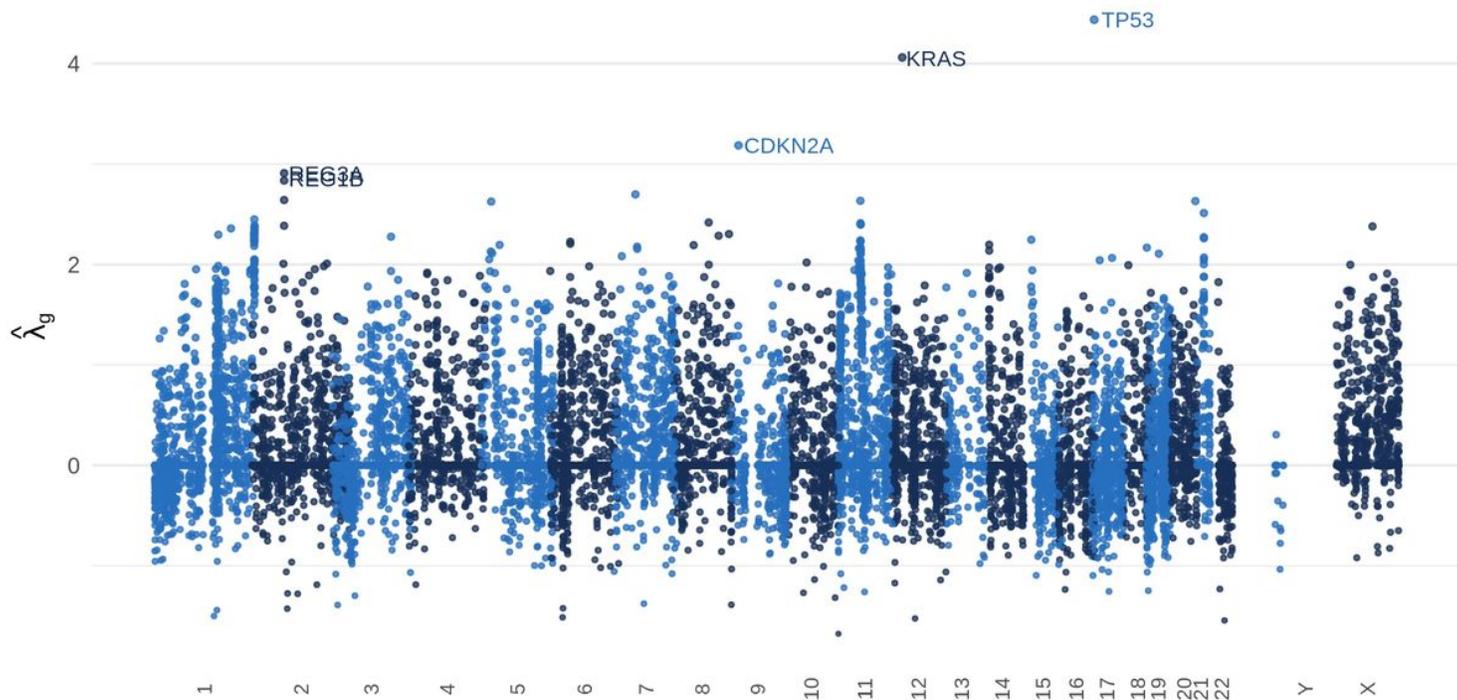


Figure 4

Manhattan plot of fitted parameters \hat{g} and their associated genes' chromosomal locations. The genes with the five largest positive parameter estimates are labelled.

