The repercussions of recognition: imprints of T cells on the tumor microenvironment

Slagter, M.

Part I

The clinical utility of T cells in cancer

CHAPTER 2

Benchmarking the foreignness of human malignancies

Maarten Slagter*, Lorenzo F. Fanchi*, Marit M. van Buuren, Jorg J.A. Calis, Philip C. Schouten, Andrew Menzies, Roel G.W. Verhaak, Arno Velds, Ron M. Kerkhoven, Gergana Bounova, Sabine C. Linn, Hendrik Veelken, Michael R. Stratton, Ludmil B. Alexandrov, Lodewyk F.A. Wessels and Ton N. Schumacher *These authors contributed equally

Cell Press preprint server: 10.2139/ssrn.3279415

Abstract

Mutational load varies widely between and within malignancies and has been used as a proxy for the immunological foreignness of human cancers. However, without well-defined reference points it is difficult to determine which human tumors can be considered sufficiently foreign to the T-cell-based immune system. We established a neoantigen prediction pipeline that processes single nucleotide variants, indels and structural variants and established its precision in identifying T-cell-recognized antigens. We subsequently used this pipeline to benchmark the immunological foreignness of human cancers against that of human pathogens for which T-cell control has been established. We demonstrate that up to 50% of tumors, spanning 25 sites of origin, are more foreign than these pathogen benchmarks, due to the presentation of foreign antigens they have accumulated. These data suggest that enhancing the activity of the endogenous tumor-specific T-cell compartment through immunotherapeutic strategies may be of value for a large fraction of human cancers.

Introduction

The ability of the T-cell-based immune system to specifically recognize and destroy human cancers has attracted strong attention. Antibodies against the T-cell checkpoint molecules CTLA-4, PD-1, and its ligand PD-L1, can induce tumor regression in a range of human malignancies[1–7]. In addition, both infusion of autologous *ex vivo*-expanded tumor-infiltrating lymphocytes (TIL) and TCR gene-modified peripheral blood lymphocytes have shown activity in melanoma[8] and HPV-associated cancers[9,10]. The anti-tumor activity of these therapies is at least partially mediated by $CD8^+$ T cells, as suggested by the predictive value of pre-treatment intratumoral $CD8^+$ infiltrates[11] and as shown by the clinical activity of purified $CD8^+$ TIL[12]. Several lines of evidence indicate that T-cell recognition of neoantigens that are formed as a consequence of somatic DNA damage are a major driving force behind the activities of these therapies. First, T-cell responses against neoantigens are observed in a large fraction of patients with highly mutated tumors[6,13] and can be boosted by immunotherapy[6,14]. Second, activity of both CTLA-4 and PD-1/PD-L1 blockade is preferentially observed in tumors with above average mutational burden in a number of tumor types[15–18]. Third, and most directly, case reports that describe the infusion of T-cell products with high levels of neoantigen reactivity showed evidence of clinical activity and associated selective pressure against neoantigenic mutations[19–21].

The availability of genomic information on different human cancers has inspired efforts to estimate the immunological foreignness of these cancers through *in silico* predictions. In a landmark paper, Allison and Vogelstein postulated that many human tumors should carry neoantigens[22]. Subsequent work has similarly used genomic data to predict the presumed neoantigen burden of different human cancers[23,24], in some cases in combination with clinical response to checkpoint inhibitors[15,25,26]. While these efforts have been very valuable to describe the relative number of predicted neoantigens in different human cancers, the data have not been well-suited to draw quantitative conclusions as a grounded reference point informing on the number of epitopes sufficient to allow the formation of protective T-cell responses has been lacking.

To advance our understanding of immunological foreignness of different human cancers, we first developed an epitope prediction pipeline (Neolution) and assessed its performance using a set of experimentally identified T-cell recognized HIV epitopes. We subsequently used this pipeline to predict neoantigens for a total of 7,290 tumors obtained from the Cancer Genome Atlas (TCGA), the International Cancer Genomics Consortium (ICGC) and the Multiple Myeloma Foundation (CoMMpass) repositories, taking into account the consequences of single nucleotide variants (SNVs), insertions-deletions (indels) and structural variants, and – where applicable – included other potential sources of immunologically foreign protein sequence encoded by non-germline DNA sequences, such as proteins encoded by oncogenic viruses and the B-cell receptor idiotype. In this manner, we transformed the entire pool of genomically novel sequences of individual tumors to the foreign antigen space (FAS), allowing for their direct comparison to a grounded, viral context. To generate such

context, we used the same prediction pipeline to analyze proteins from three pathogens for which T-cell control has been shown to be clinically relevant: the *E6* and *E7* oncogenes from the human papilloma virus (HPV)[27], the *LMP-1* and *LMP-2* oncogenes from the Epstein-Barr virus (EBV)[28], and the HIV-1 genome[29,30].

We demonstrate that a large fraction of human cancers has evolved to express a sufficient number of foreign antigens to allow T-cell recognition. Tumors positive for a viral integration are shown to be especially immunologically foreign. These data provide a strong incentive for the further development of therapeutic strategies that aim to expand (neo-)antigen-specific T-cell reactivity in not only highly, but also modestly mutated human tumor types, especially those of viral aetiology.

Results

Robust estimates of DNA damage- and virus-derived antigen loads

To generate a pan-cancer overview of the foreign antigen space of human tumors, we selected pre-treatment samples from 7,268 patients across 25 tissues of origin, covering 42 tumor subtypes, based on the availability of patient-matched DNA and RNA sequencing data (Figure 2.1, Table for list of patient identifiers and annotations). We reconstructed tumor transcripts using SNVs, indels and, for a subset of patients from whom these data were available, large-scale gene fusion events in order to obtain a set of candidate tumor-specific neoantigens. For both SNVs and indels, candidate peptides whose genomic sequences were affected by multiple mutations were modified to reflect the consequences of all variants (154,731 of 8,818,152 candidate peptides, 1.75%). Conversely, structural variants were analyzed in isolation as these were found not to substantially overlap with focal DNA damage (i.e. missense mutations and indels; only 9 out of 20,140 fusion events affected, Figure S2.1A). NMD-targetability of transcripts was predicted and targeted transcripts and their associated candidate peptides were excluded, where indicated. Finally, candidate neoantigens were annotated with the output of our four-filter epitope prediction pipeline that models the major requirements for (neo-)antigen presentation individually: expression of the mutant DNA sequence, predicted proteasomal processing, predicted HLA-binding, and self-similarity (Figure S2.1A). RNA expression was assessed using sample-matched gene expression levels from the matched tumor, requiring at least one mapped read to be considered expressed, unless indicated otherwise. HLA-binding was assessed with the frequently observed HLA-A*02:01 allele for which binding affinity predictions are most accurate[31]. Using a single HLA-allele rather than patient-matched HLA-alleles allows for uniform prediction accuracy across patients. With any of four other HLA-alleles that are also well-predicted, we did not get substantially different neoantigen loads (Figure S2.1B). From 959,792 non-synonymous DNA mutation events, we predicted 80,234 candidate peptides that passed all four filters.

