Identification and Ranking of Recurrent Neo-Epitopes in Cancer

Authors: Eric Blanc[1,5], Manuel Holtgrewe[1,5], Arunraj Dhamodaran[3], Clemens Messerschmidt[1,5], Gerald Willimsky[2,4,6], Thomas Blankenstein[2,3,4], Dieter Beule[1,3,*]

Affiliations:

[1] Core Unit Bioinformatics, Berlin Institute of Health, Berlin, Germany

[2] Institute of Immunology, Charité - Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Berlin, Germany

[3] Max Delbrück Center for Molecular Medicine in the Helmholtz Association (MDC), Berlin, Germany

[4] Berlin Institute of Health, Berlin, Germany

[5] Charité - Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Berlin, Germany

[6] German Cancer Research Center (DKFZ), Heidelberg, Germany

*To whom correspondence should be addressed: dieter.beule@bihealth.de

Abstract:

Neo-epitopes are emerging as attractive targets for cancer immunotherapy and new strategies for rapid identification of relevant candidates have become a priority. We propose a method for in silico selection of candidates which have a high potential for neo-antigen generation and are likely to appear in multiple patients. This is achieved by carefully screening 33 TCGA data sets for recurrent somatic amino acid exchanges and, for the 1,055 resulting recurrent variants, applying MHC class I binding prediction algorithms. A preliminary confirmation of epitope binding and recognition by CD8 T cells has been carried out for a couple of candidates in humanized mice. Recurrent neo-epitopes may be suitable to supplement existing personalized T cell treatment approaches with precision treatment options.

Introduction

Increasing evidence suggests that clinical efficacy of cancer immunotherapy is driven by T cell reactivity against neo-antigenes (1, 2, 3, 4). For tumors with no viral etiology these neo-antigenes are created either by aberrant expression of genes normally restricted to immuno-privileged tissues, or by tumor specific DNA alterations that result in the formation of novel protein sequences.

While not yet fully understood, immune response to any mutated peptide sequence and recognition of tumor cells containing this peptide depends critically on the ability of the MHC class I complexes to bind to the mutated peptide in order to present it to a T cell (5). A variety of machine learning algorithms have been developed to determine the MHC binding in silico, see (6) for review. Most methods are trained on Immune Epitope Database (IEDB) (7) entries and use allele specific predictors for frequent alleles, while pan-methods are applied to extrapolate to less common alleles.

With the advent of affordable short read sequencing comprehensive neo-antigen screening based on whole exome sequencing has become feasible and many cancer immune therapeutic approaches try to utilize detailed understanding of the neo-epitope spectrum to create additional or boost pre-existing T cell reactivity for therapeutic purposes (8, 9). However, in practice the selection and validation of the most promising neo-epitope candidates is a difficult and time-consuming task.

By virtue of the underlying mutational processes, the genome architecture and accessibility as well as for functional reasons within the disease process, certain somatic mutations will be present in multiple patients while still being highly specific to the tumor (10, 11). This might open the way to supplement existing personalized cancer immune treatments approaches with precision treatment options.

Analysis of neo-epitope candidates has been carried out for selected highly recurrent mutations occurring in known cancer-related genes (12, 13). Here we provide an unbiased, comprehensive study of neo-epitopes arising from recurrent mutations: we carefully screen large cohorts of cancer genomes for recurrent missense mutations, identify possible candidates for strongly binding neo-antigens in multiple HLA-1 alleles using MHC-I binding predictions, and finally rank these candidates according to the expected number of target patients. A couple of these candidates were tested for in-vivo activity.

Results

Recurrent Variants and Neo-Epitopes

From the GDC repository (14), we have collected somatic variants for 33 TGCA studies. After removing patients without clinical meta-data, and studies with less than 100 patients, we have selected 1,384,531 high-confidence missense SNPs from 9,641 patients, see methods for details. Using this data, 1,055 variants are deemed recurrent (supplementary table 1), as they can be found in more than 1% of the patients in the respective study cohort. These recurrent variants correspond to 869 unique protein changes, as some appear in multiple cancer entities. 77 of the recurrent variants occur in at least 3% of their cohort (43 unique protein changes).

From these 869 unique protein changes, we have generated neo-epitope candidates that are predicted to be strong MHC class I binders in at least one of the 11 frequent HLA-1 types that we considered for initial selection. 415 (48%) of them lead to a strong binder prediction. In total, there are 772 candidates that are recurrent in a cancer entity cohort, and predicted as binding for a considered HLA-1 type. These candidates are unique among all the 9-, 10- & 11-mers containing the variant: the selection process retains only the peptide sequence with the lowest predicted IC50. Figure 1 and table 1 provide an overview of the variant selection and neo-epitope candidates generation processes, while supplementary table 2 lists all neo-epitopes (weak and strong predicted binders) after removing redundancy.

Despite large differences between variant filtering protocols, there is notable overlap between variants deemed recurrent by the above process, and variants identified in the cancer hotspot datasets (10) (supplementary figure 1). This overlap is strongly dependent on how frequent those variants are observed: there are 54 common variants out of the 61 variants observed more than 10 times over the whole dataset (>88%). Among the 819 retained variants (see methods), only 5 appear among the variants flagged as possible false positive by Chang et al. (<1%).

Confirmation of known cancer neo-epitopes

To validate our selection mechanism we aggregated two studies (15, 16) which collect reports of spontaneous CD8+ T-cell responses in cancer patients in whom the target epitopes were subsequently discovered. Both sets together (supplementary table 3) contain 37 epitopes, 35 of which could be mapped to an ENSEMBL transcript (33 unique genes). For 27 of these epitopes our pipeline predicted strong binding with the specific HLA-1 type reported in the corresponding wet-lab investigations. Another 5 epitopes where predicted as weak binders, some of the latter are also predicted to be strong binders in other HLA-1 types. Our pipeline classified 70% of a set of known tumor neo-antigens as strong binders and another 14% as weak binders. Therefore we may assume that these 70% provide a rough estimate of the sensitivity of the screening. However we expect sensitivity to vary widely for different HLA-1 types due to the different amount of training data and resulting limitation of prediction performance in less frequent HLA-1 types.

