Accounting for proximal variants improves neoantigen prediction

Jasreet Hundal1, Susanna Kiwala1, Yang-Yang Feng1, Connor J. Liu1, Ramaswamy Govindan2,3, William C. Chapman4, Ravindra Uppaluri5, S. Joshua Swamidass6, Obi L. Griffith1,2,3,7*, Elaine R. Mardis8* and Malachi Griffith1,2,3,7*

Recent efforts to design personalized cancer immunotherapies use predicted neoantigens, but most neoantigen prediction strategies do not consider proximal (nearby) variants that alter the peptide sequence and may influence neoantigen binding. We evaluated somatic variants from 430 tumors to understand how proximal somatic and germline alterations change the neoantigenic peptide sequence and also affect neoantigen binding predictions. On average, 241 missense somatic variants were analyzed per sample. Of these somatic variants, 5% had one or more in-phase missense proximal variants. Without incorporating proximal variant correction for major histocompatibility complex class I neoantigen peptides, the overall false discovery rate (incorrect neoantigens predicted) and the false negative rate (strong-binding neoantigens missed) across peptides of lengths 8–11 were estimated as 0.069 (6.9%) and 0.026 (2.6%), respectively.

Over the past two decades, approaches to identify and screen for antigens, both self and non-self, have evolved rapidly1–2. This is due in part to advances in sequencing technologies, in the accuracy of algorithmic identification of somatic variants, and in computational modeling to predict the binding affinity of the resulting novel, tumor-specific peptides to major histocompatibility complex (MHC) molecules3. Thus, current immunogenomic approaches can identify somatic variants that give rise to tumor-specific mutant antigens, or ‘neoantigens’, and evaluate their ability to bind to MHC class I and class II molecules4.

Typically, to evaluate strong-binding neoantigens from genomic sequencing data, the raw sequencing reads from tumor and normal DNA libraries are aligned to the human reference genome, and somatic variants are identified by comparison of tumor to normal read alignments. The resulting somatic variants of interest (SVOIs) are then annotated to predict protein sequence changes and to infer possible neoantigenic peptides. Individual neoantigenic peptides are selected by sliding an amino acid window (usually 8–11-mers) across the variant position to consider each possible ‘register’. These peptides are assessed using various algorithms to predict binding affinity to MHC and determine the strongest-binding epitopes. These predicted neoantigenic peptides are prioritized as we have previously described1. The cancer vaccine design process, from read alignment to variant calling and neoantigen prediction, typically assumes the reference genome sequence surrounding each somatic variant is representative of the patient’s genome sequence.

However, any sequence variant proximal to an SVOI in the patient’s genome that differs from the human reference may alter the amino acid sequence of the resulting peptide (here we define proximal as ‘situated close to’ or ‘nearby’, rather than using the classic genetics definition of ‘closer to the centromere’). Existing pipelines that are used for computational prediction of neoantigens from sequencing data, such as MuPeXI and pVAC-Seq4, do not explicitly incorporate patient-specific nearby germline or somatic variants (collectively described hereafter as ‘proximal variants’) into the peptide sequence considered in neoantigen prediction. Some pipelines, such as Vaxrank5, infer the coding sequence from assembly of tumor RNA reads, thus accounting implicitly for both somatic and germline variants, but this is largely dependent on the availability of RNA sequencing data. Failing to account for patient-specific proximal variants could affect the efficacy of a vaccine, possibly resulting in immunization with incorrect peptides or failure to identify highly neoantigenic peptides.

To investigate these possibilities, we identified somatic and germline variants proximal to SVOIs in a dataset of tumor sequencing studies representing different tissue sites and mutational loads. For this analysis, given that the upper bound for the length of MHC-binding peptides (accounting for both class I and class II) is typically considered to be 30 amino acids6–8, we chose a nucleotide window of 89 base pairs (bp) upstream and downstream of each SVOI in which to identify relevant proximal variants (Online Methods). We limited our analysis to include only missense proximal variants and SVOIs. We then incorporated these proximal variants in the final peptide sequences (proximal variant correction (PVC)) and re-evaluated the resulting peptide set using our neoantigen prediction pipeline (pVAC-Seq)4. Our results suggest that taking individual proximal variation into account can have a significant effect on the accuracy of neoantigen selection, resulting in a more personalized vaccine design.