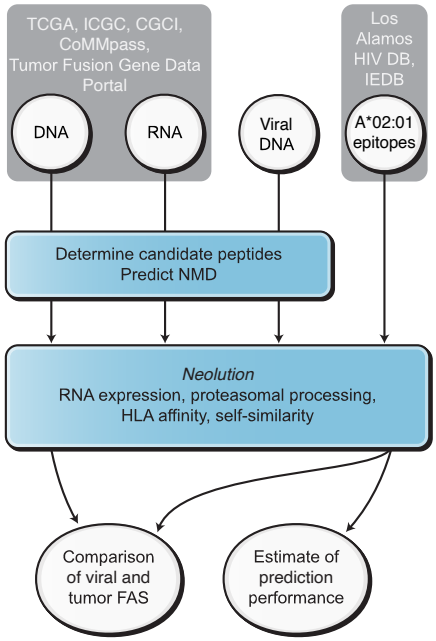


Figure 2.1: Study setup and methodology.
Tumor genome sequencing data and viral genome sequences were used as input for the indicated neo-antigen prediction pipeline, yielding tumoral and viral foreign antigen loads for direct comparison.

Focal DNA damage dominates the foreign antigen space of most tumors

Previous neoantigen identification efforts have mostly focused on SNVs rather than other classes of DNA damage like indels and gene fusions[14,23,24]. Frameshifting indels and loss-of-stop mutations can result in open reading frame extensions, where novel coding sequences ultimately become accessible to RNA translation, forming an additional source of neo-epitopes [23,32]. Similarly, while in-frame gene fusion events yield immunologically novel DNA sequences that are restricted to the fusion breakpoint, out-of-frame fusions additionally lead to the formation of downstream novel open reading frames (ORFs) until an in-frame stop codon is encountered. As can be expected, peptides derived from these novel ORFs were more likely to be dissimilar from the self-ligandome and thus have a higher probability of being T-cell recognized: of all affinity-filter-passing candidate peptides 74.0% (95% CI: [73.5%, 74.5%]) and 61.0% (95% CI: [60.8%, 61.2%]) derived from frameshifting indels and missense mutations, respectively, were classified as dissimilar from self. Next, we determined the propensity of each mutation class to generate neo-epitopes by assessing the ratio between the number of neoantigens contributed by the class and the prevalence of that class, resulting in a mutation-class-specific neoantigen yield rate. As expected, we found DNA damage types to vary significantly

in their neoantigen yield rates (Figure 2.2A) in a manner highly consistent across tumor types (Figure S2.2A) and HLA alleles. Loss-of-stop mutations generated 4.99-fold more neoantigens per mutation than missense mutations. The neoantigen yield rate of frameshifting indels only minimally exceeded that of missense mutations when accounting for NMD (1.1-fold), but yielded 3.1-fold more neoantigens when omitting the NMD-filter (Figure 2.2A). In spite of this observed variation in neoantigen yield rate between DNA damage classes, the absolute contributions of different mutation classes to the predicted antigenome is predominantly determined by their prevalence. Specifically, due to the high abundance of missense mutations, this class dominates the predicted antigenome for all investigated tumor types, despite its relatively modest neoantigen yield rate. This holds true both when NMD is not taken into account (Figure 2.2B, top) and when it is by removing predicted neoantigens derived from transcripts identified as NMD-targeted (Figure 2.2B, bottom).

Many of the mutational processes underlying tumorigenesis (e.g. UV radiation, tobacco smoke and loss of DNA repair mechanisms) are associated with unique patterns of SNVs, termed mutational signatures[33]. To assess possible differences in neoantigenicity between mutational processes, we inferred the likelihood for SNVs to have been caused by any of the identified processes and associated signatures for 6,504 tumor samples across 36 tumor types (methods, Table). This calculation was performed for each SNV individually taking into account the signature's tendency to cause the SNV and the total abundance of the signature. We observed small and statistically insignificant differences in HLA-A*02:01 neoantigen yield rates between signatures within tumor types (lowest FDR-adjusted p-value is 0.214 for lung squamous cell carcinoma, Figure S2.2B) indicating that mutational processes do not differ in their ability to yield SNV-associated HLA-A*02:01 (or other alleles, data not shown) neoantigens.

