Burden of tumor mutations, neoepitopes, and other variants are cautionary predictors of cancer immunotherapy response and overall survival

Mary A. Wood (1, 2), Benjamin R. Weeder (1, 3), Julianne K. David (1, 3), Abhinav Nellore* (1, 3, 4), Reid F. Thompson* (1, 2, 3, 5, 6, 7)

AFFILIATIONS
1. Computational Biology Program, Oregon Health & Science University
2. Portland VA Research Foundation
3. Department of Biomedical Engineering, Oregon Health & Science University
4. Department of Surgery, Oregon Health & Science University
5. Department of Radiation Medicine, Oregon Health & Science University
6. Department of Medical Informatics & Clinical Epidemiology, Oregon Health & Science University
7. Division of Hospital and Specialty Medicine, VA Portland Healthcare System

*Co-corresponding authors [nellore@ohsu.edu, thompsre@ohsu.edu]

ABSTRACT

Background: Tumor mutational burden (TMB, the quantity of aberrant nucleotide sequences a given tumor may harbor) has been associated with response to immune checkpoint inhibitor therapy and is gaining broad acceptance as a result. However, TMB harbors intrinsic variability across cancer types, and its assessment and interpretation are poorly standardized.
Methods: Using a standardized approach, we quantify the robustness of TMB as a metric and its potential as a predictor of immunotherapy response and survival among a diverse cohort of cancer patients. We also explore the additive predictive potential of RNA-derived variants and neoepitope burden, incorporating several novel metrics of immunogenic potential.
Results: We find that TMB is a partial predictor of immunotherapy response in melanoma and non-small cell lung cancer, but not renal cell carcinoma. We find that TMB is predictive of overall survival in melanoma patients receiving immunotherapy, but not in an immunotherapy-naive population. We also find that it is an unstable metric with potentially problematic repercussions for clinical cohort classification. We finally note minimal additional predictive benefit to assessing neoepitope burden or its bulk derivatives, including RNA-derived sources of neoepitopes.
Conclusions: We find sufficient cause to suggest that the predictive clinical value of TMB should not be overstated or oversimplified. While it is readily quantified, TMB is at best a suggestive surrogate biomarker of immunotherapy response. The data do not support isolated use of TMB in renal cell carcinoma.

KEYWORDS

tumor mutational burden, TMB, neoepitopes, neoepitope burden, neoantigens, splice junctions, retained introns, tumor variant burden, immunotherapy response

BACKGROUND

The advent of immunotherapy as a promising form of cancer treatment has been accompanied by a parallel effort to explore potential mechanisms and drivers of therapeutic response. For instance, tumor mutational burden (TMB, the overall quantity of aberrant nucleotide sequences a given tumor may harbor) has been associated with response to immune checkpoint inhibitor therapy (1) and overall survival (2). Similarly, the quantity of non-synonymous single nucleotide variants was shown to be associated with immunotherapy response in several independent clinical cohorts (3–6). Other sources of sequence variation such as frameshifting insertions/deletions (7) and tumor-specific alternative splicing (e.g. intron retention (8)) have also been found to correlate with immunotherapy response. These phenomena are widely accepted and appear to be particularly pronounced in patients harboring DNA repair deficiencies (9). Indeed, the checkpoint inhibitor, pembrolizumab, was granted accelerated disease-agnostic approval by the FDA on this basis for any cancer patient harboring deficiencies in their capacity to perform DNA mismatch repair (10). Moreover, an expanding cohort of clinical immunotherapy trials (e.g. NCT03668119, NCT03178552, NCT03519412) are actively utilizing TMB status as a key inclusion criterion. However, there is wide variability among techniques for measuring and interpreting TMB, raising questions of utility and reproducibility (11).

Given the perceived critical importance of TMB in the research setting and its emerging role in the oncology clinic, we sought to quantify the robustness of TMB as a metric, and explore its deeper nuances using pooled whole exome sequencing data from a variety of previously published studies. While TMB is generally correlated with downstream metrics such as neoepitope burden, we also explore the predictive capacity of neoepitope burden and its derivatives including adjustment for MHC binding robustness and peptide sequence novelty, as well as RNA-derived sources of neoepitopes.

METHODS

Variant Identification and Neoepitope Prediction

We assembled a cohort of 440 tumor samples from 414 different cancer patients from publicly available data, including 285 melanoma patients (309 tumor samples) (1,4–6,12–15), 34 non-small cell lung cancer (NSCLC) patients (34 tumor samples) (3), 10 prostate cancer patients (10 tumor samples) (16), 57 renal cell carcinoma (RCC) patients (58 tumor samples) (17), and 28 colon, endometrial, and thyroid cancer patients (29 tumor samples) (9) (see Supplementary Table 1). Whole exome sequencing (WES) reads for each sample were aligned to GRCh37d5 using the Sanger cgpmap workflow (18), which uses bwa-mem (v0.7.15-1140) (19) and biobambam2 (v2.0.69) (20) to generate genome coordinate-sorted alignments with duplicates marked and Genome Analysis Toolkit (GATK, v3.6) to realign around indels and

perform base recalibration for paired tumor and normal sequence read data. The Mbp of genome covered was determined using bedtools genomecov (v2.26.0) (21), where any base covered by a depth of at least 6 reads was considered covered, as this is twice the minimum read depth required for variant detection by SomaticSniper (22) and VarScan 2 (23). Somatic variants were called and filtered using MuSE (v0.9.9.5) (24), MuTect (v1.1.5) (25), Pindel (v0.2.5b8) (26), RADIA (v1.1.5) (27), SomaticSniper (v1.0.5.0) (22), and VarScan 2 (v2.3.9) (23) according to the mc3 variant calling protocol (28). To compare each of these somatic variant reports for consensus calling, vt (v0.5772-60f436c3) (29) was used to normalize variants, decompose biallelic/block variants, sort variants, and produce a final, unique list of variants for each caller. Somatic variants that were reported by more than one tool, or that were reported by Pindel and overlapped at least one call from another tool (as Pindel reports large insertion/deletion variants often incompatible with other callers), were retained for further analysis. Germline variants were called using GATK's HaplotypeCaller (v3.7-0-gcfedb67) (30), and we used VariantFiltration with cluster size 3 and cluster window size 15 to flag and subsequently eliminate variants with total coverage of less than 10.0 reads, quality by read depth of less than 2.0, and Phred quality of less than 100.0. We employed HapCUT2 for patient-specific haplotype phasing. To do this, germline and consensus somatic variants were combined into a single VCF using `neoepiscope`'s (31) merge functionality. HapCUT2's extractHAIRS software was run with the merged VCF and the tumor alignment file, allowing for extraction of reads spanning indels, to produce the fragment file used with HapCUT2 to predict haplotypes. Neoepitopes of 8-24 amino acids in length were predicted for this cohort using `neoepiscope`, including background germline variation and variant phasing, and enumerating neoepitopes from protein coding, nonsense mediated decay, polymorphic pseudogene, T cell receptor variable, and immunoglobulin variable transcripts. Additionally, to better understand how the choice of variant caller impacts downstream neoepitope predictions, we ran `neoepiscope` excluding background germline variation and variant phasing separately for our consensus somatic variants and variants produced by individual variant calling tools, only enumerating neoepitopes from protein coding transcripts. For patients with multiple tumor samples, the median mutation and neoepitope burdens across samples were retained. Variants that were pathogenic or likely pathogenic in cancer according to ClinVar (32) were identified using Open-CRAVAT (33), and neoepitopes deriving from these variants were flagged.

RNA Variant Identification

Among the overall cohort, 106 patients (89 melanoma patients (1,4–6) and 17 RCC patients (17)) had complementary tumor RNA-sequencing (RNA-seq) data. We aligned RNA-seq reads to both the GRCh37d5 and GRCh38 genomes using STAR (v2.6.1c) (34), using the 'intronMotif' --outSAMstrandField option and specifying NH, HI, AS, nM, and MD fields with the --outSAMattributes option. To identify putative tumor-specific splice junctions, we first downloaded called junction data including coverage and bed files for TCGA and GTEx using recount2 (35). GENCODE version 28 annotations (36) were downloaded and parsed to collect full coordinates and left and right splice sites of junctions from annotated transcripts. The TCGA phenotype file from Rail-RNA (37) was parsed to collect sample type (primary, recurrent, or metastatic tumor vs. matched normal). A new SQLite3 database was created to index all GTEx

and TCGA junctions, with linked tables containing 1) sample ids and associated junction ids; 2) sample ids and phenotype information for each sample; and 3) junction ids and junction information including GENCODE annotation status and location within protein coding gene boundaries. Junctions were extracted from the SJ.out output files generated by STAR. The known junction index was queried to collect all junctions found in normal tissue either in GTEx or in TCGA matched normal samples and these normal junctions were filtered out from the single sample set. We used the MetaSRA (38) web query interface to collect Sequence Read Archive (SRA) accession numbers for non-cancerous melanocyte cell line (39) and primary cell (40) RNA-seq experiments. The resulting accession numbers were queried against the Snaptron junction database (41,42) to download junctions from across the entire genome. All junctions found in these normal melanocyte samples as well as all fully GENCODE-annotated junctions were also eliminated from each single-sample junction set. Finally, we removed junctions where neither end was found in GENCODE-annotation, yielding a list of putative tumor-specific splice sites for each patient.

We identified tumor-specific retained introns using Keep Me Around (kma) (43). We aligned RNA-seq reads to a modified version of the GRCh37d5 using Bowtie 2 (v2.3.4.3) (44), and quantified reads using eXpress (v1.5.1) (45), as per kma recommendations. After computing intron retention, we used kma's filters to retain only transcripts that were expressed at greater than or equal to 1 transcripts per million (TPM) in at least 25% of samples, transcripts that had at least 5 unique counts in at least 25% of samples, and transcripts that had greater than 0 and less than 100 percent of introns retained. To prevent inclusion of artifacts from unprocessed transcripts, we identified outlier introns among the distribution of transcript read counts, only retaining introns with a read count greater than 3 median absolute deviations above the median intron read count for a transcript, and greater than or equal to the read count for the transcript itself. To filter out retained introns that may be expressed in normal tissues, we performed the same analysis using using publically available RNA-seq reads from melanocyte samples of 106 newborns (46). Any retained introns identified from the melanocyte RNA-seq data were then removed from the retained introns identified from the tumor RNA-seq data. Neoepitopes deriving from retained introns were predicted using the reading frame from the 5' end of the transcript of origin prior to the intron, enumerating peptides 8-24 amino acids in length.

HLA Type Prediction and Related Analyses

MHC Class I alleles for each patient were predicted from tumor WES reads using Optitype (v1.0) (47), and MHC Class II alleles for each patient were predicted from tumor WES reads using seq2hla (v2.2) (48). For each neoepitope sequence predicted (see above), a patient's predicted MHC Class I and MHC Class II alleles were used for binding affinity predictions with MHCnuggets (v2.1) (49). Neoepitopes were counted toward a patient's neoepitope burden if they bound at least one of a patient's MHC alleles with high affinity (<= 500 nM).

Modified Neoepitope Burden

To better understand how different features of tumor neoepitopes might influence response to immunotherapy, we produced several normalized neoepitope burdens. We first calculated

neoepitope burden for each patient weighted by MHC allele presentation, where a predicted neoepitope sequence counted toward the patient's neoepitope burden once for each of the patient's MHC alleles that was predicted to bind that neoepitope with high affinity (<= 500 nM). Second, neoepitope burden was calculated for each patient weighted by amino acid mismatch as follows. The closest normal peptide in the human proteome to each neoepitope was identified using blastp (v2.6.0) (50), selecting for lowest E value or, in the case of a tie among multiple peptide sequences, the selected peptide was that with the highest weighted BLOSUM62 similarity (as described previously (51)). A neoepitope sequence was counted toward the patient's neoepitope burden once for each amino acid mismatch between the neoepitope and its closest normal peptide. Third, neoepitope burden was calculated for each patient weighted by TCGA transcript expression of the transcript(s) of origin for each neoepitope. We identified expressed transcripts in matched TCGA cancer types for each disease type in our cohort (SKCM for melanoma, LUAD/LUSC for NSCLC, COAD for colon cancer, UCEC for endometrial cancer, THCA for thyroid cancer, PRAD for prostate cancer, and KIRC for RCC) from TPM values generated by the National Cancer Institute (52). A transcript was considered "expressed" for a cancer type if the 75th quantile TPM value for that transcript in that disease was greater than 1 TPM. Because these TPM values were based on GRCh38 transcripts, we used liftOver (53) to convert the coordinates of a neoepitope's mutation of origin to GRCh38 coordinates and identify overlapping transcripts. A neoepitope sequence was counted toward the patient's neoepitope burden once for each transcript of origin expressed in TCGA. Note that for patients with tumor RNA-seq data (see above), we also calculated neoepitope burden weighted by patient-specific expression of the transcript(s) of origin for each neoepitope. We used Rail-RNA (v0.2.4b) (37) on RNA-seq alignments to the GRCh37d5 genome to identify covered exons, and a transcript was considered "expressed" if at least 1 read covered any exon in the transcript. A neoepitope sequence was counted toward the patient's neoepitope burden once for each expressed transcript of origin. Finally, we multiplicatively combined these weighted burdens by multiplying scores for each epitope and totaling all epitope scores: allele presentation score by amino acid mismatch score, allele presentation score by TCGA expression score, allele presentation score by patient-specific expression score (if relevant), amino acid mismatch score by TCGA expression score, amino acid mismatch score by patient-specific expression score (if relevant), allele presentation score by amino acid mismatch score by TCGA expression score, and allele presentation score by amino acid mismatch score by patient-specific expression score (if relevant).

Statistical Analysis

Statistical analysis was performed in R (v3.5.1). The rlm function from the MASS package (v7.3-51.4) was used for robust linear model fitting, and the cor.test function was used for determining Pearson product-moment correlation values. To determine variability in TMB across variant calling tools, the median of pairwise differences in TMB between tools was divided by the median TMB across tools for each patient; the median of these values across patients was reported. The roc function from the pROC package (v1.14.0) was used to generate ROC curves for any predictors of immunotherapy response and to determine their AUC for all patients with reported immunotherapy response status (409/414, after excluding 3 colon cancer, 1 prostate

cancer, and 1 RCC patient that lacked documented response status). Logistic regression was performed using the glm function to model therapeutic response as a linear function of TMB (on log scale), and neoepitopes (log scale) on the 245 melanoma patients, 50 RCC patients, and 33 NSCLC patients with reported immunotherapy response status to either aCTLA4 or aPD1 treatment alone (excluding dual/combination checkpoint inhibitor therapy). For the subset of these patients with available RNA-seq data (see Supplementary Table 1), tumor variant burden (TVB; the sum of somatic variants, tumor-specific splice junctions, and tumor-specific retained introns; log scale) was also modeled. The fit models were subsequently used to estimate the odds of therapeutic response at the 25th and 75th TMB, TVB, and neoepitope percentiles. Each cancer type was modeled separately, with the melanoma model accounting for differences in aCTLA4 vs. aPD1 response rates. P-values were adjusted for multiple comparisons using the Benjamini-Hochberg method with the p.adjust function.

Survival Analysis

Due to the low number of observed events for some cancers, only melanoma and RCC patient cohorts were appropriate for survival analysis. Patients were included in survival analysis if they had information on both overall survival status, as well as either time to event or time to censorship data. In total, 218 melanoma patients and 56 RCC patients were selected for analysis in R (v3.5.1). The coxph function from the survival package (v2.44-1.1) was used to fit proportional hazards regression models, and the survfit function from the survival package was used to compute survival curves. For comparison with patients not treated with immunotherapy, we also performed survival analysis with SKCM and KIRC patients from TCGA. We obtained mutation annotation format (MAF) files and clinical data for these patients from the Broad Institute (54). Patients with both mutation information and survival information were used for analysis (320 SKCM patients and 415 KIRC patients). Mutational burden was determined by counting the number of somatic variants listed in each patient's MAF file, and a patient was considered to have survival information if they had information on time to death or a non-zero and non-negative value on time to last follow up.

RESULTS

Distribution of tumor variant and neoepitope burdens

We find that the median TMB (based on consensus DNA variant calls; see Methods) varies by an order of magnitude across disease types, ranging from 323 variants for RCC to 3,343 variants for MMR-deficient cancers (Supplementary Figure 1). Adjusting by genome coverage for each patient, the median TMB was 4.83 mutations/Mbp (ranging from 0.68 for RCC to 6.81 for melanoma, see Figure 1A). Almost all contributory mutations to TMB were found to be single nucleotide variants (99.98% on average, per patient), with the remainder from in-frame and frameshift deletions (see Supplementary Figure 2). Note that RNA variants such as alternative exon-exon junctions and retained introns were also assessed in the subset of patients with corresponding RNA-sequencing data (see Methods). Overall, tumor-specific junction burden appeared to be less variable across cancer types (ranging from 1301 for RCC

to 2048.5 for melanoma). While retained introns (RI) have also been described as a source of neoepitopes (8), only 27 melanoma patients with RNA-seq data had any predicted RIs, with a median RI burden in those patients of 929 introns. Integrating these tumor DNA and RNA variants (given matched RNA-seq data) into a single combined tumor variant burden (TVB) yielded a median increase of 2345 variants per patient, with RNA sources of variation accounting for an average 54.9% of overall variants (see Figure 2). Moreover, consideration of DNA variant burden alone neglects substantial somatic variation for some patients, as RNA sources of variation can constitute up to 94.3% of TVB.

As TMB and TVB are indirect assessments of cancer neoantigen load, we next calculated DNA-derived, RNA-derived and overall neoepitope burdens per patient from putative protein-level variation (see Methods). The median per-patient DNA-derived neoepitope burden (for peptides predicted to bind to at least one of a patient's MHC Class I or II alleles) was 14,131 peptides (ranging from 6,122 for RCC to 14,590 for MMR-deficient cancers, see Figure 1B) and was highly correlated with TMB itself (Pearson's product-moment correlation of 0.96, $p < 2.2 \times 10^{-16}$; see Supplementary Figure 3). While not all patients possessed RIs, the median per-patient RNA-derived neoepitope burden among the 27 melanoma patients with predicted RIs (366,843 peptides) was an order of magnitude higher than DNA-derived neoepitopes in the vast majority of cases (Supplementary Figure 4).

