Predictions of Response to Cancer Immunotherapy via Tumour Mutational Burden and Genomic Resistance Markers

Jacob Bradley[1] **and Nirmesh Patel PhD**[2]

[1]Student: Part III Systems Biology, University of Cambridge, UK
[2]Supervisor: Cambridge Cancer Genomics, Cambridge, UK

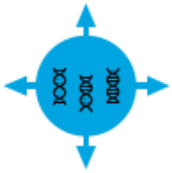
ABSTRACT

The field of immuno-oncology (IO) is making huge advances translating immunological research into successful therapies, in particular Immune Checkpoint Blockade (ICB). The best predictor of treatment effectiveness in most cancers is the metric of Tumour Mutation Burden (TMB). We consider here methods for constructing a cost-effectively concise gene panel to predict TMB. We then investigate the extent to which it is feasible to produce a single gene panel capable of predicting TMB across a range of cancer types, and methods by which one may attempt to do so. We also look into IO resistance mechanisms and genes associated with poor response to ICB, so that we can ensure the best treatment monitoring possible. Finally, we exhibit an IO monitoring panel for Non-Small Cell Lung Cancer, and analyse its performance in comparison to other commercially available assays.

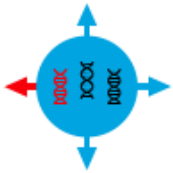
Contents

1 List of Abbreviations

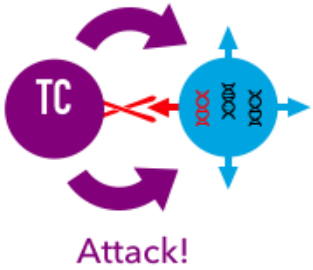
- ATC: Antigen Presenting Cell
- IO: Immuno-Oncology
- IT: Immunotherapy
- ICI: Immune Checkpoint Inhibitor
- TMB: Tumour Mutational Burden
- NAB: Neo-Antigen Burden
- MSI: Microsatellite Instability
- MSI-H: High Microsatellite Instability
- MSS: Microsatellite Stable
- MMR: Mismatch Repair
- dMMR: deficient Mismatch Repair
- ctDNA: Circulating Tumour DNA
- cfDNA: Cell-Free DNA
- tDNA: tumour DNA
- CTLA-4: Cytotoxic T-Lymphocyte Associated Protein 4
- B7: binds to CTLA-4 (also known as CD80)
- PD-1: Programmed Death Protein 1
- PD-L1: Programmed Death Ligand 1
- MHC: Major Histocompatibility Complex
- GEP: Gene Expression Profile
- DEG: Differentially Expressed Gene
- OS: Overall Survival
- CRC: Colorectal Cancer
- MM: Multiple Myeloma
- NGS: Next Generation Sequencing



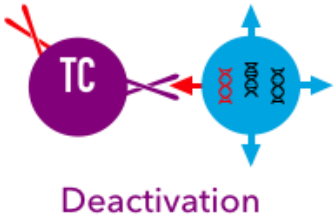
(a) Cells present small peptide antigens to their surface (blue arrows).



(b) In cancer, mutations in sequence mean alterations in antigens (red arrows).



(c) Cytotoxic T cells recognise these and become activated (T-cell receptor shown in red).



(d) To prevent autoimmunity, T cells' activation is very heavily regulated via factors such as CTLA-4 (shown in purple).

Figure 1. Mechanism of the immune checkpoint CTLA-4.

2 Introduction

2.1 Cancer is a Disease of the Genome

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 and phenotypes, is initiated and driven by the accumulation of mutations in a subset of a person's cells. In a typical cancer, mutations will accumulate (developing subclonally from a single original mutant cell) and change the behaviours of the descendant cells that will form a tumour. Normal mechanisms controlling the balance of cell growth and death are circumvented, so that tumour cells divide overly rapidly and/or die sparsely. Once decoupled from the body's control, success for cells comes in their ability to proliferate and survive. A highly competitive tumour micro-environment lends a high selective pressure[1] for even marginal fitness gains from further mutations, so that a progressed malignant tumour will have accumulated a repertoire of functional alterations, as well as developing a complex set of interactions with surrounding 'normal' tissue[2], often including shielding the tumour from immune responses and the growth of a bespoke blood supply (angiogenesis)[3].

2.2 Immune Responses to Cancer are Mediated by Checkpoints

All nucleated somatic cells produce surface antigens (Figure 1a), which are short-chain peptides derived from the proteins constituent in the cell. The immune system has a number of antigen-recognising cells, which are selected to have receptors that recognise cells displaying irregular peptide fragments. This might be because the cell's molecular machinery is being abused by a virus, because mutations in the cell's DNA are being transcribed and then translated with an altered sequence (1b), or because unaltered but normally non-coding DNA is being errantly transcribed. The identification of cells presenting foreign 'neoantigens' invokes an aggressive immune response[4], including direct cell killing by cytotoxic T-lymphocytes (1c).

It is essential for the body to protect against autoimmunity, and all responses throughout the innate and adaptive immune system are strictly regulated. Various mechanisms contribute to this control, including immune checkpoints. These function via alternate receptors on the surface of immune system cells (such as cytotoxic T-cells) which can also bind with sites on the surface of the antigen presenting cell (APC). By binding they downregulate T-cell activation. The sites to which they bind can be the class I major histocompatibility complex of the APC itself, in which case they act in competition with the activating T cell receptor[5]. This is the case for the CTLA-4 checkpoint (Figure 1d), which competes with CD28 for binding with the B7 site on the APC. Alternately, they can bind to a separate APC surface ligand, such as in the case of the PD-1/PDL-1 checkpoint inhibition axis.

2.3 Immunotherapy is an Emerging Field

There has been an explosion of interest in cancer therapies targeting immune response, with immune checkpoint blockade therapy winning the Nobel Prize in Physiology or Medicine in 2018. Immune Checkpoint Inhibitors/Bloackades (ICIs/Bs) are monoclonal antibodies that bind to and block immune checkpoints such as CTLA-4 and PD-1 (described above). Binding to immune checkpoints 'inhibits the inhibitors', allowing the activating T-cell receptor to bind successfully (Figure 2) and in principle releasing a fuller immune response against developing and advanced tumours. Therapies targeting both of the checkpoint mechanisms described have recently been approved for a small set of cancers, with strong and sustained response in a subset of patients.

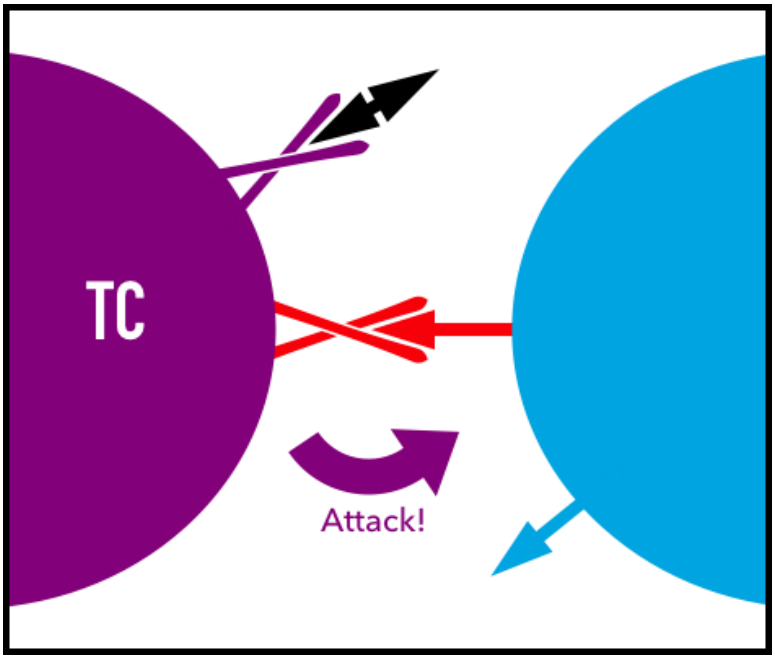


Figure 2. Checkpoint inhibitors are monoclonal antibodies that bind to checkpoint proteins such as CTLA-4 and PD-1 (shown in black).

2.4 Tumour Mutational Burden is a Genomic Biomarker for Immunnotherapy Response

The success of an aggressive immune response to a tumour depends on various factors, including sufficient infiltration of lymphocytes to the tumour micro-environment. Even with a large lymphocyte presence, however, it is necessary for tumours to present foreign-recognisable neoantigens. The neoantigen burden (NAB) is defined as the number of of neoantigens present on cells in a tumour that may be recognised as foreign by immune cells. While various computational methods to predict the structure of errantly translated peptides have been developed, it is difficult to predict accurately without a large amount of information including proteomic and transcriptomic data. As a proxy for NAB, tumour mutational burden (TMB) is used as a purely genomic biomarker. TMB is defined simply as the total number of simple mutations detected throughout a tumour sample compared to normal tissue. While not every mutation may produce a neoantigen with high binding affinity to T-cell receptors, in a tumour with high TMB it is more likely that many neoantigens will be present. TMB has been shown to correlate well with NAB in various tumour types.

2.5 TMB Varies Across and Within Cancer Types

We calculated TMB for tumour samples across a number of cancer types. Median tumour mutation burden ranges by 2-3 orders of magnitude across tumour types, but also varies by more than an order of magnitude within almost every tumour type, and for some types the interquartile range of mutation burden spans more than an order of magnitude (e.g. squamous-cell skin cancer, colorectal cancer) (Figure 3).

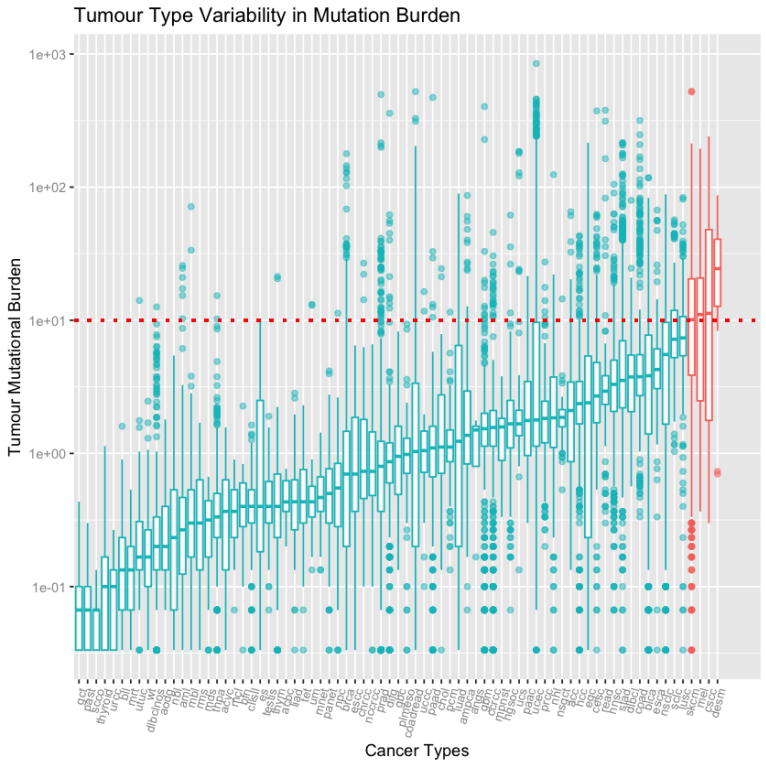


Figure 3. There is large variation between but also within tumour types. An (arbitrary) threshold of ten Mut/Mb is given, and the cancers with a median higher than that are highlighted. They are mostly cancers of the skin.

3 Results

3.1 TMB Estimation in a Single Cancer Type

3.1.1 Genome-Wide Association

We first looked at methods which are genome-wide and gene-agnostic, i.e based purely on the chromosomal position of mutations in a variant-called cancer dataset, without reference to the gene that contains them. We used a two-step method, first selecting a set of loci based on a sliding window algorithm (detailed in Section 5.2) which ranks regions based on the association between their local mutation density and genome-wide TMB score. We then used the set of locations generated as the basis for training a variety of models for estimating TMB. A one-step method combining location selection and model fitting would be ideal, but the high-dimensional nature of our dataset makes this computationally impractical.