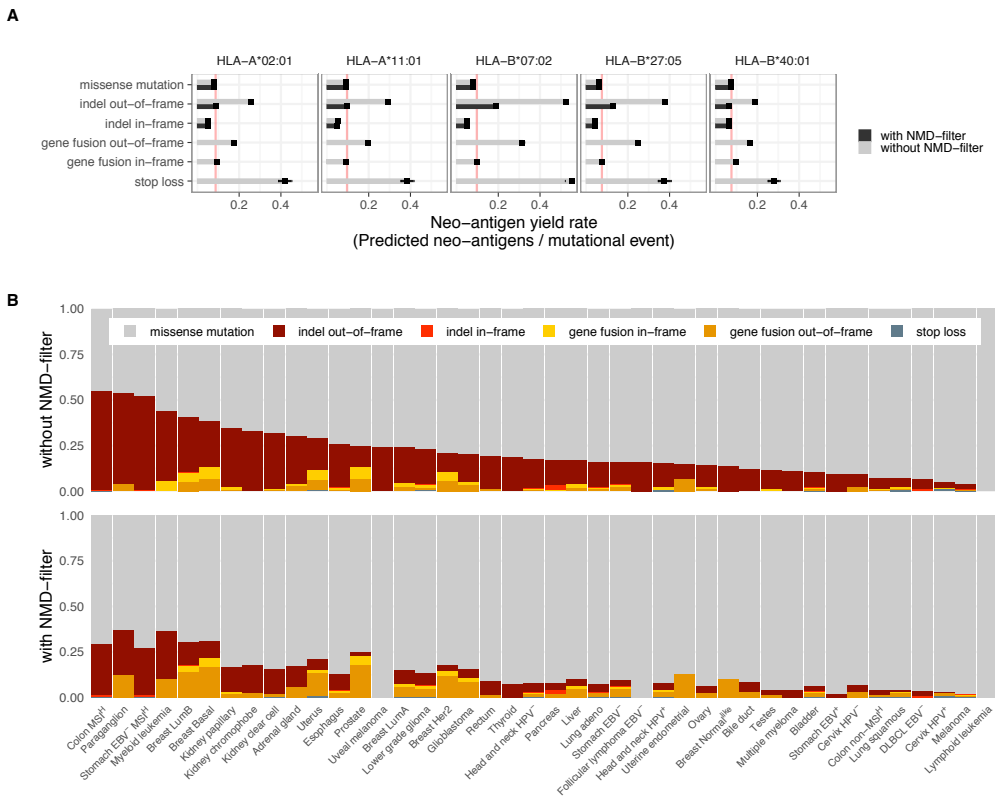


Figure 2.2: Neo-antigen contribution and neo-antigen yield rates of different types of DNA damage
A Neo-antigen yield rates (the mean number of predicted neoantigens per mutation) for five different HLA alleles and different classes of DNA damage. Weighted mean yield rates indicated by squares, surrounding lines denote 95% CIs.
B Proportional contribution of different types of DNA damage to the DNA damage-associated foreign antigen loads across human tumors, both without (top) and with (bottom) the NMD-filter applied. In the bottom panel, gene fusion and stop loss variants were assumed not to be affected by NMD.

Viral benchmarks to compare DNA-damage derived antigen loads with

Prior analyses have demonstrated that the number of predicted neoantigens correlates well with the number of genomic alterations[6,24,26]. However, in absence of a benchmark for the FAS known to suffice for T-cell recognition, it has been difficult to determine for which human tumors the foreign antigen pool is sufficiently large to elicit T-cell reactivity. To provide such a benchmarked analysis of the foreign antigen loads in human tumors, we selected three viral references that have been shown to provide sufficient genomically foreign T-cell antigens to yield clinically meaningful responses. First, the E6 and E7 oncoproteins from human papilloma virus (HPV) which are expressed

in cervical cancer, head and neck cancer, and ano-genital cancers, but also in premalignant neoplastic lesions. Vaccination of patients with premalignant vulvar intraepithelial neoplasia with pools of overlapping HPV16 E6 and E7 peptides has previously been demonstrated to lead to regression of lesions in more than 80% of patients[27]. Thus, clinically meaningful T-cell reactivity against epitopes within the E6 and E7 oncoproteins can apparently be induced in a large fraction of patients. Second, the EBV LMP-1 and LMP-2 oncogenes which are causally involved in EBV-induced transformation of a number of tissue types and the expression of these genes continues after initial cellular transformation. Evidence for clinically relevant immunogenicity of LMP-1/LMP-2 is provided by the observation that T-cell products generated against these two proteins mediate regression of a large fraction of EBV-positive relapsed or refractory Hodgkin and non-Hodgkin lymphomas (OR of 62%, CR of 52%; (Bollard et al., 2014; Heslop et al., 2010). Third, as a non-cancer related viral reference, numerous CTL responses have been identified in patients infected with HIV[34]. Furthermore, reports on both humans and non-human primates have shown a correlation between HIV-specific $CD8^+$ T cells and control of viremia[29,30], and evidence for CTL-mediated control of HIV-1 is also provided by the association of slow disease progression with certain HLA alleles and presence of conserved T-cell epitopes[34–37].

To use these viral gene sets as benchmarks, their predicted foreign antigen loads are required to scale similarly to viral immunogenicity as those from human genes. As gene expression influences cell surface antigen abundance[38–40] and as such likely T-cell recognition probability, structural differences between viral and human gene expression would complicate the intercomparison of foreign antigen loads. We compared the expression levels of the benchmark viral oncogenes and human protein-coding genes in tissue samples maximally similar to those in the studies we used to define our benchmarks. To the best of our knowledge, RNA sequencing data of VIN lesions used for the HPV E6/E7 benchmark is not available, so we analyzed HPV-positive high-grade cervical intraepithelial neoplasia (CIN) lesions. Here we found the average expression of HPV E6/7 to be slightly lower (0.77-fold difference in medians, $p = 0.83$; Figure S2.3A) than human genes, which is unlikely to be of discernible effect on cell surface antigen abundance as RNA expression levels vary over 4 orders of magnitude (Fig S3A). In an analogous analysis of EBV^+ diffuse large B-cell lymphomas, average expression of EBV *LMP-1* and *LMP-2* was found to be ~7.7-fold lower (0.13-fold difference in medians, $p = 0.2$; Figure S2.3B) than human genes. The trend towards somewhat lower expression of the viral oncogenes, renders these oncogenes a reasonable and possibly slightly conservative benchmark for the foreign antigen load known to suffice for the induction of clinically relevant T cell reactivity.

Benchmarking the foreign antigen space of human malignancies against human pathogens

To generate a comprehensive estimate of immunological foreignness across cancer types, we combined the predicted HLA*A:02:01-neo-antigen loads resulting from somatic DNA damage, with those

of non-germline sequences in the rearranged B-cell receptor in case of B-cell malignancies and of expressed viral genes for virus-positive cancers. For the latter class of genes, viral contributions were restricted to genes expressed within the top three quartiles of the human transcriptome. As we previously found neoantigen depletion to be incomplete at best (Chapter), we analyzed HLA-A*02:01 neo-antigen predictions as a proxy for all class I neo-antigens for all patients, regardless of their HLA repertoires.