In addition to reporting the bulk number of neoepitopes per patient, we also analyzed the distribution of peptide presentation by patient-specific HLA types. Overall, a median of 34% of possible peptides are presented by one or more patient-specific MHC Class I or II alleles. Among these, any given neoepitope is, on average, only presented by a single MHC allele (Figure 3A, Supplementary Figure 5A). There are many additional degrees of freedom to surveil the peptide-level consequences of an individual variant (e.g. individual single nucleotide variants may give rise to as many as 272 different peptides of 8-24aa lengths, any of which might be presented via one or more MHC Class I or II alleles). As such, we find that 98.2% of all DNA variants resulting in peptide-level change(s) have at least one neoepitope putatively presented by at least one HLA allele, with a median of 4 different HLA alleles able to present one or more neoepitopes from each individual variant (Figure 3B, Supplementary Figure 5B). Moreover, the percentage of variants presented increases with increasing MHC heterozygosity (Figure 3C, Supplementary Figure 6). Within the cohort, 251 patients had pathogenic cancer-related mutations (see Methods), with an average of 1.6 such variants per patient. Consistent with prior work demonstrating a relative paucity of peptide presentation from cancer driver mutations (55), we find that a smaller number (approximately 84%) of driver variants in this cohort yielded neoepitopes, with only 12.6% of neopeptides from these variants on average being predicted to bind to any of a patient's HLA alleles (Figure 3).

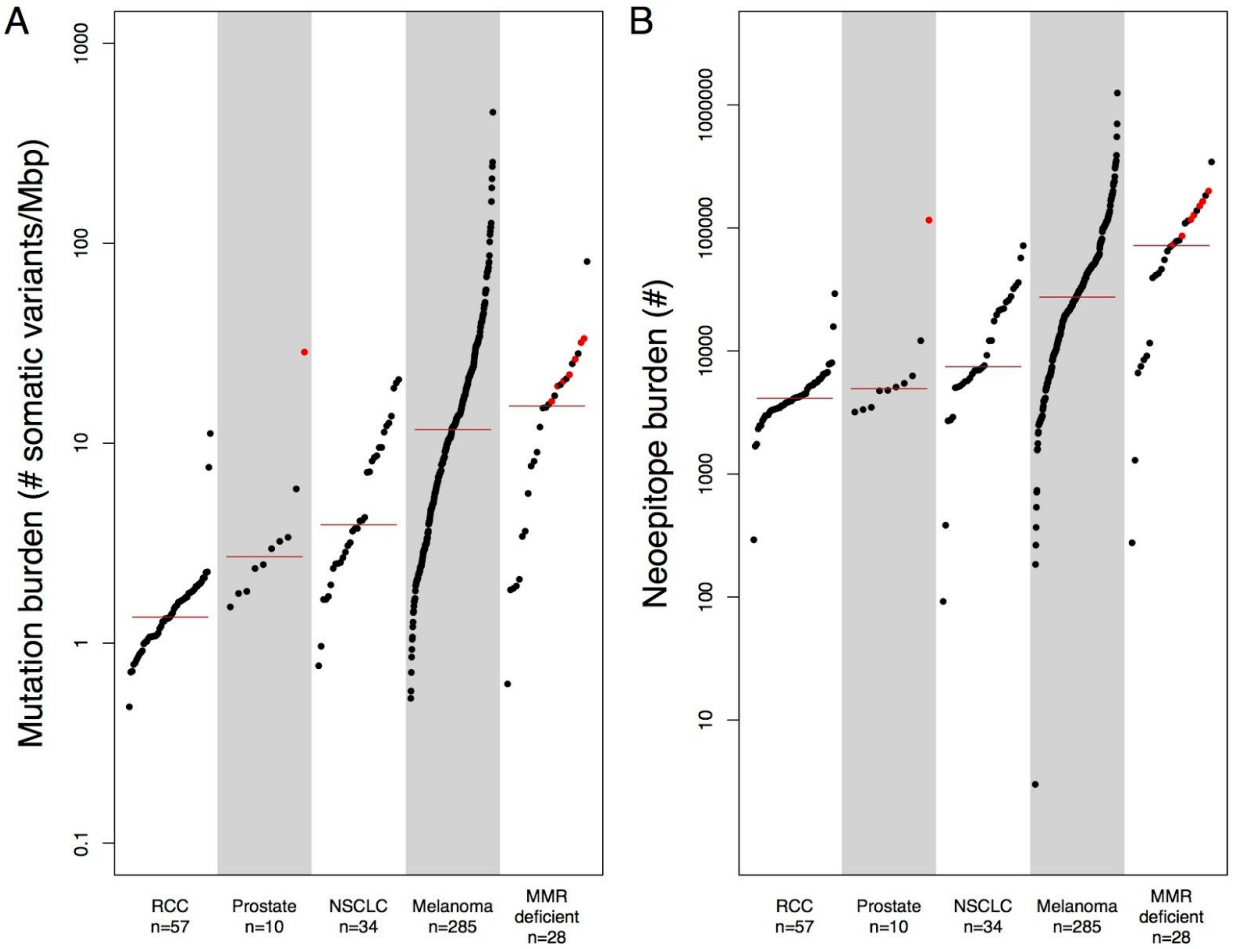


Figure 1: *Per-patient distribution of mutation and neoepitope burdens across 7 cancer types. A) The number of somatic DNA variants per patient (scaled for sequence coverage) are shown along the y-axis, with each dot representing an individual cancer patient (cancer types shown along the x-axis). Note that MMR-deficient cancers here represent a cohort of 3 different cancer types including colon, endometrial, and thyroid. Red colored dots correspond to patients with microsatellite instability (see Methods). B) The number of putative neoepitopes per patient are shown along the y-axis, with each dot representing an individual cancer patient (cancer types shown along the x-axis). Abbreviations as follows: MMR=mismatch repair.*

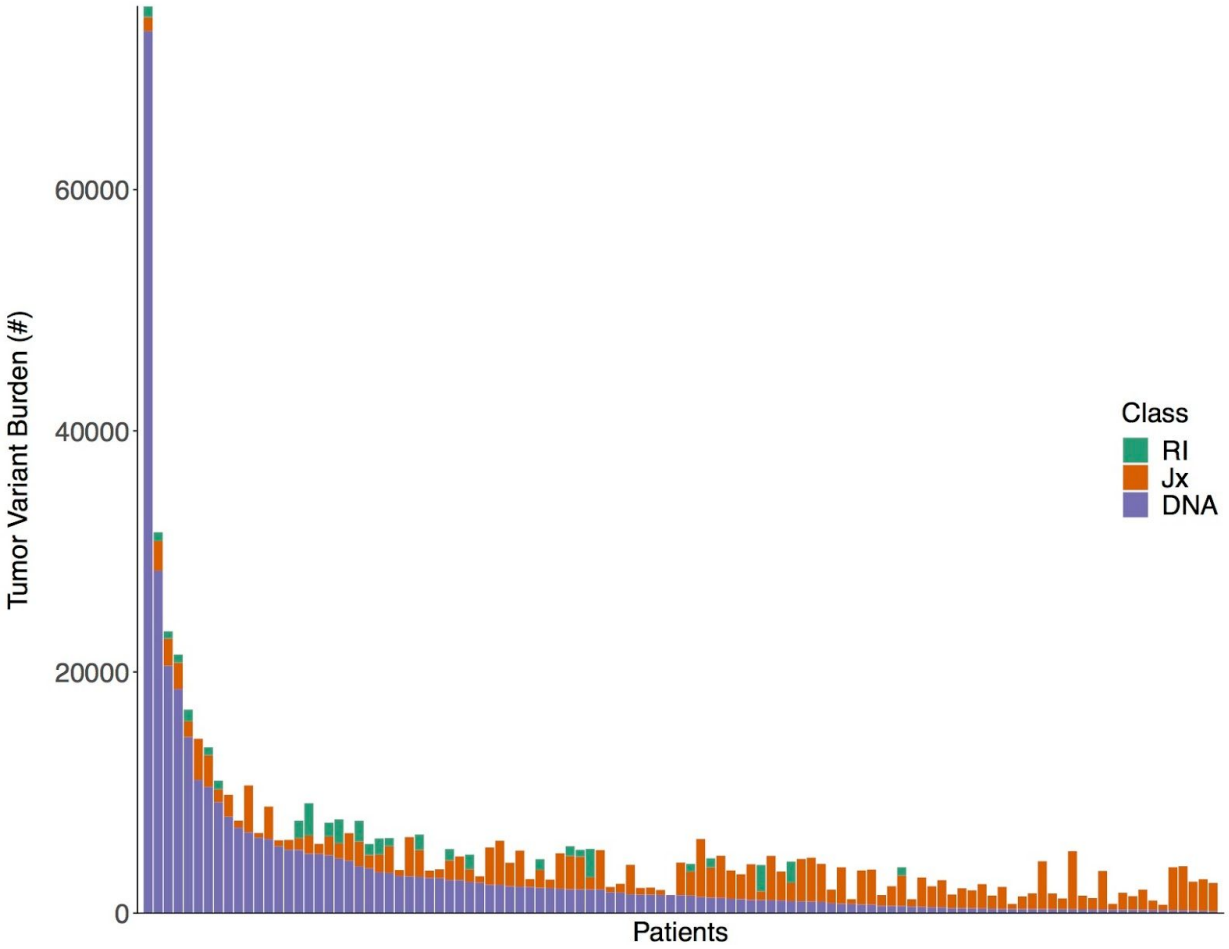


Figure 2: *Per-patient distribution of overall tumor variant burden and its components. The number of total tumor variants per patient is shown along the y-axis, with the numbers of retained introns (RI), tumor-specific exon-exon junctions (Jx), and somatic DNA variants (DNA) shown in green, red, and purple, respectively. The data for each individual patient is displayed as stacked bars along the x-axis, sorted from left to right by the number of somatic DNA variants (from highest to lowest).*

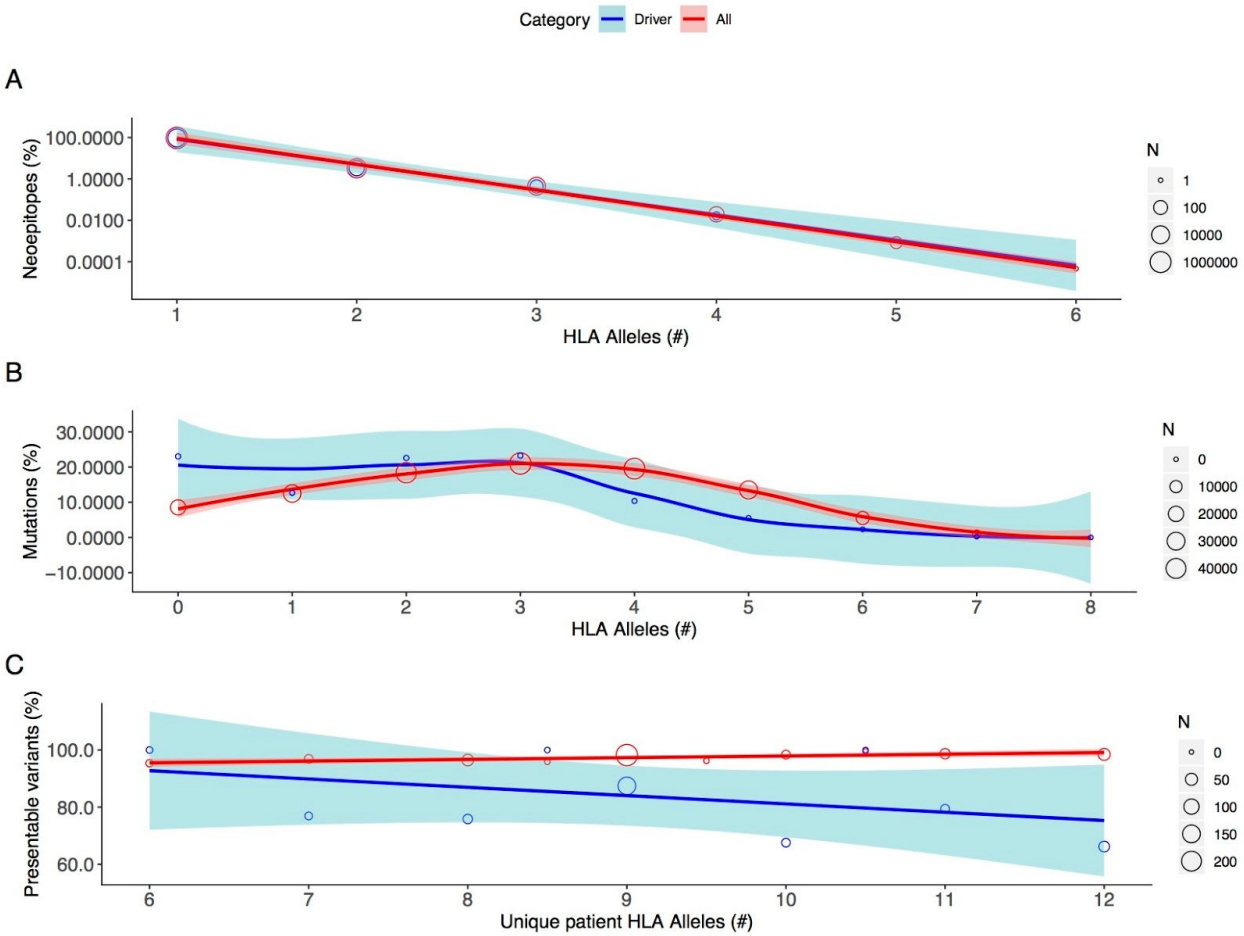


Figure 3: *Robustness of putative neoepitope presentation. A) The number of unique patient-matched HLA alleles that are predicted to present an individual neoepitope is shown along the x-axis, with the y-axis (log-scale) corresponding to the overall percent of neoepitopes sharing that same robustness of HLA presentation. Red and blue curves denote the best fit line based on linear regression for all neoepitopes and those resulting from cancer driver mutations, respectively. The surrounding red and light blue shading denotes the 95% confidence interval for all and driver-derived neoepitopes, respectively. Individual data points are shown as open circles, whose diameter corresponds to the number of neoepitopes as shown by the corresponding scale at right. B) The total number of unique patient-matched HLA alleles that are predicted to present one or more neoepitopes arising from a single DNA mutation is shown along the x-axis, with the y-axis corresponding to the overall percent of mutations sharing that same robustness of HLA presentation. Red and blue curves denote the best fit line based on local polynomial regression for all mutations and cancer driver mutations, respectively. The surrounding red and light blue shading denotes the 95% confidence interval for all and driver mutations, respectively. Individual data points are shown as open circles, whose diameter corresponds to the number of mutations as shown by the corresponding scale at right. C) The percentage of total variants that are predicted to be presented by one or more patient-matched HLA alleles is shown along the y-axis, with the x-axis corresponding to the number of unique*

HLA alleles for that patient. Red and blue curves denote the best fit line based on linear regression for all mutations and cancer driver mutations, respectively. The surrounding red and light blue shading denotes the 95% confidence interval for all and driver mutations, respectively. Individual data points are shown as open circles, whose diameter corresponds to the number of mutations as shown by the corresponding scale at right. Note that a predicted HLA binding affinity threshold of ≤500nM was used in all cases (see Methods).

Tumor variant and neoepitope burdens as predictors of response and survival

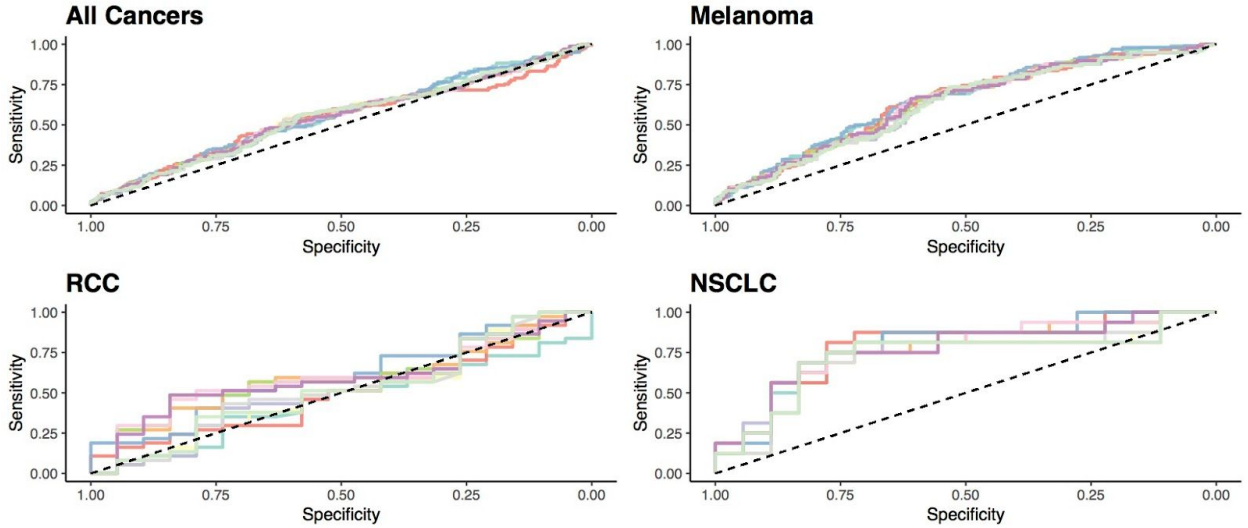
We next sought to quantify immunotherapy response rate as a function of TMB, TVB, and neoepitope burden. Using disease-specific logistic regression models, we find that every log fold increase in TMB increases the odds of immunotherapy response in metastatic melanoma patients by 54%, independent of targeted immune checkpoint ($p < 0.001$). However, consistent with clinical observations, the response rates for αPD1 therapies are at least 10% greater than for comparable patients with metastatic melanoma receiving αCTLA4 (Table 1). Similarly, we find a 265% increase in the odds of response for non-small cell lung cancer patients per log fold change in TMB ($p = 0.019$), though we find no significant association between odds of response and TMB for patients with renal cell carcinoma ($p = 0.642$).

	αPD1			αCTLA4			
Cancer type & Metric	N	Response (25th %ile)	Response (75th %ile)	N	Response (25th %ile)	Response (75th %ile)	p-value
Melanoma							
TMB	50	0.331	0.528	195	0.238	0.415	< 0.001
Neoepitopes	50	0.388	0.572	195	0.246	0.407	< 0.001
TVB	27	0.393	0.485	61	0.323	0.407	0.302
NSCLC							
TMB	33	0.282	0.682				0.019
Neoepitopes	33	0.303	0.712				0.021
RCC							
TMB	50	0.677	0.647				0.642
Neoepitopes	50	0.647	0.674				0.693
TVB	17	0.66	0.542				0.369

Table 1: *Immunotherapy (αPD1 and αCTLA4) response probability based on logistic models of tumor mutational burden (TMB), neoepitope burden (Neoepitopes), and combined tumor DNA- and RNA- variant burden (TVB) for melanoma, non-small cell lung cancer (NSCLC) and renal cell carcinoma (RCC).*

Neoepitope burden alone predicted response to immunotherapy comparatively well as TMB, calculated using both raw and coverage-adjusted counts (see Figure 4). Additionally, we weighted neoepitope burden by several criteria hypothesized to be related to increased immunogenicity, including: number of amino acid mismatches per peptide, number of MHC alleles predicted to bind each peptide, and number of TCGA-expressed transcripts of origin for the peptide (see Methods). In all cases, these weighted burdens yielded similar predictive capabilities to TMB or unadjusted neoepitope burden, though mismatch- and allele-weighted neoepitope burdens incrementally improved predictive capacity for RCC patients (see Figure 4). Interestingly, global assessment of HLA presentation (unique HLA allele count per patient)

added slight predictive capacity to TMB in all cases (see Figure 4). However, the capacity for any of these metrics to predict patient-level immunotherapy response varied substantially by cancer type, with the highest predictive power for the non-small cell lung cancer cohort, but a very limited predictive capability in RCC or when pooled across all cancer types (see Figure 4). Indeed, TMB as calculated by consensus variant calls predicts immunotherapy response more poorly than the experimental noise of the breadth of genomic coverage (Mbp) obtained via DNA sequencing (see Supplementary Figure 7).