Results

To determine how frequently proximal variants occur within the vicinity of an SVOI, we assessed 430 tumors with varying mutational

1McDonnell Genome Institute, Washington University School of Medicine, St. Louis, MO, USA. 2Siteman Cancer Center, Washington University School of Medicine, St. Louis, MO, USA. 3Division of Oncology, Department of Internal Medicine, Washington University School of Medicine, St. Louis, MO, USA. 4Department of Surgery, Washington University School of Medicine, St. Louis, MO, USA. 5Department of Surgery/Otolaryngology, Brigham and Women’s Hospital and Dana-Farber Cancer Institute, Boston, MA, USA. 6Department of Pathology and Immunology, Washington University School of Medicine, St. Louis, MO, USA. 7Department of Genetics, Washington University School of Medicine, St. Louis, MO, USA. 8Institute for Genomic Medicine, Nationwide Children’s Hospital, Columbus, OH, USA. *e-mail: elaine.mardis@nationwidechildrens.org; mgriffit@wustl.edu
loads identified from whole-genome/exome sequencing data from matched normal and tumor tissue (Fig. 1; Online Methods). Specifically, data from 100 cases each of melanoma, hepatocellular carcinoma, and lung squamous cell carcinoma were obtained from The Cancer Genome Atlas (TCGA). We also evaluated data from 48 cases of HER2+ breast cancer, 34 cases of small cell lung cancer, 30 cases of hepatocellular carcinoma, 15 cases of oral squamous cell carcinoma, and 1 case of hypermutated glioblastoma (1 primary and 2 metastatic samples) from in-house studies. After performing alignment and variant calling, we confirmed the linkage of SVOIs and proximal (somatic or germline) variants by phasing the variants using the Genome Analysis Toolkit (GATK) (Fig. 1b; Online Methods). Then, the list of SVOIs from each of the samples was intersected with the respective lists of in-phase amino acid-altering proximal variants to assess their presence within the chosen nucleotide window.

**Missense variants overlap with missense proximal variants.** Of 430 tumor samples analyzed, 380 samples (88.3%) had at least one (range, 1–377) missense SVOI in phase with a proximal missense variant. Of a total of 103,673 missense variants identified in these tumors, there were 7,783 SVOIs (7.5%) with a proximal missense variant (somatic or germline) within 89 nucleotides on either side. Of these missense SVOIs, 5,344 (5.1%) were also in phase with their respective proximal variants. In most cases (93.8%), the SVOIs had a single proximal germline or somatic variant in phase, but occasionally multiple (range, 2–6) variants were proximal to the SVOI. An average of 241 missense somatic variants were analyzed per sample. Per subject, an average of 6.5% of SVOIs had a proximal missense variant, and 5% had one or more proximal missense variants in phase with the SVOI. On average, 62.2% of these proximal variants were germline missense variants, and 37.7% were somatic missense variants. The majority (68.0%) of proximal somatic variants were contributed by dinucleotide polymorphisms. Most variant callers (including those used for the harmonized analysis of TCGA data in the Genomic Data Commons) report dinucleotide polymorphisms as two separate single-nucleotide variants (SNVs). Excluding the dinucleotide polymorphisms, on average, 88.4% of the proximal variants were germline missense SNPs, and 11.6% were somatic missense SNVs. Supplementary Table 1 shows, for each sample, the percentage of SVOIs harboring any neighboring variants within the specified 89 bp window and the percentage of the total SVOIs that had any proximal variants in phase. It also shows the breakdown of numbers of somatic versus germline proximal variants for each sample, along with the numbers of variants contributed by dinucleotide polymorphisms.

**Predicted binding affinity changes with PVC.** To identify neoantigens capable of eliciting an effective antitumor T cell response, it is critical both to determine the correct tumor-specific peptide sequence and to assess its ability to bind to MHC. To do this, we quantified the impact of missing or incorrectly selected strong-binding neoantigens when proximal variants to assess their presence within the chosen nucleotide window.