Comparing the foreign antigen loads of human tumors to those of our viral benchmarks, we found large overall consistency with observed clinical successes of cancer immunotherapy obtained so far[41], while simultaneously also highlighting tumor types worthy of further exploration in terms of immunotherapeutic targeting of foreign antigens. 16 out of the 42 evaluated tumor types had a median foreign antigen load below the lowest of the 3 benchmarks, the HPV E6/E7 oncogenes, which were enriched for hematological and neurological tumors (Figure 2.3A). Our current analysis provides no evidence for the average foreign antigen repertoire of these tumors to be sufficiently sized for clinical actionability. The upper side of the foreignness spectrum is composed of cancers with foreign antigen score that are larger than the HPV E6/E7 benchmark, comprising 53% of all assayed samples. Here, 4/26 tumor types and 9% of analyzed samples had higher median foreign antigen loads than the EBV LMP-1/LMP-2 proteins. In view of the profound clinical activity of T cell products directed against these antigens, these data provide strong evidence that the repertoire of neo-antigens present in these tumors should commonly suffice to allow strong T cell reactivity. Notably, virus-positive tumors, including HPV^+-tumors, all displayed median foreign antigen loads higher than that of the EBV LMP-1/LMP-2 oncoprotein-benchmark. Foreignness benchmarks were robust to variation of the MHC-affinity thresholds of the prediction pipeline, keeping the other pipeline parameters constant (Figure 2.3B).

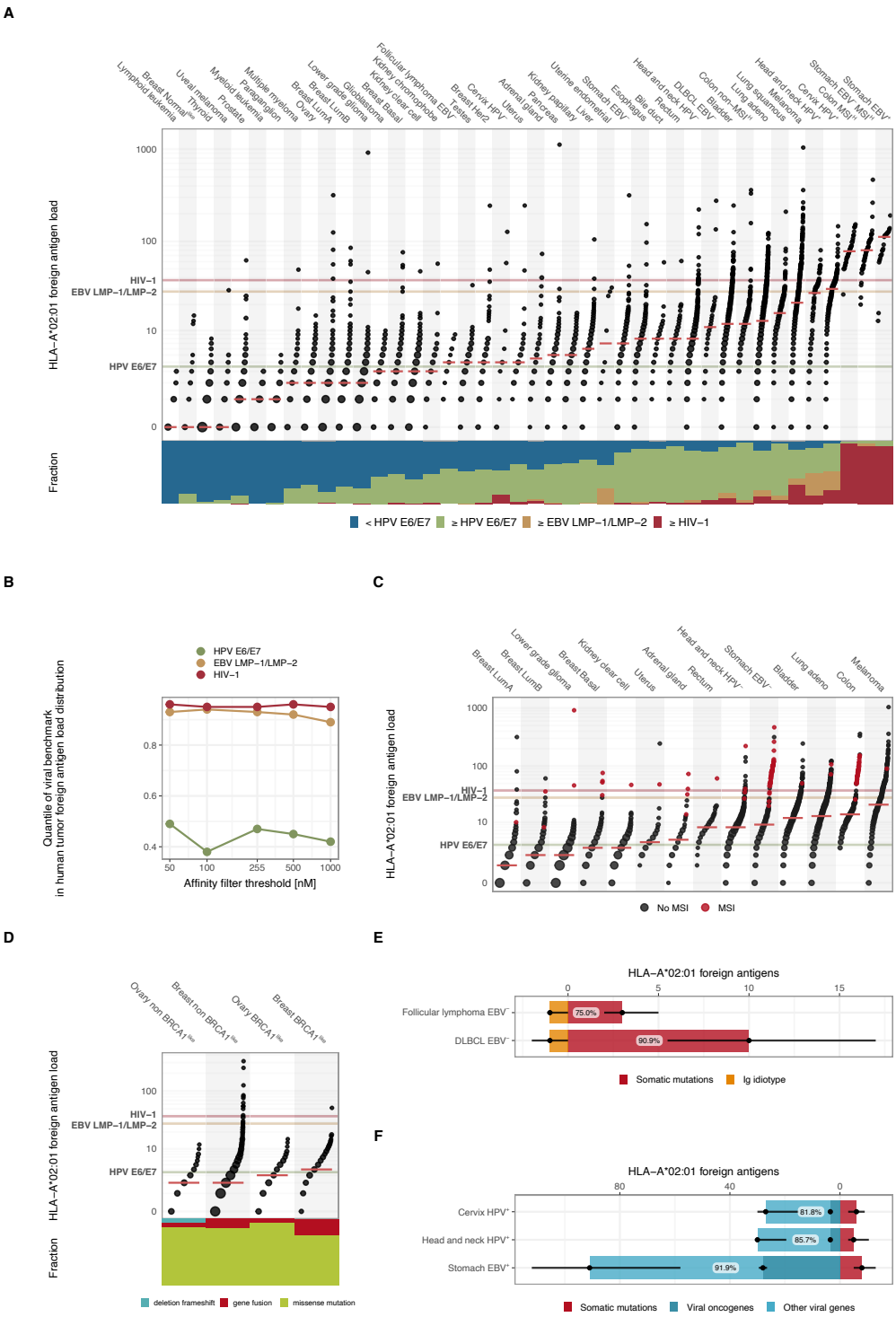

A
HLA–A*02:01 foreign antigen load
HIV–1
EBV LMP–1/LMP–2
HPV E6/E7
Fraction
< HPV E6/E7
≥ HPV E6/E7
≥ EBV LMP–1/LMP–2
≥ HIV–1
B
Quantile of viral benchmark in human tumor foreign antigen load distribution
HPV E6/E7
EBV LMP–1/LMP–2
HIV–1
Affinity filter threshold [nM]
C
HLA–A*02:01 foreign antigen load
No MSI
MSI
D
HLA–A*02:01 foreign antigen load
Fraction
deletion frameshift
gene fusion
missense mutation
E
HLA–A*02:01 foreign antigens
Follicular lymphoma EBV⁻
DLBCL EBV⁻
75.0%
90.9%
Somatic mutations
Ig idiotype
F
HLA–A*02:01 foreign antigens
Cervix HPV⁺
Head and neck HPV⁺
Stomach EBV⁺
81.8%
85.7%
91.9%
Somatic mutations
Viral oncogenes
Other viral genes