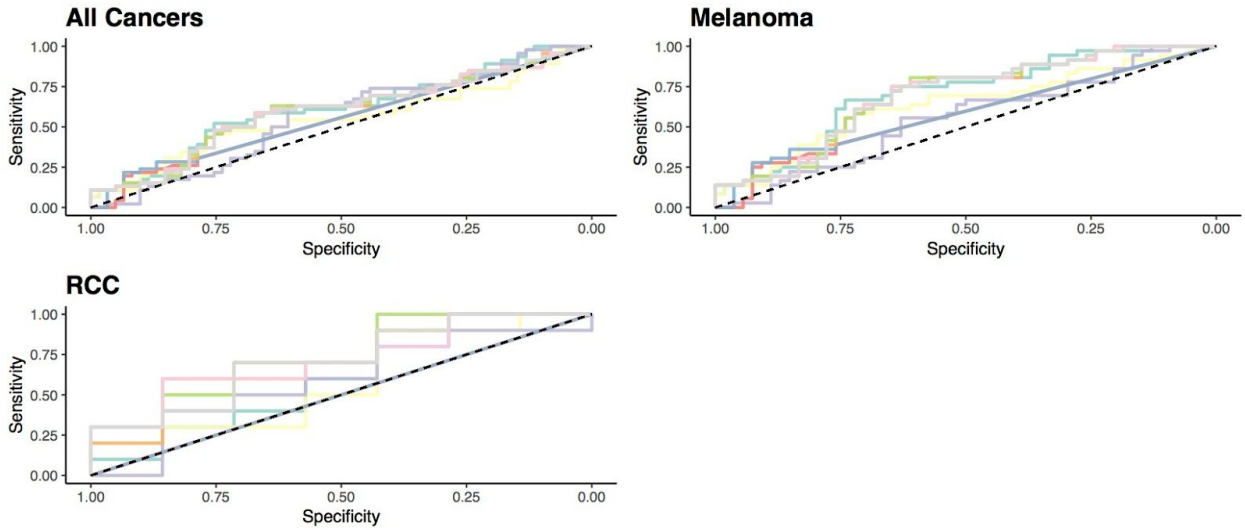


AUC by Burden and Cancer Type

Cancer type	N	TMB (raw)	TMB (adj.)	TMB*HLA	Neoepitopes	Mismatch (M)	Allele (A)	TCGA (T)	M*A	M*T	A*T	M*A*T
All	409	0.556	0.539	**0.559**	0.548	0.547	0.546	0.550	0.547	0.548	0.550	0.548
Melanoma	285	0.652	0.647	**0.660**	0.641	0.639	0.640	0.630	0.639	0.627	0.630	0.628
RCC	56	0.454	0.501	**0.603**	0.586	0.593	0.593	0.519	0.592	0.515	0.506	0.515
NSCLC	34	0.771	**0.788**	**0.788**	0.771	0.767	0.771	0.726	0.771	0.729	0.722	0.726

Figure 4: *Receiver operating characteristic curves of predictive capacity of 11 different mutation/neoepitope burden metrics. The upper panels depict the true positive rate (sensitivity, y-axis) and false positive rate (1-specificity, x-axis) for each metric across all probability thresholds. The table in the lower panel reports the area-under-the-curve (AUC) for each metric (columns) applied to a different cancer cohort (rows), with colors above the methods indicating the color of the corresponding curve in the upper panels. TMB is used as a predictor in both raw and coverage-adjusted forms, as well as in a multiplicative combination with patient HLA allele count. Neoepitope burden is used as predictor in both raw and extended formats (see Methods). Extended neoepitope burden metrics include number of amino acid mismatches (M), number of HLA alleles predicted to bind each epitope (A), and number of transcripts expressing each epitope in TCGA (T), along with their multiplicative combinations. Bold-faced values indicate the best value for each cancer cohort.*

For patients with tumor RNA-seq data, we also investigated how TVB and RNA-derived neoepitopes predicted response to immunotherapy (see Figure 5). We specifically considered tumor-specific junction burden, retained intron burden, retained intron neoepitope burden, and patient-specific expression-weighted neoepitope burdens (see Methods). As before, the vast majority of metrics (e.g. TMB, TVB) were all comparable in terms of predictive performance; however, the incorporation of amino acid mismatches and gene expression information substantially improved predictive capability for RCC patients, acknowledging the small cohort size (n=17, see Figure 5).



AUC by RNA Feature Burden and Cancer Type

Cancer type	N	TMB	TVB	Jx	RI	RI epitopes	Expression (E)	AA mismatches (M) * E	Alleles (A) * E	M*A*E
All	106	**0.610**	0.550	0.562	0.549	0.555	0.598	0.597	0.598	0.596
Melanoma	89	**0.701**	0.637	0.551	0.578	0.588	0.697	0.697	0.697	0.694
RCC	17	0.614	0.557	0.586	0.500	0.500	0.714	**0.743**	0.714	0.714

Figure 5: *Receiver operating characteristic curves of predictive capacity of 9 different variant/neoepitope burden metrics. The upper panels depict the true positive rate (sensitivity, y-axis) and false positive rate (1-specificity, x-axis) for each metric across all probability thresholds. The table in the lower panel reports the area-under-the-curve (AUC) for each metric (columns) applied to a different cancer cohort (rows), with colors above the methods indicating the color of the corresponding curve in the upper panels. TMB and TVB are used as predictors in the raw formats. Jx represents the number of tumor-specific junctions per patient, and RI represents the number of retained introns per patient, with RI epitopes representing neoepitopes derived from those retained introns. Neoepitope burden is used as predictor in its RNA-feature-extended formats (see Methods). Extended neoepitope burden metrics include number of expressed transcripts for each epitope (E), number of amino acid mismatches (M), number of HLA alleles predicted to bind each epitope (A), and number of transcripts expressing each epitope in TCGA (T), along with their multiplicative combinations.*

Using an established threshold for identifying tumors with “high” TMB, namely TMB that exceeds the disease-matched 80th percentile (2), we investigated the metric’s capacity to predict overall survival in the context of immune checkpoint blockade therapy. While not statistically significant ($p > 0.05$, based on Cox PH model), we saw a clear trend towards improved overall survival among individuals with metastatic melanoma and a high TMB (Figure 6A). Additionally, model comparisons using different TMB percentile cutoffs suggest that differences in overall survival for high and low TMB groups may be threshold dependent and alter model significance (see Supplementary Figure 8). In contrast, the same trend is not seen between TMB and overall survival among a separate cohort of patients (TCGA) in the absence of immunotherapy (Figure 6A). We also observed no differences in survival among individuals with renal cell carcinoma (Figure 6B). In both cases, TVB and neoepitope burden demonstrate comparable capacities to stratify overall survival as TMB (Supplementary Figures 9 and 10).

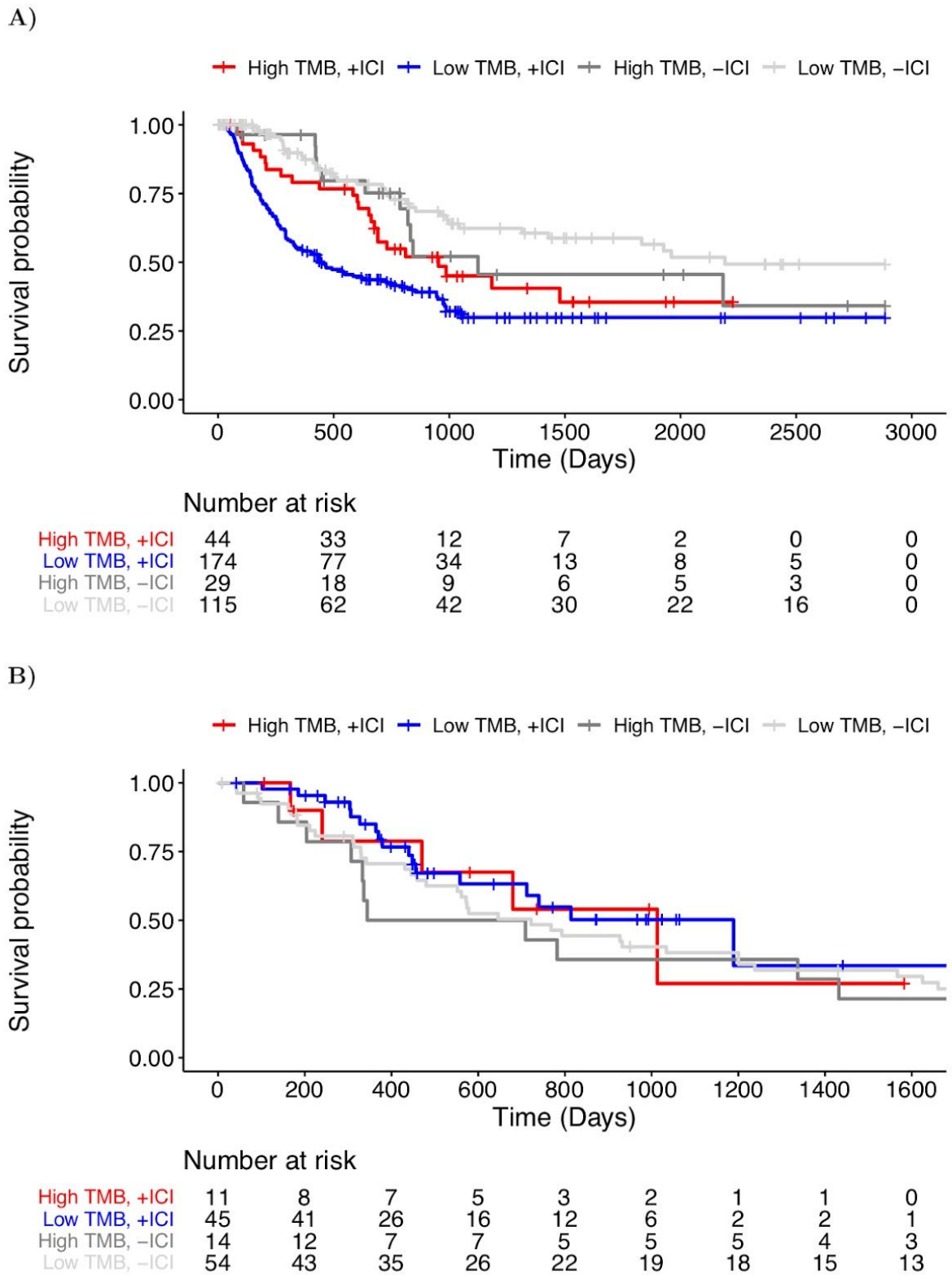


Figure 6: *Overall survival among cancer patients with high and low TMB. A) Kaplan-meier curves for immunotherapy-treated (+ICI) and immunotherapy-naive (-ICI) Stage III-IV melanoma*

patients with high TMB (>80th percentile) are shown in red, and dark gray, respectively, while immunotherapy-treated (+ICI) and immunotherapy-naive (-ICI) patients with low TMB (≤80th percentile) are shown in blue and light gray, respectively. The underlying table corresponds to the number of patients at risk of death at each timepoint. Note: TCGA SKCM patient data (-ICI) is censored at 2,885 days (maximal follow-up in immunotherapy-treated cohort) to emphasize comparable time-scales. B) Kaplan-meier curves for the immunotherapy-treated (+ICI) and immunotherapy-naive (-ICI) metastatic (Stage IV) renal cell carcinoma patients with high TMB (>80th percentile) are shown in red, and dark gray, respectively, while immunotherapy-treated (+ICI) and immunotherapy-naive (-ICI) patients with low TMB (≤80th percentile) are shown in blue and light gray, respectively. The underlying table corresponds to the number of patients at risk of death at each timepoint. Note: TCGA KIRC patient data is censored at 1,724 days (maximal follow-up in immunotherapy-treated cohort) to emphasize comparable time-scales.

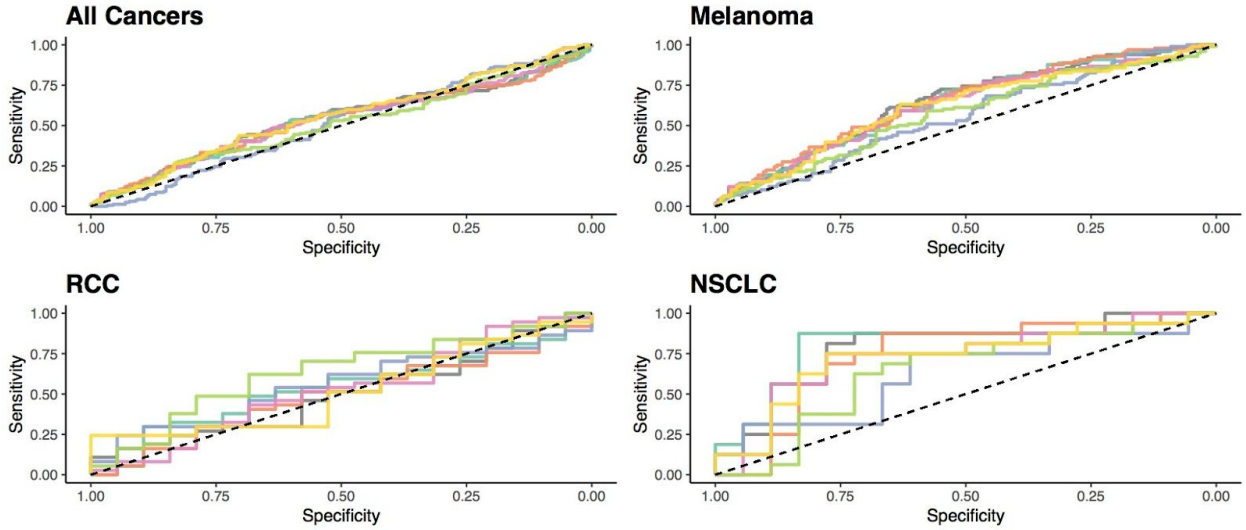
Metric instability of tumor variant and neoepitope burdens

We find, however, that TMB is not robust across variant calling methods. TMB as reported by individual variant calling tools was moderately similar to that reported by consensus calls (see Supplementary Figure 11), but variability in per-caller TMB increased with increasing number of variants (see Supplementary Figure 12). Additionally, the difference in TMB between the highest and lowest counts from individual callers per patient (median difference of 1,817 variants per patient) reflects a substantial fraction of the overall TMB, accounting for a median 59.3% of the value in the metric overall (see Methods).

We also compared tumor mutational burden as reported by the authors of the original manuscripts from which our cohort originated with our standardized consensus approach. While author-reported and consensus values for TMB were significantly correlated (Pearson's product-moment correlation of 0.61, $p < 2.2x10^{-16}$; see Supplementary Figure 13), we note that author-reported values have a universally higher predictive capacity than we observe using consensus data (Supplementary Figure 14). We also find important discrepancies in per-patient classification. Approximately 4.8% of patients are incongruously determined to be TMB "high" or "low" (using a TMB threshold >80th percentile as per (2)), however as many as 13.6% of patients may be dubiously classified using alternative thresholds (e.g. 22nd-36th percentiles; see Supplementary Figure 15). Consensus and author-reported nonsynonymous mutation burdens exhibited a similar extent of correlation as well as per-patient instability of classification (Pearson's product-moment correlation of 0.65, $p < 2.2x10^{-16}$; see Supplementary Figures 14 and 16). The correlation between consensus-derived neoepitope burden and that reported by the original manuscripts was significant, but not strong (Pearson's product-moment correlation of 0.23, $p = 0.00073$; see Supplementary Figure 17). Moreover, we find that response hazard ratios are not stable based on TMB thresholds, a phenomenon especially dramatic in the RCC cohort (see Supplementary Figure 8), and consistent with prior findings that a single TMB threshold is inappropriate to apply across different cancer types (2).

Finally, we find that the predictive performance of TMB is sensitive to the method(s) used to perform variant calling (see Figure 7). Note that the same phenomenon holds true for raw TMB

counts (see Supplementary Figure 18). While outside the scope of the current manuscript, note also that the identity of resulting neoepitopes is also highly sensitive to variant calling method (see Supplementary Figure 19).



AUC by Coverage–adjusted Mutation Burden and Cancer Type

Cancer type	N	Consensus	MuSE	MuTect	Pindel	RADIA	SomaticSniper	VarScan
All	409	0.539	0.531	0.529	0.508	0.537	0.513	**0.553**
Melanoma	285	0.647	0.639	**0.650**	0.560	0.625	0.574	0.620
RCC	56	0.501	0.538	0.491	0.555	0.528	**0.633**	0.519
NSCLC	34	0.788	**0.806**	0.740	0.632	0.733	0.639	0.729

Figure 7: *Receiver operating characteristic curves of predictive capacity of coverage-adjusted TMB from 7 different variant calling methods. The upper panels depict the true positive rate (sensitivity, y-axis) and false positive rate (1-specificity, x-axis) for each method across all probability thresholds. The table in the lower panel reports the area-under-the-curve (AUC) for each method (columns) applied to a different cancer cohort (rows), with colors above the methods indicating the color of the corresponding curve in the upper panels. TMB as determined by consensus calling (see Methods) is compared to the individual variant calling tools used in consensus calling. RCC=renal cell carcinoma, NSCLC=non-small cell lung cancer.*

DISCUSSION

To the best of our knowledge, this is the first study to evaluate TMB and correlated downstream metrics such as neoepitope burden from whole exome sequencing data using a gold standard ensemble approach (28,56) applied to a meta-cohort of immunotherapy-treated

cancer patients across multiple studies and disease types. This study also introduces the concept of tumor variant burden, incorporating potential RNA-derived sources of variants where available, and is the first study to estimate immunotherapy response rate as a function of TMB, TVB, and neoepitope burden. Moreover, this study is the first to quantitatively evaluate the stability of TMB as a metric, and the first to directly compare the predictive capacities of multiple TMB and related metrics.