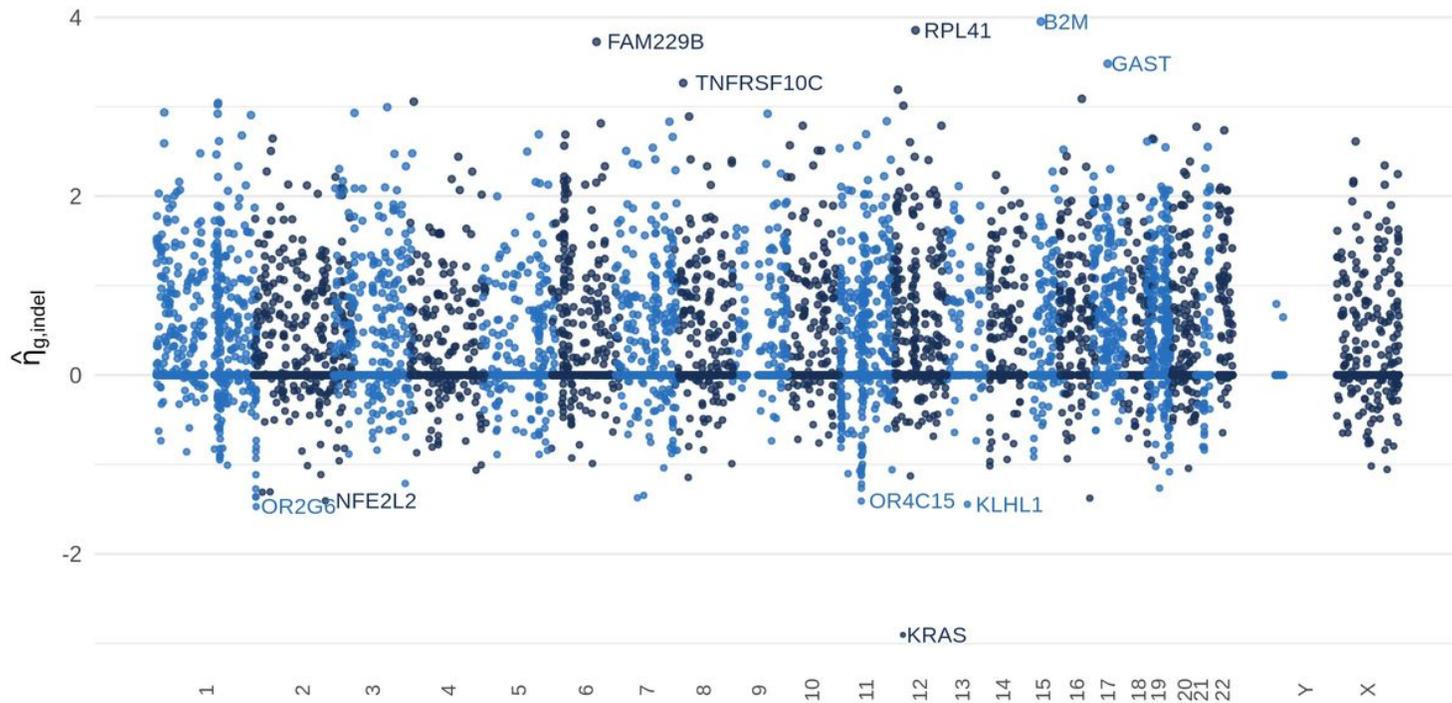


Figure 5

Manhattan plot of fitted parameters $\hat{ng}_{j,indel}$ and their associated genes' chromosomal locations. The five largest positive and negative genes are labelled.

