1. Supplementary Information

Computational identification of neoantigens

Neoantigens from the three datasets were inferred using a consistent pipeline established at Memorial Sloan Kettering Cancer Center. Raw sequence data reads were aligned to the reference human genome (hg19) using the Burrows-Wheeler Alignment tool. Base-quality score recalibration, and duplicate-read removal were performed, with exclusion of germline variants, annotation of mutations, and indels as previously described[4]. Local realignment and quality score recalibration were conducted using the Genome Analysis Toolkit (GATK) according to GATK best practices[41,42]. For sequence alignment and mutation identification, the FASTQ files were processed to remove any adapter sequences at the end of the reads using cutadapt (v1.6)[43]. The files were then mapped using the BWA mapper (bwa mem v0.7.12)[44], the SAM files sorted, and read group tags added using the PICARD tools. After sorting in coordinate order, the BAM's were processed with PICARD MarkDuplicates. First realignment was carried out using the InDel realigner followed by base quality value recalibration with the Base-QRecalibrator.

A combination of 4 different mutation callers (Mutect 1.1.4, Somatic Sniper 1.0.4, Varscan 2.3.7, and Strelka 1.013) were used to identify single nucleotide variants (SNVs)[45,46,47]. As previously described, SNVs with an allele read count of less than 4 or with corresponding normal coverage of less than 7 reads were filtered out[48].

The assignment of a somatic mutation to a neoantigen was estimated using a previously described bioinformatics tool called NASeek[4]. Briefly, NASeek is a computational algorithm that first translates all mutations in exomes to strings of 17 amino acids, for both the wild type and mutated sequences, with the amino acid resulting from the mutation centrally situated. Secondly, it evaluates putative MHC Class I binding for both wild type and mutant nonamers using a sliding window method using NetMHC3.4[19] (http://www.cbs.dtu.dk/services/NetMHC-3.4/) for patient-specific HLA types, to generate predicted binding affinities for both peptides. NetMHC3.4 predicts binding of peptides to HLA alleles using artificial neural networks. Prediction values are given in nM IC50 values and are trained on nonamer peptides like those used in our analysis. NASeek finally assesses for similarity between nonamers that predicted to be presented by patient-specific MHC Class I. All nonamers with inferred affinities below 500 nM are defined as neoantigens.

Clonal tree reconstruction

Tumor clones are reconstructed using the PhyloWGS software package (https://github.com/morrislab/phylowgs)[31]. The input data for the algorithm is extracted from exome sequencing data: (1) mutation reads obtained with the pipeline described above, and (2) allele-specific copy-number variant data, obtained with FACETS v0.5.0[49]. Briefly, the package clusters mutations into clones by the frequency of their reads and it infers possible nesting of clones (ancestral relations) between pairs of clones. Intuitively, an ancestral clone needs to have higher frequency then its derived clone. From this information PhyloWGS reconstructs high likelihood tumor geneological trees.

Sequence alignments

For each patient, we perform sequence alignments of IEDB sequences and the patient's neoantigen sequences. We use BLAST and a blastp program with BLOSUM62 matrix and a strong gap penalty -11 to prevent gapped alignments. The gap extension cost is set to the default value -1. We do not impose any threshold on alignment E-values and consider all alignments. Alignment scores for these identified alignments are then computed with Biopython Bio.paiwise2 package.

HLA-permutation testing

We randomize patient HLA-types by randomly reshuffling the HLA labels within each patient cohort. We then recall neoantigens (using NASeek as described above). HLA randomization changes MHC binding properties of neoantigens and their predicted dissociation constants. Some peptides identified as neoantigens for the original HLA type no longer have inferred affinities below 500 nM; some peptides that did not meet this criterion before meet it under randomized HLA assignment. We then repeat the fitness model evaluation for each patient, using the consistent set of parameters trained on the original cohorts (Extended Data Fig. 3, $a = 26$, $k = 4.87$, $\tau = 0.09$). We follow with survival analysis and evaluate the log-rank test score for patient separation as given by the fitness model. For each cohort, we perform 10 iterations of this analysis, the result is reported in Extended Data Fig. 6.

Amino acid diversity

We define the amino acid diversity at i-th position in a neoantigen as e^{H_i}, where H_i is entropy of amino acid usage at this position, i.e.

$$H_i = -\sum_{j=1}^{20} f(a_{ij}) \log(f(a_{ij})),$$

where $f(a_{ij})$ is frequency of the i-th position in all neoantigens in a group. Inferred neoantigens are nonamers, so i ranges in value from 1 to 9. The diversity of neoantigens at a given site were compared to the values found in the human proteome in Lehman, et al.[23].

To calculate the expected number of words in the proteome we utilize the frequency of amino acids from Lehman, et al. We compute the entropy associated with the frequency of amino acids in the human genome:

$$H(a) = -\sum_{j=1}^{20} f(a_j) \log(f(a_j)),$$

where $f(a_j)$ is the frequency of the j-th amino acid in the human genome. The expected number of words of length n is therefore $e^{nH(a)}$. This value is compared to the observed number of words of length n in the reference proteome for GRCh38.p7 using an entropy of 2.90[23]. Finite genome size exhausts word usage between 5 and 6-mers. By 9-mer length words the ratio of observed to expected words is approximately 0.000052.

Calculation of TCR discrimination length

There are approximately 10^8 unique T-cell receptors in a given human[50], and, moreover, the genome wide entropy of amino acid usage is approximately 2.90[23]. Therefore, one expects the length, L, of words TCRs can typically discriminate to be given by $10^8 \approx e^{2.90L}$ on average (as opposed to say 20^L if one assumed uniform genome amino acid usage). Solving for this length yields $L \approx 6.35$.