Ultimately, we show that TMB is a cautionary predictor of immunotherapy response, with substantial caveats regarding: 1) predictive capacity differences among different cancer types, with RCC being no better than random chance, 2) sensitivity of TMB and downstream metrics to variant calling methodology, and 3) stability of TMB thresholds and their ability to classify patients in a population. This suggests that the prospective clinical utilization of TMB is likely subject to many of these same issues, and may result in unintended harms, whether due to omission of therapy for individuals with “low” TMB who might nonetheless benefit, or due to increased risk of toxicity in a “high” TMB population subject to overuse of immunotherapy.

With rare exception, we find no added predictive benefit to evaluating more complex bulk metrics downstream of TMB. Akin to prior observations, incorporation of HLA genotype diversity adds slightly to the predictive capacity of TMB (57). Given the added technical effort and costs required to perform these analyses, we conclude that TMB is likely the optimal bulk assessment of tumor variation among those tested, though inclusion of HLA diversity data may marginally improve estimates. However, such bulk measurements neglect the potential importance of individual cancer neoantigens.

This study has several limitations. First, numerous sampling based assays have also been used to assess TMB (e.g. (2,58,59)), however, we did not evaluate these data in this study, instead focusing on whole exome sequencing data as the prevailing gold standard for accurate mutational assessment. Note that these targeted assays would not enable incorporation of HLA allelic diversity data into a predictive model. Note also that there is wide variability among TMB assay design, analysis, and performance, with the potential for overestimation of TMB when using gene-targeted assays (60). Ultimately, along with the substantial variability among widely-used targeted assays (11), and the futility of expecting universal adoption of a single technique, this study highlights the need for increased standardization of TMB interpretation, a subject of active pursuit by the TMB Harmonization Project (61). Second, we did not compare TMB in this dataset with other potential predictors of immunotherapy response (e.g. based on gene expression (62) or copy number instability (63)), however it is possible that TMB could be synergistic with such orthogonal metrics. Third, by virtue of the retrospective nature of these data and limited availability of whole exome sequencing cohorts, this study cannot be assumed to translate to emerging immunotherapies and instead is interpretable exclusively for αPD1 and αCTLA4 therapy.

While this study is consistent with multiple prior reports demonstrating the importance of TMB in predicting immunotherapy response (e.g. (2,64)), the caveats raised herein are of high concern for the field overall. Our collective emphasis on TMB is understandable given its relative ease of quantification using various techniques, however it is indeed a cautionary and indirect predictor. Tumors with higher TMB have been hypothesized to have more neoantigens that can be recognized by the immune system in response to checkpoint inhibition, yet the data

presented here and data previously published (2) support the use of substantially different "absolute" TMB thresholds for immunotherapy response prediction across different diseases. This suggests an added layer of as-of-yet undefined complexity not captured in the current bulk metrics, and likely related to disease-specific biology.

CONCLUSIONS

In conclusion, we find sufficient cause to suggest that the predictive clinical value of TMB should not be overstated or oversimplified. While it is readily quantified, TMB is at best a suggestive surrogate biomarker of immunotherapy response. The data confirms TMB as a reasonable predictor in non-small cell lung cancer, and a weak predictor in melanoma. The data does not support TMB in isolation as a predictive biomarker for RCC, though it may be feasibly combined with HLA allelic diversity to achieve marginal performance.

LIST OF ABBREVIATIONS

αCTLA4: Anti-cytotoxic T-lymphocyte-associated protein 4, αPD1: Anti-programmed cell death protein 1, AUC: Area Under the Curve, COAD: Colon Adenocarcinoma, GATK: Genome Analysis Toolkit, GTEx: Genotype Tissue Expression, HLA: Human Leukocyte Antigen, Indel: Insertion or deletion, KIRC: Kidney Renal Clear Cell Carcinoma, kma: Keep Me Around, LUAD: Lung Adenocarcinoma, LUSC: Lung Squamous Cell Carcinoma, MHC: Major Histocompatiblity Complex, MMR: Mismatch Repair, NSCLC: Non Small Cell Lung Cancer, PRAD: Prostate Adenocarcinoma, RCC: Renal Cell Carcinoma, RI: Retained intron, RNA-seq: RNA sequencing, ROC: Receiver Operating Characteristic, SKCM: Skin Cutaneous Melanoma, SRA: Sequence Read Archive, TCGA: The Cancer Genome Atlas, THCA: Thyroid carcinoma, TMB: Tumor Mutational Burden, TPM: Transcripts Per Million, TVB: Tumor Variant Burden, UCEC: Uterine Corpus Endometrial Carcinoma, VCF: Variant Call Format, WES: Whole Exome Sequencing

DECLARATIONS

Availability of data and materials: Data used in these analyses is available on the Sequence Read Archive under accessions PRJNA278450, PRJNA293912, PRJNA305077, PRJNA306070, PRJNA307199, PRJNA312948, PRJNA324705, PRJNA343789, PRJNA357321, PRJNA369259, PRJNA420786, and PRJNA82745 and on the European Genome-Phenome Archive under the accession EGAD00001004352. Data from Le et al. (9) and Graff et al. (16) are available from those authors upon reasonable request. The results of our analyses of these data are available as Supplemental Tables 2-6 in Additional File 1.

Competing interests: The authors have no competing interests to declare.

Disclaimer: The contents do not represent the views of the U.S. Department of Veterans Affairs or the United States Government.

Funding: Funding for this study was generously provided by the Sunlin & Priscilla Chou Foundation.

Author's contributions: Study and conceptual design was by MAW, AN, and RFT. Data analysis was by MAW, BRW, and JKD. Manuscript writing was predominantly by MAW, AN, and RFT, with contributions from BRW and JKD. Final manuscript review and approval was by MAW, BRW, JKD, AN, and RFT.

Acknowledgments: Thank you to Adam Struck and Kyle Ellrott for their support implementing somatic variant calling pipelines. Thank you to Julie Graff and Rachel Slottke for providing prostate cancer cohort data.

REFERENCES

1. Van Allen EM, Miao D, Schilling B, Shukla SA, Blank C, Zimmer L, et al. Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. Science. 2015;350(6257):207–11.

2. Samstein RM, Lee C-H, Shoushtari AN, Hellmann MD, Shen R, Janjigian YY, et al. Tumor mutational load predicts survival after immunotherapy across multiple cancer types. Nat Genet. 2019 Feb;51(2):202–6.

3. Rizvi NA, Hellmann MD, Snyder A, Kvistborg P, Makarov V, Havel JJ, et al. Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. Science. 2015 Apr 3;348(6230):124–8.

4. Carreno BM, Magrini V, Becker-Hapak M, Kaabinejadian S, Hundal J, Petti AA, et al. A dendritic cell vaccine increases the breadth and diversity of melanoma neoantigen-specific T cells. Science. 2015;348(6236):803–8.

5. Hugo W, Zaretsky JM, Sun L, Song C, Moreno BH, Hu-Lieskovan S, et al. Genomic and Transcriptomic Features of Response to Anti-PD-1 Therapy in Metastatic Melanoma. Cell. 2017 Jan 26;168(3):542.

6. Snyder A, Makarov V, Merghoub T, Yuan J, Zaretsky JM, Desrichard A, et al. Genetic basis for clinical response to CTLA-4 blockade in melanoma. N Engl J Med. 2014 Dec 4;371(23):2189–99.

7. Turajlic S, Litchfield K, Xu H, Rosenthal R, McGranahan N, Reading JL, et al. Insertion-and-deletion-derived tumour-specific neoantigens and the immunogenic phenotype: a pan-cancer analysis. Lancet Oncol. 2017 Aug;18(8):1009–21.

8. Smart AC, Margolis CA, Pimentel H, He MX, Miao D, Adeegbe D, et al. Intron retention is a source of neoepitopes in cancer. Nat Biotechnol. 2018;36(11):1056–8.

9. Le DT, Durham JN, Smith KN, Wang H, Bartlett BR, Aulakh LK, et al. Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade. Science. 2017 Jul

28;357(6349):409–13.

10. Boyiadzis MM, Kirkwood JM, Marshall JL, Pritchard CC, Azad NS, Gulley JL. Significance and implications of FDA approval of pembrolizumab for biomarker-defined disease. J Immunother Cancer. 2018 May 14;6(1):35.

11. Meléndez B, Van Campenhout C, Rorive S, Remmelink M, Salmon I, D'Haene N. Methods of measurement for tumor mutational burden in tumor tissue. Transl Lung Cancer Res. 2018 Dec;7(6):661–7.

12. Zaretsky JM, Garcia-Diaz A, Shin DS, Escuin-Ordinas H, Hugo W, Hu-Lieskovan S, et al. Mutations Associated with Acquired Resistance to PD-1 Blockade in Melanoma. N Engl J Med. 2016 Sep 1;375(9):819–29.

13. Gao J, Shi LZ, Zhao H, Chen J, Xiong L, He Q, et al. Loss of IFN-γ Pathway Genes in Tumor Cells as a Mechanism of Resistance to Anti-CTLA-4 Therapy. Cell. 2016 Oct 6;167(2):397–404.e9.

14. Roh W, Chen P-L, Reuben A, Spencer CN, Prieto PA, Miller JP, et al. Integrated molecular analysis of tumor biopsies on sequential CTLA-4 and PD-1 blockade reveals markers of response and resistance. Sci Transl Med [Internet]. 2017 Mar 1;9(379). Available from: http://dx.doi.org/10.1126/scitranslmed.aah3560

15. Amaria RN, Reddy SM, Tawbi HA, Davies MA, Ross MI, Glitza IC, et al. Publisher Correction: Neoadjuvant immune checkpoint blockade in high-risk resectable melanoma. Nat Med. 2018 Dec;24(12):1942.

16. Graff JN, Alumkal JJ, Drake CG, Thomas GV, Redmond WL, Farhad M, et al. Early evidence of anti-PD-1 activity in enzalutamide-resistant prostate cancer. Oncotarget. 2016 Aug 16;7(33):52810–7.

17. Miao D, Margolis CA, Gao W, Voss MH, Li W, Martini DJ, et al. Genomic correlates of response to immune checkpoint therapies in clear cell renal cell carcinoma. Science. 2018 Feb 16;359(6377):801–6.

18. cancerit. cancerit/dockstore-cgpmap [Internet]. GitHub. [cited 2018 Sep 12]. Available from: https://github.com/cancerit/dockstore-cgpmap

19. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754–60.

20. gt. gt1/biobambam2 [Internet]. GitHub. [cited 2018 Sep 12]. Available from: https://github.com/gt1/biobambam2

21. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010 Mar 15;26(6):841–2.

22. Larson DE, Harris CC, Chen K, Koboldt DC, Abbott TE, Dooling DJ, et al. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. Bioinformatics. 2012 Feb 1;28(3):311–7.

23. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 2012 Mar;22(3):568–76.

24. Fan Y, Xi L, Hughes DST, Zhang J, Zhang J, Futreal PA, et al. MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. Genome Biol. 2016 Aug 24;17(1):178.

25. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol. 2013 Mar;31(3):213–9.

26. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics. 2009;25(21):2865–71.

27. Radenbaugh AJ, Ma S, Ewing A, Stuart JM, Collisson EA, Zhu J, et al. RADIA: RNA and DNA integrated analysis for somatic mutation detection. PLoS One. 2014 Nov 18;9(11):e111516.

28. Ellrott K, Bailey MH, Saksena G, Covington KR, Kandoth C, Stewart C, et al. Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. Cell Syst. 2018 Mar 28;6(3):271–81.e7.

29. Tan A, Abecasis GR, Kang HM. Unified representation of genetic variants. Bioinformatics. 2015 Jul 1;31(13):2202–4.

30. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010 Sep;20(9):1297–303.

31. Wood MA, Nguyen A, Struck A, Ellrott K, Nellore A, Thompson RF. neoepiscope improves neoepitope prediction with multi-variant phasing [Internet]. bioRxiv. 2018 [cited 2019 May 13]. p. 418129. Available from: https://www.biorxiv.org/content/10.1101/418129v1.abstract

32. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. 2018 Jan 4;46(D1):D1062–7.

33. open-cravat [Internet]. Github; [cited 2019 May 30]. Available from: https://github.com/KarchinLab/open-cravat

34. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013 Jan 1;29(1):15–21.

35. recount2: analysis-ready RNA-seq gene and exon counts datasets [Internet]. [cited 2019 May 15]. Available from: https://jhubiostatistics.shinyapps.io/recount/

36. GENCODE - Human Release 28 [Internet]. [cited 2019 May 15]. Available from: https://www.gencodegenes.org/human/release_28.html

37. Nellore A, Collado-Torres L, Jaffe AE, Alquicira-Hernández J, Wilks C, Pritt J, et al. Rail-RNA: scalable analysis of RNA-seq splicing and coverage. Bioinformatics. 2017 Dec 15;33(24):4033–40.

38. Bernstein MN, Doan A, Dewey CN. MetaSRA: normalized human sample-specific metadata for the Sequence Read Archive. Bioinformatics. 2017 Sep 15;33(18):2914–23.

39. MetaSRA [Internet]. [cited 2019 May 21]. Available from: http://metasra.biostat.wisc.edu/?and=CL:0000148¬=DOID:162&sampletype=cell%20line

40. MetaSRA [Internet]. [cited 2019 May 21]. Available from: http://metasra.biostat.wisc.edu/?and=CL:0000148¬=DOID:162&sampletype=primary%20cells

41. Wilks C, Gaddipati P, Nellore A, Langmead B. Snaptron: querying splicing patterns across tens of thousands of RNA-seq samples. Bioinformatics. 2018 Jan 1;34(1):114–6.

42. Snaptron User Guide — Snaptron 1.6 documentation [Internet]. [cited 2019 May 21]. Available from: http://snaptron.cs.jhu.edu/

43. kma [Internet]. Github; [cited 2019 May 13]. Available from: https://github.com/pachterlab/kma

44. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012 Mar 4;9(4):357–9.

45. Roberts A, Pachter L. Streaming fragment assignment for real-time analysis of sequencing experiments. Nat Methods. 2013 Jan;10(1):71–3.

46. NCI Primary Human Melanocyte QTL Study (ID 421623) - BioProject - NCBI [Internet]. [cited 2019 May 13]. Available from: https://www.ncbi.nlm.nih.gov/bioproject/PRJNA421623/

47. Szolek A, Schubert B, Mohr C, Sturm M, Feldhahn M, Kohlbacher O. OptiType: precision HLA typing from next-generation sequencing data. Bioinformatics. 2014 Dec 1;30(23):3310–6.

48. Boegel S, Löwer M, Schäfer M, Bukur T, de Graaf J, Boisguérin V, et al. HLA typing from RNA-Seq sequence reads. Genome Med. 2012 Dec 22;4(12):102.

49. Bhattacharya R, Sivakumar A, Tokheim C, Guthrie VB, Anagnostou V, Velculescu VE, et al. Evaluation of machine learning methods to predict peptide binding to MHC Class I proteins [Internet]. Available from: http://dx.doi.org/10.1101/154757

50. Altschul SF, Gish W, Miller W, Myers EW. Basic local alignment search tool. Journal of molecular [Internet]. 1990; Available from: https://www.sciencedirect.com/science/article/pii/S0022283605803602

51. Wood MA, Paralkar M, Paralkar MP, Nguyen A, Struck AJ, Ellrott K, et al. Population-level distribution and putative immunogenicity of cancer neoepitopes. BMC Cancer. 2018 Apr

13;18(1):414.

52. Tatlow PJ, Piccolo SR. A cloud-based workflow to quantify transcript-expression levels in public cancer compendia. Sci Rep. 2016 Dec 16;6:39259.

53. Lift Genome Annotations [Internet]. [cited 2019 May 13]. Available from: https://genome.ucsc.edu/cgi-bin/hgLiftOver

54. Broad GDAC Firehose [Internet]. [cited 2019 May 17]. Available from: http://gdac.broadinstitute.org/

55. Marty R, Kaabinejadian S, Rossell D, Slifker MJ, van de Haar J, Engin HB, et al. MHC-I Genotype Restricts the Oncogenic Mutational Landscape. Cell. 2017 Nov 30;171(6):1272–83.e15.

56. Ewing AD, Houlahan KE, Hu Y, Ellrott K, Caloian C, Yamaguchi TN, et al. Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. Nat Methods. 2015 Jul;12(7):623–30.

57. Chowell D, Morris LGT, Grigg CM, Weber JK, Samstein RM, Makarov V, et al. Patient HLA class I genotype influences cancer response to checkpoint blockade immunotherapy [Internet]. Vol. 359, Science. 2018. p. 582–7. Available from: http://dx.doi.org/10.1126/science.aao4572

58. Chalmers ZR, Connelly CF, Fabrizio D, Gay L, Ali SM, Ennis R, et al. Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. Genome Med. 2017 Apr 19;9(1):34.

59. Lyu G-Y, Yeh Y-H, Yeh Y-C, Wang Y-C. Mutation load estimation model as a predictor of the response to cancer immunotherapy. NPJ Genom Med. 2018 Apr 30;3:12.

60. Nguyen A, Garner C, Reddy SK, Sanborn JZ, Benz SC, Seery TE, et al. Three-fold overestimation of tumor mutation burden using 248 gene panel versus whole exome. J Clin Orthod. 2018 May 20;36(15_suppl):12117–12117.

61. Tumor Mutational Burden (TMB) [Internet]. Friends of Cancer Research. 2018 [cited 2019 May 29]. Available from: https://www.focr.org/tmb

62. Auslander N, Zhang G, Lee JS, Frederick DT, Miao B, Moll T, et al. Publisher Correction: Robust prediction of response to immune checkpoint blockade therapy in metastatic melanoma. Nat Med. 2018 Dec;24(12):1942.