4 out of 34 unique identifiable variants studied by van Buuren et al (16) and Fritsch (15) are found among our set of high confidence missense variants. CDK12:E928K occurs in one out of 530 Uterine Carcinoma donors, LPGAT1:D266N in one out of 985 Breast carcinoma donors, and CDK4:R24C occurs in 1 out of 467 of melanoma-patients. None of these epitopes have been assessed in our analysis, as their recurrence is below the 1% threshold. Only CTNNB1:S37F fulfills the 1% recurrence threshold (9 uterine carcinoma patients), and the peptide is predicted to be a strong binder for HLA-C*07:02. A longer peptide is predicted to be a strong binder for HLA-B*15:01.

The CDK4:R24C peptide (sequence ACDPHSGHFV, see supplementary table 1) is not predicted to bind to HLA-A*02:01, even tough it leads to confirmed T cell response (17), and has been related to cutaneous malignant melanoma and hereditary cutaneous melanoma (18, 19). Therefore the list of recurrent amino acids exchanges for which no binding epitopes are predicted might still be a valuable resource for future research, see supplementary tables 1 (all recurrent variants) and 2 (includes candidates predicted as weak binders).

Enrichment in known cancer related genes

We observe that recurrent neo-epitope candidates are substantially enriched in known cancer-related genes (figure 1C). Initially approximatively one percent of all observed variants are found in genes that have been described (20, 21) as oncogenes or tumor suppressor genes. When recurrent unique protein changes are considered, the fraction of known oncogenes or tumor suppressor genes is substantially increased to 13% and 6.5% respectively. These fractions only marginally increase to 14% and 7% when only the unique protein changes leading to predicted strong binders for our 11 HLA-1 types are considered. Supplementary table 4 shows a similar enrichment of known cancer-related genes per cohort. We observe that the enrichment is stronger for oncogenes than for tumor suppressors. This might be expected, as activating mutations in oncogenes are mainly distributed on a few protein positions, while loss of function mutations in tumor suppressors are generally distributed more broadly along the protein sequence.

It is interesting to observe that several of the highly prevalent neo-epitope candidates occur in genes that are involved in known immune escape mechanisms: RAC1:P29S is recurrent in study SKCM (melanoma), is predicted to lead to strong binding neo-epitopes for HLA-A*01:01 and HLA-A*02:01, and is reported to up-regulate PD-L1 in melanoma (22). CTNNB1:S33C is recurrent in studies LIHC (liver hepatocellular carcinoma) and UCEC (uterine corpus endometrial carcinoma), is predicted to lead to strong binding neo-epitopes for HLA-A*02:01, and has been shown to increase the expression of the Wnt-signalling pathway in hepatocellular carcinoma (23), leading to modulation of the immune response (24) and ultimately to tumor immune escape (25). In a separate study, Cho et al (26) show that this mutation confers acquired resistance to the drug imatinib in metastatic melanoma. Finally, FLT3:D835Y recurrent in study LAML (acute myeloid leukemia), is predicted to lead to a strong binding neo-epitope for HLA-A*01:01, HLA-A*02:01 and HLA-C*06:02, and following Reiter et al. (27), Tyrosine Kinase Inhibitors promote the surface expression of the mutated FLT3,

enhancing FLT3-directed immunotherapy options, as its surface expression is negatively correlated with proliferation.

While the described mechanisms are probably sufficient to explain immune escape in tumor evolution, the candidates could nevertheless be viable targets for adoptive T cell therapy or TCR gene therapy.

Confirmation of MHC binding and T cell reactivity

In order to further test predicted epitopes for recognition in vivo we utilized transgenic mice that harbor the human TCRαβ gene loci, a chimeric HLA-A2 gene and are deficient for mouse TCRαβ and mouse MHC I genes (termed ABabDII). These mice have been shown to express a diverse human TCR repertoire (28, 29) and thus mimic human T cell response. Validation was carried out for candidate neo-epitopes RAC1:P29S & TRRAP:S722F (figure 2). ABabDII mice were immunized at least twice with mutant peptides and IFNγ producing CD8+ T cells were monitored in ex vivo ICS analysis 7 days after the last immunization. CD8+ T cells were purified from spleen cell cultures of reactive mice using either IFNγ-capture or tetramer-guided FACSort. Sequencing of specific TCR α and β chain amplicons that were obtained by RACE-PCR revealed that this procedure yields an almost monoclonal CD8+ T cell population (not shown). In both cases, tested neo-antigen candidates lead to T cell reactivity, confirming not only predicted MHC binding by our pipeline but also immunogenicity in vivo in human TCR transgenic mice. Therefore this workflow also allows to generate potentially therapeutic relevant TCRs to be used in the clinics for cancer immunotherapy.

Recurrent neo-epitopes in patient populations

Upon assumption of statistical independence, the product of the frequency of a recurrent neo-epitope with the frequency of class I alleles in the population and the incidence rates of cancer types provides an estimate for the number of patients that carry that specific neo-epitope. Using the number of newly diagnosed patients per year and HLA-1 frequency in the US population, we are able to compute the expected number of patients for 18 cancer entities for which both cancer census data and a TCGA study are available. The occurrence numbers for individual neo-epitope candidates range from 0 to 2,254 for PIK3CA:H1047R in breast cancer patients of type HLA-C*07:01; table 2 presents a summary of expected patient numbers for the complete set of candidates . We estimate that the previously discussed RAC1:P29S is present in 628 patients with the HLA-A*02:01 allele per year in the US alone: in 556 melanoma patients and in 72 lung small cell, head & neck or uterine carcinomas patients (see supplementary table 5 for details). For the CTNNB1:S33C variant, the total number of HLA-A*02:01 patients in the US is expected to be 364, from uterine corpus, prostate and liver cancer types. As another example, 115 myeloid leukemia patients in the US are expected to be of type HLA-A*02:01 and harbor the FLT3:D835Y variant.