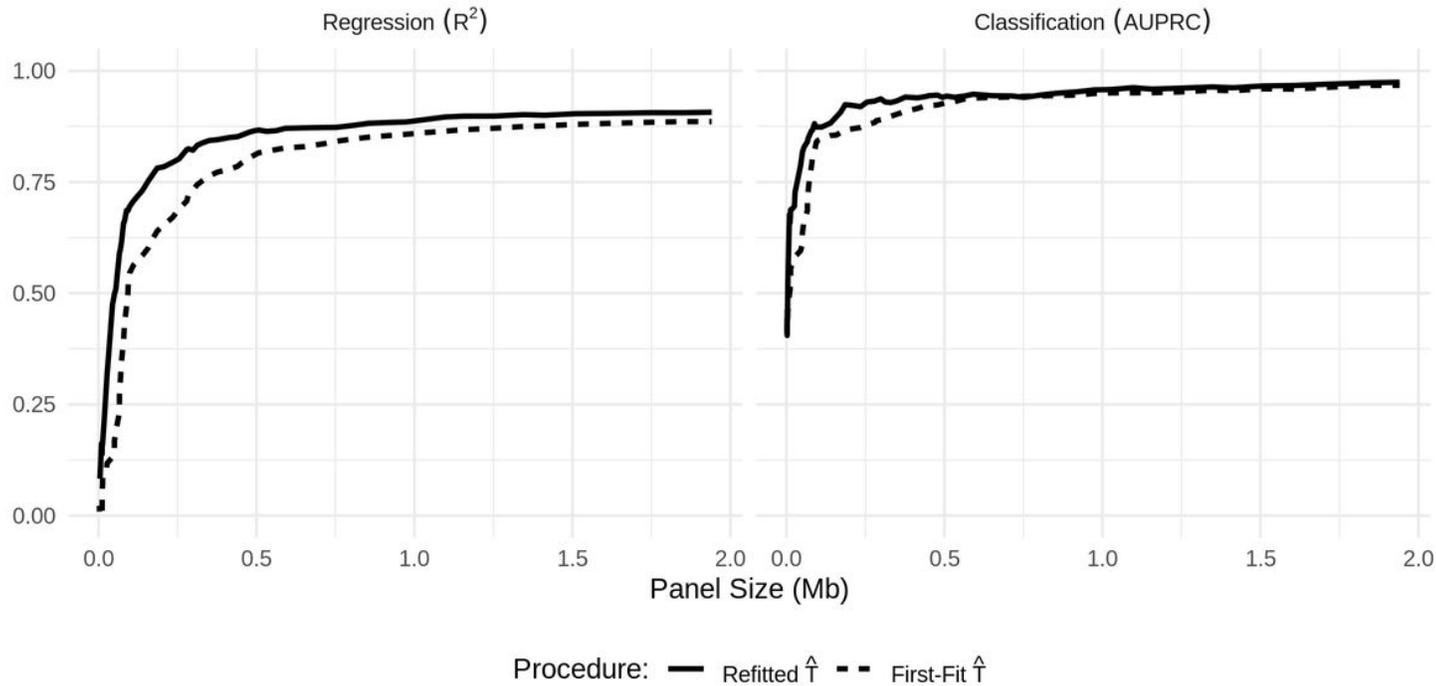


Figure 6

Performance of our first-fit and refitted estimators of TMB as the selected panel size varies. Left: R2, Right: AUPRC.