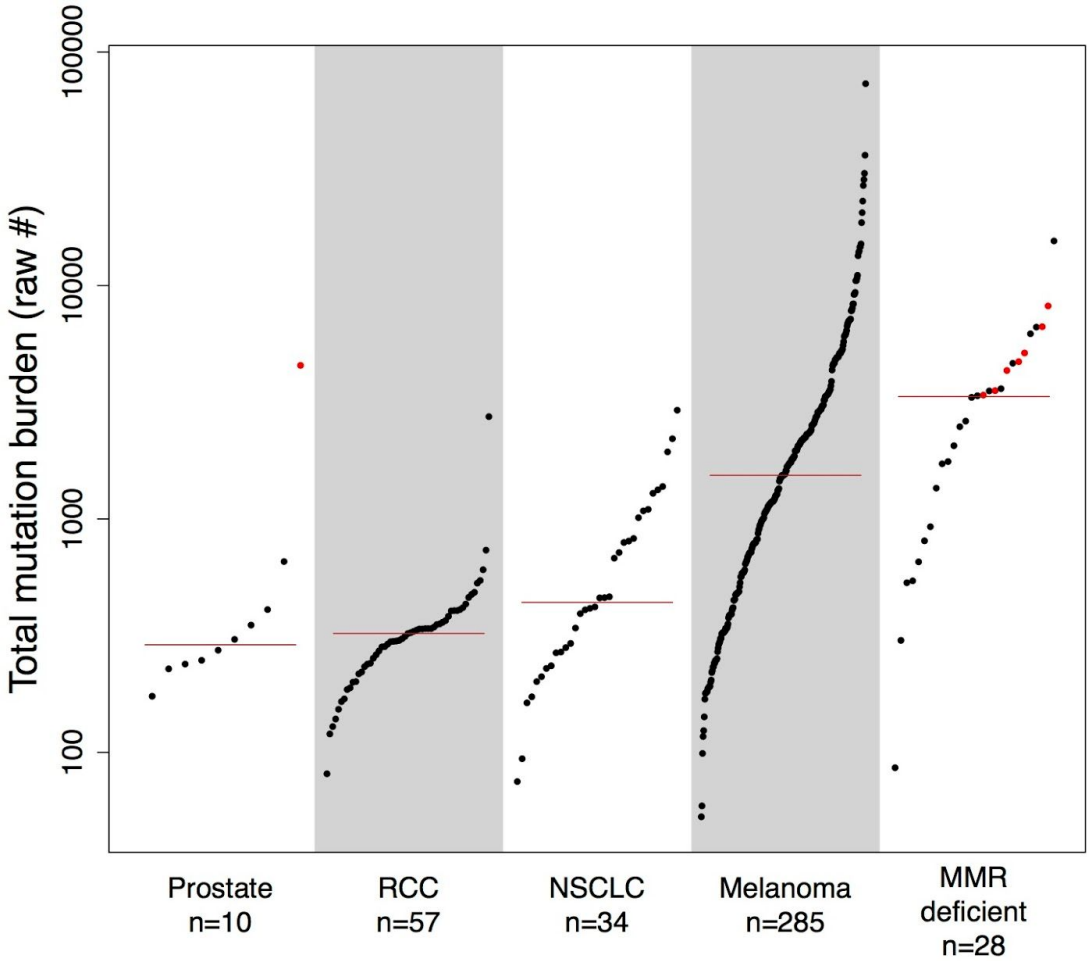
63. Weiss GJ, Beck J, Braun DP, Bornemann-Kolatzki K, Barilla H, Cubello R, et al. Tumor Cell-Free DNA Copy Number Instability Predicts Therapeutic Response to Immunotherapy. Clin Cancer Res. 2017 Sep 1;23(17):5074–81.

64. Goodman AM, Kato S, Bazhenova L, Patel SP, Frampton GM, Miller V, et al. Tumor Mutational Burden as an Independent Predictor of Response to Immunotherapy in Diverse Cancers. Mol Cancer Ther. 2017 Nov;16(11):2598–608.

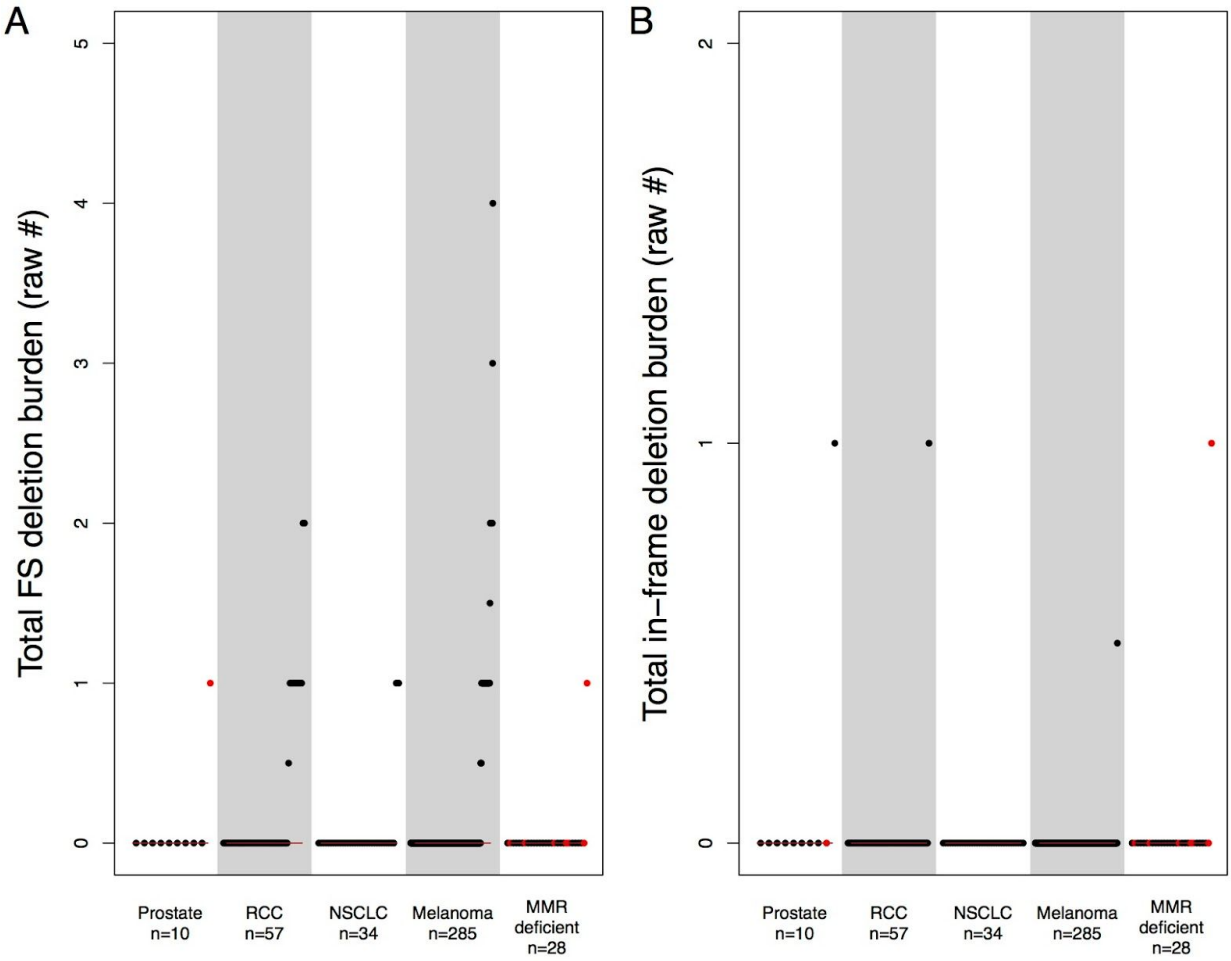
SUPPLEMENTARY DATA

Cancer type	Number of patients	Number of tumor samples	Reference
Melanoma	15	15	Amaria et al.
Melanoma	3	6	Carreno et al.*
Melanoma	16	16	Gao et al.
Prostate	10	10	Graff et al.
Melanoma	38 (27)	39 (28)	Hugo et al.*
Colon, endometrial, thyroid	28	29	Le et al.
RCC	57 (17)	58 (17)	Miao et al.
NSCLC	34	34	Rizvi et al.
Melanoma	35	53	Roh et al.
Melanoma	64 (20)	64 (20)	Snyder et al.*
Melanoma	110 (40)	110 (40)	Van Allen et al.*
Melanoma	4	6	Zaretsky et al.

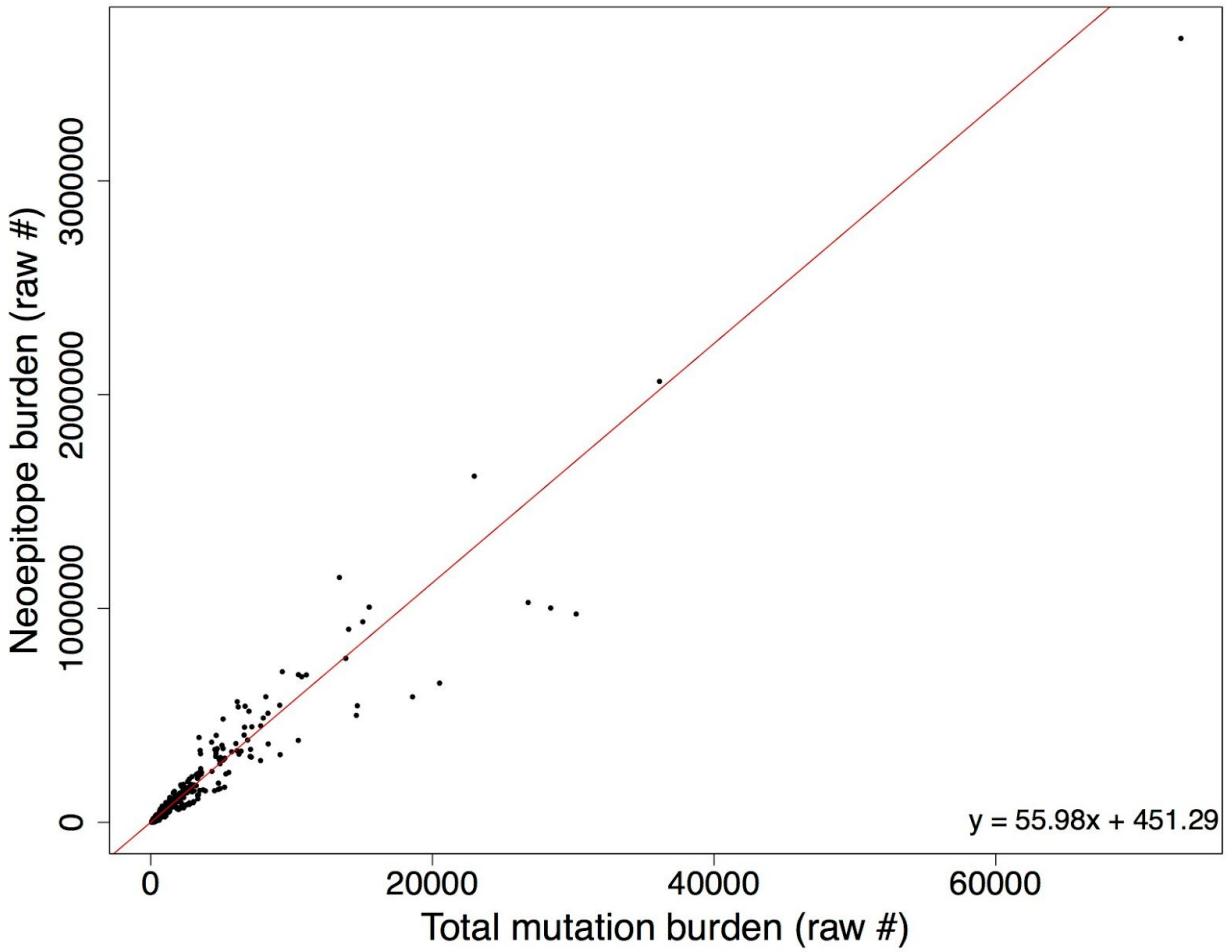
Supplementary Table 1: *Summary of patients samples used for analysis. Publicly available WES data from 12 studies was used to determine TMB (see Materials and Methods). We summarize the study which produced each data set, the cancer types represented, and the number of patients/tumor samples sequenced. Studies that had complementary RNA sequencing reads available for at least a subset of patients are indicated by an asterisk in the "Reference" column, and the number of samples with complementary RNA sequencing data are indicated in parentheses in the "Number of patients" and "Number of tumor samples" columns if different than the number of samples with WES.*



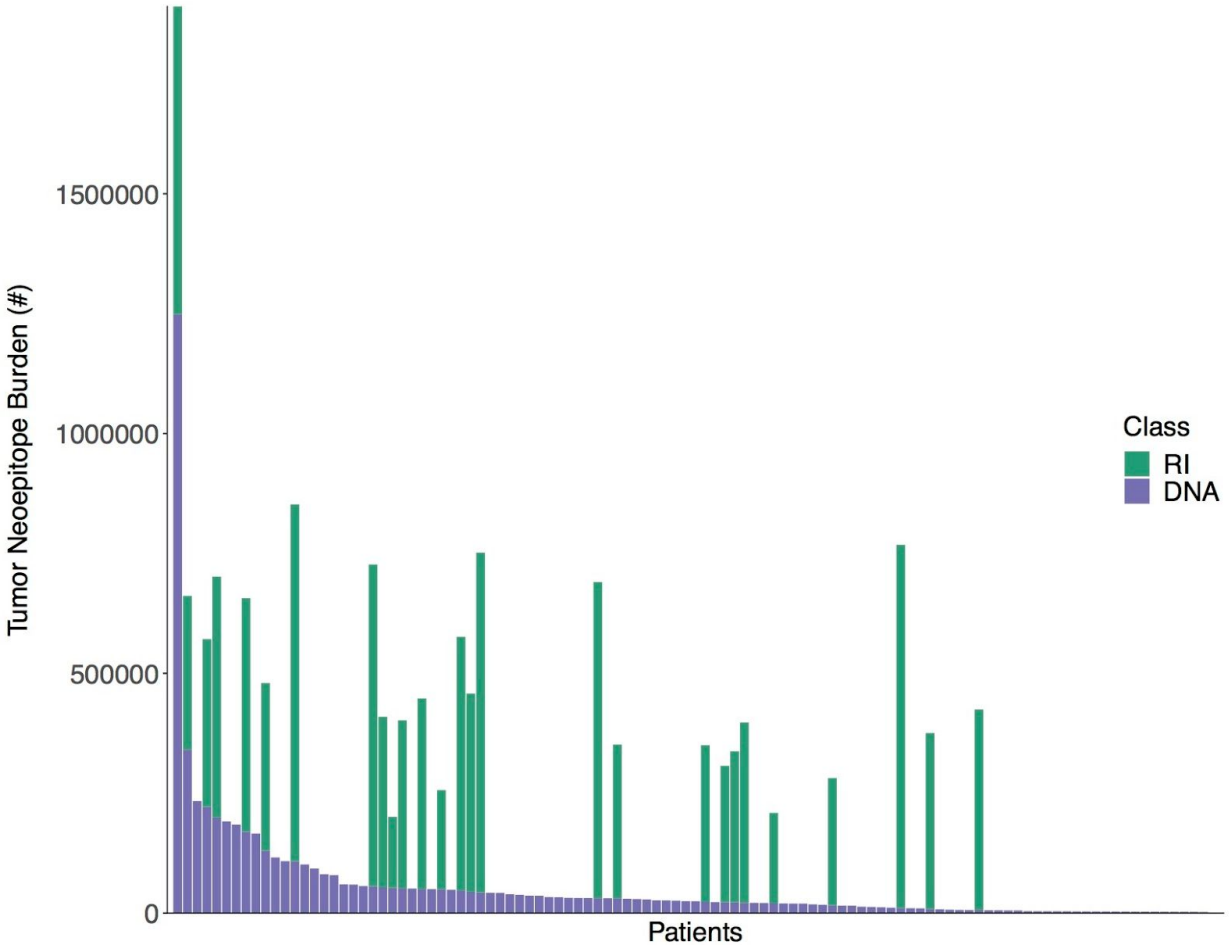
Supplementary Figure 1: *Per-patient distribution of raw mutation burdens across 7 cancer types. The raw number of somatic DNA variants per patient are shown along the y-axis, with each dot representing an individual cancer patient (cancer types shown along the x-axis). Note that MMR-deficient cancers here represent a cohort of 3 different cancer types including colon, endometrial, and thyroid. Red colored dots correspond to patients with microsatellite instability (see Methods). Abbreviations as follows: RCC=renal cell carcinoma, NSCLC=non-small cell lung cancer, MMR=mismatch repair.*



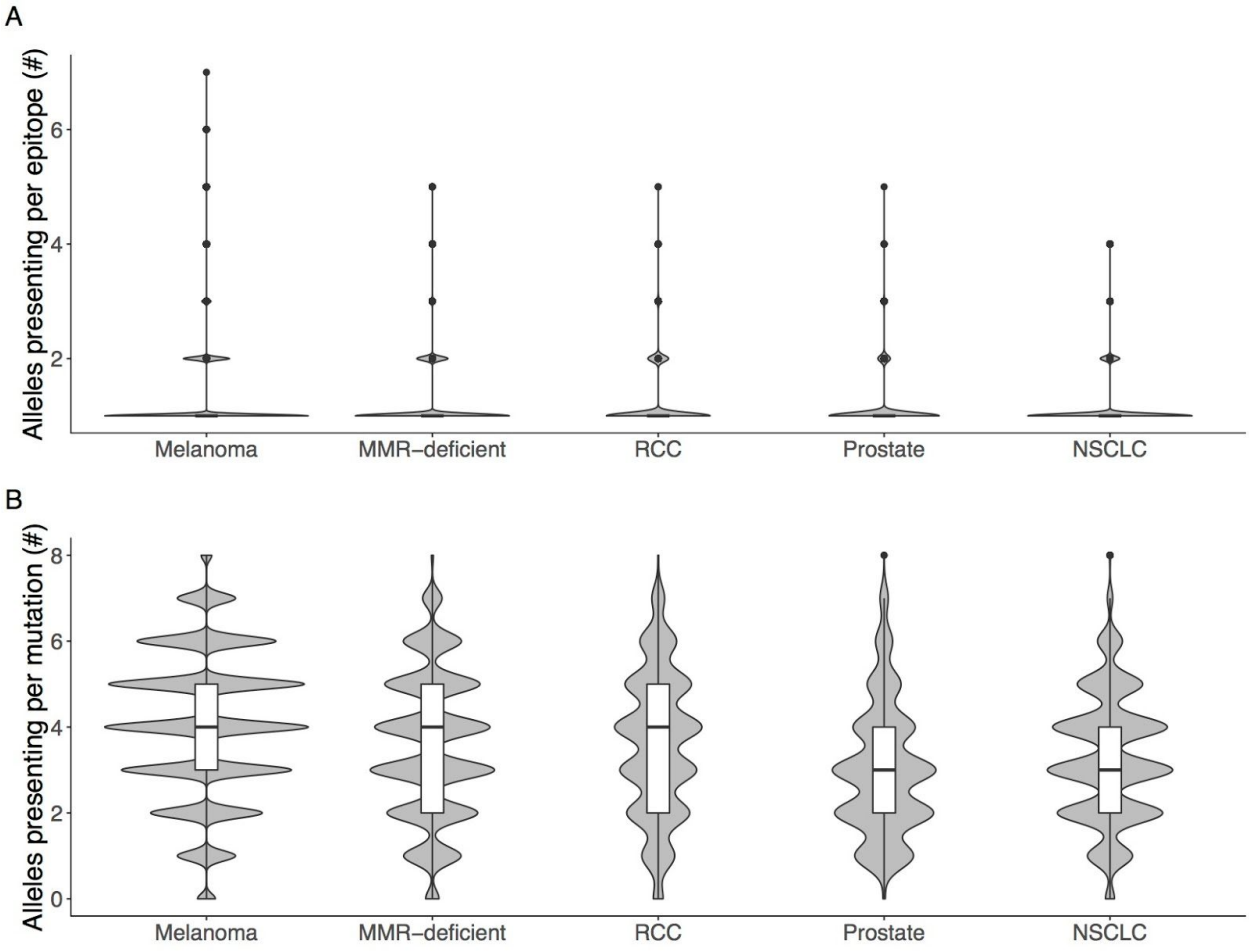
Supplementary Figure 2: *Per-patient distribution of deletion burdens across 7 cancer types. A) The number of somatic frameshift (FS) deletions per patient are shown along the y-axis, with each dot representing an individual cancer patient (cancer types shown along the x-axis). Note that MMR-deficient cancers here represent a cohort of 3 different cancer types including colon, endometrial, and thyroid. Red colored dots correspond to patients with microsatellite instability (see Methods). B) The number of somatic in-frame deletions per patient are shown along the y-axis, with each dot representing an individual cancer patient (cancer types shown along the x-axis). Abbreviations as follows: RCC=renal cell carcinoma, NSCLC=non-small cell lung cancer, MMR=mismatch repair.*