Figure 2.3: Pan-cancer overview of immunological foreignness
A (Top) Pan-cancer foreign antigen loads contrasted to the foreign antigen load of three viral gene sets known to allow the formation of protective T-cell responses. Tumoral foreign antigen loads are composed of neoantigens generated through focal DNA damage, virus-derived antigens (virus-positive tumors only) and immunoglobulin idiotype-derived antigens (B-cell malignancies only). Fusion gene contributions are omitted as these could not be evaluated for all patients and their contribution was modest (see Figure 2.2). Dot size reflects the number of samples on a coordinate, red lines indicate median foreign antigen load per tumor type. (Bottom) Distributions of tumoral foreignness relative to indicated viral benchmark gene sets.
B Robustness of the benchmarking of tumor foreignness to the peptide MHC-affinity threshold used in neoantigen predictions. Quantiles of the three viral benchmarks within the distributions of tumoral foreign antigen loads are depicted.
C Distribution of microsatellite instable tumors across indicated tumor types. Only those TCGA-tumor types in which at least one MSI^{H}-case was detected (shown as red dots) are depicted.
D Foreign antigen loads and compositions of breast and ovarian cancers classified by presence or absence of BRCA1-like signature characteristics in copy number profiles (Schouten et al., 2015).
E Foreign antigen contributions of focal DNA damage and immunoglobulin idiotype sequences in B-cell lymphomas. Black dots denote medians, black bars span the 25% and 75% percentiles. Percentage labels denote the contribution of viral genes to the overall foreign antigen load.
F Foreign antigen contributions of focal DNA damage and either expressed HPV genes in HPV^{+} head and neck and cervical tumors, or expressed EBV genes in EBV^{+} stomach adenocarcinomas. Black dots denote medians, black bars span the 25% and 75% percentiles. Percentage labels denote the contribution of viral genes to the overall foreign antigen load.

Mismatch repair deficiency enlarges tumor foreignness

Consistent with the high response rates to checkpoint blockade[15,16], mismatch-repair deficient colorectal and non-colorectal cancers are predicted to have very high neoantigen loads, deriving from somatic mutations in repetitive regions of the genome, termed microsatellites (Figure 2.3C). We assessed the microsatellite instability (MSI)-status of all TCGA samples in our cohort with a classifier trained on mutation type frequencies[42] and found MSI in 12/29 assayed tumor types (1.8% of assayed samples) and that the foreign antigen loads of 91% of MSI^{H}-tumors are higher than that of our upper benchmark, the full HIV genome (Figure 2.3C). This indicates that MSI-tests can be used to directly identify highly foreign tumors across tumor types. Although mismatch-repair proficient colon cancers had lower neoantigen loads than their mismatch-repair deficient counterparts, virtually all tested exceeded our lower HPV E6/E7 benchmark. This suggests that the low response rate to PD-1 blockade that has been observed in this tumor type[15] cannot solely be explained by a lack of immunological foreignness, suggesting an enrichment of immune evasion mechanisms there.

Defects in homology-mediated DNA repair augment foreignness

Defects in homology-mediated DNA repair (e.g. by loss of function of the *BRCA1/2* genes) have been associated with large-scale genomic rearrangements such as inter- and intrachromosomal gene fusion events in breast and ovarian cancers[43–45]. These rearrangements can yield protein-coding tran-

scripts with fusions either occurring in-frame, leading to novel junctional sequence at the boundary of the two fused exons, or out-of-frame, generating complete stretches of novel coding sequences. To assess whether defects in homologous recombination DNA repair are associated to the complete foreign antigen load, we stratified patients of the TCGA breast and ovarian cancer cohorts according to their 'BRCA-like' status, as assessed with a 'BRCA1-like' classifier trained to recognize DNA copy number aberration profiles characteristic for *BRCA1*-mutants[46]. BRCA1-like breast cancer was significantly more foreign than non-BRCA1-like breast cancer (1.75-fold difference in medians, $p = 5 \times 10^{-10}$), reaching median foreign antigen loads equal to our lower benchmark HPV E6/E7 (Figure 2.3D, top), partially due to a higher presence of gene fusions (Figure 2.3D, bottom). Similarly, BRCA1-like patients were significantly more foreign than non-BRCA1-likes in the ovarian carcinoma cohort (1.5-fold difference in medians, $p = 6 \times 10^{-3}$), but neither group's median reached the lower HPV E6/7 benchmark. Consistent with our observations, an association has been found between *BRCA1/2*-mutation status and elevated neo-antigen loads in high-grade serous ovarian cancer[47].

Composition of foreign antigens informs on novel potential therapeutic targets

As vaccination strategies in cancers with viral etiology have primarily focused on the targeting of viral antigens, we wished to explore how the foreign antigen loads of viral oncogenes in these tumors compares to that formed as a consequence of somatic mutations. Follicular B-cell lymphomas, despite being very low in exonic non-synonymous mutations (median: 30, range: [9, 79]), get the primary contribution to their foreignness from neoantigens originating from somatic mutations (Figure 2.3E). Due to their low numbers, idiotype-derived antigens (median: 1, range: [0, 4]) are less attractive therapeutic targets, compared to the more abundant DNA damage-derived neoantigens. Similarly, the neoantigens generated by somatic mutations in the highly mutated diffuse large B-cell lymphomas (DLBCL) easily surpass those of idiotypic sequences in these tumors, and it will be interesting to establish whether long-term responses of DLBCL following CD19 CAR T-cell therapy[48] are in some cases accompanied by induction of neoantigen-specific T-cell reactivity. On the contrary, for HPV^+ cervix and head & neck and EBV^+-stomach cancers, the majority of the foreign antigen repertoire is made up by antigens of viral origin (Figure 2.3F). even when restricting the viral contribution to the LMP-1/LMP-2 oncogenes

Gene fusions have been suggested as another potentially important source of foreign antigens[49,50]. We detected 11426 gene fusion events in 3359 out of 6049 investigated tumor samples across 33 malignancies. Likely due to their low prevalence, the overall contribution of fusion genes to foreignness was modest (6.8% of total foreignness; 95% CI: [6.3%, 7.2%]). This suggests low prioritization of immunotherapeutic solely and specifically targeting fusion gene antigens. Notable exceptions to this include prostate, uterine and some breast cancers (Luminal B and Basal subtypes) for which fusion-derived neoantigens increased the foreignness to surpass that of our lower HPV E6/E7 benchmark

(Figure S2.4A).