In some rare cases, a proximal variant may translate to the same amino acid sequence as the SVOI; alternatively, the SVOI and proximal variant may both lead to amino acid changes if considered in isolation, but if they are in phase and considered together, they result in no change to the amino acid sequence. To take into account these cases and accurately assess the effect of amino acid changes due to proximal variants on binding predictions, we considered only those registers that contained both the proximal variant and the SVOI amino acid changes, when translated together. Across 8–11-mers, on average 45.95% of all neoantigen peptide registers contained both. Figure 2 summarizes the effect of proximal variants on neoantigen binding affinity. Although the effect is less pronounced for 8-mers, the smallest length we examined, we see drastic changes in binding affinity due to PVC across all four peptide lengths (represented as log$_{10}$ of mutant (MT) epitope fold change (MT$_{uncorrected}$/MT$_{corrected}$), with ranges spanning from −3.0 to 3.1 for 8–11-mers (Fig. 2a). Figure 2c,d shows the distribution of log$_{10}$[MT fold change] scores for 9-mer and 10-mer peptides, respectively. For both peptide lengths, most weak binders stay within the same range before and after PVC, but very few strong binders remain unchanged after PVC. We chose 500 nM as the binding affinity cut-off for a potential binder, as most known T cell epitopes have an affinity value of less than 500 nM. For the binding prediction changes, we considered a call as erroneous only if PVC yielded at least a 10% change in predicted binding affinity.

**Impact of PVC on false discovery rate (FDR) and false negative rate (FNR).** In addition to the effect a proximal amino acid substitution may have on a neoantigen’s binding potential, it is also important to consider whether the peptide sequence of the selected neoantigen is correct and representative of the sequence in the tumor. Failure to do so may affect the immunogenic potential of the neoantigen being selected, as the uncorrected neoantigen will not produce tumor-specific T cells even if it binds well and is presented by the MHC.

To determine how many neoantigens were being erroneously predicted, and the effect that mischaracterization of neoantigens due to proximal variants would have on candidate selection, we calculated the FNR and FDR after applying PVC. The FNR represents the probability of potential MHC binders (binding affinity <500 nM) being discarded (false negatives), and the FDR represents the probability of erroneous peptides being mistaken for potential binders (false positives).

An average of 9 SVOIs and 10 neoantigenic peptides were mischaracterized per case. As a consequence, 1,165 potential binders (MT$_{corrected}$ <500 nM) were erroneously rejected, and 3,305 peptides that were strong binders before PVC were misidentified across all 430 individuals investigated here. Overall, FNR and FDR across 8–11-mers were 0.026 and 0.069, respectively (Fig. 2b).

As a representative example, Supplementary Fig. 2 illustrates data from one of the TCGA melanoma samples with a heterozygous missense SNV in the reverse-strand gene MARCH10 that overlaps an in-phase heterozygous germline SNP 21 nucleotides upstream. When translated, this germline SNP results in an S357F (NP_001275708.1:p.Phe357Ser) alteration that is 7 amino acids downstream of the missense somatic variant F350S (NP_001275708.1:p.Ser350Phe). This variant directly affects the final neoantigen sequence for a peptide of any length (>8-mer). To evaluate the effect of this germline SNP on the binding affinity of the neoantigen peptide, we calculated the binding affinities of the uncorrected and the PVC neoantigenic peptides. The binding affinity of the best register for a 10-mer peptide using the uncorrected approach (MT$_{uncorrected}$ = 55.44 nM) is within the range for a good binder (<500 nM). However, after including this individual’s proximal germline variant, the binding affinity for the same register decreases almost 70-fold (MT$_{corrected}$ = 3,766.72 nM), thus predicting
Neoepitope binding prediction with proximal variant ‘corrected’ peptides

**Fig. 1 | Overview of the pipeline.** The steps required for incorporating and assessing the impact of proximal variants on neoantigen binding prediction are depicted as a flow diagram. There are three main steps. 

*Alignement and variant calling of matched tumor (pink) and normal (green) sequencing data.*

- **SVOI**
  - In-phase proximal SNV
  - Out-of-phase proximal SNV

- Scan 89 bp upstream and downstream of SVOI
- Find overlapping amino acid-altering variants (germline or somatic)

- Phase somatic and germline variants to determine linkage (GATK ReadBackedPhasing)

- Consider only phased proximal variants for downstream neoantigen prediction

**Neoepitope binding prediction with proximal variant ‘corrected’ peptides**

- **Uncorrected wild-type**
  - KLPEPCPS
  - 439.03 nM
- **Uncorrected mutant**
  - KLPEPCPSM
  - 119.19 nM

- **Corrected wild-type**
  - KLPEPCPS
  - 6,198.21 nM
- **Corrected mutant**
  - KLPEPCPSM
  - 2,485.05 nM

**a**
- 430 tumor–normal pairs
  - TCGA LUSC (100 cases)
  - TCGA MEL (100 cases)
  - TCGA HCC (100 cases)
  - Breast cancer (48 cases)
  - SCLC (34 cases)
  - HCC (30 cases)
  - OSCC (15 cases)
  - GBM (1 case; 3 mets)