Associations with TMB Vary Widely by Genomic Region We performed a sliding window algorithm with a very wide window across a set of non-small cell lung cancer data from cBioPortal (Figure 4). We saw a wide variety in association between local and global TMB at different points in the genome, indicating that we could expect a set of regions chosen on the basis of their score would perform far better at TMB estimation than an arbitrary set of locations of the same overall size.

Applying Linear Models to Panels Generated by an Iterative Sliding Window Procedure Show a Panel-Size Independent Peak in Performance We used cBioPortal data for non-small cell lung cancer to generate a training and test set of tumour samples. On the training set we applied the iterative "zoom and repeat" sliding window procedure described in Methods (Section 5.3). Specifying a desired panel size, for each iteration of the algorithm we selected the best set of windows (by p-value) whose widths summed to the given panel size. We then used the same training set to produce training matrices with columns given by windows selected as part of the panel and rows given by individual tumour samples. We then applied an ordinary linear model to each of our training matrices. Finally, we used the reserved test data to test the performance of the models generated on unseen samples.

We found improved performance on training (based on proportion of variability explained) and test sets for larger panel sizes, and a sharp peak in test set performance at around iteration six, regardless of panel size (Figure 5). We considered various techniques for addressing this overfitting at later stages of the algorithm, including cross-validation and ridge regression, but at

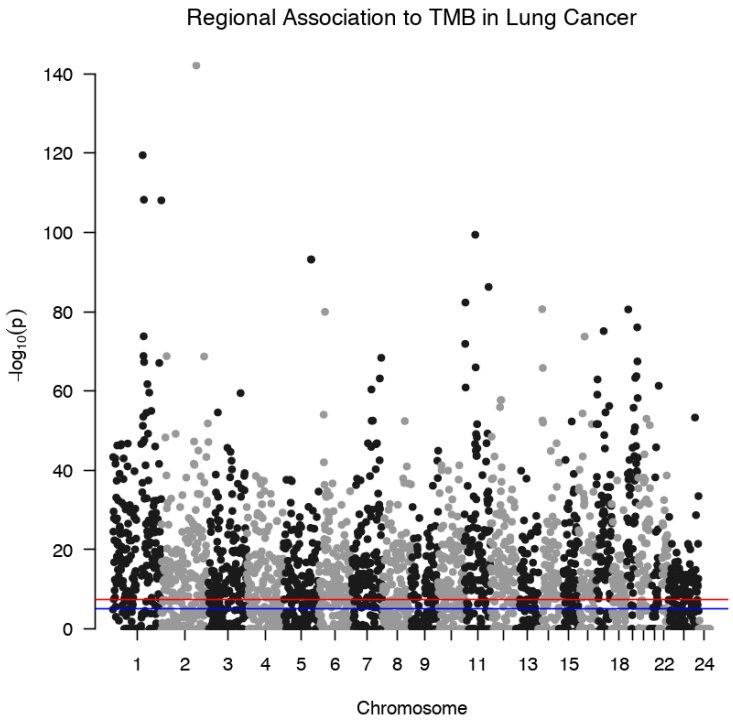


Figure 4. Genome-wide association for local TMB to global TMB. Here window size is 10^6 base pairs and the window moves with jumps of 10^5 base pairs.

the time of investigation found more promising results via gene-based methods so swtiched our focus to these.

Naively Applying Random Forest or Support Vector Machine Learning Gives No Immediate Improvement on TMB High/Low Classification In a clinical setting, what is often needed practically is a high/low classification, to recommend to a clinician whether it might be sensible to proceed with ICB treatment or not, with some sensible threshold. While there has been a recent international effort towards standardising what is considered 'high' TMB, in most studies the figure is decided upon independently and retrospectively. We used a threshold value of 10 Mut/Mb, as this is consistent with other recent research and also close to the 80% quantile value for TMB Score in our NSCLC data (this is consistent with the within-cancer 80% quantile that is been used as a threshold in other sections).

Given a classification problem, other machine learning techniques such as Random Forests (RF) and Support Vector Machines (SVM) are more appropriate than in the regression case. We applied these methods, alongside our previous linear regression used as a classifier, to the training and test sets described above. The results are given in Figure 6. We found that neither random forests nor support vector machines gave marked improvement on linear regression. SVMs were more resilient at later iterations of the sliding window procedure, continuing to increase in sensitivity but suffering lower specificity than either of the other two methods. In an unbalanced population (weighted towards TMB Low tumours), it is particularly important that specificity remain high, in order to avoid an untenable number of false positive cases.

3.1.2 Gene Oriented Methods

We next investigated an alternate approach, which was to group exomic mutations by gene, and use these instead of genomic windows of a defined length as features. There are disadvantages and advantages to this approach.

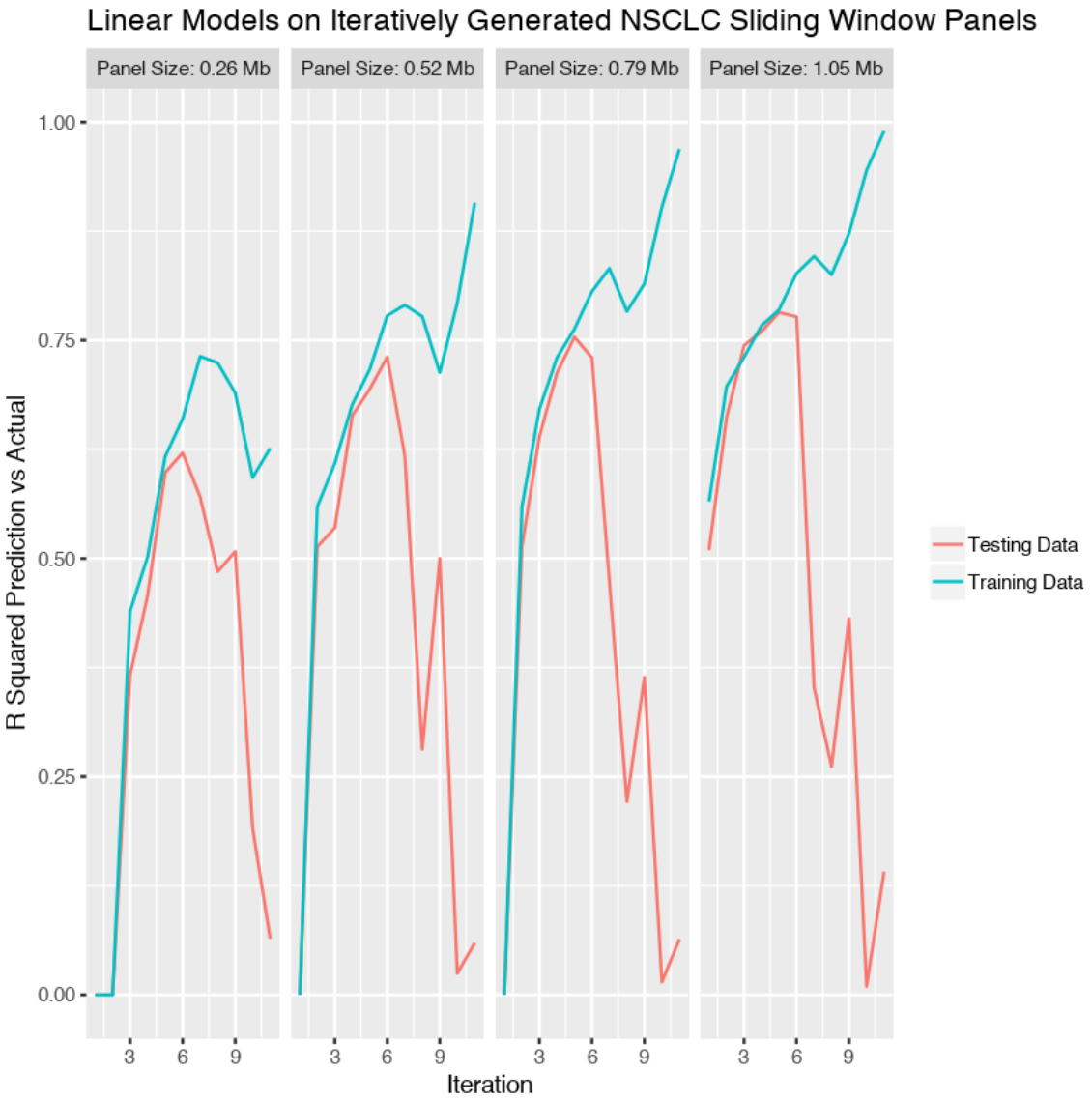


Figure 5. Performance of linear models across four panel sizes and across iterations of the zoom, repeat window size optimisation algorithm. For the first iteration of the algorithm we use windows of size 2^{20} (approximately 1 megabase), zoom by a factor of two at each iteration and iterate 10 further times, so that the final iteration is working on window sizes of 2^{10} (approximately 1 kilobase).