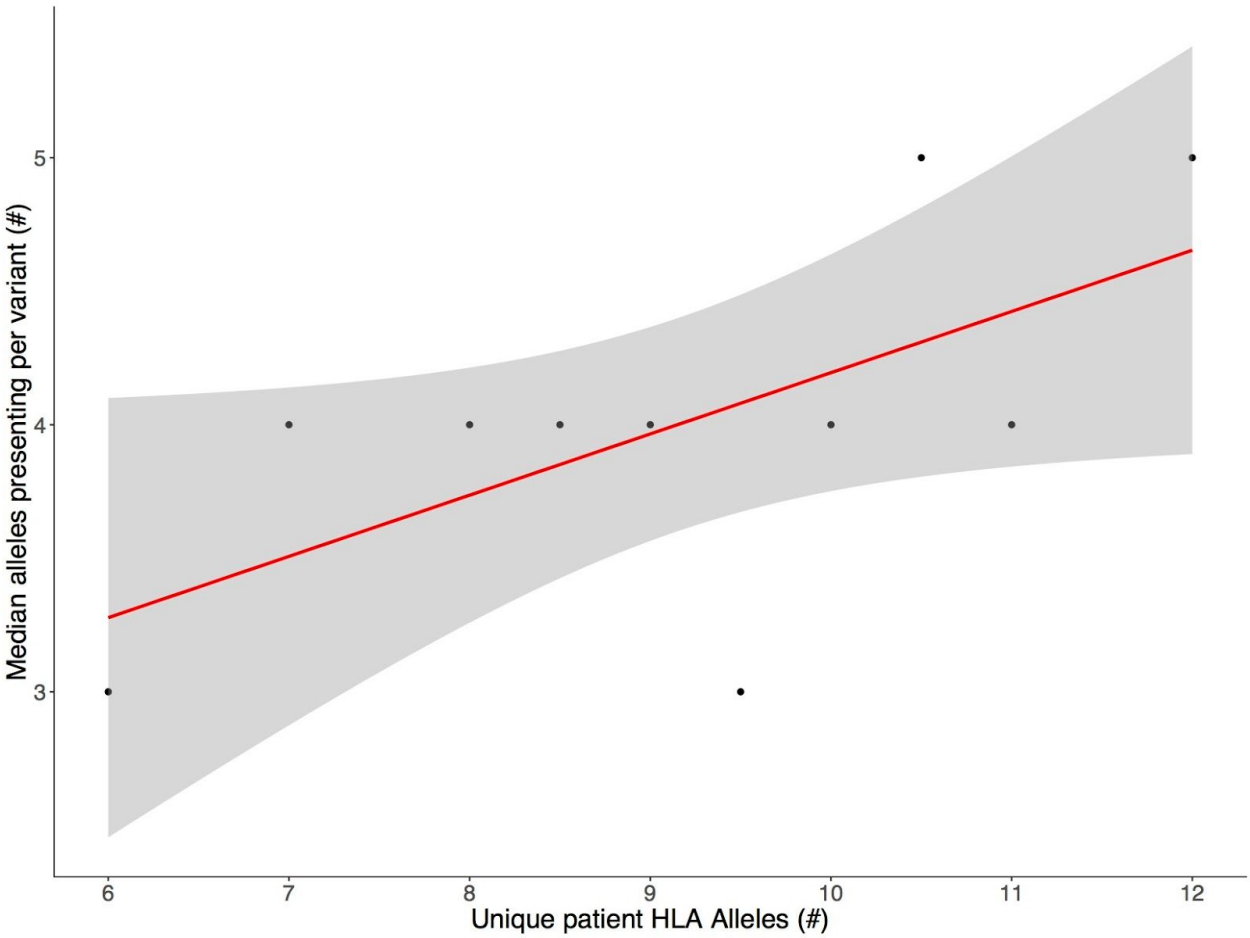
Supplementary Figure 3: *TMB correlates with neoepitope burden. Tumor mutational burden (x-axis) and neoepitope burden (y-axis) are strongly correlated (Pearson product-moment correlation of 0.96, $p < 2.2x10^{-16}$). The best fit line as determined by linear regression is shown in red, with its equation in the bottom right corner.*



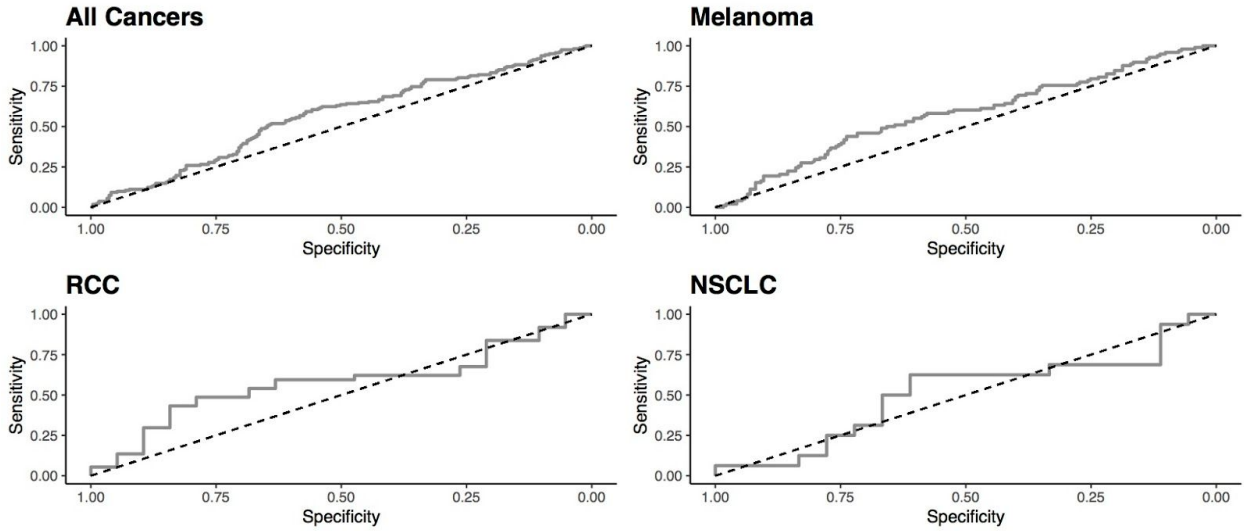
Supplementary Figure 4: *Per-patient distribution of overall tumor neoepitope burden and its components. The number of total tumor neoepitopes per patient is shown along the y-axis, with the numbers of neoepitopes derived from retained introns (RI) and somatic DNA variants (DNA) shown in green and purple, respectively. The data for each individual patient is displayed as stacked bars along the x-axis, sorted from left to right by the number of neoepitopes derived from somatic DNA variants (from highest to lowest).*



Supplementary Figure 5: *Robustness of putative neoepitope presentation among 5 different cancer groups. A) The number of unique patient-matched HLA alleles that are predicted to present an individual neoepitope is shown along the y-axis, with each violin plot distribution corresponding to a different cancer group along the x-axis, as labeled. B) The total number of unique patient-matched HLA alleles that are predicted to present one or more neoepitopes arising from a single DNA mutation is shown along the y-axis, with each violin plot distribution corresponding to a different cancer group along the x-axis, as labeled. Note that the width of each violin plot at each point along the y-axis corresponds to the relative quantity of data points in that group for that value of the y-axis. Furthermore, the lower and upper borders of the box within each violin plot corresponds to the 25th and 75th percent quantiles of the dataset for that group, respectively, with the median value shown as a horizontal black line within the box. Note that a predicted HLA binding affinity threshold of ≤500nM was used in all cases (see Methods).*



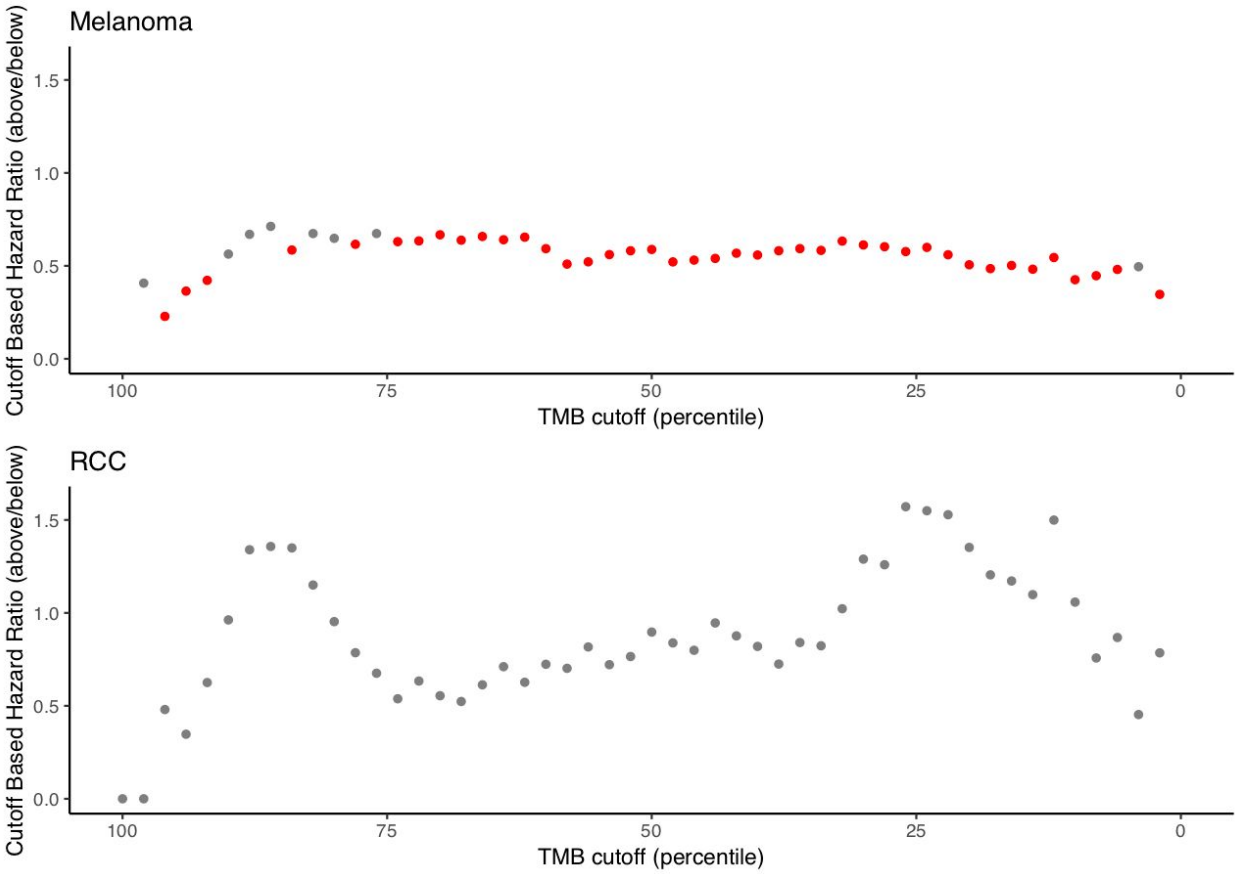
Supplementary Figure 6: *Robustness of putative neoepitope presentation. The median number of unique patient-matched HLA alleles that are predicted to present one or more neoepitopes arising from a single DNA mutation is shown along the y-axis, with the x-axis corresponding to patient-specific HLA heterozygosity (as the number of unique MHC I and II alleles per patient). Red curve denotes the best fit line based on linear regression, with surrounding gray shading denoting the 95% confidence interval. Note that a predicted HLA binding affinity threshold of ≤500nM was used in all cases (see Methods).*



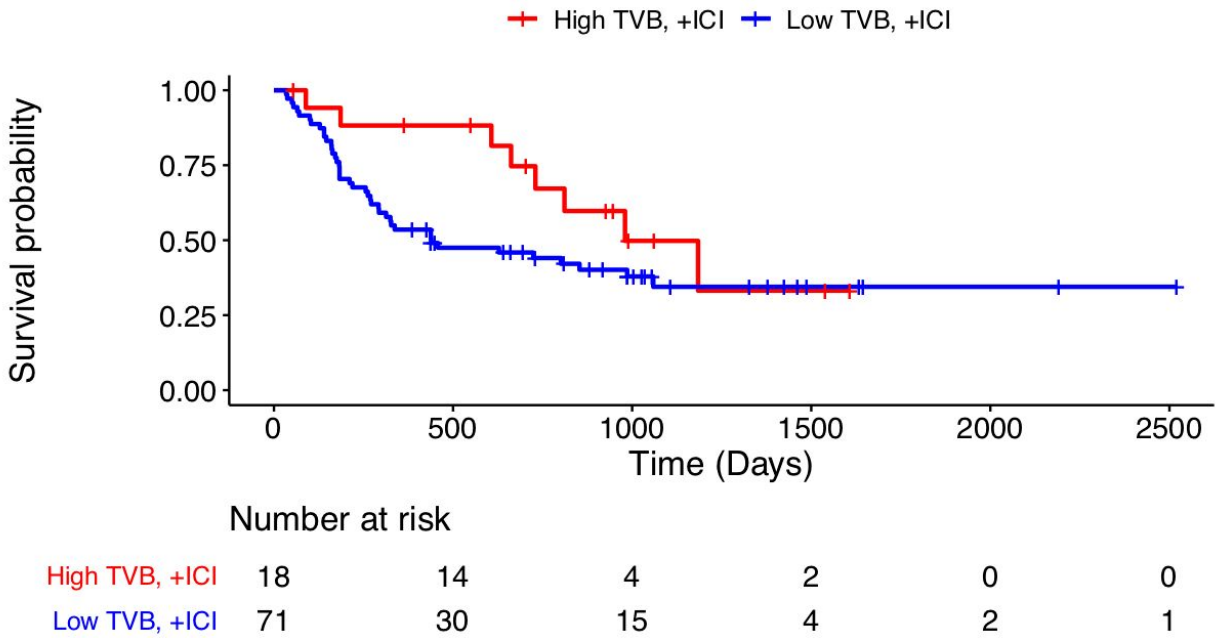
AUC by Coverage and Cancer Type

Cancer type	N	Coverage
All	409	0.567
Melanoma	285	0.578
RCC	56	0.578
NSCLC	34	0.510

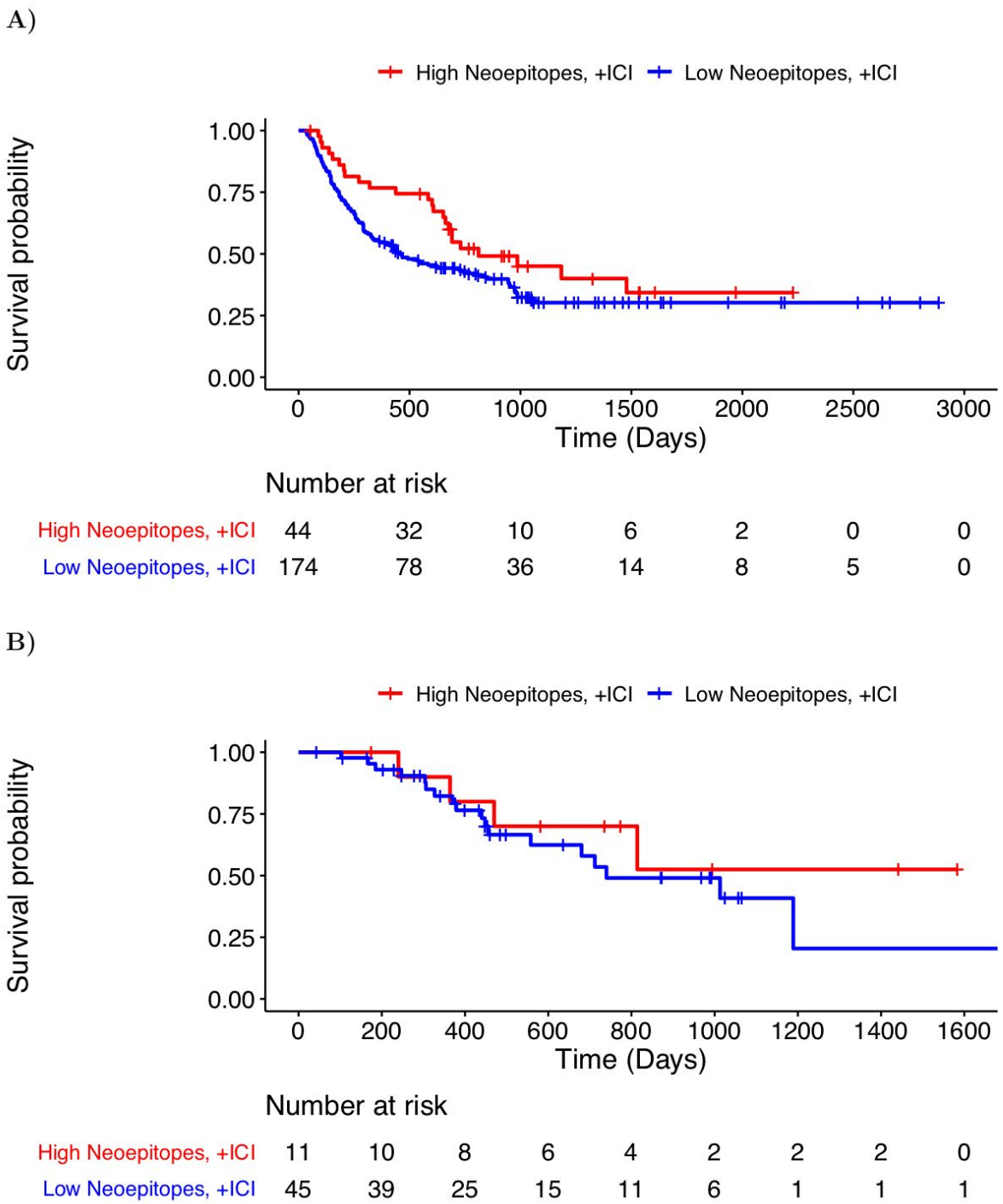
Supplementary Figure 7: *Receiver operating characteristic curves of predictive capacity of Mbp og genomic coverage. The upper panels depict the true positive rate (sensitivity, y-axis) and false positive rate (1-specificity, x-axis) for genomic coverage across all probability thresholds. The table in the lower panel reports the area-under-the-curve (AUC) for coverage (right column) applied to a different cancer cohort (rows). RCC=renal cell carcinoma, NSCLC=non-small cell lung cancer.*