**b**
- Exome sequencing (tumor and normal)

- Normal DNA
- Tumor DNA

**c**
- Phasing of proximal variants

**Neoantigen binding predictions**

- **Uncorrected wild-type**
  - AAG CTA CCA GAG CCA TGT CCT TCA ACG
  - KLPEPCPS
  - 439.03 nM
- **Uncorrected mutant**
  - AAG CTA CCA GAG CCA TGT CCT TCA A
  - KLPEPCPSM
  - 119.19 nM

- **Corrected wild-type**
  - AAG GTA CCA GAG CCA TGT CCT TCA A
  - KLPEPCPS
  - 6,198.21 nM
- **Corrected mutant**
  - AAG GTA CCA GAG CCA TGT CCT TCA A
  - KLPEPCPSM
  - 2,485.05 nM
a very weak binder. Using the uncorrected analysis approach, one might have selected this neoantigenic peptide for a vaccine, but after PVC, the candidate peptide is determined to be unsuitable. This result illustrates the importance of using the individual variation of the germline genome while selecting and designing neoantigens for personalized immunotherapy.

Discussion

There are some caveats and limitations to our approach. First, the analysis was restricted to single-nucleotide changes (that is, mis-sense somatic SNVs that are near another germline or somatic SNV) and did not seek to evaluate whether other, potentially relevant types of variants were found nearby. These include insertions and deletions (indels, both somatic and germline) as well as different types of structural variants that often have a more significant impact on peptide sequences but are rarer than SNVs. Phasing of indels and structural variants is also not currently handled by software such as GATK’s ReadBackedPhasing. Second, our analysis ‘window’ (89 bp) was defined in genomic coordinates. It is substantially more complicated to consider this window size in the context of transcriptome coordinates, since intronic coordinates must be ignored when scanning upstream and downstream. This is further complicated in genes with alternative transcripts and hence multiple combinations of introns and exons to consider. Our ability to determine phase for variants separated by an intron would be limited in whole-genome/exome sequencing data (although this could be evaluated in RNA sequencing data with sufficient read lengths). Lastly, for this study, we considered neoantigen binding predictions only to MHC class I molecules. MHC class II peptides are much longer due to an open binding groove, and the subsequent impact of proximal variants on the peptide sequence would therefore be even more pronounced. Due to these limitations, our results likely underestimate the impact of PVC.

Moreover, even with seemingly small FDR and FNR, the importance of accounting for the effect of proximal variants is clear when we consider clinical vaccine design scenarios. For example, ten or fewer peptides are usually selected for the final vaccine from a larger number of initial candidates. Given this scenario, we calculated the...
probability of choosing at least one weak binder or of omitting one strong binder in the final vaccine without PVC. For the first probability, we calculated \(1 - (1 - \text{FDR})^{10} = 0.513\), and for the second we calculated \(1 - (1 - \text{FNR})^{10} = 0.228\). The probability that at least one of these errors occurs for each individual evaluated is \(1 - (1 - \text{FDR})^{10} \times (1 - \text{FNR})^{10} = 0.624\). Thus, for neoantigen identification without PVC in 100 individuals, we can expect that approximately 51 individuals would receive a suboptimal vaccine specifically due to receiving a neoantigen with an incorrect peptide sequence, 23 would receive a suboptimal vaccine specifically due to missing a strong-binding neoantigen, and 62 would receive a suboptimal vaccine due to at least one of these causes.

Design of personalized cancer vaccines is complex, time-consuming, and expensive. Previous work has shown that only about 16–43% of the predicted neoantigenic peptides included in a vaccine formulation yield a CD8+ T cell response\(^{13-15}\). Our study demonstrates the importance of ensuring the selected neoantigens correctly represent the individual’s genome and therefore maximize the likelihood of eliciting an immune response. PVC based on the patient’s genome can eliminate errors during neoantigen candidate selection, potentially increasing the efficacy of personalized vaccines. Further studies may also demonstrate the importance of considering proximal variants when using neoantigen load to predict response to checkpoint blockade inhibition therapies.

URLs. Picard, https://broadinstitute.github.io/picard/; pVACtools, http://pvactools.org/.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41588-018-0283-9.