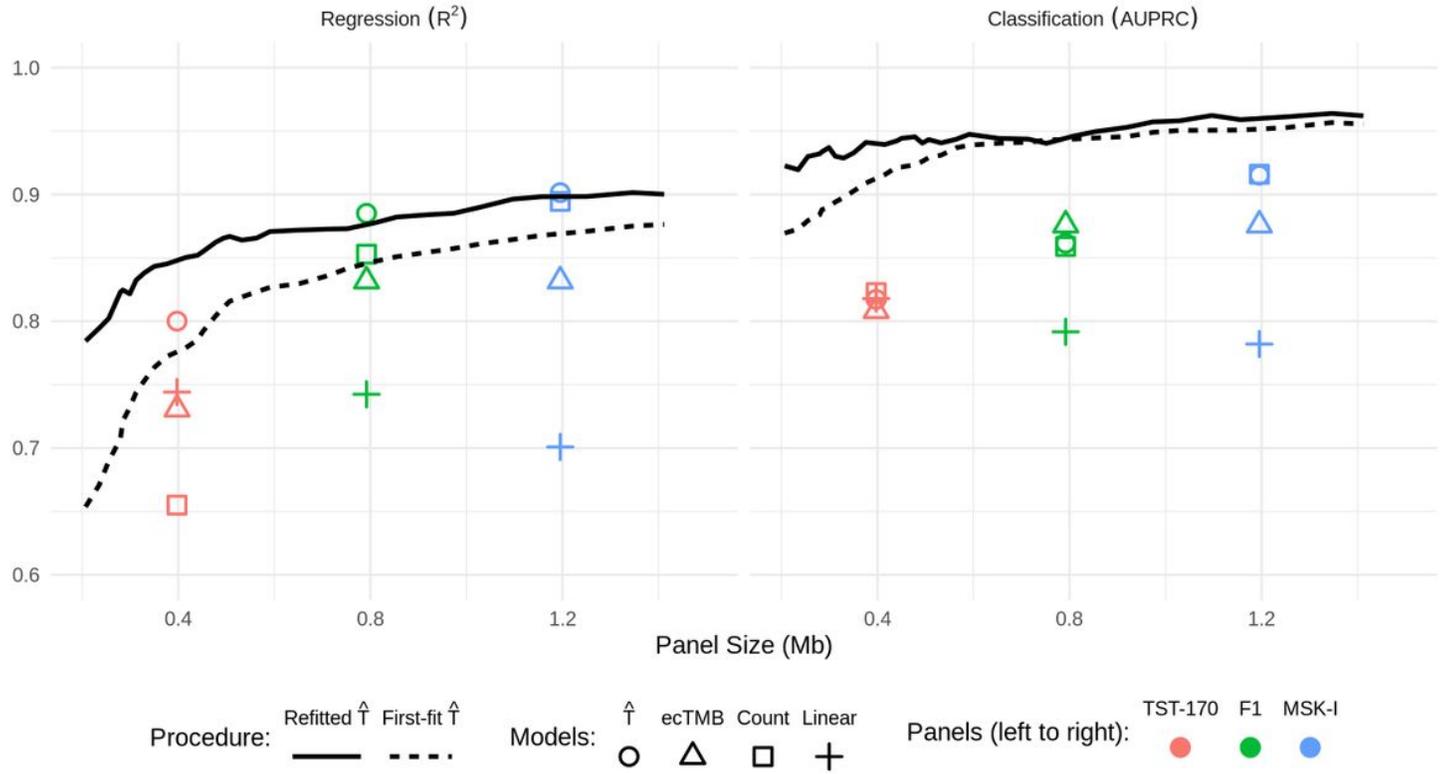


Figure 7

The performance of our TMB estimator in comparison to existing approaches. Left: R2, Right: AUPRC.