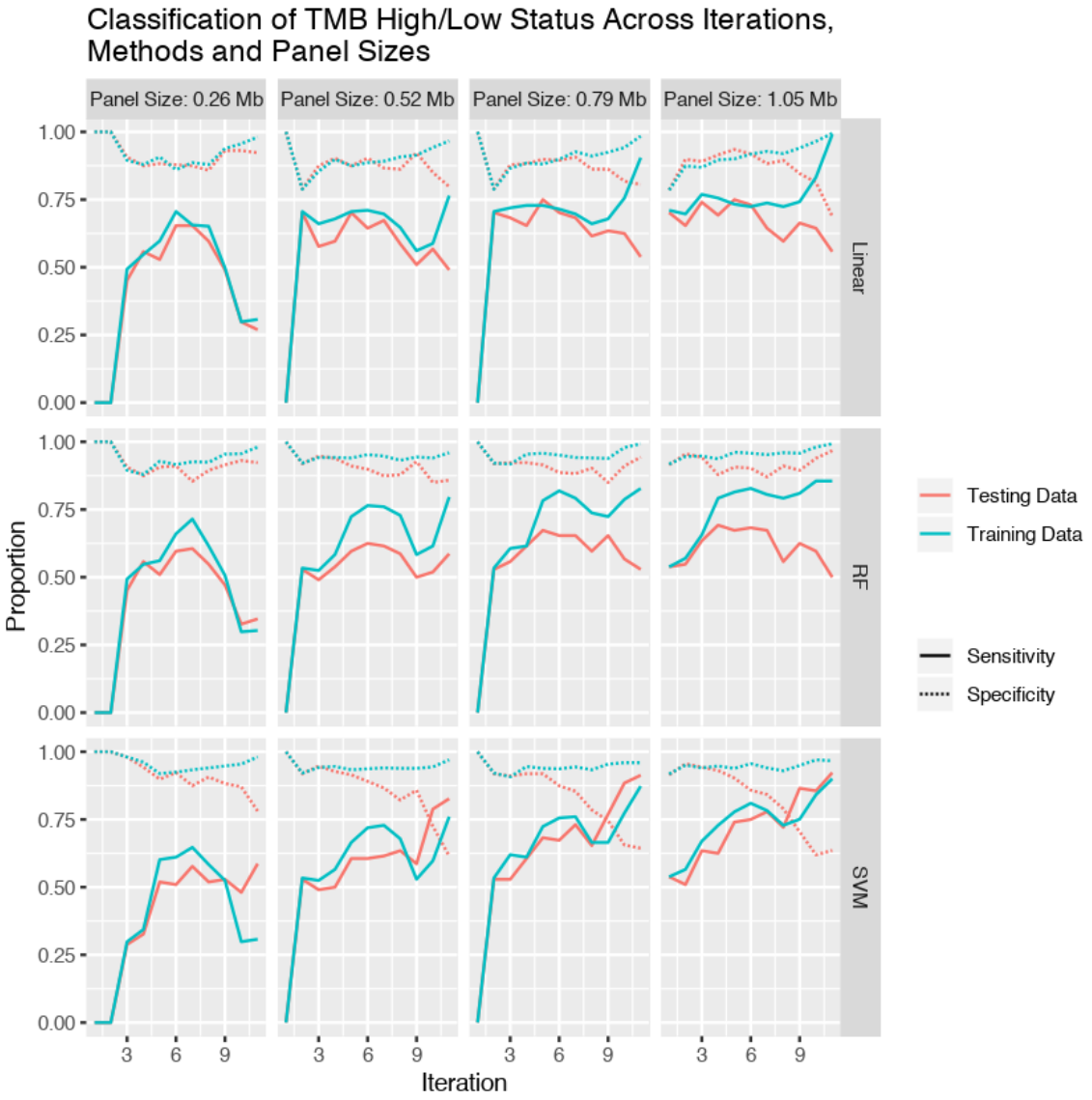
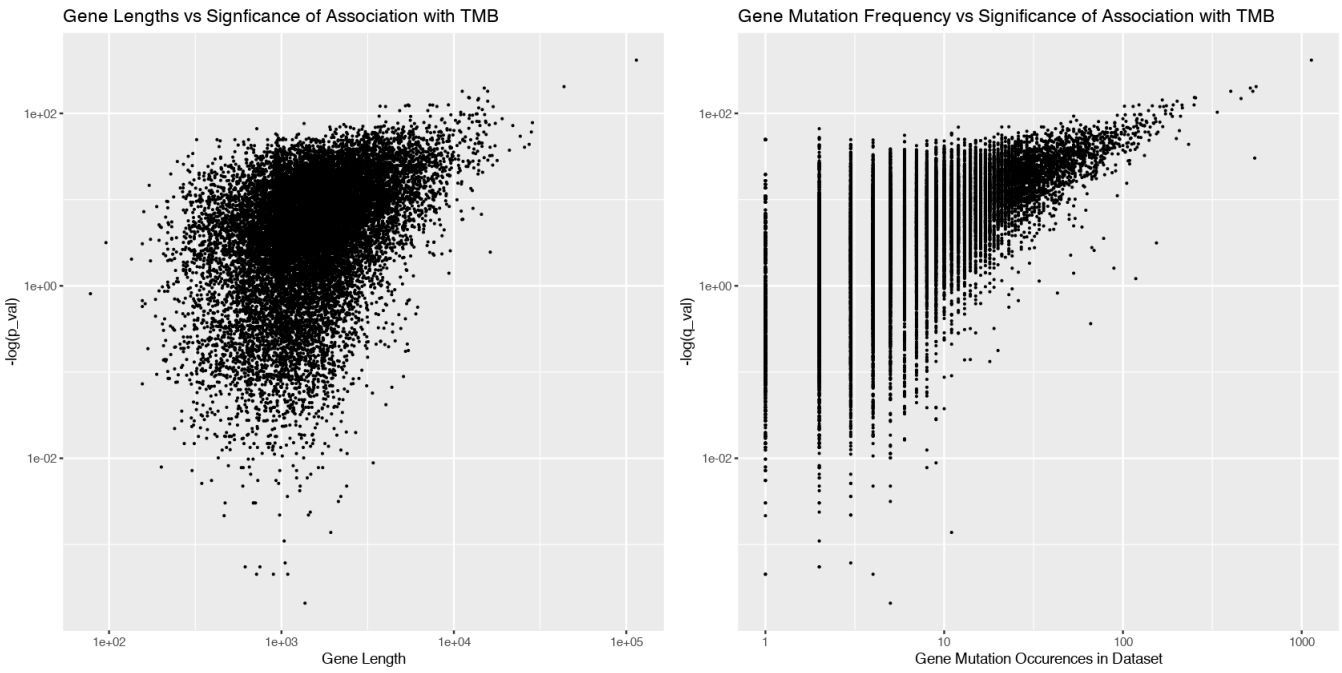


Figure 6. Sensitivity and specificity are reported for applying three methods (Linear, RF, SVM) across iterations of our sliding window algorithm. Sensitivity = proportion of true TMB High samples that are predicted as TMB High, specificity = proportion of of TMB Low samples that are predicted as TMB Low.



(a) Longer genes tend to have higher association with TMB.

(b) Genes more often mutated tend to have higher association with TMB.

Figure 7. Mutation rate and length influence a gene's likelihood to be significantly predictive of TMB.

Advantages:

- Improved confidence that there is not (potentially) wasteful or misleading non-coding material in our genomic windows that have been 'carried along for the ride' by being in close proximity to important regions
- Utilisation of known biology
- Computationally less expensive
- No need to optimise 'window size' parameter

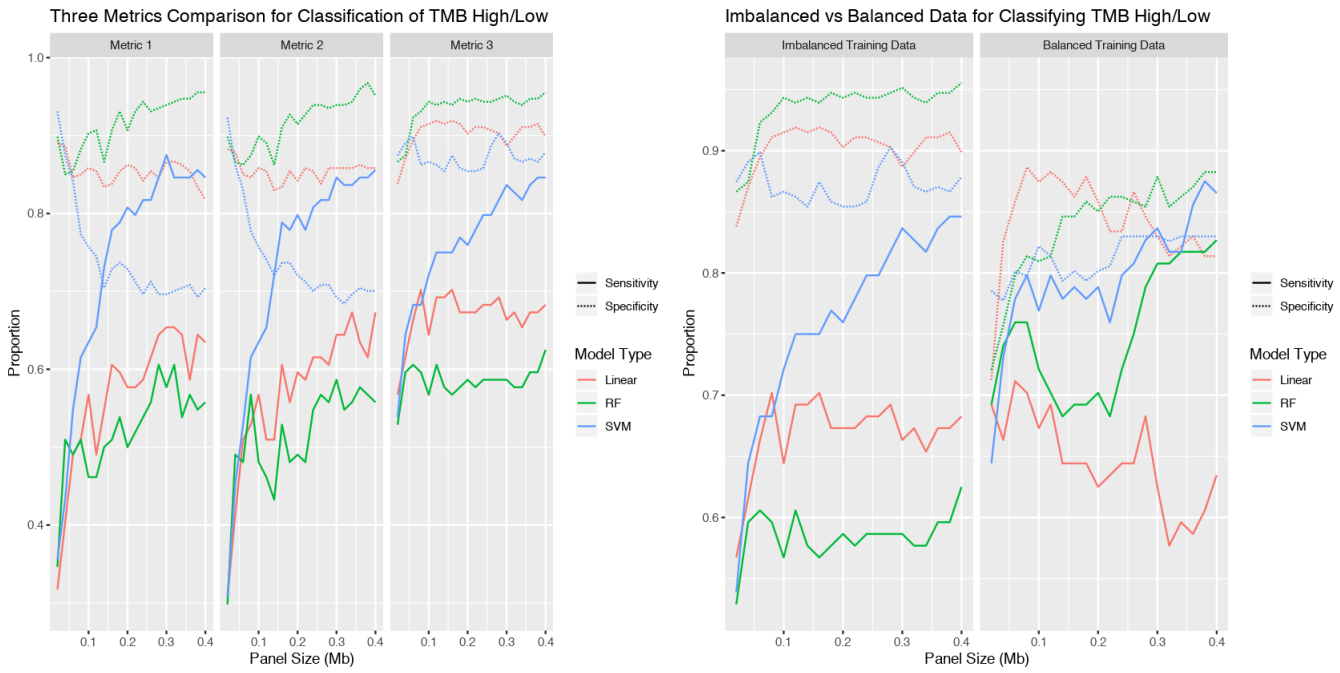
Disadvantages:

- Loss of information if different regions of a gene are of differing importance
- We now have to account for differences in gene size altering the statistical importance of their local mutation burden

Gene Mutation vs TMB Association Scores Correlate Well With Gene Length and Mutation Frequency For each of the genes in our training dataset, we computed a p-value for that gene having non-zero effect in a linear regression of TMB against mutation count in that gene. Figure 7a shows a strong association between gene length and p-value. We also found, unsurprisingly, that the frequency with which a gene is mutated in our dataset is associated with p-value. It was clear that simply selecting genes to include on the basis of association with TMB would not be the most efficient way to produce a concise NGS panel. This provided motivation to develop further methods for weighting association with TMB against gene length.

Adjusting Gene vs TMB Correlation Scores to Account for Gene Length Improves Performance of TMB Estimation Panels We proposed and tested three metrics for gene selection, $\mathbf{M}_1$, $\mathbf{M}_2$ and $\mathbf{M}_3$, which are defined in Methods (Section 5.7) and derived/explained in Appendix A . To construct a panel of a certain size using a given metric, we ordered genes in descending order according to that metric, and selected the largest number from the top of the ordering whose lengths did not sum to greater than the required panel size. $\mathbf{M}_1$ was a crude and only loosely motivated metric that we used for benchmarking the other metrics. $\mathbf{M}_2$ was derived by assuming that mutations occur as Bernoulli random events for each nucleotide pair, while $\mathbf{M}_3$ was adjusted from $\mathbf{M}_2$ to acknowledge that mutations do not occur uniformly across the genome, and to weight more favourably on genes that are mutated at high rate.

We found the best performing metric/model combination to be SVM models applied to panels generated by our third metric. For all three metrics SVMs were more sensitive for almost all panel sizes, but for $\mathbf{M}_1$ and $\mathbf{M}_2$ this came at the cost of distinctly



(a) Performance of Linear, RF and SVM models across a range of panel sizes, with panels created using the M1, M2 and M3 metrics.

(b) Performance of Linear, RF and SVM models across a range of panel sizes, with panels created using the M3 metric, and models trained on imbalanced (#(TMB High Samples) < #(TMB Low Samples)) and balanced (#(TMB High Samples) = #(TMB Low Samples) datasets.

Figure 8. Testing a variety of methods for selecting gene panels and training classification models.

reduced specificity (Figure 8a). A panel size of 0.25Mb was required for specifity and sensitivity both over 80% for SVM trained, $\mathbf{M}_3$-chosen panels. As a comparison, selecting genes purely on the basis of their association with TMB (with no adjustment for gene length) and restricting to a panel size 0.25Mb gave a panel of eight genes. We trained linear, RF and SVM models for TMB High/Low on these genes, none of which classified with a sensitivity of above 60% on test data.

As a further comparison, we used LASSO regression (see Methods Section **??**) to create a series of panels of increasing size. LASSO is a one-step method, incorporating model selection and linear model fitting into a single step, via a penalising weight factor. We used a modified weight factor to incorporate the lengths of genes. We then additionally applied RF and SVM models to the panels created by each iteration of the LASSO process, with the TMB classification performance shown in Figure 9. We found that the LASSO underperformed compared to our own method.

Balancing Training Dataset Closes Sensitivity/Specificity Gap A common problem when applying machine learning techniques to binary classification problems is imbalanced classes. This is such a situation, where by definition the population of TMB Low samples will be much larger (approximately 20% of sample in this case). In vastly imbalanced cases, a naively applied model will often find the optimum accuracy is obtained by predicting that any input belongs to the larger class. We therefore trained linear, RF and SVM models on panels chosen using the $\mathbf{M}_3$ metric on synthetically balanced training data, which we produced by randomly deleting TMB Low samples until the classes were equally sized. This had a great impact on the sensitivity of RF and SVM models (Figure 8b), particularly at small panel sizes, but was offset by poorer specificity.

A more direct way to balance sensitivity and specificity was to train regression models and adjust the threshold of predicted TMB at which a sample was predicted to be TMB High (while holding constant the threshold to be actually TMB High). We discuss this in our analysis of an example panel's performance in Section 3.4.2.