Received: 2 August 2017; Accepted: 19 October 2018; Published online: 3 December 2018

Acknowledgements

We are grateful to the research participants and their families, without whom this study would not be possible. We thank G. Dunn for early access to raw data for the published glioblastoma hypermutator case included in our analysis. We also thank R. Schreiber and R. Carreno for the initial discussions that inspired the study, and for their expertise and guidance during the study. R.G. was supported by the National Institutes of Health (NIH) National Cancer Institute (U01CA213844). S.J.S. was supported by the NIH National Library of Medicine (R01LM012222 and R01LM012482). O.L.G. was supported by the NIH National Human Genome Research Institute (R00HG007940) and the NIH National Cancer Institute (U01CA209936).

Author contributions

J.H. was involved in all aspects of this study, including designing and developing the methodology, analyzing and interpreting data, and writing the manuscript, with input from C.J.L., S.J.S., O.L.G., E.R.M., and M.G. S.K. was involved in development of neoantigen prediction software and participated in the data analysis and writing the manuscript. Y.-Y.F. contributed to data analysis, interpretation, and writing the manuscript. R.G., W.C.C., and R.U. provided unpublished tumor datasets and provided critical feedback on the manuscript; E.R.M. and M.G. supervised the study. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-018-0283-9. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to E.R.M. or M.G.

Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2018
Methods

Sequence data alignment and variant calling. To investigate the prevalence of proximal variants (germline SNPs or somatic variants), we analyzed publicly available sequencing data from TCGA as well as datasets generated in-house, together representing seven different tissue sites. These datasets were chosen to adequately represent tumors with low, medium, and high mutational burdens.

Analysis of in-house whole-genome/exome sequencing datasets was performed as previously described\(^{25,30}\). Briefly, raw sequencing reads from both the tumor and normal tissue were aligned to the human reference genome sequence (either GRCh37 or GRCh38) using the Burrows–Wheeler Aligner\(^{22}\), then merged and de-duplicated using Picard (see URLs). A combination of three or four different variant callers was used to identify somatic variants by comparison of tumor and normal variant calls: SAMtools\(^{21}\), Sniper\(^{1}\), Strelka\(^{22}\), and VarScan\(^{23}\). These variants were filtered as previously described\(^{23,25}\) and then manually reviewed using the Integrated Genomics Viewer according to the standard operating procedures\(^{29}\) to obtain a list of high-confidence variant calls. On average, 80% of the filtered variants passed manual review. Germline variant analysis of the normal samples was performed using SAMtools.

For the TCGA data, aligned tumor and normal binary alignment maps (BAM files) from the Burrows–Wheeler Aligner (version 0.7.12-r1039) as well as somatic variant calls (VCFs in VCF format) were downloaded from the Genomic Data Commons. We restricted our analysis to consider only ‘PASS’ variants in these VCFs that are higher-confidence than the raw set. Since TCGA does not provide germline variants, we used GATK’s HaplotypeCaller to perform germline variant calling with default parameters. These calls were refined using VariantRecalibrator in accordance with GATK best practices\(^{2}\).

For this study, we restricted the variant calls to include only missense SNVs in both TCGA and in-house datasets.

Phasing of variants to assess linkage. Somatic and germline missense variant calls from each sample were combined using GATK’s CombineVariants, and the variants were subsequently phased using GATK’s ReadBackedPhasing algorithm.

In silico HLA typing. OptiType\(^{1}\) was used to perform in silico HLA typing for the in-house samples. For the datasets downloaded from TCGA, existing in silico HLA typing information was obtained from The Cancer ImmunoGenome Atlas database\(^{34}\).

Choosing an appropriate window for neoantigen analysis. Due to the absence of subject-specific HLA class II typing information, we limited our neoantigen binding prediction analysis to MHC class I, though we believe that the HLA class II peptides are also important in contributing to immunogenicity. Our nucleotide window was therefore chosen such that it encompasses both MHC class I and class II peptide lengths, to demonstrate the prevalence of proximal variants within that genomic region. Most MHC class I peptides are around 8–11 amino acids in length. There is no length restriction on MHC class II peptides due to an open binding groove, and longer peptide lengths are much more common—typically 13–25 amino acids\(^{33}\), but peptides as long as 30 amino acids have been reported\(^{7}\). The majority (99.2%) of human linear T cell epitopes with MHC class II restriction currently reported in the Immune Epitope Database\(^{34}\) are 8–30-mers. To identify the best-binding 30-mer around a missense variant of interest, one would ideally scan 29 amino acids upstream and downstream of the mutant amino acid, hence a window of 59 amino acids. At the nucleotide level, this corresponds to 87 nucleotides on each side. Given that the frame of the missense mutation is not always known, we allow for 2 extra bases, leading to a window size of 89 nucleotides on each side of the SVOI.