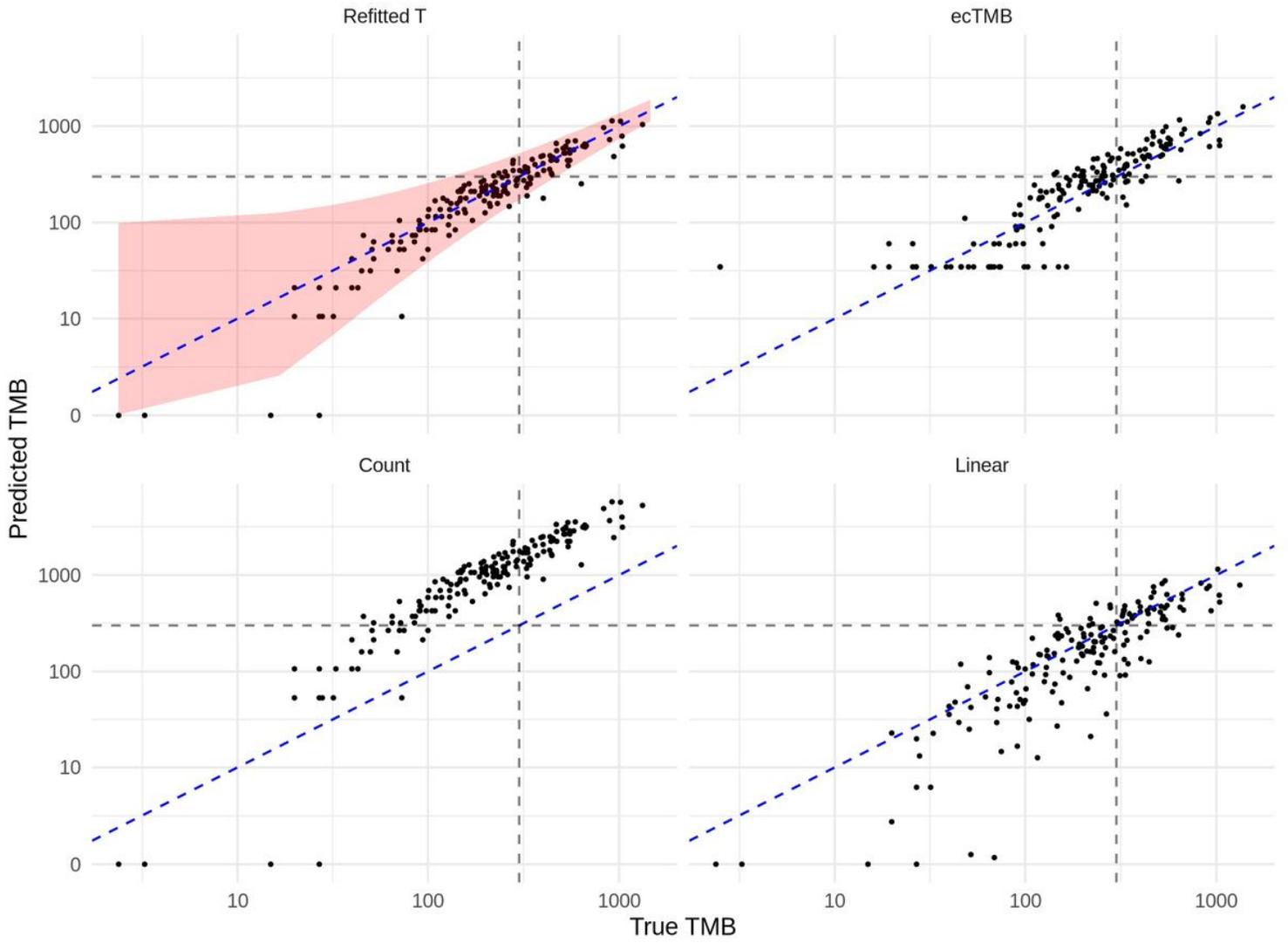


Figure 8

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.

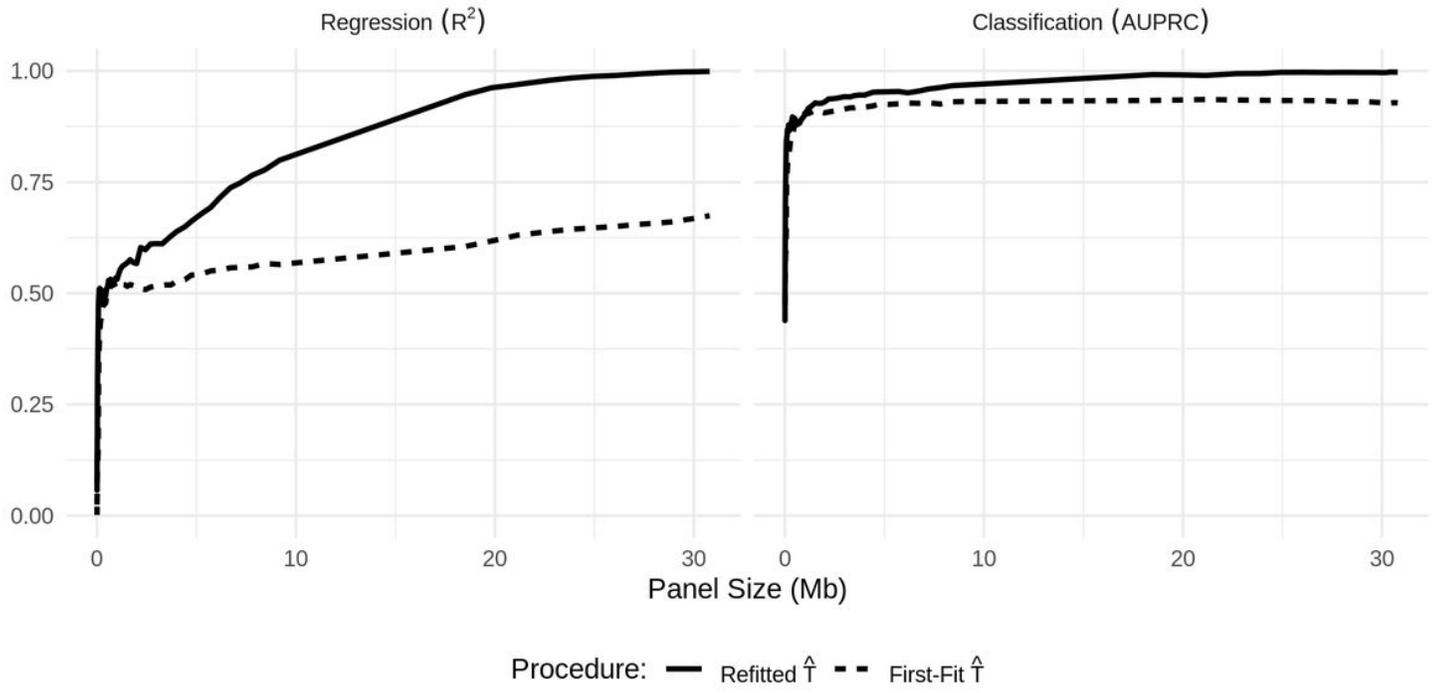


Figure 9

Performance of our first-t and refitted estimators of TIB as the selected panel size varies. Left: R2, Right: AUPRC.