Figure 3 shows the cumulative expected number of patients that carry a specific epitope, and with matching HLA-1 type, for the 50 candidates with the highest expected patients number. The number of patients is derived from the sum over all

cancer entities, including those in which the candidate is not recurrent according to our criteria. For example, among newly diagnosed patients of type HLA-C*04:02, 88 prostate cancer patients are expected to carry the mutation PIK3CA:R88Q, even though its observed frequency in the PRAD study is as low as 0.2%.

Discussion

Using existing cancer studies and neo-epitope binding predictions to MHC class I proteins, we propose a ranking of neo-epitopes that occur frequently in observed cancer patient cohorts, and that are potentially accessible to T cell immunotherapy treatments. This ranking is based on the expected number of patients of a particular HLA-1 type, who carry the recurrent mutation.

Despite numerous mechanisms of immune evasion, neo-epitopes are important targets of endogenous immunity (30): in some cases at least, it has been shown that they contribute to tumor recognition (31), achieve high objective response (in melanoma, 32), and a single of them is presumably sufficient for tumor regression (33). Moreover, positive association has been shown between antigen load and cytolytic activity (34), activated T cells (35) and high levels of the PD-1 ligand (36). Taken together, these results suggest that neo-epitopes occupy a central role in regulating immune response to cancer, and that this role can be exploited for cancer immunotherapy.

Targeting neo-epitopes based on non-recurrent, “private” somatic variants requires isolation of TCRs for each individual patient, which is currently still challenging (37). However, successful application of treatments based on genetically engineered lymphocytes has already been shown for epitopes arising from unmutated proteins (“public” epitopes): the MART-1 and gp100 proteins have been targeted in melanoma cases (38), as these proteins are expressed in the tumor cells. In another trial, Robbins et al. (39) have studied long-term follow-up of patients who were treated with TCR-transduced T cells against NY-ESO-1, a protein whose expression is normally restricted to testis, but which is frequently aberrantly expressed in tumor cells. The authors show that the treatment may be effective for some patients. These results show that immune treatments can be beneficial, even when the selected epitopes are not obtained from sequencing the patient's tumor, but originate from ”off-the-shelf” peptides, whose sequence is known prior to the sequencing of the patient's tumor.

However, targeting such unmutated epitopes presents safety and efficacy concerns (2): the administration of T cells transduced with MART-1 specific T-cell receptor have led to fatal outcomes (40), and cross-reactivity of TCR against MAGE-A3 (a protein normally restricted to testis and placenta) caused cardiovascular toxicity (41). Neo-epitopes based on recurrent somatic variants potentially alleviate such problems (as the target sequences are truly restricted to tumor cells), while retaining the benefits of “public” epitopes, for example regarding regulatory hurdles.

The neo-epitope landscape is diverse and sparse (33), with few neo-epitopes that are both predicted strong binders and present in multiple patients. In their analysis, Hartmaier et al. (13) estimate that neo-epitopes suitable used for precision immuno-therapy might be relevant for about 0.3% of the population, in broad agreement with this study. However, the absolute number of patients is still considerable, even for neo-epitopes arising from less frequent mutations (between 1 and 3% of the cohort).

Taking into account the fact that MHC binding is a necessary but not sufficient condition for T cell activity, and the limitations of MHC binding prediction algorithms,

we provide an objective ranking of neo-epitopes based on recurrent variants, as a basis for the development of off-the-shelf immunotherapy treatments. We experimentally confirm T cell reactivity, thus immunogenicity, for a couple of these neo-epitopes.

Methods

Data Sets

Somatic variants for different cancer entities have been determined using matched pairs of tumor and blood whole exome or whole genome sequencing in the TCGA consortium. We downloaded the open-access somatic variants from GDC data release 7.0, consisting of 33 TCGA projects and 10,182 donors in total from (14). We excluded patients without corresponding entries in the clinical information tables, and 7 projects with less than 100 samples, yielding 9,641 samples covering 26 cancer studies.

Variant Selection

For each sample we selected all single nucleotide variants obtained by the “mutect2” pipeline, that had a “Variant_Type” equal to “SNP”, a valid ENSEMBL transcript ID and a valid protein mutation in “HGVSp_Short”. From these variants, we selected those with a “Variant_Classification” equal to “Missense_Mutation”. We checked that all variants had a “Mutation_Status” equal to (up to capitalisation) “Somatic”, that the total depth “t_depth” was the sum of the reference “t_ref_count” and the alternate “t_alt_count” alleles counts, and that the genomics variant length is one nucleotide. To avoid high number of false positives we consider only variants that are supported by at least 5 reads and have a VAF of at least 10%. Furthermore we removed any variant that occurs with more than 1% in any population contained in the ExAC database version 0.31 (42), by coordinates liftover from the GRCh38 to hg19 human genome versions. This way we obtained 26 cancer entity data sets containing a total of 9,641 samples with an overall 1,384,531 variants.

Recurrent Protein Variant Filtering

We define recurrence strictly on the protein/amino acid exchange level, i.e. different nucleotide acid variants leading to the same amino acid exchange due to code redundancy will be counted together. Recurrent protein variants are defined within each TCGA study. A protein variant is deemed recurrent when it appears in at least 1% of all the patients in the cohort. As cancer types are only considered when the number of patients involved in the studies is greater than 100, this threshold ensures that every recurrent variant has been observed in at least 2 patients for a given cancer type. To be conservative, the recurrence frequency has been computed using, for the denominator, all patients with clinical information in the study, including those without high-confidence missense SNVs. Using this definition, the total number of recurrent amino acid changes is 1,055. A variant recurrent in multiple cancer types is counted multiple times in the above number, the number of unique recurrent variants regardless of the cancer is 869. Supplementary table 1 shows the most frequent amino acid exchanges across 25 cancer entities, as no variant from project TCGA-KIRC's donors is labeled as recurrent.