3.2 TMB Estimation Across Cancer Types

Genomic Hotspot Distributions Across Three Cancer Types Motivate Pan-Cancer Biased Data Pooling Procedures We performed genome wide assocation analyses across three cancer types (breast, colorectal and lung) using cBioPortal data,

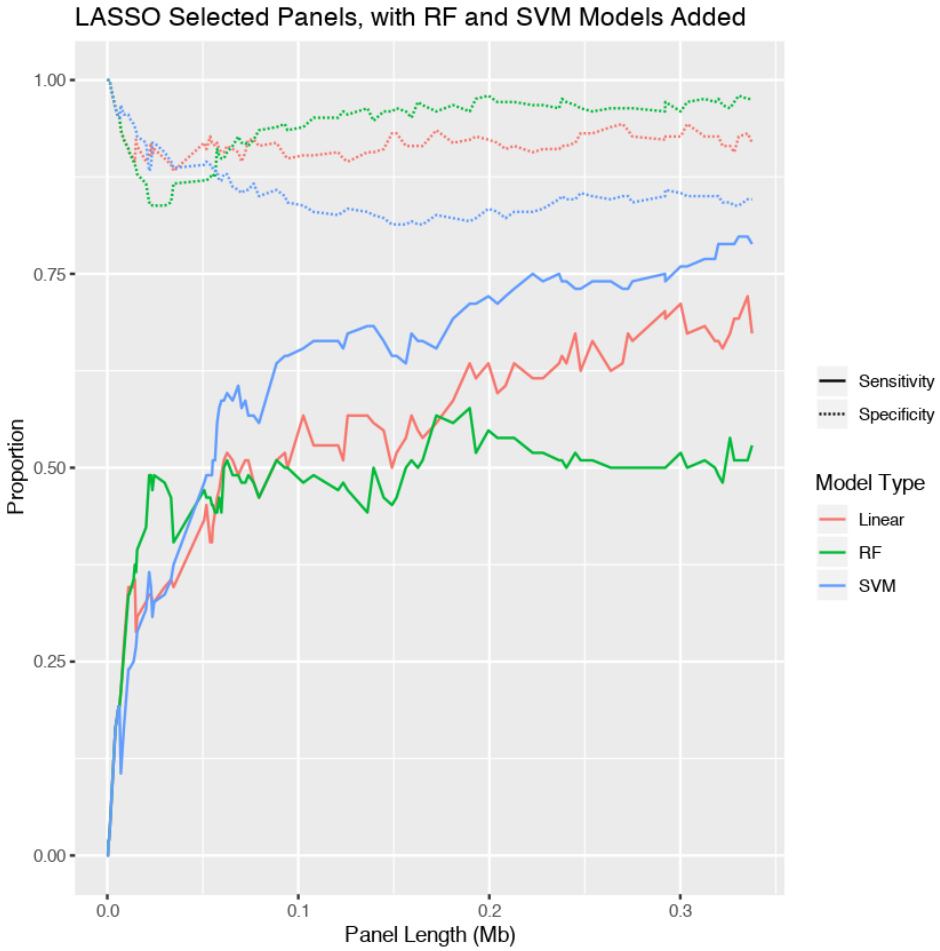
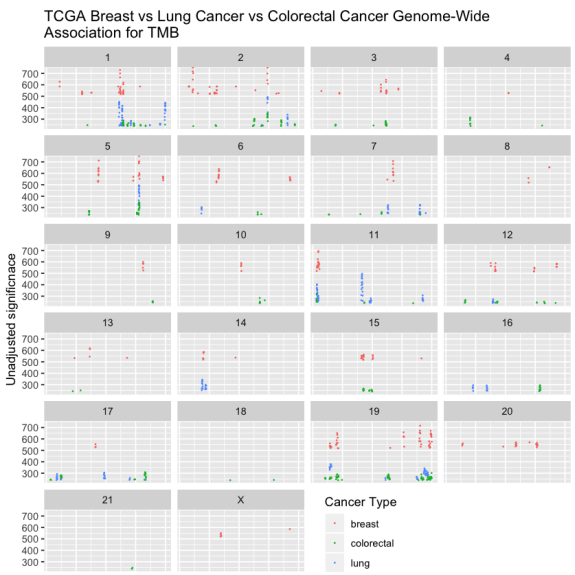
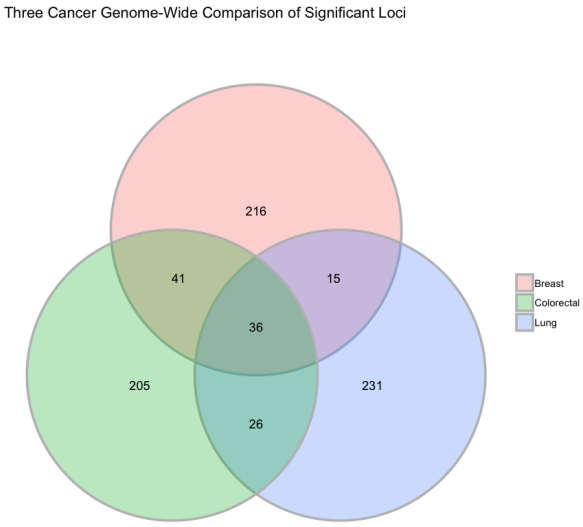


Figure 9. Sensitivity/specificty performance of LASSO generated models while increasing panel size.



(a) Different regions are significant in determining TMB for different regions, and show various levels of overlap between disease types.



(b) The number of hotspots shared between cancer types.

Figure 10. A comparison across three cancer types of the genomic loci whos local mutation density is best correlated with global mutation density.

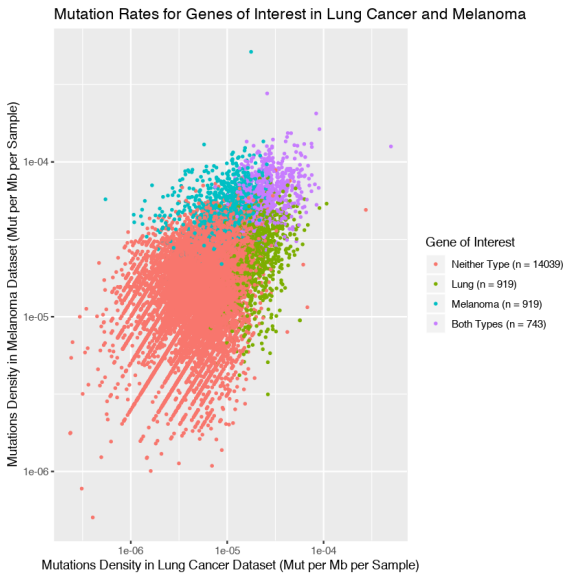
and for each cancer type selected a set of regions of interest. We defined a region of interest to be one whose p-value from our sliding window procedure was in the top 1% of its cancer type. When we compared these loci across cancer types (displayed in their chromosomal locations in Figure 10a), we found a mixture of hotspots shared between all three cancer types, two cancer types, or unique to a single cancer type. We report the number of hotspots in each category in Figure 10b. While the majority of hotspots were unique to their cancer type, the disproportionately high (as opposed to random chance) number of shared loci motivated approaches to panel selection biased towards pan-cancer applicability.

Where Frequently Mutated Regions Between Two Cancer Types Are Similar, Joint Cancer Panels Can Perform Comparably to Individual Panels Given the successes of gene-oriented methods in the previous section, we began with the same selection metric, and attempted to extend our previous work towards panel selection for two cancer types simultaneously. We considered non-small cell lung cancer and skin cutaneous melanoma. We found that simply pooling data from the two cancer types and treating them as if they were one reduced the accuracy of TMB status classification in both cancer types. After considering several pooling procedures, the most successful we trialled was to compute $\mathbf{M}_3$ metrics for the genes in each dataset, order the genes by rank, and then take a 'pooled rank' given by the minimum of each gene's two ranks (one from each dataset). We then selected genes to include in our panels via ordering by pooled rank, and trained SVMs separately on the two populations. Under this procedure applied to a test set, we found comparable results in both cancer types to panels developed separately for each type via the $\mathbf{M}_3$/SVM method (Figure 11a). We also found in this case that there was significant overlap between genes scoring well for metric $\mathbf{M}_3$ between the two types, due to a strong correlation in genes' mutation density (see Figure 11b). When we applied the same techniques to cancer pairs with more differing mutation profiles, we observed far poorer performance relative to individually generated panels. We hypothesise that distance between two cancers' mutation profiles will be a good predictor for how easily a panels may be developed able to accurately class TMB in both tumour types.

3.3 Genomic Markers of Resistance

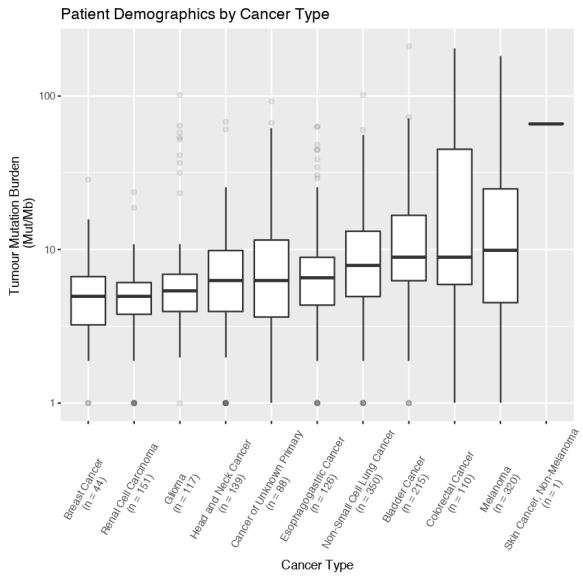
To study resistance against immune checkpoint blockade, we used a publicly available dataset[6] of patients ($n = 1661$) across a variety of cancer types and three ICB drug regimens (PDL-1, CTLA4 and a combination). Study demographics, including TMB distribution across cancer types and overall survival (OS) Kaplan-Meier curves stratified by drug regime, are shown in Figure 12. TMB was measured for each patient, and sequencing data was available for the MSK-IMPACT gene panel[7], comprising 474 known cancer-related genes.

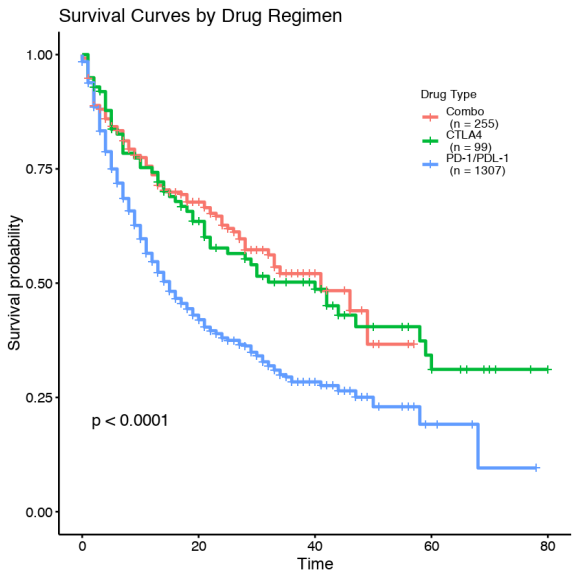
(a) Performance of panels produced for each cancer, versus jointly produced panels.



(b) Assocation between gene mutation density across all genes that occured in NSCLC + SKCM studies. Genes of interest in one or both cancer types highlighted. Gene of interest was defined as having an $\mathbf{M}_3$ score in the top 10% within cancer type.

Figure 11. Comparable performance to individual panels achieved for joint panels between non-small cell lung cancer and skin cutaneous melanoma.





(a) TMB distribution across the 11 cancer types included in the study. **(b)** Survival curves for the three drug regimens included in the study.

Figure 12. Demographics of TMB and response to treatment across cancer type and drug regimen respectively.

3.3.1 Four Genes Show Significant Decrease in Overall Survivall for ICB Patients

For each gene in the IMPACT panel we used a Cox proportional-hazards model with two cofactors: TMB and the presence of at least one mutation in that gene. We also applied a log-rank test on difference between survival curves for mutations in each gene, with TMB-H status (top 20% quantile within cancer type) as a stratifying variable. The inclusion of TMB as a cofactor/stratifying variable was to mitigate the effects of high TMB being known as a hazard reducing factor. We said that a gene was a significant resistance gene if it satisfies one of the following:

- Gene is mutated in at least 30 samples (to ensure applicability of parametric test), with a (Benjamini-Hochberg corrected) Cox proportional-hazards p-value ≤ 0.05 (to ensure significance) and a positive z-value (to ensure it is deleterious)

- Gene has a (Benjamini-Hochberg corrected) log-rank p-value ≤ 0.05 and a positive z-value.