As we found NMD to have a substantial impact on neoantigen load (Figure 2.2A), especially for antigens derived from indels, a potential immunostimulatory measure would be to inhibit NMD to augment the targetable antigenome[51,52]. By including epitopes from NMD-flagged transcripts in our foreignness estimates, we find that an average of 10% (95% CI: [10%, 11%]) of HLA-A*02:01 neoantigens could be gained (Figure S2.4B). As expected, the highest gains were in indel-rich, mismatch-repair deficient tumor types (colon MSI^H: 32.2%, stomach MSI^H: 31.2%, Figure S2.4B). The evaluated kidney cancers, through their elevated indel loads, additionally showed gains of 16% (95% CI: [14%, 18%]) in foreignness when accounting for NMD-blockade, averaged over all three evaluated subtypes. We found up to 14% of clear cell kidney tumors to have potentially clinically impactful increases in foreignness, as they met or surpassed the HPV E6/E7 benchmark after NMD-blockade (Figure S2.4B).

Accounting for intratumoral heterogeneity consistently lowers foreign antigen space

Clonal antigens that are presented by all tumor cells in a lesion can reasonably be expected to have a larger contribution to tumor regression than subclonal antigens, and the observed inverse relationship between tumor heterogeneity and immunotherapy outcome provides indirect support for the superior value of clonal antigens as T-cell targets[53]. We aimed to address how our foreigness estimates would be affected when accounting for intratumoral heterogeneity.

To account for tumor heterogeneity in our antigenome estimates, we devised two ways of incorporating cellularity-estimates in our predicted neoantigen loads. In a first, conservative approach, we fully restricted predicted antigenomes to those peptides that are derived from clonal mutations (i.e. with an inferred cellularity $\geq .95$[53,54]). In this, we assumed the contribution of viral genes to be shared by all cells in HPV^+ and EBV^+ cancers due to their role in cellular transformation. We observed an average 51.6% (95% CI: [50.5%, 52.7%]) decrease in these heterogeneity-aware foreign antigen loads as compared to the heterogeneity-unaware foreign antigen loads presented earlier (Figure 2.4A), decreasing the number of tumor types with median foreign antigen loads equal or exceeding our lower benchmark HPV E6/E7 from 12/18 to 8/18 (Figure S2.5A). The above analysis ignores that subclonal mutations can be immunogenic and could contribute to tumor rejection, limited to the subclones harboring them. Indeed, T-cell recognition of subclonal mutations has been reported[55]. This inspires an approach in which neoantigens are weighted by the cellularity of their associated somatic mutations. We found cellularity-weighted estimates of the foreign antigen loads to be 23.5% lower (95% CI: [22.7%, 24.3%], when compared to heterogeneity-unaware foreign antigen loads (Figure 2.4A,B), more mildly decreasing the number of tumor types with median foreign antigen loads equal or exceeding our lower benchmark HPV E6/E7 from 12/18 to 10/18. Interestingly, the decrease in foreign antigen load when accounting for ITH is not constant across tumor types. Whereas

the antigenic mutanomes of all breast cancer subtypes and melanoma were relatively rich in nearly clonal mutations and hence experienced only mild decreases in foreign antigen loads (median decreases ranging between 8.2% and 13.6%), thyroid cancer on the other side of the spectrum underwent a median decrease of 42.2% (Figure 2.1A). Read coverage depth in next-generation sequencing is known to influence the sensitivity of variant callers to (subclonal) mutations[56]. However, the variation in effect of cellularity weighting was not explained by differences in median read coverage of variant loci (Figure S2.5B).

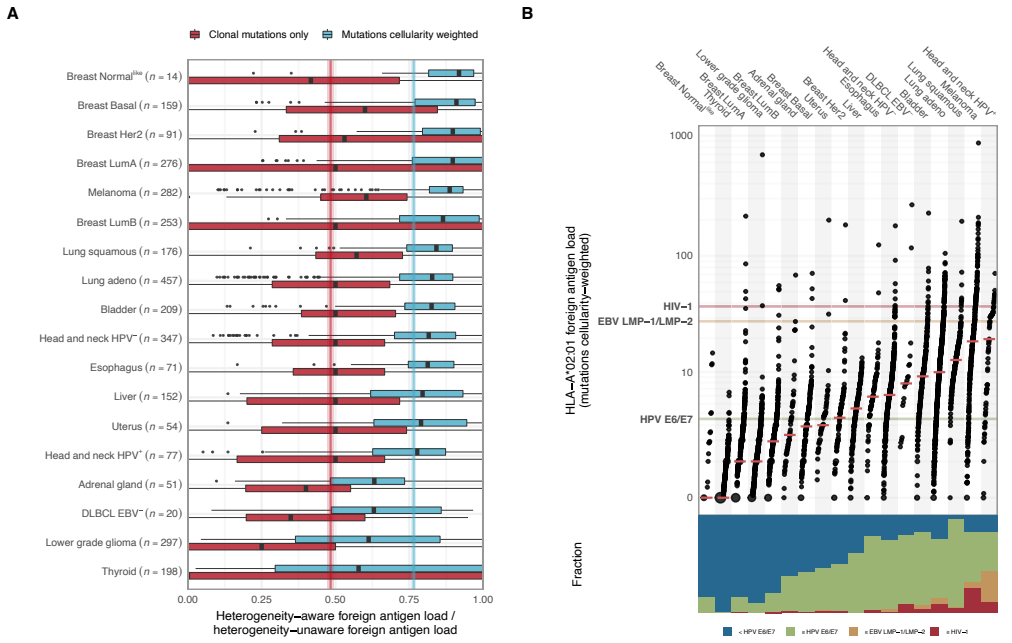


Figure 2.4: Effect of tumor heterogeneity on foreign antigen loads
A. Distributions of the ratio between heterogeneity-aware and -unaware somatic mutation-derived foreign antigen estimates. Colored lines reflect the average decrease in foreign antigen loads when either restricting the computation of the foreign antigen loads to clonal mutations (52%, 95% CI: [50%, 53%, red), or when weighting mutations according to their cellularity (23%, 95% CI: [23%, 24%], blue line).
B. Pan-cancer overview of foreign antigen loads in which neo-antigens contributed by focal DNA damage types are weighted by their estimated cellularity and combined with viral and idiotype antigen contributions.

Discussion

We devised an unbiased method to estimate the immunological foreignness resulting from different forms of DNA damage and oncogenic viral insertions in human cancers, and then benchmarked these

results against the foreignness of human pathogens for which clinically relevant T-cell recognition has convincingly been demonstrated. The main conclusion of this work is that a substantial fraction of human cancers expresses a sufficient number of neo-antigens to allow T cell recognition.