Recurrent variants occurring at the same positions (for example when gene's IDH1 codon R132 is mutated to amino acid H, C, G or S) have been merged into 819 variants suitable for comparisons with the cancer hot spots lists (10). 122 out of the 819 merged variants belong to the set of 470 cancer hotspot variants, and 5

(PCBP1:L100, SPTLC3:R97, EEF1A1:T432, BCLAF1:E163 & TTN:S3271) to the set of presumptive false positives hotspots listed in the supplementary material of (10).

Epitopes Selection and MHC Class I Binding Prediction

In the next step all peptide stretches (9-, 10, or 11-mers) containing any of the identified recurrent amino acid exchanges are generated in silico. For MHC class I binding prediction we selected 11 frequent HLA-1 types: HLA-A*02:01, HLA-A*03:01, HLA-A*11:01, HLA-B*07:02, HLA-B*15:01, HLA-C*04:01, HLA-C*07:02, HLA-A*01:01, HLA-B*08:01, HLA-C*06:02, HLA-C*07:01. We predicted the MHC class I binding using NetMHCcons (43) v1.1, which predicts peptides IC50 binding, and classifies these predictions as non-binder, weak and strong binders. In the study, we have used these classes to filter our neo-epitope candidate.

For a given recurrent variant and a given HLA-1 type, the epitope prediction pipeline can produce multiple overlapping epitopes candidates, differing only by their length. To remove such size redundancy, only the epitope with the lowest predicted mutant sequence IC50 is retained. This procedure also removes non-overlapping epitopes, to keep only at most one epitope per recurrent protein variant, cancer type and HLA-1 type. For comparison we also compute the IC50 for the respective wild type peptide.

This way, we obtain 769 strong binding recurrent peptides and 1829 weak binders. Their complete list is in supplementary table 2, where each candidate is listed with the HLA-1 type it is preticted to bind to.

Data QC

To ensure that the proportion of variants caused by technical artifacts is small, we have computed the proportion of SNVs called in poly-A, poly-C, poly-G or poly-T repeats of length greater than 6 have been computed for each data study (44), for unique variants (that occur in only one patient across a project cohort), and for variants that are observed more than once in a cohort (supplementary figure 1). For comparison, we have computed the expected frequency of such events, assuming that all possible 11-mers (the mutated nucleotide at the center, flanked by 5 nucleotides on each side) are equiprobable, regardless of their sequence.

Generation of mutation-specific T cells in ABabDII mice

For immunisation 8-12-week old ABabDII mice were injected subcutaneously with 100 µg of mutant short peptide (9-10mers, JPT) supplemented with 50 µg CpG 1826 (TIB Molbiol), emulsified in incomplete Freund’s adjuvant (Sigma). Repetitive immunizations were performed with the same mixture at least three weeks apart. Mutation-specific CD8+ T cells in the peripheral blood of immunized animals were assessed by intracellular cytokine staining (ICS) for IFNγ 7 days after each boost. All animal experiments were performed according to institutional and national guidelines and regulations after approval by the governmental authority (Landesamt für Gesundheit und Soziales, Berlin).

Patient Number Estimates and HLA-1 Frequencies

HLA-1 frequency data f for the U.S. population was retrieved from the Allele Frequency Net Database (AFND) (45). Frequency data were estimated by averaging the allele frequencies* of multiple population datasets from the North American (NAM) geographical region. The major U.S. ethnic groups were included and sampled under the NAM category.

Cancer incidence data for the U.S. population (N_d) was retrieved from the GLOBOCAN 2012 project of the International Agency for Research on Cancer, WHO (46).

We assume that the fraction of a recurrent variant in the U.S. population affected by cancer entity *d* (r_d) is identical to the observed ratio of that variant in the corresponding TCGA study.

Given a variant leading to a neo-epitope predicted to be binding strongly to MHC class I proteins of HLA-1 type *h*, the number of patients of HLA-1 type *h* whose tumor contain that variant is expected to be

$$n_h = f_h \sum_d r_d N_d$$

The summation runs over 18 diseases *d* for which both the TCGA projects and the cancer incidence data are available.