Four genes were identified as significant resistance genes: *TP53, STK11, KEAP1* and *SMARCA4* (Figure 13). All four were significant by Cox test and had sufficient mutant sample size ($n = 98, 132, 98$ and 738 respectively) (Figure 14). *TP53* and *STK11* were also significant by the non-parametric log-rank test. Given that all four genes producing significant results had a good mutated sample size, it is not surprising that both genes identified by log-rank were also identified by Cox, as a non-parametric test will typically have lesser power.

Of the patients treated with immune checkpoint blockade in the study, 51.8% had a mutation in one of the four resistance genes identified.

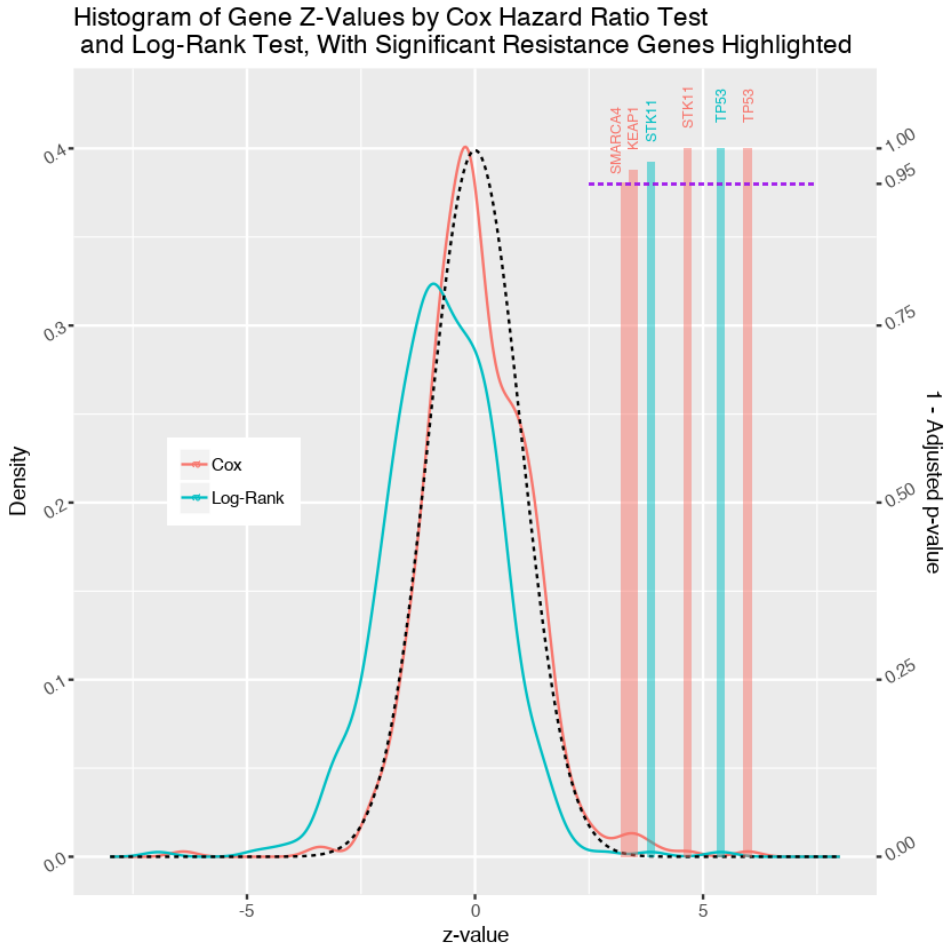
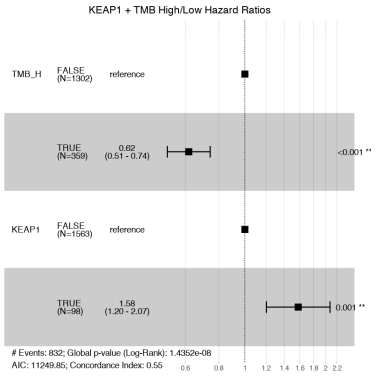


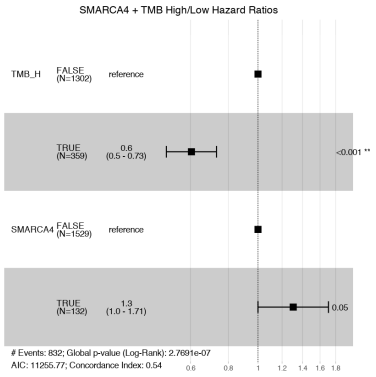
Figure 13. Z-values produced by Cox proportional-hazard tests follow a standard normal distribution. Z-values from the log-rank test appear normal, but with a slight negative shift and higher dispersion. Bars denote the z-value locations of the genes identified as resistance markers.

3.3.2 KEAP1 and STK11 Show Greater Relative Resistance Effects for TMB-H Patients

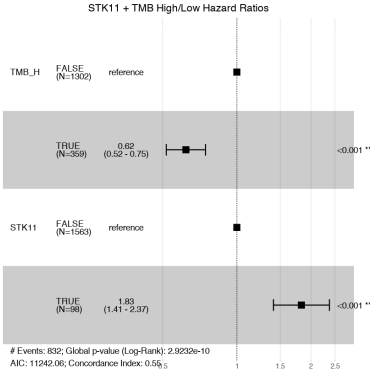
To further investigate the interaction of Tumour Mutation Burden and mutations in resistance genes, we sampled the available data to only include patients with higher TMB. Specifically, for the range of quantiles $x = 0, 0.01, \ldots, 0.9$, we removed all patients with TMB in the bottom $x\%$ within their cancer type. We then, for each of our four resistance genes, fitted a Cox model on the remaining data, and computed the hazard ratio for mutations in that gene (without TMB as a cofactor). We see in Figure 15, for genes *KEAP1* and *STK11*, a noticeable peak in estimated hazard ratio when we only consider the top 40% of samples



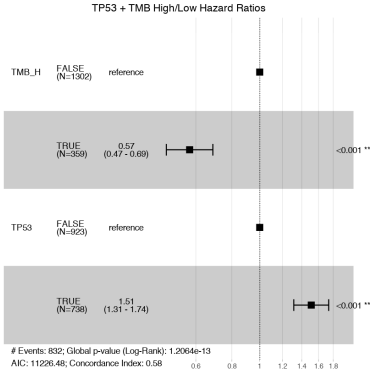
(a) Mutation to the KEAP1 gene shows deleterious effect with hazard ratio 1.62.



(b) Mutation to the SMARCA4 gene shows deleterious effect with hazard ratio 1.58.



(c) Mutation to the STK11 gene shows deleterious effect with hazard ratio 1.86.



(d) Mutation to the TP53 gene shows deleterious effect with hazard ratio 1.53.

Figure 14. Forest plots of hazard ratio for our four significant resistance genes, with non-mutant status as a reference and Tumour Mutation Burden High/Low Status as a cofactor.

by TMB score. While *SMARCA4* is a significant resistance gene when considered with TMB as a cofactor, it never achieved significance when considered as a single variable, no matter how TMB is thresholded.

3.3.3 Genes Particularly Influential for Immunotherapy

For a gene panel to predict prognosis of a particular patient under immunotherapy, simply those genes associated with the largest change in survival will be crucial. However, some of these genes may be important for prognosis of a patient in general, and functionally have relatively little direct interaction with ICB. We therefore compare results from two studies, both in Non-Small Cell Lung Cancer, with one cohort treated by immunotherapy and the other not. We look for genes whose mutations have elevated hazard in the ICB cohort. Ideally these trial groups would be randomised and planned in one study, but unfortunately no such trial data is currently available.

For cohorts A (ICB treated, $n = 350$, sequenced on MSK-IMPACT) and B (non-ICB treated, $n = 982$, sequences with WXS), we restricted attention to the 463 genes for which data was available and patients with mutations in both samples. We then, for each gene, produced Cox proportional-hazards models of survival in each cohort with mutations in that gene as a single factor. We selected genes for which the following were true:

- In both cohorts, at least 10 tumours carried mutations (we relaxed the restriction from the previous 30 as A- our data were fewer in NSCLC, B- we were simply looking for genes showing potential to inform further study, and C- having a large proportion of patients carrying the gene for widespread actionability was less of a concern).
- Hazard ratio for mutation carriers was greater in cohort A than cohort B (to show specific ICB relevance).
- Hazard ratio for mutation carriers in cohort A was > 1 (so that we were identifying resistance genes).

Figure 15. Estimated hazard ratios across a range of TMB cutoffs, for mutation in four resistance genes. For significance, a p-value $leq 0.05$ and samples of both mutated/non-mutated patients of size > 30 were required.

Gene	N^o Mutation Carriers (Cohort A)	N^o Mutation Carriers (Cohort B)	Hazard Ratio (Cohort A)	Hazard Ratio (Cohort B)	Intersection of Confidence Intervals
TP53	217	667	1.19	1.01	0.400
KEAP1	70	140	1.28	1.09	0.578
PTEN	16	62	1.38	0.862	0.586
KDR	14	77	1.35	0.899	0.719
STAG2	11	30	1.59	0.735	0.746
SMARCA4	35	61	1.08	0.919	0.838
NTRK1	13	22	1.25	0.613	0.848
CTNNB1	13	20	1.14	0.455	0.856
ATRX	18	58	1.12	0.874	0.861
PBRM1	18	22	2.10	0.938	0.871
STK11	70	89	1.28	1.21	0.880
ATR	12	44	1.16	0.870	0.919
BRAF	24	57	1.15	1.05	1.10

Table 1. Genes with elevated hazard in ICB-treated patients.

We hoped to see genes for which 95% confident intervals of hazard ratio were disjoint between the two cohorts. Unfortunately, at this cohort size, none were so dramatically differentiated between the two cohorts. However, there were several genes satisfying the condition above, including some with hazard ratios ≤ 1 for cohort B and > 1 for cohort A. We list them in table 1, in ascending order of the size of intersection between their 95% confidence intervals for Hazard ratio. All four resistance genes identified above appear (bolded in table).

As a first investigative step to analyse the biological function of the genes we identified, we performed a gene ontology enrichment analysis, the top ten results of which are shown in Table 2 as ordered by the classic Fisher method. We found elevated terms pertaining to general drug response, regulation of cell growth, and signalling. The most surprising elevated term was GO:0051173, "positive regulation of nitrogen compound metabolic process". While it is unwise to make immediate conclusions from a gene ontology analysis, we considered this an interesting result, and given more time would have delved deeper into which genes were labelled with this term and what significance it might play.

3.4 A Concise Prognostic Panel for Response to Immunotherapy in Non-Small Cell Lung Cancer: Comparison with MSK-IMPACT Panel

We combine work from previous sections to compose a prognostic panel for NSCLC patients being treated with ICB therapy. Our panel consists of two components: a set of genes for estimating TMB, and a set of key resistance genes. We use metric $\mathbf{M}_3$ from section 3.1.2 to select a set of genes with total length approximately 0.25Mb. This incorporates 199 genes, which are listed in Appendix B . We also include the four resistance genes identified in section 3.3.1, also listed in Appendix B . *TP53* appears in both groups, so that in total we have a panel of 202 genes, with associated genomic coding regions of total length 0.26Mb. This is less than a quarter of the size of the *MSK-IMPACT* panel (1.18Mb), granting large reductions in resources necessary to perform sequencing on tumour samples, particularly when sequencing with high coverage to infer subclonal tumour structure.

