Supplementary Methods

Characterization of neoantigens

In the TCIA project, HLA alleles were called from RNA-sequencing FASTQ files using Optitype [1], selected for its high performance and for its applicability to RNA-sequencing data. For each subject, the HLA alleles estimated from the sample with the highest coverage over the HLA locus were considered. To estimate the mutated proteins, authors focused on non-synonymous missense mutations, and selected mutations associated to Uniprot protein identifiers. The protein sequence retrieved from Uniprot was changed according to the non-synonymous, missense mutations reported in the MAF file, and truncated in case of stop codons. Authors removed candidate proteins affected by annotation inconsistencies between protein identifiers and predicted effect. Peptides of 8-11 amino acids in length, covering the mutated region of the protein, were analyzed with NetMHCpan (Version 2.8) [2] to estimate their binding affinity to the HLA alleles. Self-antigens mapping to human Uniprot proteins were identified with BLAST and filtered out. Amongst the candidate antigenic peptides, authors selected strong binders with binding affinity < 500 nM as described in Rooney *et al.* study [3], and considered peptides arising from expressed genes. Authors identified expressed genes as those having median TPM greater than 2 in a given cancer type, as previously shown [4].

References:

[1] Szolek A, Schubert B, Mohr C, Sturm M, Feldhahn M and Kohlbacher O. OptiType: precision HLA typing from next-generation sequencing data. Bioinformatics. 2014; 30(23):3310-3316.

[2] Nielsen M, Lundegaard C, Blicher T, Lamberth K, Harndahl M, Justesen S, Roder G, Peters B, Sette A, Lund O and Buus S. NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and -B locus protein of known sequence. PLoS One. 2007; 2(8):e796.

[3] Rooney MS, Shukla SA, Wu CJ, Getz G and Hacohen N. Molecular and genetic properties of tumors associated with local immune cytolytic activity. Cell. 2015;

160(1-2):48-61.

[4] Wagner GP, Kin K and Lynch VJ. A model based criterion for gene expression calls using RNA-seq data. Theory Biosci. 2013; 132(3):159-164.