References

1. Schumacher, T. N. & Schreiber, R. D. Neoantigens in cancer immunotherapy. Science 348, 69–74 (2015).
2. Blankenstein, T., Leisegang, M., Uckert, W. & Schreiber, H. Targeting cancer-specific mutations by T cell receptor gene therapy. Curr. Opin. Immunol. 33, 112–119 (2015).
3. Rosenberg, S. A. & Restifo, N. P. Adoptive cell transfer as personalized immunotherapy for human cancer. Science 348, 62–8 (2015).
4. Wirth, T. C. & Kühnel, F. Neoantigen Targeting - Dawn of a New Era in Cancer Immunotherapy? Front. Immunol. 8, 1848 (2017).
5. Engels, B. et al. Relapse or Eradication of Cancer Is Predicted by Peptide-Major Histocompatibility Complex Affinity. Cancer Cell 23, 516–526 (2013).
6. Snyder, A. & Chan, T. A. Immunogenic peptide discovery in cancer genomes. Curr. Opin. Genet. Dev. 30, 7–16 (2015).
7. Vita, R. et al. The immune epitope database (IEDB) 3.0. Nucleic. Acids Res. 43, D405–D412 (2015).
8. Leisegang, M. et al. Eradication of Large Solid Tumors by Gene Therapy with a T-Cell Receptor Targeting a Single Cancer-Specific Point Mutation. Clin. Cancer Res. 22, 2734–43 (2016).
9. Li, F. et al. Rapid tumor regression in an Asian lung cancer patient following personalized neo-epitope peptide vaccination. Oncoimmunology. 5, e1238539 (2016).
10. Chang, M. T. et al. Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. Nat. Biotech. 34, 155–163 (2016).
11. Martincorena, I. et al. Universal Patterns of Selection in Cancer and Somatic Tissues. Cell 171, 1029–1041.e21 (2017).
12. Warren, R.L. & Holt, R.A. A census of predicted mutational epitopes suitable for immunologic cancer control. Human Immunology 71 245-254 (2010).
13. Hartmaier, R. J. et al. Genomic analysis of 63,220 tumors reveals insights into tumor uniqueness and targeted cancer immunotherapy strategies. Genome Medi. 9, 16 (2017).
14. Grossman, R. L. et al. Toward a Shared Vision for Cancer Genomic Data. N. Engl. J. Medi. 375, 1109–1112 (2016).
15. Fritsch, E. F. et al. HLA-binding properties of tumor neoepitopes in humans. Cancer Immunol. Res. 2, 522–9 (2014).
16. Van Buuren, M. M., Calis, J. J. & Schumacher, T. N. High sensitivity of cancer exome-based CD8 T cell neo-antigen identification. Oncoimmunology. 3, e28836 (2014).
17. Wölfel, T et al. A p16INK4a-insensitive CDK4 mutant targeted by cytolytic T lymphocytes in a human melanoma. Science 269, 1281–4 (1995).
18. Landsberg, J. et al. Autochthonous primary and metastatic melanomas in Hgf-Cdk4R24C mice evade T-cell-mediated immune surveillance. Pigment Cell Melanoma Res. 23, 649–660 (2010).

19. Platz, A., Ringborg, U. & Hansson, J. Hereditary cutaneous melanoma. Semin. Cancer Biol. 10, 319–326 (2000).
20. Vogelstein, B. et al. Cancer genome landscapes. Science 339, 1546–58 (2013).
21. Rubio-Perez, C. et al. In Silico Prescription of Anticancer Drugs to Cohorts of 28 Tumor Types Reveals Targeting Opportunities. Cancer Cell 27, 382–396 (2015).
22. Vu, H. L., Rosenbaum, S., Purwin, T. J., Davies, M. A. & Aplin, A. E. RAC1 P29S regulates PD-L1 expression in melanoma. Pigment Cell Melanoma Res. 28, 590–598 (2015).
23. Austinat, M. et al. Correlation between β-catenin mutations and expression of Wnt-signaling target genes in hepatocellular carcinoma. Mol. Cancer 7, 21 (2008).
24. Pai, S. G. et al. Wnt/beta-catenin pathway: modulating anticancer immune response. J. Hematol. Oncol. 10, 101 (2017).
25. Spranger, S. & Gajewski, T. F. A new paradigm for tumor immune escape: β-catenin-driven immune exclusion. J. Immunotherapy Cancer 3, 43 (2015).
26. Cho, J. et al. Emergence of CTNNB1 mutation at acquired resistance to KIT inhibitor in metastatic melanoma. Clin. Transl. Oncol. 19, 1247–1252 (2017).
27. Reiter, K et al. Tyrosine kinase inhibition increases the cell surface localization of FLT3-ITD and enhances FLT3-directed immunotherapy of acute myeloid leukemia. Leukemia 32, 313–322 (2018).
28. Li, L.-P. et al. Transgenic mice with a diverse human T cell antigen receptor repertoire. Nat. Med. 16, 1029–1034 (2010).
29. Li, L. & Blankenstein, T. Generation of transgenic mice with megabase-sized human yeast artificial chromosomes by yeast spheroplastembryonic stem cell fusion. Nat. Prot. 8, 1567–1582 (2013).
30. Bethune, M. T. & Joglekar, A. V. Personalized T cell-mediated cancer immunotherapy: progress and challenges. Curr. Opini. Biotech. 48, 142–152 (2017).
31. Robbins, P. F. et al. Mining exomic sequencing data to identify mutated antigens recognized by adoptively transferred tumor-reactive T cells. Nat. Med. 19, 747–52 (2013).
32. Rosenberg, S. A. & Dudley, M. E. Adoptive cell therapy for the treatment of patients with metastatic melanoma. Curr. Opini. Immunol. 21, 233–240 (2009).
33. Tran, E. et al. Cancer immunotherapy based on mutation-specific CD4+ T cells in a patient with epithelial cancer. Science 344, 641–5 (2014).
34. Rooney, M., Shukla, S., Wu, C., Getz, G. & Hacohen, N. Molecular and Genetic Properties of Tumors Associated with Local Immune Cytolytic Activity. Cell 160, 48–61 (2015).
35. Charoentong, P. et al. Pan-cancer Immunogenomic Analyses Reveal Genotype-Immunophenotype Relationships and Predictors of Response to Checkpoint Blockade. Cell Reports 18, 248–262 (2017).
36. McGranahan, N. et al. Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. Science 351, 1463–9 (2016).
37. Strønen, E. et al. Targeting of cancer neoantigens with donor-derived T cell receptor repertoires. Science 352, 1337–41 (2016).

38. Johnson, L. A. et al. Gene therapy with human and mouse T-cell receptors mediates cancer regression and targets normal tissues expressing cognate antigen. Blood 114, 535–46 (2009).
39. Robbins, P. F. et al. A pilot trial using lymphocytes genetically engineered with an NY-ESO-1-reactive T-cell receptor: long-term follow-up and correlates with response. Clin. Cancer. Res. 21, 1019–27 (2015).
40. Van den Berg, J. H. et al. Case Report of a Fatal Serious Adverse Event Upon Administration of T Cells Transduced With a MART-1-specific T-cell Receptor. Mol. Ther. 23, 1541–1550 (2015).
41. Linette, G. P. et al. Cardiovascular toxicity and titin cross-reactivity of affinity-enhanced T cells in myeloma and melanoma. Blood 122, 863–71 (2013).
42. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature 536, 285–291 (2016).
43. Karosiene, E., Lundegaard, C., Lund, O. & Nielsen, M. NetMHCcons: a consensus method for the major histocompatibility complex class I predictions. Immunogenetics 64, 177–186 (2012).
44. Laehnemann, D., Borkhardt, A. & McHardy, A. C. Denoising DNA deep sequencing datahigh-throughput sequencing errors and their correction. Brief. Bioinform. 17, 154–179 (2016).
45. González-Galarza, F. et al. Allele frequency net 2015 update: new features for HLA epitopes, KIR and disease and HLA adverse drug reaction associations. Nucleic. Acids Res. 43, D784–D788 (2015).
46. Ferlay, J. et al. Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. Int. J. Cancer 136, E359–E386 (2015).