3.4.1 TMB Estimation

An advantage of our panel being comprised of a relatively small number of genes is that we can more easily use statistical learning techniques for regression and classification tasks. For example, we see that on a reasonably sized publicly available lung cancer dataset[8] ($n = 1144$, split into $n_{\text{training}} = 793$, $n_{\text{test}} = 351$), we can achieve a comparable R^2 value simple using an ordinary linear regression model for TMB against genes in our panel, than estimates of TMB calculated using mutation density across the *MSK-IMPACT* panel (Figure 16). When we applied a similar linear model to the *IMPACT* panel, we found that the larger number of covariates meant that on the size of dataset we had available, the regression performed worse than both our own panel with a linear regression applied and the larger panel with a TMB estimation procedure based purely on panel-wide mutation density.

3.4.2 Classifying TMB High/Low Status

Using the same NSCLC training and test sets as in the previous section, we use as a benchmark the regression generated by the *IMPACT* panel as a classifier. A confusion matrix for its performance at classification is given in Figure 17. We may also alter the threshold for predicting TMB (while keeping the true TMB High threshold fixed) to slide between over-sensitivity and

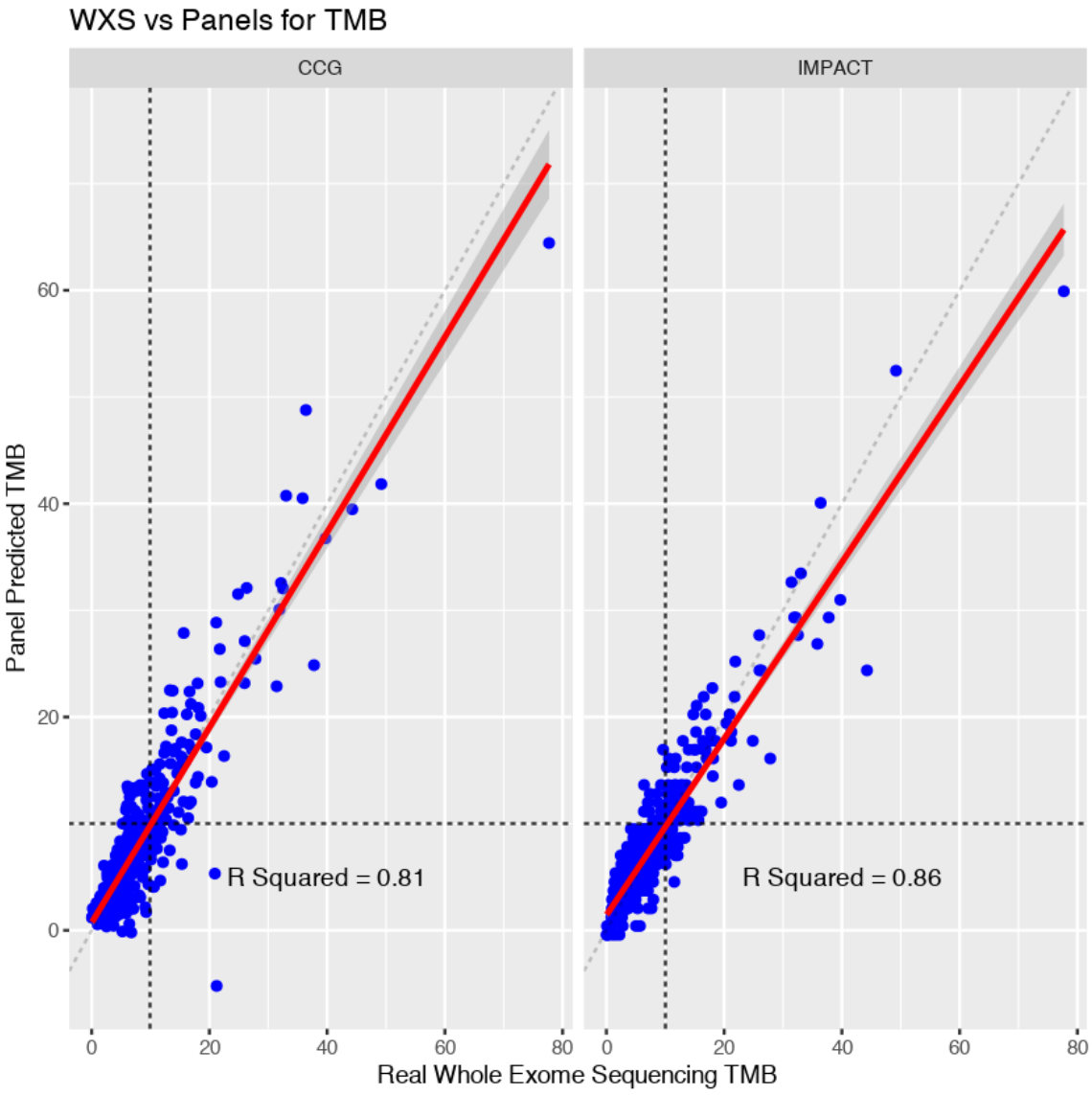


Figure 16. Predicting TMB using a linear regression model with CCG panel genes as covariates, versus mutation density estimates given by the larger *MSK-IMPACT* panel.

Rank	GO ID	Term	Annotated	Expected	classicFisher	classicKS	elimKS
1	GO:0035690	cellular response to drug	316	0.28	2.8	0.0071	0.0071
2	GO:0051128	regulation of cellular component organization	2227	1.95	4.9	5.1	0.1126
3	GO:0045786	negative regulation of cell cycle	518	0.45	8.6	1.1	0.7086
4	GO:0044087	regulation of cellular component biogenesis	824	0.72	8.7	9.0	0.0014
5	GO:0042493	response to drug	917	0.80	2.0	0.0092	0.0283
6	GO:0048519	negative regulation of biological processes	4606	4.03	2.4	4.4	0.7272
7	GO:0009967	positive regulation of signal transduction	1456	1.27	4.0	2.1	0.0731
8	GO:0051173	positive regulation of nitrogen compound metabolic process	2837	2.48	6.4	6.9	0.5487
9	GO:0051130	positive regulation of cellular component organization	1097	0.96	7.9	8.7	0.0017
10	GO:0010647	positive regulation of cell communication	1606	1.40	9.4	1.0	0.0710

Table 2. Enriched Gene Ontology terms for lung cancer immunotherapy resistance genes.

over-specificity. The effect of making such changes is elucidated, also in Figure 17. We find an AUC for receiver operation characteristic of 0.95. We may apply a variety of the techniques discussed in Section 3.1 to the classification problem. We may use the linear regression from above to generate a classification, as well as balanced or imbalanced SVMs. A breakdown of performance metrics for these methods are given in Table 4. We also present ROC curves for these methods against the *IMPACT* benchmark in Figure 18. We find that an SVM trained on our panel has an AUC of 0.94, compared to 0.95 for *IMPACT*.

4 Discussion

Our original goals were to investigate TMB estimation in a single cancer type, extended this to a pan-cancer context, identify resistance markers to ICB, and produce an immunotherapy monitoring panel. We identified methods that performed well for the single-cancer scenario, and had limited success extending these across cancers. We identified targets for further research into mechanisms of resistance, and demonstrated comparable performance across a variety of metrics to a much larger generic cancer monitoring panel.

4.1 Principal Conclusions

Gene-Based Methods Show Greatest Initial Promise Despite the advantages highlighted of proceeding in a gene-agnostic manner, we found that our gene-oriented methods consistently outperformed the former. A large part of this may have been due to computational power restraints. These partly necessitated the iterative 'zoom, repeat' procedure we employed, whose drawbacks included the potential to miss key sites buried in unimportant regions, as well as unnecessarily including unimportant sites adjacent to important ones. Finally, the overfitting problems we experienced (particularly employing linear and RF models) may well have been in part due to the exponential increase in number of covariates throughout the progress of the algorithm. We made attempts to counter this via automated cluster of loci into pooled groups and ridge regression, which showed some initial promise before we moved on.

Intelligent Panel Selection and Model Training Both Play Key Roles Improving Performance We saw in Sections 3.1 and 3.4 convincing evidence that it is worth investing time and thought into intelligently designing panels, both at the stage of loci selection and model training. Altering the procedures employed at both steps greatly affected the performance of our panels. Our final choice of panel selection method, metric $\mathbf{M}_3$, tested better than gene-agnostic procedures, length-agnostic gene based procedures, and LASSO methods, the last of which would be seen as a standard technique for addressing such problems.
By employing intelligent, data-driven methods for panel selection and prediction, we produced a panel that produces a comparable performance to a far larger commercially available panel (TMB regression R^2 0.81 vs 0.86, classification AUC 0.94 vs 0.95), on a far smaller section of genome. This will allow for faster, deeper and cheaper sequencing and processing of results for clinicians and patients.

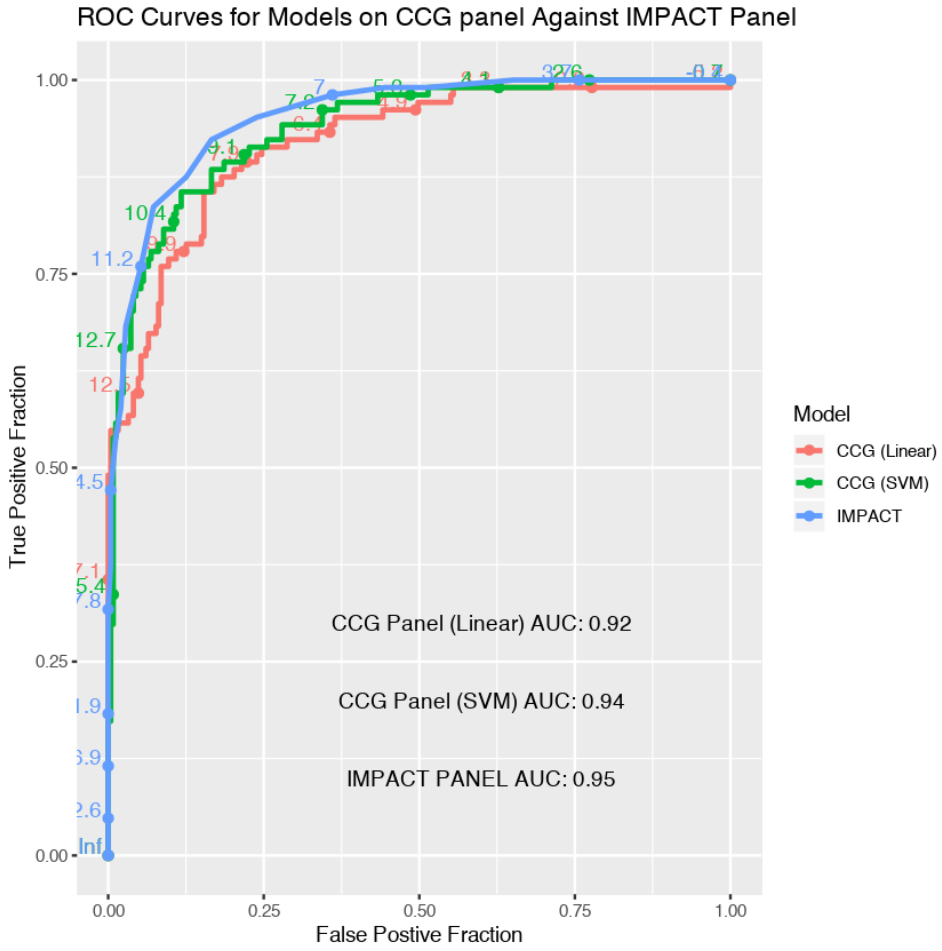
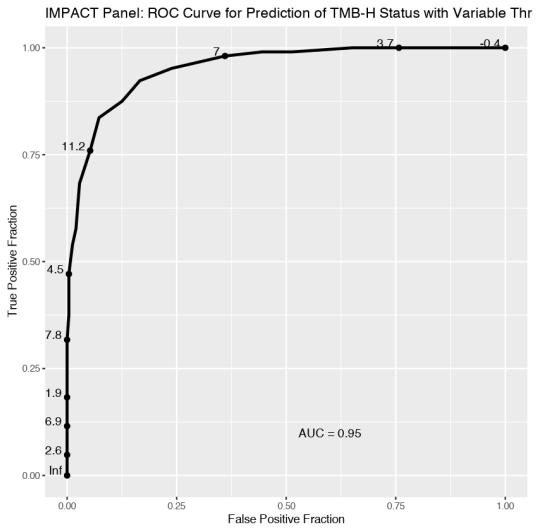


Figure 18. ROC Curves for IMPACT and CCG panels, with linear and SVM models applied to the latter.