By establishing the precision of our epitope prediction pipeline, and by using the same prediction pipeline on foreign genomes known to contain T cell targets, we removed concerns around the unknown false positive rates that have complicated the interpretation of earlier antigenome landscape efforts. One striking outcome of this benchmarking is that a substantial set of human cancers for which no forms of immunotherapy have been approved to date (e.g. MSS colon, bile duct, MSS endometrial, pancreatic, adrenocortical, and uterine cancers) do contain a larger number of predicted neo-antigens than the HPV E6 and E7 oncogenes that have successfully been used as therapeutic targets in premalignant disease. One interpretation of these data is that defects at other stages in the cancer immunity cycle/cancer immunogram[57,58] are interfering with T cell-mediated tumor control in these tumor types. For tumor types, such as MSS colon cancer, that do commonly encode a number of neo-antigens that lies above that of the HPV E6 and E7 oncogenes, it will be of interest to determine whether inefficient T cell priming, or the immunosuppressive molecules such as TGF-β, may for instance explain the poor responsiveness to PD-1/ PD-L1 blockade.

In future efforts, the accuracy of the estimates of foreignness that we provide may be further improved in several ways. First, the contribution of multi-basepair indels and gene fusion events is potentially underestimated as sequencing and the employed variant calling methodologies to detect these events were less robust than for SNVs[59–61]. Effect size of such optimization would be expected to be most significant for tumor types that are characterized by chromosomal instability and by relatively low degrees of focal DNA damage, such as *BRCA1/2* deficient cancers. A second point of improvement will be to include the weakly compensatory interaction between MHC-binding affinity and transcription levels in cell surface antigen abundance, as shown by mass-spectrometry of HLA-eluted peptides[38,40,62,63]. However, as gene expression levels between the viral HPV and EBV benchmarks and the tumor-associated transcripts differ only minimally, we expect the effect of inclusion of a weighted measure of RNA expression/MHC affinity to be small. Analysis of the contribution of subclonal versus clonal antigens to the total pool of predicted neo-antigens revealed that for tumor types such as breast and melanoma accounting for intratumoral heterogeneity only minimally alters the size of this pool whereas for thyroid carcinoma more than half of patients completely lost their foreign antigen repertoire. In particular for this tumor type it will be important to understand to what extent subclonal T cell antigens can contribute to tumor control, for instance by development of mouse models with a fixed level of clonal and increasing levels of subclonal neo-antigens.

Based on the data from the Melief and Bollard/Rooney groups it is apparent that induction of T cell reactivity against HPV E6/ E7 and EBV LMP1/LMP2 frequently results in regression of premalignant HPV$^+$ lesions and EBV$^+$ lymphomas, respectively[27,28]. An important and incompletely resolved question is whether these therapy-induced T cell responses are by themselves sufficient, or whether the observed clinical responses are at least partially due to epitope spreading. In case of the HPV anti-

gens, such a dominant role of novel T cell responses as a consequence of epitope spreading appears less likely, as the premalignant lesions that respond to vaccination[27] are unlikely to have accumulated substantial numbers of somatic mutations[64]. Nevertheless, a better understanding of epitope spreading in therapy-induced tumor control would be of value.

In earlier work by Turajlic and colleagues, the neo-antigen yield rate of frameshifting indels was found to be 9-fold higher than that of non-synonymous SNVs[32], whereas in our analyses these DNA mutation types differed only minimally in this respect. The discrepancy between these data likely stems from the omission of a model for NMD and the use of a model for thymic negative selection of neo-antigen-reactive T cells that is strict on SNV-derived peptides but not applied to peptides originating from indels, whereas SNV and indel-derived epitopes in our work were both passed through the similarity-to-self filter.

We have provided a pan-cancer overview of immunological foreignness that predicts a potential value of cancer immunotherapies that goes substantially beyond the tumor types for which T cell checkpoint currently forms an approved strategy, and that includes common cancers such as MSS CRC. Analysis of the mechanisms that hold back T cell reactivity in these tumors and strategies to circumvent these mechanisms will be of major interest.

Bibliography

1. Brahmer, J., Reckamp, K.L., Baas, P., Crino, L., Eberhardt, W.E., Poddubskaya, E., Antonia, S., Pluzanski, A., Vokes, E.E., Holgado, E., et al. (2015). Nivolumab versus Docetaxel in Advanced Squamous-Cell Non-Small-Cell Lung Cancer. N Engl J Med *373*, 123–135.

2. Ferris, R.L., Blumenschein, Jr., G., Fayette, J., Guigay, J., Colevas, A.D., Licitra, L., Harrington, K., Kasper, S., Vokes, E.E., Even, C., et al. (2016). Nivolumab for Recurrent Squamous-Cell Carcinoma of the Head and Neck. N Engl J Med *375*, 1856–1867.

3. Topalian, S.L., Sznol, M., McDermott, D.F., Kluger, H.M., Carvajal, R.D., Sharfman, W.H., Brahmer, J.R., Lawrence, D.P., Atkins, M.B., Powderly, J.D., et al. (2014). Survival, durable tumor remission, and long-term safety in patients with advanced melanoma receiving nivolumab. J Clin Oncol *32*, 1020–1030.

4. Gettinger, S.N., Horn, L., Gandhi, L., Spigel, D.R., Antonia, S.J., Rizvi, N.A., Powderly, J.D., Heist, R.S., Carvajal, R.D., Jackman, D.M., et al. (2015). Overall Survival and Long-Term Safety of Nivolumab (Anti-Programmed Death 1 Antibody, BMS-936558, ONO-4538) in Patients With Previously Treated Advanced Non-Small-Cell Lung Cancer. J Clin Oncol *33*, 2004–2012.

5. 5. Motzer, R.J., Escudier, B., McDermott, D.F., George, S., Hammers, H.J., Srinivas, S., Tykodi, S.S., Sosman, J.A., Procopio, G., Plimack, E.R., et al. (2015). Nivolumab versus Everolimus in Advanced Renal-Cell Carcinoma. N Engl J Med *373*, 1803–1813.

6. Rizvi, N.A., Hellmann, M.D., Snyder, A., Kvistborg, P., Makarov, V., Havel, J.J., Lee, W., Yuan, J., Wong, P., Ho, T.S., et al. (2015). Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. Science *348*, 124–128.