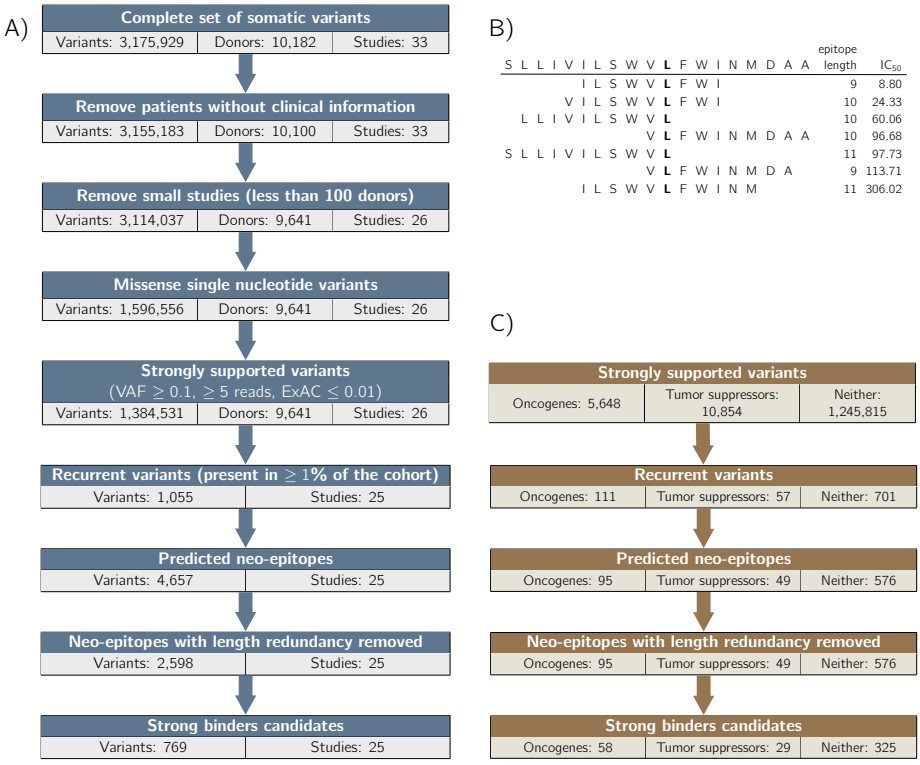

Figure 1: (A) Overview of the recurrent neo-epitope candidates generation process: TCGA studies are selected for at least 100 donors with clinical annotations. For each of these studies, recurrent strongly supported missense Single-Nucleotide Variants are collected. Neo-epitopes binding to 11 HLA-1 types are predicted, redundancy is removed from that set (see B) and strong binders are retained. (B) Example of epitope redundancy: the 18 amino-acids long sequence surrounding recurrent variant GLRA3:S274L generates 7 binding neo-epitopes for the type HLA-A*02:01. Our pipeline retains only the strongest predicted binder for a given variant and HLA-1 type pair (the first, with an IC50 of 8.8 nM in the example). (C) Number of SNVs classified as Oncogenes or Tumor Suppressors by Vogelstein et al. (20), at various point of the variant selection and neo-epitope filtering process.