The appropriate nucleotide window for any peptide length can be calculated using the following formula: 

\[(\text{peptide length} - 1) \times 3 + 2\]

Corrected neoantigen binding prediction using pVACtools. For each sample, the phased variant calls and the somatic variant calls were annotated using Variant Effect Predictor (VEP)\(^{4}\), specifically using the Downstream plugin and the custom Wildtype plugin, available via pVACtools (see URLs). To evaluate the effect of relevant nearby variants on neoantigen identification, we re-assessed the binding affinities of the neoantigens with the corrected mutant peptide sequence (Fig. 1c) using NetMHCV4.0\(^{42}\) via an updated version of the pVACtools software. This version takes as input the VEP-annotated phased VCF file of somatic and germline variants, in addition to the existing VEP-annotated somatic VCF file.

Calculating FNR and FDR. To calculate FNR and FDR, we first determined the number of weak binders before PVC that were falsely omitted (false negatives (FN)), as well as the number of peptides that were identified as strong binders before PVC, but whose sequence (MTpeptide) was altered due to a proximal variant and which were thus incorrectly considered during neoantigen selection (false positives (FP)). We also calculated the number of peptides that were identified strong binders before correction and remained unaltered by proximal variants (true positives (TP)).

\[
\text{FN} = (\text{MT}_{\text{uncorrected}} > 500 \text{ nM}) \land \text{MT}_{\text{corrected}} \leq 500 \text{ nM}
\]

\[
\text{FP} = (\text{MT}_{\text{uncorrected}} < 500 \text{ nM}) \land (\text{MT}_{\text{corrected}} \neq \text{MT}_{\text{uncorrected}})
\]

\[
\text{TP} = (\text{MT}_{\text{uncorrected}} < 500 \text{ nM}) \land (\text{MT}_{\text{corrected}} = \text{MT}_{\text{uncorrected}})
\]

The FNR is then defined as the number of false negatives divided by the number of false positives plus the number of true positives. The FDR is defined as the number of false positives divided by the number of all positive calls, including both true positives and false positives.

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

Code availability. The proximal variant analysis code has been merged into the master branch of the pVACtools GitHub repository (https://github.com/griffithlab/pVACtools/tree/master). It is part of pVACtools package version 1.1 and can be installed by running ‘pip install pvactools --upgrade’ on the command line.

Data availability

Several of the in-house sequencing datasets used in the study have been previously published and deposited in various databases. All sequence data for the HER2 breast cancer samples can be accessed via the Database of Genotypes and Phenotypes (dbGaP; study accession phs001291)\(^{3}\). Data for the oral squamous cell carcinoma project and hepatocellular carcinoma samples are part of other manuscripts currently in preparation and can be accessed under dbGaP study accessions phs001623 and phs001106, respectively. Results for the glioblastoma and small cell lung cancer cases\(^{1}\) have been published and can be accessed under dbGaP study accessions phs001665 and phs001049, respectively. TCGA data can be accessed under dbGaP study accession phs000178.