Supplementary Figure 8: *Variation in estimated hazard ratio based on TMB threshold selection. For melanoma and RCC separately, cox proportional hazard models were fit comparing patients above and below each TMB percentile cutoff at 2% intervals. The relative hazard ratio for those above the threshold compared to those below the threshold was plotted, with red representing models with corresponding unadjusted p-values < 0.05.*

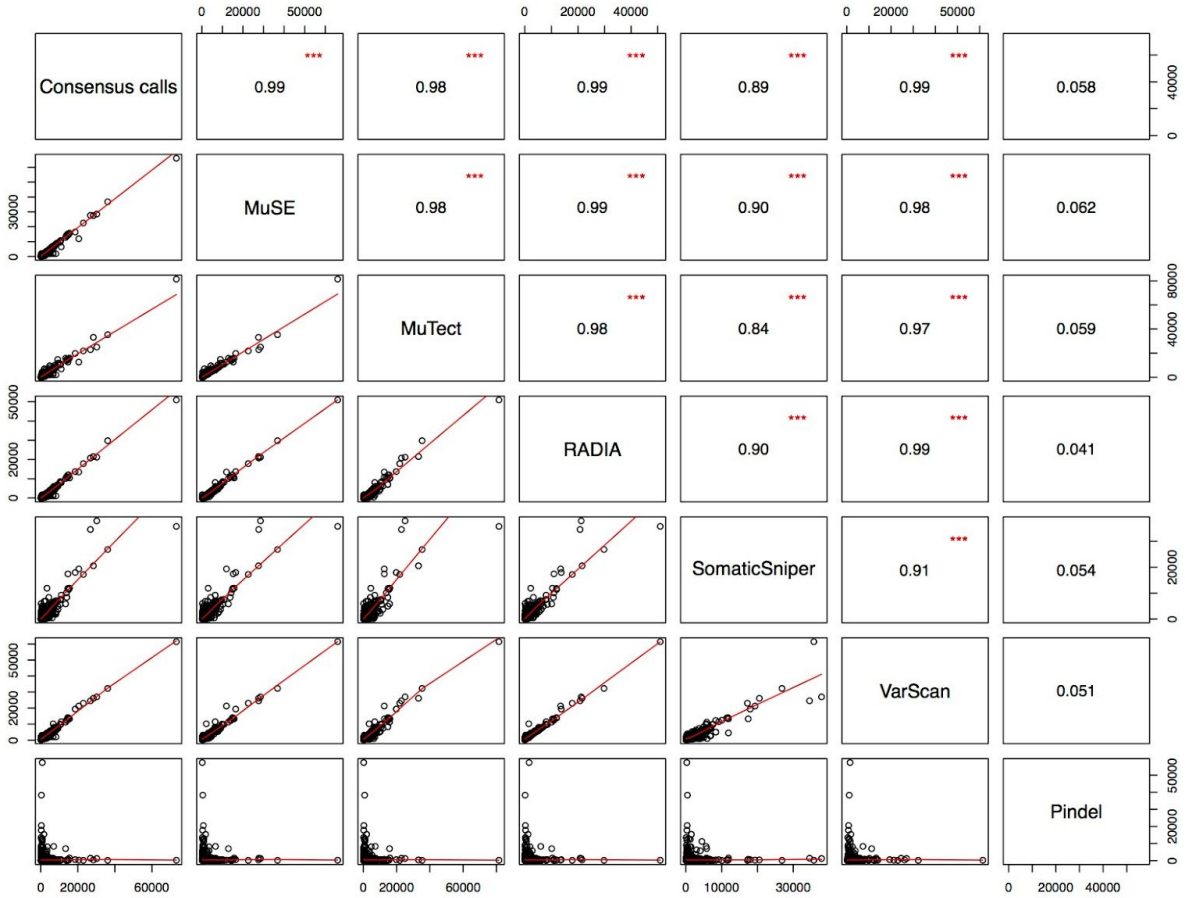


Supplementary Figure 9: *Overall survival among melanoma patients with high and low tumor variant burden (TVB). Kaplan-meier curves for the immunotherapy-treated patients with high TVB (≥80th percentile) and TVB burden (<80th percentile) are shown in red and blue, respectively. The underlying table corresponds to the number of patients at risk of death at each timepoint.*

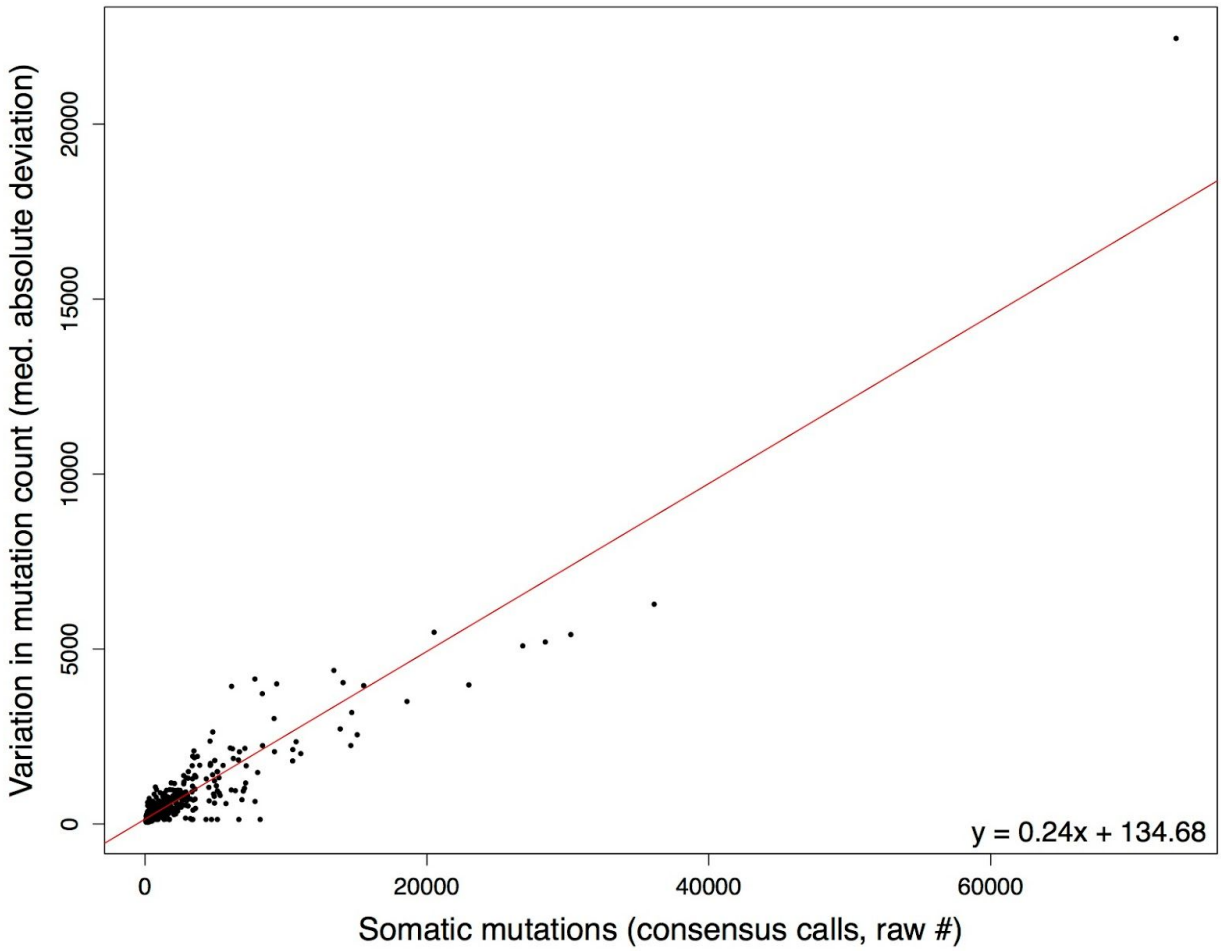


Supplementary Figure 10: *Overall survival among melanoma and renal cell carcinoma patients with high and low neoepitope burden. A) Overall survival among melanoma patients with high and low neoepitope burden. Kaplan-meier curves for the immunotherapy-treated patients with high neoepitope burden (≥80th percentile) and low neoepitope burden (<80th percentile) are shown in red and blue, respectively. The underlying table corresponds to the number of patients at risk of death at each timepoint. B) Overall survival among metastatic renal cell carcinoma patients with high and low neoepitope burdens. Kaplan-meier curves for the*

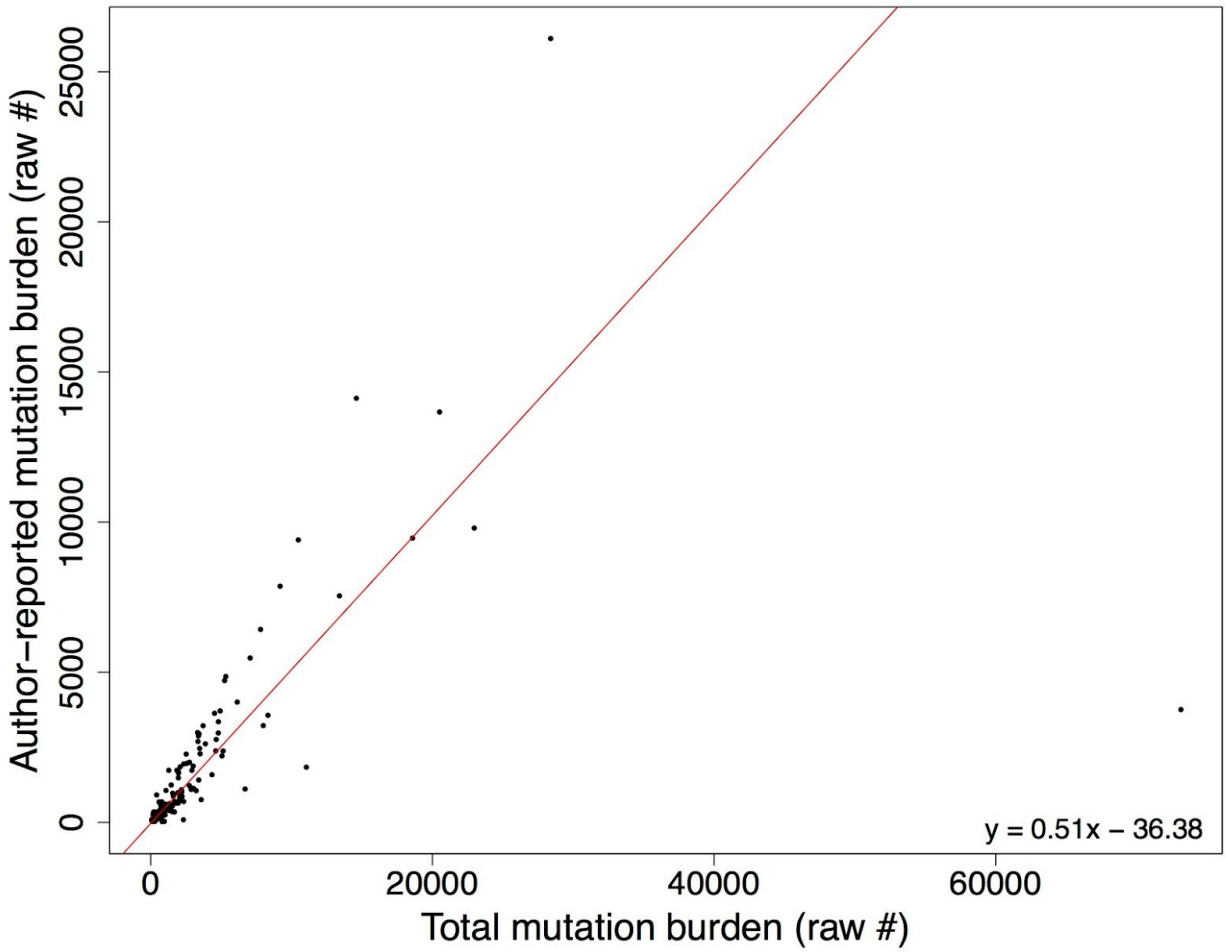
immunotherapy-treated patients with high neoepitope burden (≥80th percentile) and low neoepitope burden (<80th percentile) are shown in red and blue, respectively. The underlying table corresponds to the number of patients at risk of death at each timepoint.



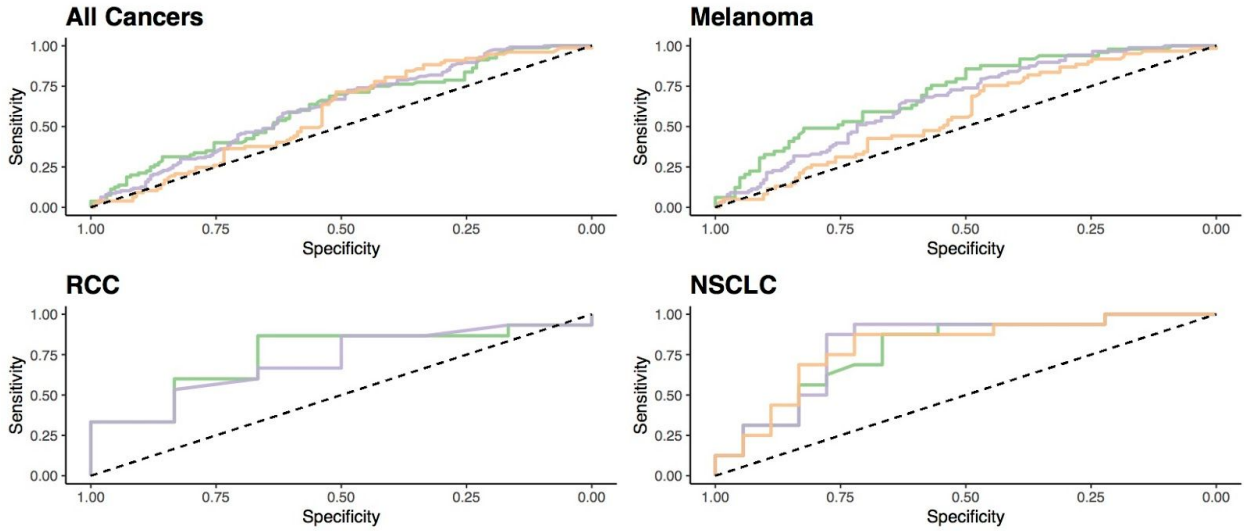
Supplementary Figure 11: *Pairwise differences in normalized total mutation burden as determined by 7 different computational approaches (see Methods). Each computational approach is identified along the diagonal panels, while the values in the upper panels denote the Pearson correlation coefficients between every pairwise combination of computational approaches (identified by corresponding row and column). The three red asterisks denote significant correlation at the p < 0.001 level. The scatterplots in the lower panels denote the TMB as calculated by each pairwise combination of computational approaches, with the x- and y-axes corresponding to the TMB calculated by the approach identified by the corresponding column and row, respectively; each open circle represents a single patient datapoint. Note that the red lines correspond to the best fit linear model.*



Supplementary Figure 12: *Variation in somatic mutation count increases with increased TMB from consensus variant calls. The median absolute deviation (MAD) in variant count across 6 variant calling tools used to determine consensus variant calls (y-axis, see Methods) increases with increasing TMB as determined by consensus calling (x-axis). The best fit line as determined by linear regression is shown in red, with its equation in the bottom right corner.*



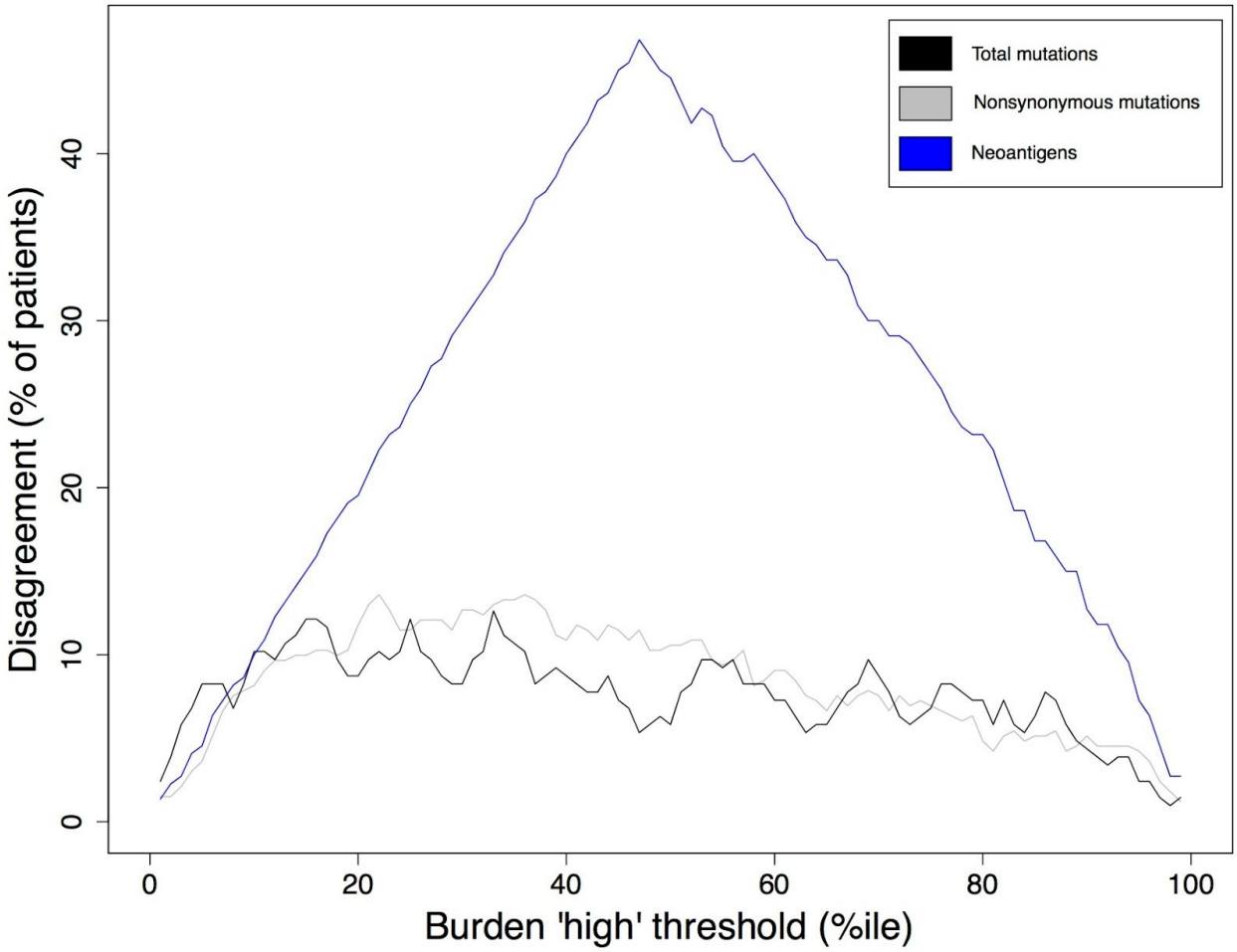
Supplementary Figure 13: *Author-reported total mutation burden correlates with consensus TMB. The total mutational burden as described by the authors of the original manuscripts from which the cohort derives (y-axis) correlates with our TMB derived from consensus variant calling (x-axis, Pearson product-moment correlation of 0.61, $p < 2.2x10^{-16}$). The best fit line as determined by linear regression is shown in red, with its equation in the bottom right corner.*