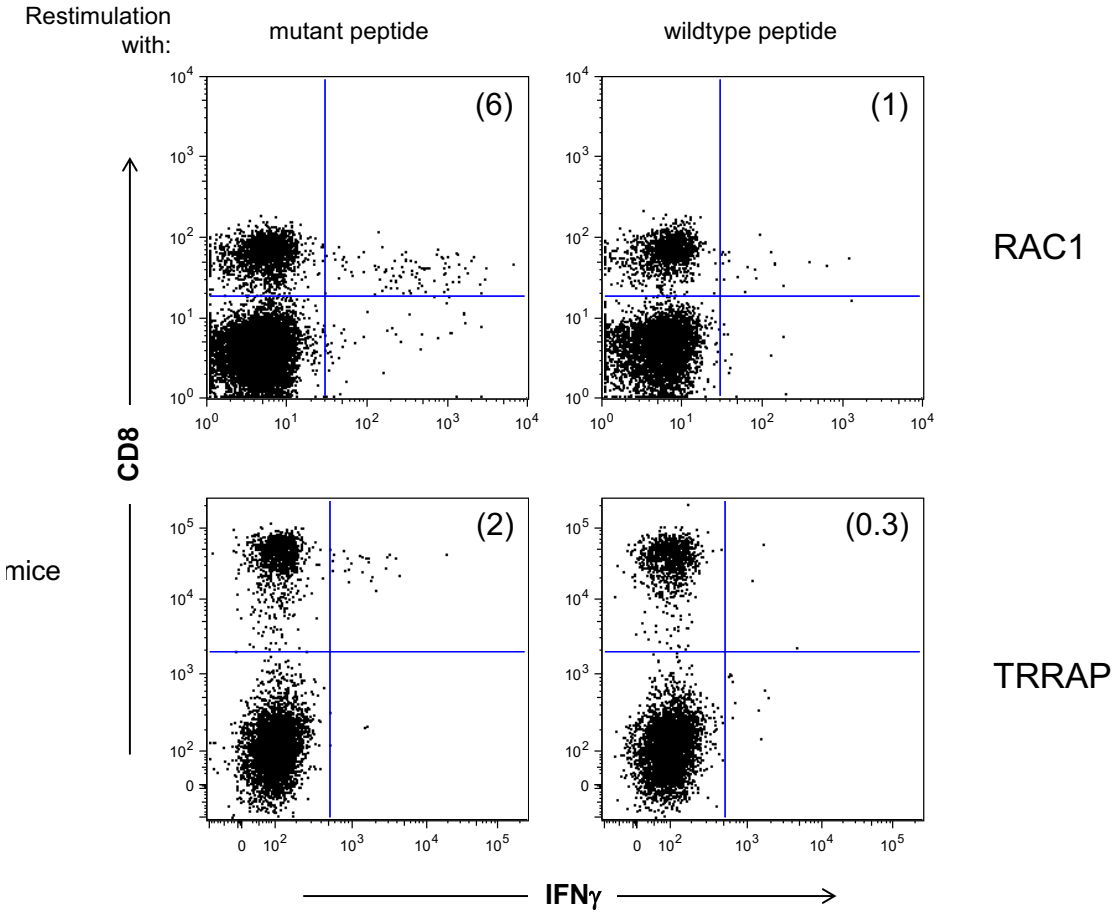

Figure 2: Recognition of predicted epitopes by $CD8^+$ T cells: Epitopes for recurrent mutations that have been identified *i*n silico to bind to HLA-A*02:01 using our pipeline were synthesized and used for immunization of human TCR transgenic ABabDII mice. Examples (RAC1:P29S and TRRAP:S722F) of ex vivo ICS analysis of mutant peptide immunized ABabDII mice 7 days after the last immunization are shown. Polyclonal stimulation with CD3/CD28 dynabeads was used as positive control, stimulation with an irrelevant peptide served as negative control (data not shown).

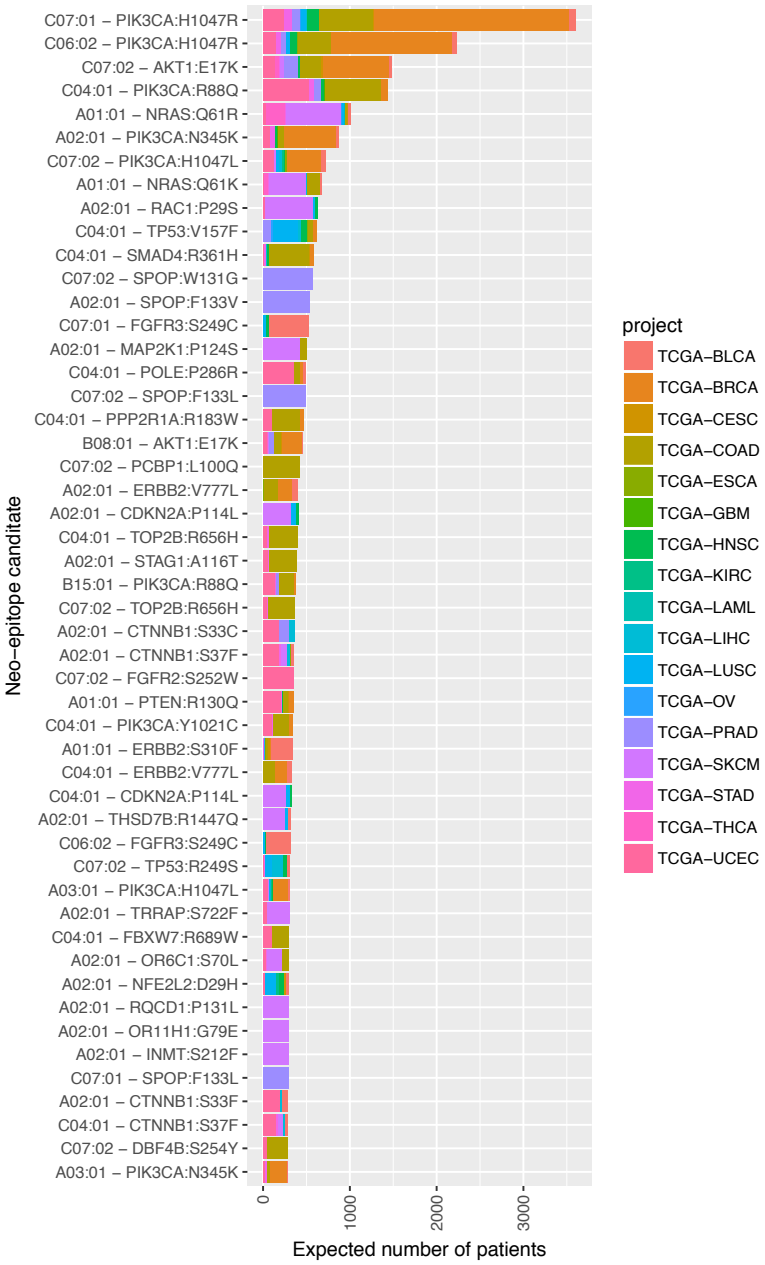

Figure 3: 50 most frequent neo-epitope and HLA-1 combinations in patients for which strong MHC I binding is predicted: for each candidate, the expected number of patients is obtained by summing over the 18 cancer entities for which the number of newly diagnosed patients in the US is available, and for which a corresponding TCGA study has been included in our analysis.