(a) Confusion matrix for TMB High/Low status predicted with Impact Panel. Sensitivity = 85.6%, Sensitivity = 87.9%.

	Predicted High	Predicted Low
Real High	87 (24.8%)	17 (4.84%)
Real Low	18 (5.13%)	229 (65.2%)



(b) Receiver Operator Characteristic curve for prediction of TMB High/Low status. TMB High prediction thresholds are annotated.

Figure 17. Analysing the performance of *MSK-IMPACT* as a benchmark classifier.

Panel (Model)	IMPACT	CCG (Linear)	CCG (SVM)	CCG (Balanced SVM)
Sensitivity	83.7%	77.9%	85.6%	90.4%
Specificity	92.7%	88.7%	87.9%	82.1%

Table 4. Comparison of sensitivity/specificity performance across panels and models.

Differences in Mutation Profile Limit Use of Single Panels Across Cancer Types Despite limited successes in particular cancer pairs, overall we struggled to produce predictive results on test data comparable to cancer-type specific panels. We have indicative evidence that overlap in regions of high mutation density could be a predictor of how easily two cancer types may submit to TMB estimation/classification from a single panel, and with more time would have investigated this more formally.

Genes Relating to Known Immune Mechanisms are Highlighted as Resistance Markers Of the four genes identified as significant resistance markers (*TP53*, *KEAP1*, *STK11* and *SMARCA4*), we were pleased that each has known associations with cancer development, and at least two are associated with known immune evasion mechanisms. *TP53* is widely acknowledged as a crucial tumour suppressor[9] and so it would have been very surprising had it not appeared. *KEAP1* is involved in Class I MHC mediated antigen processing and presentation[10], an essential part of foreign cell recognition by the immune system, and so we can sensibly explain the association between loss of function in this pathway with specific resistance to immunotherapy in the context of current immunology. *STK11* has a known role in DNA damage response[11], with clear deleterious reprecussions for immune action against a tumour. *SMARCA4* is involved in both the Integrated Breast Cancer Pathway and cytokine signalling[12], so the precise nature of its involvement in immunotherapeutic resistance could be a topic for further study.

4.2 Study Limitations

Availability of Sequencing Data The wealth of publicly available data is a key strength cancer genomics, through resources such as The Cancer Genome Atlas and cBioPortal. That said, we encountered some key hindrances in available datasets. Firstly, inconsistencies between the sequencing methods and targeted regions limited our ability to fully integrate data from differing sources. Secondly, an additional aim of our study was to specifically identify any non-coding regions that showed a strong relationship with TMB. Unfortunately, too little data in a useful variant called format was available for WGS samples.

Computational Power Several of the techniques we developed (in particular the iterative sliding window algorithm) may not have been necessary if more computational power had been available, and adopting more brute force approach may have produced better results.

Pooling of Separate Studies for Resistance Analysis When identifying genes whose resistance potential was particularly elevated in an immunotherapy context, we pooled two separate studies, one with patients on ICB treatment, and the other with

patients on conventional cancer treatments. In terms of trial design, the ideal would have been to have patients in a single cohort randomly allocated to one of the two pools, to avoid confounding factors coming from the methods by which patients were selected to take part in either trial.

Longitudinal Clinical Information In studying resistance, in particular acquired and adaptive resistance, longitudinal data with good quality clinical information attached (e.g RECIST scores) would have been ideal. In general, we worked with a single data point per patient and overall survival scores.

4.3 Future Directions

Non-Coding Regions May Also Contribute to Neoantigenic Burden Recent studies have suggested that neo-antigenic burden originates in part from errantly transcribed non-coding regions[13]. Further investigation into the role of non-coding regions of DNA could give a fuller picture of a tumour's likely response to immunotherapy.

Incorporation of Mutation Information Beyond Count May Allow More Sophisticated Machine Learning Techniques In our study we never incorporated mutation information beyond non-silent mutation count (e.g. amino acid change, position in gene coding sequence). Incorporation of larger amounts of information without increasing the size of training set would require more sophisticated machine learning methods, such as convolutional neural networks, with which progress has been reported[14] in other areas of genomics.

Further Study of the Interaction of TMB and Resistance Markers May Shed Light on Immune Mechanisms We found suggestions in Section 3.3.2 of a more subtle relationship between TMB and hazard for some genes. While this was an exploratory and tentative finding, we were not able to produce a simple statistical explanation for the phenomenon. We would therefore recommend it for further exploration.

For Maximum Clinical Utility, Selection Criteria Based on Cancer Relevance and Estimation Relevance May Allow Greatest Patient Benefit We developed a variety of methods focussing on efficiency of TMB calculation. We did not explicitly weight our gene selection methods towards choosing genes with known clinical relevance to cancer, and doing so in the future may allow for accurate TMB estimation while providing a wider portfolio of known actionable sites for clinicians.

5 Materials and Methods

5.1 TMB Determination via Exome Data

We calculated TMB using exome data, including only non-silent mutations, as defined below. We quoted standard human exome length as 30Mb.

We only consider as somatic mutations:

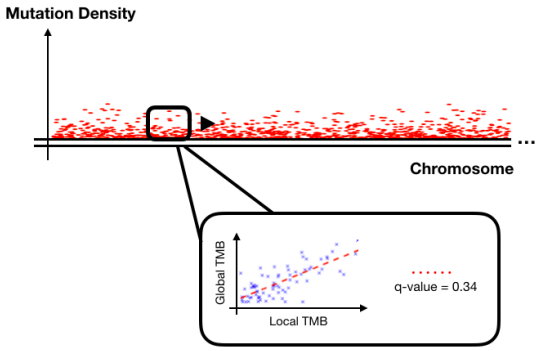
- Frameshift Deletions
- Frameshift Insertons
- In-Frame Deletions
- In-Frame Insertions
- Missense Mutations
- Nonsense Mutations
- Nonstop Mutations
- Splice Sites
- Translation Start Sites

and we ignore the following 'silent' variants:

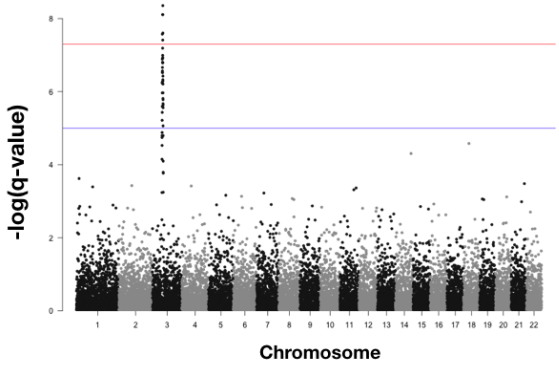
- 3'Flanks
- 3'UTRs
- 5'Flanks
- 5'UTRs
- IGRs
- Introns
- RNAs
- Splice Regions

5.2 Sliding Window Algorithm

In order to establish a pool of candidate genetic loci, we needed to measure which genome regions were most strongly associated with TMB. To determine regions whose mutation density was strongly indicative of overall burden, we used a sliding window algorithm (Figure 19a). Letting J be a jump length parameter and W be a window size parameter, the procedure followed as described below:



(a) Sliding window correlates local mutation burden with global mutation burden.



(b) Manhattan plot showing regions of high association (example figure from R package qqman documentation[15]).

Figure 19. Genome-wide association with TMB using a sliding window algorithm.

1. For a base pair, consider an interval of width W around it in the genome.
2. For this window, calculate the local mutation burden (for each patient separately).
3. Calculate a correlation metric across patients between the local burden at the interval around that base pair, and the patient's whole exome TMB score. Convert this metric to a p-value.
4. Move along J base pairs, and repeat.

From this data we were able to conduct genome-wide association analysis of regions significantly associated with TMB (Figure 19b). We converted p-values generated by sliding windows to q-values via a Benjamini-Hochberg correction.

5.3 Iteratively Applying Sliding Windows

When performing a sliding windows algorithm, there is an important parameter "Window Size" that requires optimisation. Additionally, the size of the human genome makes performing a sliding window algorithm with small window size computationally untenable. We aimed to solve both problems simultaneously by opting for an iterative scheme, with a zoom parameter Z:

1. Run a sliding windows procedure across the genome(/subset of genome inherited from previous iteration)
2. Select regions of genome with p-values in the top $\frac{100}{Z}\%$ of the regions considered. Discard all others.
3. Decrease jump length and window size by a factor of Z.

5.4 Random Forests

Random Forests are an ensemble learning technique based on decision trees. They are capable of both regression and classification learning. One of their advantages is a high level of interpretability- there are various techniques for extracting explanatory information about the internal processing of a trained RF model. We implemented our RF models in R using the package "randomForest"[16].

5.5 Support Vector Machines

Support Vector Machines are a machine learning technique, more often used in classification problems (although regression is also possible). They are based on embedding predictors into a more high-dimensional space via some mapping, known as a kernel, which is often non-linear. They then attempt to adjust weights iteratively to find the most efficient partition of this space. We implemented our SVM models in R using the package "e1071"[17].

5.6 Bioinformatics

We used a variety of Bioconductor packages in R. In particular for annotating gene lengths we used BiomaRt[18]. For gene length we used the attributes "genome_coding_start" and "genome_coding_end", and summed the differences between each instance of the two for a given gene.

5.7 Gene Length Weighted Metrics of Association

Given a set of samples $i = 1, 2, \ldots, N$, and a given gene, let:

- TMB denote a vector of length I containing TMB scores for each sample.
- TMB_{gene} be a vector of the same length, with each entry $TMB_{gene}[i]$ giving the local mutation burden (number of nonsilent mutations) for sample i within the coding region for the gene.
- p_{gene} be the p-value given after performing a linear regression of TMB against TMB_{gene}.
- n_{gene} be the coding length of the gene.
- $\mathbf{1}$ be a vector of length N with each element equal to one.

Then we define

$$\mathbf{M}_1 := \frac{-\log(p_{gene})}{n_{gene}}$$

$$\mathbf{M}_2 := \frac{\mathrm{Cor}(TMB, TMB_{gene})}{\sqrt{n_{gene}}}$$

$$\mathbf{M}_3 := \frac{\mathrm{Cor}(TMB, TMB_{gene})(TMB_{gene} \cdot \mathbf{1})}{(\sqrt{n_{gene}})^3}$$

5.8 LASSO Methods

LASSO methods proceed by fitting linear regression models, using an altered loss function, typically augmented from the standard least squares loss function $L(\beta; X, y)$ to

$$L(\beta; X, y) + \lambda |\beta|_1$$

where λ is a variable parameter that may be altered to choose model size, and $|\cdot|_1 = \sum_{gene} |\cdot|$ is the L_1 norm. We used an altered L_1 norm taking into account gene length, so that our full loss function was

$$L(\beta; X, y) + \sum_{gene} n_{gene} \beta_{gene}$$

We used the R package "glmnet"[19] to implement LASSO methods.

5.9 Survival Analysis

We implemented Cox Regression models, which fit a generalised linear model against individuals' estimated hazard ratio. We used Overall Survival as our measure of success, and implemented our models using the R packages "survival"[20] and "survminer"[21]. Log-Rank tests are a non-parametric test for distinguishing difference in survival curve, and we impleneted them using the R package "coin"[22].

5.10 Gene Ontology

We implemented our GO enrichment analysis using the R package "topGO"[23]. We ranked gene ontology terms according to the classic Fisher test.

Acknowledgements

Many thanks to everyone at CCG, in particular Nirmesh Patel, Harry Clifford and Hannah Thompson- also Morton. Additional thanks to the entire Systems Biology 2019 cohort and teaching staff.