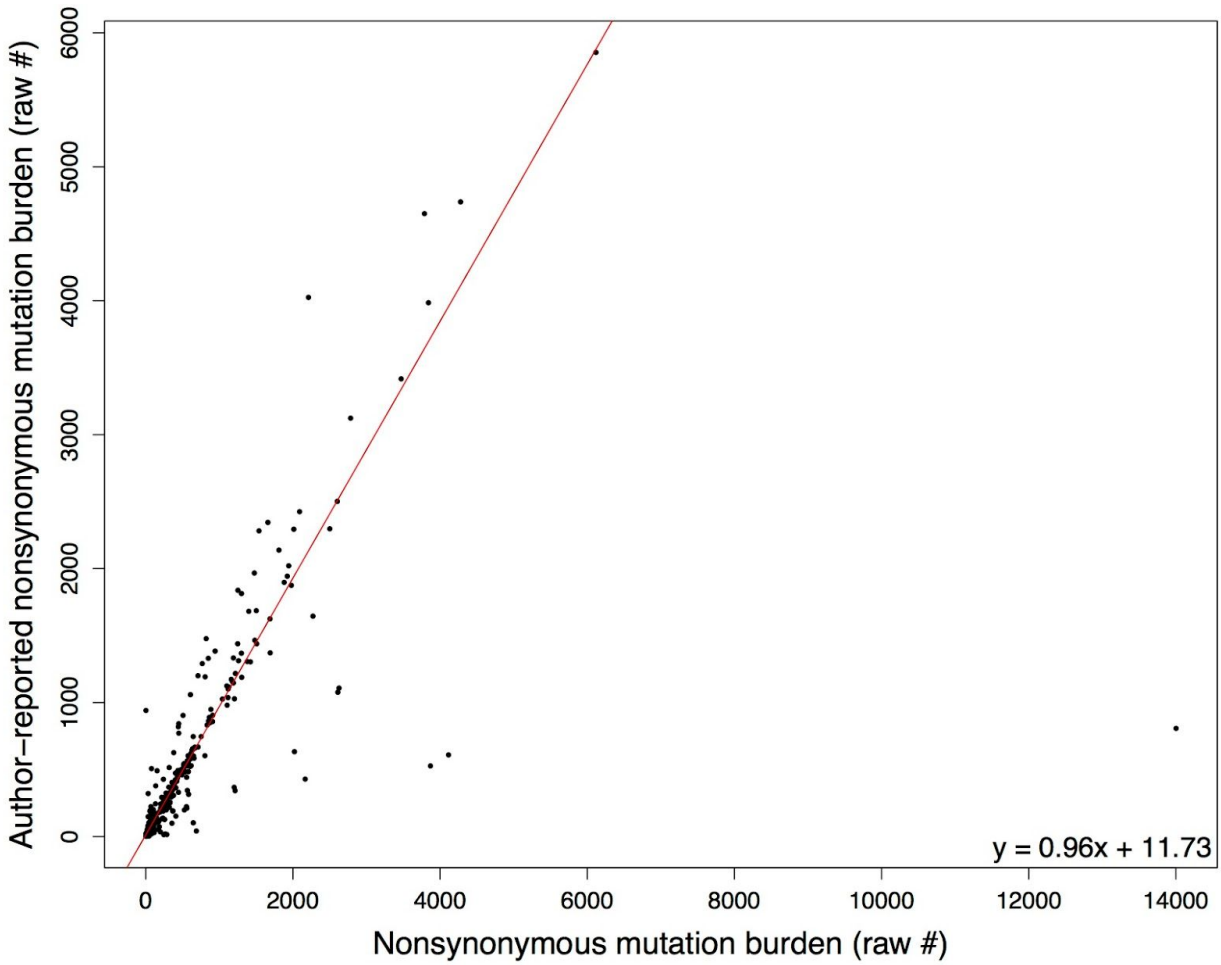
AUC by Author–reported Burden and Cancer Type

Cancer type	Total mutations	Nonsynonymous mutations	Neoantigens
All	0.620	**0.621**	0.587
Melanoma	**0.714**	0.667	0.582
RCC	**0.744**	0.711	NA
NSCLC	0.780	**0.809**	0.799

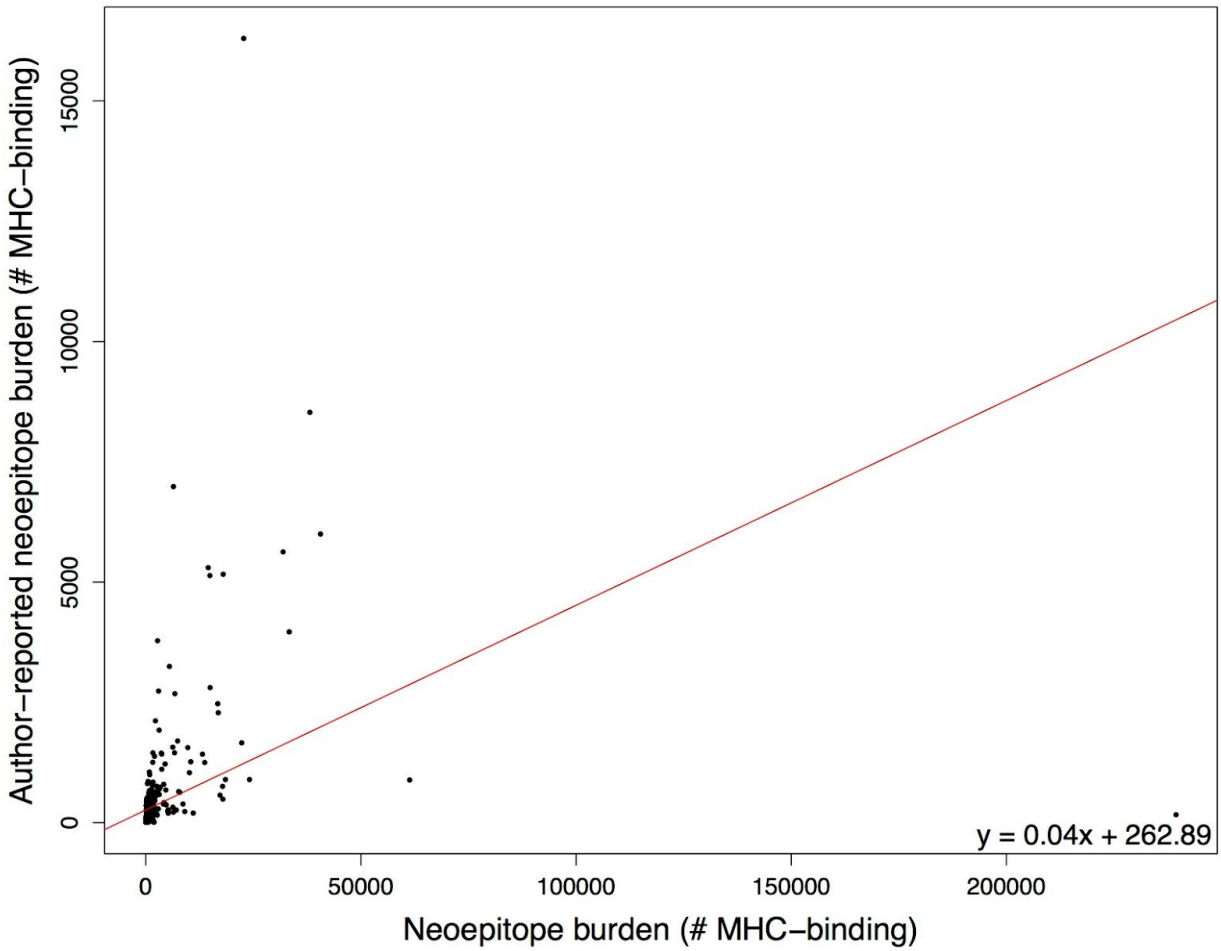
Supplementary Figure 14: *Receiver operating characteristic curves of predictive capacity of author-reported mutation and neoepitope burdens. The upper panels depict the true positive rate (sensitivity, y-axis) and false positive rate (1-specificity, x-axis) for each method across all probability thresholds. The table in the lower panel reports the area-under-the-curve (AUC) for each method (columns) applied to a different cancer cohort (rows), with colors above the methods indicating the color of the corresponding curve in the upper panels. Bold-faced values indicate the best value for each cancer cohort. RCC=renal cell carcinoma, NSCLC=non-small cell lung cancer.*



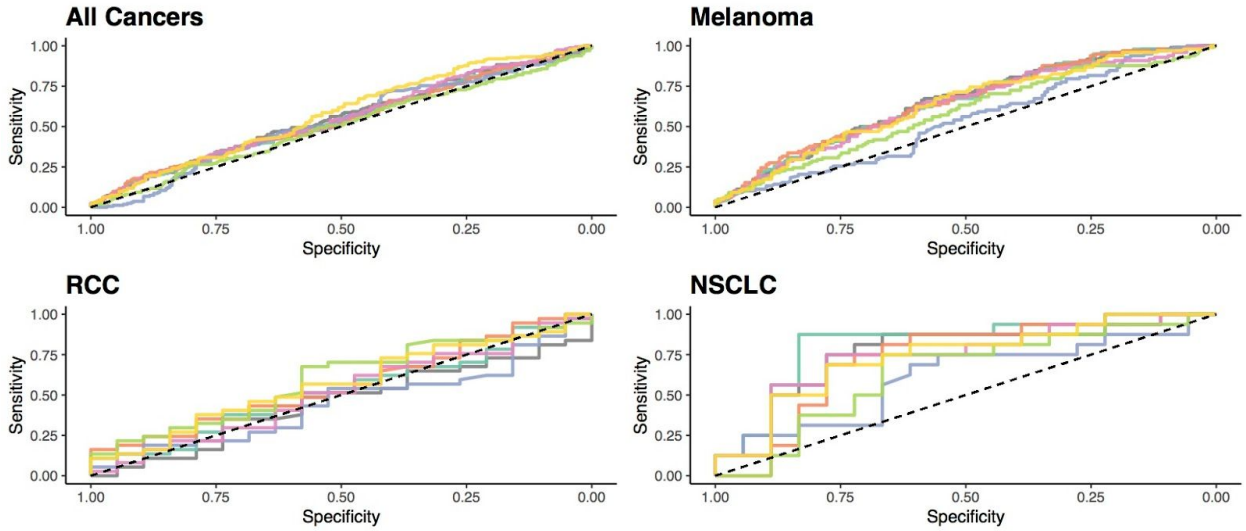
Supplementary Figure 15: *Cohort-level disagreement in classification of individual patients as TMB or neoepitope burden "high" v. "low". TMB and neoepitope burdens were calculated using a standardized consensus approach (see Methods) and were compared with author-reported values from the original cohort source studies. The overall disagreement between classifications of consensus and author-reported data (y-axis) was calculated using different percentile thresholds (x-axis) to classify each individual as e.g. TMB "high" or "low". This process was repeated for all mutations (black line), nonsynonymous mutations (gray line), and putative neoantigens (blue line).*



Supplementary Figure 16: *Author-reported nonsynonymous mutation burden correlates with nonsynoymous variants from consensus calling. The nonsynonymous mutational burden as described by the authors of the original manuscripts from which the cohort derives (y-axis) correlates with our consensus variant calling-derived nonsynonymous mutation burden (x-axis, Pearson product-moment correlation of 0.65, $p < 2.2x10^{-16}$). The best fit line as determined by linear regression is shown in red, with its equation in the bottom right corner.*



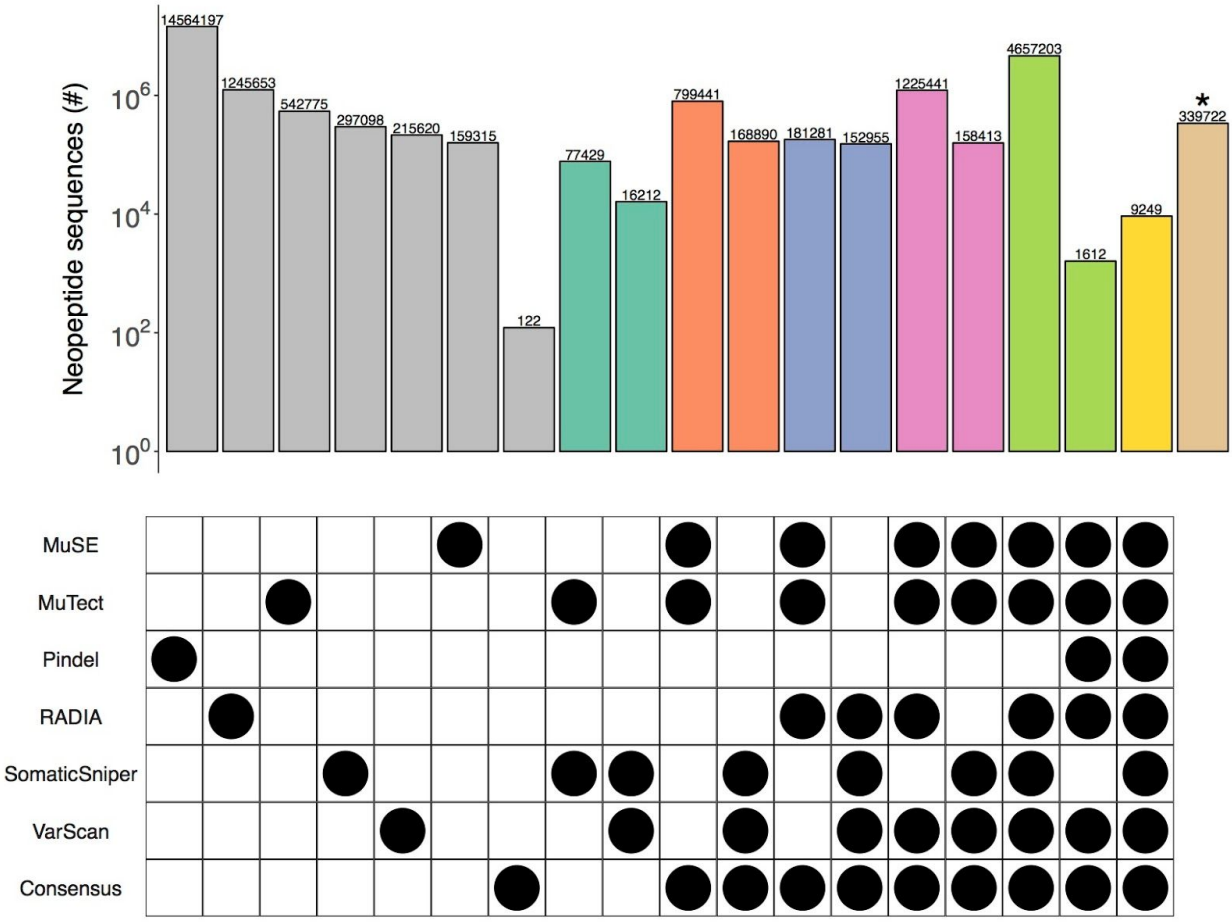
Supplementary Figure 17: *Author-reported neoepitope burden correlates with neoepitopes derived from variants from consensus calling. The neoepitope burden as described by the authors of the original manuscripts from which the cohort derives (y-axis) correlates with our consensus variant calling-derived neoepitope burden (x-axis, Pearson product-moment correlation of 0.65, $p < 2.2x10^{-16}$). The best fit line as determined by linear regression is shown in red, with its equation in the bottom right corner.*



AUC by Mutation Burden and Cancer Type

Cancer type	N	Consensus	MuSE	MuTect	Pindel	RADIA	SomaticSniper	VarScan
All	409	0.556	0.544	0.542	0.529	0.553	0.504	**0.581**
Melanoma	285	0.652	0.648	**0.654**	0.536	0.633	0.584	0.637
RCC	56	0.454	0.522	0.561	0.454	0.515	**0.596**	0.569
NSCLC	34	0.771	**0.816**	0.726	0.613	0.743	0.642	0.740

Supplementary Figure 18: *Receiver operating characteristic curves of predictive capacity of TMB from 7 different variant calling methods. The upper panels depict the true positive rate (sensitivity, y-axis) and false positive rate (1-specificity, x-axis) for each method across all probability thresholds. The table in the lower panel reports the area-under-the-curve (AUC) for each method (columns) applied to a different cancer cohort (rows), with colors above the methods indicating the color of the corresponding curve in the upper panels. TMB as determined by consensus calling (see Methods) is compared to the individual variant calling tools used in consensus calling. Bold-faced values indicate the best value for each cancer cohort. RCC=renal cell carcinoma, NSCLC=non-small cell lung cancer.*



Supplementary Figure 19: *Detailed comparison of the complete set of neopeptide sequences predictions from MuSE, Mutect, Pindel, RADIA, SomaticSniper, VarScan, and consensus variant calling. Patterns of agreement or disagreement among groups of neopeptide sequences predicted from variants derived from different combinations of tools across all patients are shown along each column, and each row indicates the neopeptide predictions associated with variants from the indicated tool (e.g. the first column corresponds to neopeptides predicted only from Pindel variants). The number of neopeptides in each column (bar in upper pane) corresponds to the size of the subset predicted for variants from the indicated combination of tools (black circles in the bottom panel). Columns with gray bars represent neopeptides predicted from variants derived from only a single tool while columns with teal, orange, blue, pink, or green bars represent neopeptides predicted from variants derived from the most common two combinations of 2, 3, 4, 5, or 6 variant calling tools. The column with the yellow bar represents neopeptides predicted from variants deriving from all tools. The column with the brown bar (indicated by an asterisk) represents variants derived from less common combinations of 2-6 variant calling tools.*