	Number of patients		Variants per patient		Missense variant per patient		Recurrent	Strong
Project name	**Total**	**With clinical data**	**Average**	**Median**	**Average**	**Median**	**Variants**	**Binders**
TCGA-BLCA	412	412	326	226	157	109	22	14
TCGA-BRCA	986	986	123	62	50	25	8	10
TCGA-CESC	289	289	358	157	143	62	17	16
TCGA-COAD	399	397	666	176	288	82	41	34
TCGA-ESCA	184	184	246	187	95	73	80	72
TCGA-GBM	393	390	212	70	93	36	15	5
TCGA-HNSC	508	508	201	139	97	66	14	3
TCGA-KIRC	336	336	79	69	33	31	0	0
TCGA-KIRP	281	281	85	82	39	38	5	2
TCGA-LAML	143	143	69	15	16	6	14	7
TCGA-LGG	508	507	70	36	33	16	14	0
TCGA-LIHC	364	364	149	120	70	58	11	16
TCGA-LUAD	567	515	367	242	180	113	7	0
TCGA-LUSC	492	492	368	301	187	153	20	19
TCGA-OV	436	435	173	121	58	47	10	7
TCGA-PAAD	178	178	168	50	77	19	24	12
TCGA-PCPG	179	179	13	12	5	4	8	5
TCGA-PRAD	495	495	59	35	27	15	3	7
TCGA-READ	137	136	475	148	232	70	320	186
TCGA-SARC	237	237	119	70	45	26	2	0
TCGA-SKCM	467	467	841	472	413	229	266	220
TCGA-STAD	437	437	488	157	211	74	17	14
TCGA-TGCT	144	128	23	21	9	8	9	6
TCGA-THCA	492	492	22	12	6	5	4	3
TCGA-THYM	123	123	39	24	10	4	6	2
TCGA-UCEC	530	530	1,672	149	708	54	118	109
TCGA-ACC	*92*	*92*	*117*	*36*	*0*	*0*	*0*	*0*
TCGA-CHOL	*51*	*45*	*110*	*62*	*0*	*0*	*0*	*0*
TCGA-DLBC	*37*	*37*	*173*	*157*	*0*	*0*	*0*	*0*
TCGA-KICH	*66*	*66*	*44*	*25*	*0*	*0*	*0*	*0*
TCGA-MESO	*82*	*82*	*47*	*44*	*0*	*0*	*0*	*0*
TCGA-UCS	*57*	*57*	*183*	*67*	*0*	*0*	*0*	*0*
TCGA-UVM	*80*	*80*	*23*	*16*	*0*	*0*	*0*	*0*
Total	**10,182**	**10,100**	**Total number:**	**3,155,183**	**Total number:**	**1,384,531**	**1,055**	**769**

Table 1: Overview of the 33 TCGA studies used in this analysis: the studies displayed in italics have not been used for the determination of recurrent variants, as the number of patients is less than 100. The number of strong binders includes all occurrences of recurrent neo-epitopes, so a variant may be counted multiple times when it is predicted to be binding several HLA-1 types.

A) Cancer entity	Study	Number of patients	HLA-A01:01 (7.61%)	HLA-A02:01 (20.36%)	HLA-A03:01 (6.60%)	HLA-A11:01 (4.37%)	HLA-B07:02 (6.51%)	HLA-B08:01 (4.80%)	HLA-B15:01 (4.46%)	HLA-C04:01 (16.69%)	HLA-C06:02 (5.72%)	HLA-C07:01 (9.28%)	HLA-C07:02 (15.39%)
Bladder Urothelial Carcinoma	BLCA	69300	340	638	197	152	142	112	89	223	376	596	383
Invasive Breast Carcinoma	BRCA	204800	110	1090	461	199	108	375	120	482	1411	2305	1234
Cervical Squamous Cell Carcinoma	CESC	14000	58	135	16	17	13	14	23	142	27	66	152
Colon Adenocarcinoma	COAD	154840	861	1628	770	760	450	222	375	2961	1019	1620	1800
Esophageal Adenocarcinoma	ESCA	4750	23	151	26	16	20	5	13	101	23	26	83
Glioblastoma Multiforme	GBM	3204	7	17	3	7	0	0	3	8	5	3	3
Head & Neck Squamous Cell Carcinoma	HNSC	58000	43	208	112	45	81	11	81	301	136	220	175
Renal Clear Cell Carcinoma	KIRC	57600	13	70	23	0	0	8	0	0	10	0	26
Papilliary Renal Cell Carcinoma	KIRP	8064	0	6	0	3	0	0	0	29	0	0	0
Acute Myeloid Leukemia	LAML	13500	77	115	0	8	49	0	0	0	48	9	73
Hepatocellular Carcinoma	LIHC	29700	12	496	131	84	69	12	7	68	69	112	174
Lung Squamous Cell Carcinoma	LUSC	66000	181	642	328	162	282	13	77	593	151	172	366
Serous Ovarian Cancer	OV	16800	26	24	33	10	15	2	14	39	9	14	36
Prostate Adenocarcinoma	PRAD	260000	120	852	69	69	34	50	70	175	387	628	1202
Melanoma	SKCM	75000	2649	7890	1817	936	861	530	203	2186	438	1000	2457
Stomach Adenocarcinoma	STAD	25000	47	172	56	66	26	8	30	206	114	166	130
Thyroid Cancer	THCA	46400	394	0	0	0	0	14	0	0	0	0	44
Endometrial Carcinoma	UCEC	55000	942	2804	817	501	369	290	493	2222	855	1602	1779
Total		**1161958**	**5904**	**16936**	**4858**	**3033**	**2517**	**1666**	**1598**	**9736**	**5080**	**8539**	**10116**
B) Number of candidates			55	91	68	64	33	24	24	48	48	50	55

Table 2: Expected number of newly diagnosed US patients by HLA-1 type and cancer entity. A) Expected number of patients of a given HLA-1 type who harbor at least one potentially immunogenic neo-epitope candidate for that HLA-1 type. Both the cancer incidence (46) and the allele frequency (45) are estimated for the US population. The probability that a patient harbors at least one variant from the set of neo-epitope candidates is computed using the assumption that the occurrence of variants in a cancer patient are statistically independent events. B) Number of neo-epitope candidates identified in the 18 studies shown in A, predicted to be a strong binder to the corresponding HLA-1 type.