References

1. Anderson, A. R., Weaver, A. M., Cummings, P. T. & Quaranta, V. Tumor Morphology and Phenotypic Evolution Driven by Selective Pressure from the Microenvironment. *Cell* **127**, 905–915, DOI: 10.1016/j.cell.2006.09.042 (2006).
2. Lim, B., Woodward, W. A., Wang, X., Reuben, J. M. & Ueno, N. T. Inflammatory breast cancer biology: the tumour microenvironment is key. *Nat. Rev. Cancer* **18**, 485–499, DOI: 10.1038/s41568-018-0010-y (2018).
3. De Palma, M., Biziato, D. & Petrova, T. V. Microenvironmental regulation of tumour angiogenesis. *Nat. Rev. Cancer* **17**, 457 (2017).

4. Lu, Y. C. & Robbins, P. F. Targeting neoantigens for cancer immunotherapy. *Int. Immunol.* **28**, 365–370, DOI: 10.1093/intimm/dxw026 (2016).

5. Palucka, A. K. & Coussens, L. M. The Basis of Oncoimmunology. *Cell* **164**, 1233–1247, DOI: 10.1016/j.cell.2016.01.049 (2016). 15334406.

6. Samstein, R. M. *et al.* Tumor mutational load predicts survival after immunotherapy across multiple cancer types. *Nat. Genet.* **51**, 202–206, DOI: 10.1038/s41588-018-0312-8 (2019).

7. Cheng, D. T. *et al.* Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT). *The J. Mol. Diagn.* **17**, 251–264, DOI: 10.1016/j.jmoldx.2014.12.006 (2015).

8. Campbell, J. D., Alexandrov, A. & Kim, J. Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. **48**, 607–616, DOI: 10.1038/ng.3564.Distinct (2016). 15334406.

9. Kastenhuber, E. R. & Lowe, S. W. Putting p53 in Context. *Cell* **170**, 1062–1078, DOI: 10.1016/j.cell.2017.08.028 (2017).

10. Wijdeven, R. *et al.* *Chemical and genetic control of IFNγ-induced MHCII expression*, vol. 19 (2018).

11. Wang, Y.-S. *et al.* LKB1 is a DNA damage response protein that regulates cellular sensitivity to PARP inhibitors. *Oncotarget* **7**, 4–7, DOI: 10.18632/oncotarget.12334 (2016).

12. Koh, A. S. *et al.* Rapid chromatin repression by Aire provides precise control of immune tolerance. **19**, 162–172, DOI: 10.1038/s41590-017-0032-8.Rapid (2018).

13. Laumont, C. M. *et al.* Noncoding regions are the main source of targetable tumor-specific antigens. *Sci. Transl. Medicine* **10**, eaau5516, DOI: 10.1126/scitranslmed.aau5516 (2018).

14. Harries, L., Shawe-taylor, S. P. J. & Clifford, H. SomaticNet : a novel deep learning approach to detecting cancer mutations. (2018).

15. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The Am. J. Hum. Genet.* **81**, 559–575, DOI: 10.1086/519795 (2007). arXiv:1011.1669v3.

16. Meyer, D. *et al.* Package 'randomForest'. 193–194, DOI: 10.1023/A (2018).

17. Dimitriadou, E. *et al.* *Package 'e1071'* (2009).

18. Steffen Durinck1, Paul T. Spellman1, Ewan Birney2 & Huber2, W. Mapping Identifiers for the Integration of Genomic Datasets with the R/Bioconductor package biomaRt Steffen. **4**, 1184–1191, DOI: 10.1038/nprot.2009.97.Mapping (2011).

19. Friedman, Hastie, Simon & Tibshirani. Lasso and elastic-net regularized generalized linear models. R Package Version 2.0-13. (2018).

20. Therneau, T. *A package for survival analysis. R package 2.37-2* (2012).

21. Kassambara, A. Package ' survminer '. (2018).

22. Hornik, K., Hothorn, T., Van de Weil, M. A. & Zeileis, A. Implementing a Class of Permutation Tests :. *J. Stat. Softw.* **28** (2008).

23. Alexa, A. & Rahnenfuhrer, J. topGO: enrichment analysis for Gene Ontology. R Packag. version 2.26.0. *R Packag. version 2.26.0* DOI: 10.1038/ncomms8832 (2016).

6 Appendices

6.1 Appendix A

Deriving Gene Length Weighted Metrics of Association Previously we have used p-value as our selection metric for genomic loci. More specifically, we have usually used the metric $-\log(p_{gene})$ so that a higher score will indicate a more significant gene. As our first length-adjusted metric, we divide by the length n_{gene} so that for genes of the same significance, a shorter gene will have a higher score. Likewise, for genes of the same length a more significant one will have a higher score. Thus we define:

$$\mathbf{M}_1 := \frac{-\log(p_{gene})}{n_{gene}}$$

When proceeding more formally, p-value is actually slightly cumbersome to work with, so we switch to using correlation, defined as

$$\mathrm{Cor}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var(X)Var(Y)}}} = \frac{\mathbb{E}\big[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])\big]}{\sqrt{\mathbb{E}\big[(X-\mathbb{E}[X])^2\big]\mathbb{E}\big[(Y-\mathbb{E}[Y])^2\big]}}$$

with expected values replaced by appropriate averages in the case of estimating from data. We see that the rankings of genes according to correlation very closely correspond to the ranking of genes via p-value (Figure 20).

To construct a sensible weighting factor for gene length, we consider an extremely simplified model of mutation: a genome of length N, with a gene of length n_{gene}, and mutations occurring randomly at a rate p per base pair. Then, with X_i being independent identically distributed bernoulli variables ($i = 1,\ldots,N$), we have

$$TMB = \sum_{i=1}^{N} X_i \sim \mathrm{Bin}(N,p)$$

$$TMB_{gene} = \sum_{i=1}^{n_{gene}} X_i \sim \mathrm{Bin}(n_{gene},p)$$

where TMB and TMB_{gene} are global and gene-local mutation burdens respectively. It can be shown easily that

$$\mathrm{Cor}(TMB, TMB_{gene}) = \sqrt{\frac{n_{gene}}{N}}.$$

Therefore our second proposal for a metric of a gene's importance for TMB determination is

$$\mathbf{M}_2 := \frac{\mathrm{Cor}(TMB, TMB_{gene})}{\sqrt{n_{gene}}}$$

In practice we find that the genes selected via this metric tend to be very small and very rarely mutated, making them of limited use for a gene panel. This makes sense, as our initial assumption that mutations occur uniformly throughout the genome is clearly not realistic. We therefore add a weighting factor so that our third metric is given by something like $\mathbf{M}_3 = \hat{p}\mathbf{M}_2$, where $\hat{p}$ is an estimate of mutation rate in that gene. We could calculate $\hat{p}$ by dividing the total number of times the gene in question is mutated in our dataset by the number of samples in our dataset, then further dividing by gene length to geta mutation rate. However, the number of samples in the dataset is invariant across the dataset, so we see that

$$\hat{p}_{gene} \propto \frac{TMB_{gene}\cdot\mathbf{1}}{n_{gene}}$$

and thus define

$$\mathbf{M}_3 := \frac{\mathrm{Cor}(TMB, TMB_{gene})(TMB_{gene}\cdot\mathbf{1})}{(\sqrt{n_{gene}})^3}$$

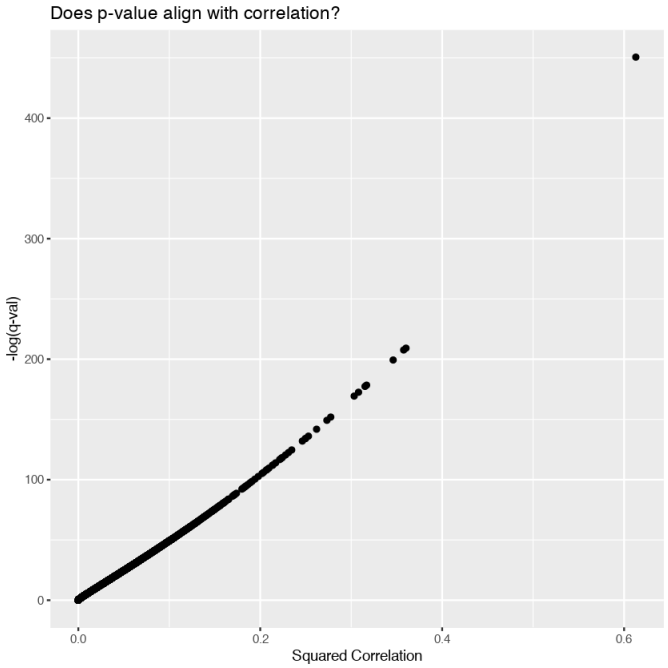


Figure 20. p-value (in this case adjusted via Benjami-Hochberg to q-value) corresponds very well to correlation.

6.2 Appendix B

List of Genes for Calculating TMB TP53, REG3A, POM121L12, REG1B, CDKN2A, TPTE, REG1A, OR8H2, OR2T3, OR2T33, DEFB110, OR1C1, CDH10, DEFB115, OR2L13, LCE2B, OR5L2, OR8J3, OR4A15, OR2W3, REG3G, OR2G6, RPL10L, OR6F1, OR5L1, HBB, BRINP3, SPATA8, OR4A5, ZNF536, OR8I2, OR2T12, OR2AK2, MYF5, OR4N2, LCE1C, OR2T4, OR5D18, ZP4, OR10G8, HTN1, OR4M2, ZNF716, OR5B12, OR5D14, OR11L1, OR5F1, HIST1H3B, OR2G3, OR4C6, PRAC1, OR4Q3, GCSAML, DCAF4L2, OR4L1, NTM, OR2T10, ST6GAL2, OR14A16, OR2L3, SMR3B, RGS7, OR2M2, SGCZ, OR5AS1, OR2M5, OR2L8, MPPED2, OR8K1, OR2G2, TGIF2LX, OR5D13, CDH18, CDH12, OR5T1, OR2M7, ZNF479, FCRL1, OR4C15, OR5W2, OR2T34, OR8H3, TMEM207, CSMD3, CDH9, TAS2R1, OR4K15, MBL2, OR8D2, LCE1E, LCE3D, TRIM51, OR4C46, PAPPA2, FAM24A, OR4K2, GABRA2, SORCS1, OR6K2, ZNF804A, CETN1, PLN, OR2M3, OR8K3, FAM47A, KIF2B, OR4P4, OR8K5, DEFB112, FTMT, FBXL7, LCE2D, OR13G1, LHX8, PYHIN1, OR10K2, HIST1H2BF, HCN1, NPAP1, NLRP3, ST6GALNAC3, CALN1, SNTG1, LCE2C, OR10W1, OR10T2, CARD18, OR5M9, LRRC4C, UQCR11, GALNT13, OR10AG1, SLN, KCNJ12, TRIM58, DKK2, GRM8, RGS18, TRIM48, DCAF12L1, LRRTM4, PDHA2, SPANXN1, OR5T3, CHRM2, FAM135B, HIST1H2BH, CD1A, PGK2, OR2L2, TECRL, OR1S2, GNGT1, OR5H15, CNBD1, PCDH15, LCE3A, SPRR2G, MMP16, OR2B11, HIST1H2AC, OR10K1, CTSG, CYYR1, LRRC30, OR5R1, OR6K6, ASB5, COL11A1, PSG8, CCER1, SGCD, OR5D16, ACSM2A, SLITRK3, MYCT1, BCHE, OR14C36, LCE1F, MS4A13, CST5, OR4D5, HGF, HIST1H2BC, TUBA3C, CA10, RUNX1T1, LRFN5, OR10Q1, KCNJ3, RHOH, SLITRK1, OR4A16, OR8H1, ADAMTS12, CNTNAP2, PRDM9, SLC39A12, TBX22

List of Genes for Monitoring Resistance TP53, STK11, KEAP1, SMARCA4