Identification of closest nonamers in human proteome to neoantigens

We have mapped the WT and MT 9-mer peptides to all proteins in the current human reference genome (GRCh38.p7) with at least 8 out of 9 matches and no gaps (allowing only mismatches). For this we used LAST[51] (version 819) with the following parameters:

```
lastal -f BlastTab -j1 -r2 -q1 -e15 -y2 -m100000000 -l4 -L4 -P0
```

(9-mer mapping with at most one mismatch is guaranteed to have a matching 4-mer word).

One expects the mutated peptide to only map to the same location as the WT peptide, WT mapping exactly (9 matches) and MT mapping with one mismatch (8 matches). The expected case is that the WT peptide maps to the proteome exactly and the MT peptide maps to the proteome with one mismatch and only to the loci WT peptide maps to.

This rule can be violated in the following cases, sorted from the most to the least severe:

1. WT peptide does not map to the proteome exactly. Some possible reasons are: a difference in the reference assemblies used for mutation calling and peptide mapping,

a germline mutation mistakenly identified as somatic, or a difference between the patient genome and the reference genome used for alignments.
2. WT peptide maps to the proteome exactly (9 matches), MT peptide maps to the proteome exactly (9 matches) but to a different locus.
3. WT peptide maps to the proteome exactly, MT peptide maps to the proteome with one mismatch; however, MT peptide maps with one mismatch to the subjects WT does not map exactly.
4. WT peptide maps to the proteome exactly, MT peptide maps to the proteome with one mismatch; however, MT peptide maps with one mismatch to a different locus on the gene WT maps to.

We have examined each peptide for the worst possible scenario. We have gone from category 1 to 4 in the list. Category 1 indicates a difference in the reference genome. Categories 2-4 typically are due to mutations that occur in repetitive gene families with many paralogs. Once we identified that a peptide belongs to any category, we excluded it from further considerations. This way the numbers of peptides in each category add up to the total number of peptides. Below is a summary for the different datasets utilized in this study:

Snyder, et al[4].:
29781 total peptides, (1) 35 WT unmapped, leaving 29746
27674 expected peptides (92.93%), (2) 361 have 9 matches in MT, (3) 1644 have other alignments, (4) 67 have other alignments to the same subject.

Van Allen, et al.[5]:
39373 total peptides, (1) 42 WT unmapped, leaving 39331
36783 expected peptides (93.42%), (2) 387 have 9 matches in MT, (3) 2076 have other alignments, (4) 85 have other alignments to the same subject.
Rizvi, et al.[6]:
5581 total peptides, (1) 6 WT unmapped, leaving 5575
5125 expected peptides (91.83%), (2) 105 have 9 matches in MT, (3) 323 have other alignments, (4) 22 have other alignments to the same subject.

Additional supplementary files for each dataset are included as Supplementary Data:

mt-with-9.tsv – list of peptides from category 2 and the subjects each one aligns to.

peptides-with-extra-aln.tsv – peptides from group 3 and the subjects each one aligns to.

peptides-multima,pping-same-subj.tsv – peptides from group 4 and their alignments including the start and end coordinates

Set of epitopes used for positive T cell assays

We utilized a set of epitopes from the Immune Epitope Database (IEDB), a repository of over 120,000 immune epitopes[20]. IEDB is a collection of epitope-specific experimental assays – the nature of which can be accessed by various fields (www.iedb.org). Every T cell assay reflects the binding of an epitope-specific TCR to an experimentally tested antigen[20]. In our cases we restricted our analysis to linear epitopes from human infectious diseases studies presented by class I MHC molecules for which there were positive T cell assays. For negative controls, we downloaded epitopes associated with assays satisfying all of the above fields, except we did not restrict for positive assays. We then excluded those for which we assigned positive by IEDB to create a negative assay list. As the database changes from time to time, we included both lists as Supplementary Data.

Error bars for Kaplan-Meier curves

Error bars on Kaplan-Meier curves were calculated in GraphPad Prism 7. The error bars are defined by the standard error of the Kaplan-Meier estimator using Greenwood's formula[52].

Additional References

41. DePristo, M. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genet.* **43**, 491-498 (2011).
42. Van der Auwera, G.A. et al. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Curr. Prot. in Bioinformatics* **43**, 11.10.1-11.10.33 (2013).
43. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10-12 (2011).
44. Li, H., & Durbin R. Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics* **25**, 1754-1760 (2009).
45. Wei, L. et al. MAC: identifying and correcting annotation for multi-nucleotide variations. *BMC Genomics* **16**, 569 (2015).
46. Snyder, A. & Chan, T.A. Immunogenic peptide discovery in cancer genomes. *Curr Opin Genet Dev* **30**, 7-16 (2015).
47. Nielsen, M. et al. Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. *Protein Sci* **12**, 1007-1017 (2003).
48. Riaz, N. et al. Recurrent SERPINB3 and SERPINB4 mutations in patients who respond to anti-CTLA4 immunotherapy. *Nat. Genet.* **48**, 1327-1329 (2016).
49. Shen, R. & Seshan, V.E. FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. *Nucleic Acids Res.* **44**, e131 (2016).
50. Arstila, T. P. et al. A direct estimate of the human αβ T cell receptor diversity. *Science* **286**, 958–961 (1999).
51. Kiełbasa, S.M., Wan, R., Sato, K., Horton, P., & Frith, M.C. Adaptive seeds tame genomic sequence comparison. *Genome Res.* **21**, 487-493 (2011).
52. Greenwood Jr., M. The natural duration of cancer. *Rep Public Health and Related Subjects.* **33** (1926).