References

20. Griffith, M. et al. Genome Modeling System: a knowledge management platform for genomics. PLoS Comput. Biol. 11, e1004274 (2015).
21. Griffith, M. et al. Optimizing cancer genome sequencing and analysis. Cell Syst. 1, 210–223 (2015).
22. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics 26, 589–595 (2010).
23. Li, H. et al. The Sequence Alignment/Map format and SAMTools. Bioinformatics 25, 2078–2079 (2009).
24. Larson, D. E. et al. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. Bioinformatics 28, 311–317 (2012).
25. Saunders, C. T. et al. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics 28, 1811–1817 (2012).
26. Koboldt, D. C. et al. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics 25, 2283–2285 (2009).
27. Koboldt, D. C. et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 22, 568–576 (2012).
28. Griffith, M. et al. Comprehensive genomic analysis reveals FLT3 activation and a therapeutic strategy for a patient with relapsed adult B-lymphoblastic leukemia. Exp. Hematol. 44, 603–613 (2016).
29. Barnell, E. K. et al. Standard operating procedure for somatic variant refinement of tumor sequencing data. Genet. Med. https://doi.org/10.1038/s41436-018-0278-z (2018).
30. Van der Auwera, G. A. et al. From FastQ to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr. Protoc. Bioinformatics 43, 11.10.1–11.10.33 (2013).
31. Sozek, A. et al. OptiType: precision HLA typing from next-generation sequencing data. Bioinformatics 30, 3310–3316 (2014).
32. Charoeintong, P. et al. Pan-cancer immunogenomic analyses reveal genotype-immunophenotype relationships and predictors of response to checkpoint blockade. Cell Rep. 18, 248–262 (2017).
33. Chicz, R. M. et al. Predominant naturally processed peptides bound to HLA-DR1 are derived from MHC class II-related molecules and are heterogeneous in size. Nature 358, 764–768 (1992).
34. Vita, R. et al. The immune epitope database (IEDB) 3.0. Nucl. Acids Res. 43, D405–D412 (2014).
35. McLaren, W. et al. The Ensembl Variant Effect Predictor. Genome Biol. 17, 122 (2016).
36. Andreatta, M. & Nielsen, M. Gapped sequence alignment using artificial neural networks: application to the MHC class I system. *Bioinformatics* **32**, 511–517 (2016).

37. Nielsen, M. et al. Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. *Protein Sci.* **12**, 1007–1017 (2003).
Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

- [x] The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- [x] An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- [x] The statistical test(s) used AND whether they are one- or two-sided
- [x] Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- [x] A description of all covariates tested
- [x] A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- [x] A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- [x] For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted. Give P values as exact values whenever suitable.
- [x] For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- [x] For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- [x] Estimates of effect sizes (e.g. Cohen’s d, Pearson’s r), indicating how they were calculated
- [x] Clearly defined error bars
  - State explicitly what error bars represent (e.g. SD, SE, CI)

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection

TCGA data was downloaded via the Genomics Data Commons (GDC). dbGAP study accession IDs have been included for all datasets under the 'Data availability section.

Data analysis

All software has been cited throughout the manuscript where used. Genome Modeling System (GMS) was used for alignment and variant calling of in-house datasets. Specific softwares used under the GMS were BWA, Picard, Samtools, Sniper, Strelka, and VarScan. Germline variant calling and phasing was done using Samtools for in-house samples and GATK for TCGA samples. HLA-typing was performed via OptiType. Variant annotation was done using Variant Effect Predictor using custom plugins. These plugins are available for download as part of the pVACtools suite. pVACtools was used to assess for the presence of proximal variants, and for predicting corrected neoantigen binding scores via NetMHCV 4.0. Github repository for proximal variant analysis code:

https://github.com/griffithlab/pVACtools/tree/proximal_variants

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.
Data

Policy information about availability of data
All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:
- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Several of the in-house sequencing datasets used in the study have been previously published and deposited in various databases. All sequence data for the HER2+ breast cancer samples can be accessed via the Database of Genotypes and Phenotypes (dbGAP; study accession: phs001291). Data for oral squamous cell carcinoma project and hepatocellular carcinoma samples are part of other manuscripts currently in preparation, and can be accessed under dbGAP study accession phs001623 and phs001106, respectively. Results for the glioblastoma case and small cell lung cancer have been published and can be accessed under dbGAP study accessions phs001663 and phs001049, respectively. TCGA data can be accessed under dbGAP study accession phs000178.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences
☐ Behavioural & social sciences
☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size
Samples were chosen such that they represented low, medium and high mutational burden tumors. For TCGA, we chose a sample size that could be stored and further analyzed on our local compute clusters. No statistical methods were used to predetermine sample size as power analysis was not relevant to this study. The study is primarily observational in nature, no statistical tests were performed and no p-values are reported.

Data exclusions
No data was excluded from the analysis.

Replication
For reproducibility of our results, details of relevant software and code have added in the code availability section. Conventional biological replicates are not relevant to this study.

Randomization
N/A. We did not group the samples into different cohorts, hence randomization was not applicable to this study.

Blinding
N/A. We did not compare cases and controls to study difference in effects, hence blinding was not applicable to this study.

Reporting for specific materials, systems and methods

Materials & experimental systems
n/a Involved in the study
☐ Unique biological materials
☐ Antibodies
☐ Eukaryotic cell lines
☐ Palaeontology
☐ Animals and other organisms
☐ Human research participants

Methods
n/a Involved in the study
☐ ChIP-seq
☐ Flow cytometry
☐ MRI-based neuroimaging