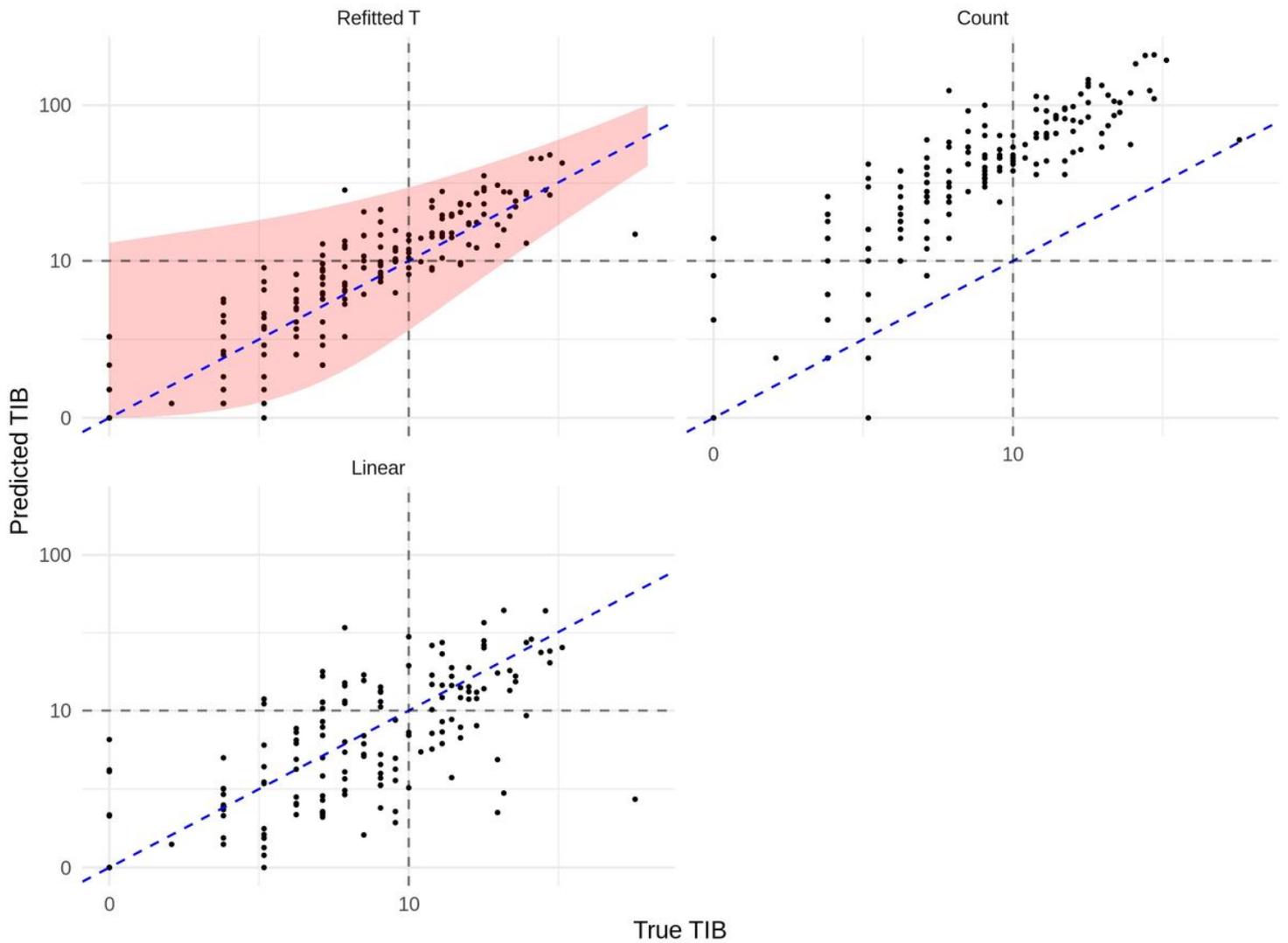


Figure 10

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.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [CanningsEPCflat.pdf](#)