7. Robert, C., Long, G.V., Brady, B., Dutriaux, C., Maio, M., Mortier, L., Hassel, J.C., Rutkowski, P., McNeil, C., Kalinka-Warzocha, E., et al. (2015). Nivolumab in previously untreated melanoma without BRAF mutation. N Engl J Med *372*, 320–330.

8. Rosenberg, S.A., and Restifo, N.P. (2015). Adoptive cell transfer as personalized immunotherapy for human cancer. Science *348*, 62–68.

9. Draper, L.M., Kwong, M.L., Gros, A., Stevanovic, S., Tran, E., Kerkar, S., Raffeld, M., Rosenberg, S.A., and Hinrichs, C.S. (2015). Targeting of HPV-16+ Epithelial Cancer Cells by TCR Gene Engineered T Cells Directed against E6. Clin Cancer Res *21*, 4431–4439.

10. Stevanovic, S., Draper, L.M., Langhan, M.M., Campbell, T.E., Kwong, M.L., Wunderlich, J.R., Dudley, M.E., Yang, J.C., Sherry, R.M., Kammula, U.S., et al. (2015). Complete regression of metastatic cervical cancer after treatment with human papillomavirus-targeted tumor-infiltrating T cells. J Clin Oncol *33*, 1543–1550.

11. Tumeh, P.C., Harview, C.L., Yearley, J.H., Shintaku, I.P., Taylor, E.J., Robert, L., Chmielowski, B., Spasic, M., Henry, G., Ciobanu, V., et al. (2014). PD-1 blockade induces responses by inhibiting adaptive immune resistance. Nature *515*, 568–571.

12. Dudley, M.E., Gross, C.A., Somerville, R.P., Hong, Y., Schaub, N.P., Rosati, S.F., White, D.E., Nathan, D., Restifo, N.P., Steinberg, S.M., et al. (2013). Randomized selection design trial evaluating CD8+-enriched versus unselected tumor-infiltrating lymphocytes for adoptive cell therapy for patients with melanoma. J Clin Oncol *31*, 2152–2159.

13. Tran, E., Ahmadzadeh, M., Lu, Y.C., Gros, A., Turcotte, S., Robbins, P.F., Gartner, J.J., Zheng, Z., Li, Y.F., Ray, S., et al. (2015). Immunogenicity of somatic mutations in human gastrointestinal cancers. Science *350*, 1387–1390.

14. van Rooij, N., van Buuren, M.M., Philips, D., Velds, A., Toebes, M., Heemskerk, B., van Dijk, L.J., Behjati, S., Hilkmann, H., El Atmioui, D., et al. (2013). Tumor exome analysis reveals neoantigen-specific T-cell reactivity in an ipilimumab-responsive melanoma. J Clin Oncol *31*, e439-42.

15. Le, D.T., Uram, J.N., Wang, H., Bartlett, B.R., Kemberling, H., Eyring, A.D., Skora, A.D., Luber, B.S., Azad, N.S., Laheru, D., et al. (2015). PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. N Engl J Med *372*, 2509–2520.

16. Le, D.T., Durham, J.N., Smith, K.N., Wang, H., Bartlett, B.R., Aulakh, L.K., Lu, S., Kemberling, H., Wilt, C., Luber, B.S., et al. (2017). Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade. Science *357*, 409–413.

17. Powles, T., Eder, J.P., Fine, G.D., Braiteh, F.S., Loriot, Y., Cruz, C., Bellmunt, J., Burris, H.A., Petrylak, D.P., Teng, S.L., et al. (2014). MPDL3280A (anti-PD-L1) treatment leads to clinical activity in metastatic bladder cancer. Nature *515*, 558–562.

18. Wolchok, J.D., Chiarion-Sileni, V., Gonzalez, R., Rutkowski, P., Grob, J.-J., Cowey, C.L., Lao, C.D., Wagstaff, J., Schadendorf, D., Ferrucci, P.F., et al. (2017). Overall Survival with Combined Nivolumab and Ipilimumab in Advanced Melanoma. N. Engl. J. Med. *377*, 1345–1356.

19. Tran, E., Robbins, P.F., Lu, Y.C., Prickett, T.D., Gartner, J.J., Jia, L., Pasetto, A., Zheng, Z., Ray, S., Groh, E.M., et al. (2016). T-Cell Transfer Therapy Targeting Mutant KRAS in Cancer. N Engl J Med *375*, 2255–2262.

20. Tran, E., Turcotte, S., Gros, A., Robbins, P.F., Lu, Y.C., Dudley, M.E., Wunderlich, J.R., Somerville, R.P., Hogan, K., Hinrichs, C.S., et al. (2014). Cancer immunotherapy based on mutation-specific CD4+ T cells in a patient with epithelial cancer. Science *344*, 641–645.

21. Verdegaal, E.M., de Miranda, N.F., Visser, M., Harryvan, T., van Buuren, M.M., Andersen, R.S., Hadrup, S.R., van der Minne, C.E., Schotte, R., Spits, H., et al. (2016). Neoantigen landscape dynamics during human melanoma-T cell interactions. Nature *536*, 91–95.

22. Segal, N.H., Parsons, D.W., Peggs, K.S., Velculescu, V., Kinzler, K.W., Vogelstein, B., and Allison, J.P. (2008). Epitope landscape in breast and colorectal cancer. Cancer Res *68*, 889–892.

23. Rajasagi, M., Shukla, S.A., Fritsch, E.F., Keskin, D.B., DeLuca, D., Carmona, E., Zhang, W., Sougnez, C., Cibulskis, K., Sidney, J., et al. (2014). Systematic identification of personal tumor-specific neoantigens in chronic lymphocytic leukemia. Blood *124*, 453–462.

24. Rooney, M.S., Shukla, S.A., Wu, C.J., Getz, G., and Hacohen, N. (2015). Molecular and genetic properties of tumors associated with local immune cytolytic activity. Cell *160*, 48–61.

25. Hugo, W., Zaretsky, J.M., Sun, L., Song, C., Moreno, B.H., Hu-Lieskovan, S., Berent-Maoz, B., Pang, J., Chmielowski, B., Cherry, G., et al. (2016). Genomic and Transcriptomic Features of Response to Anti-PD-1 Therapy in Metastatic Melanoma. Cell *165*, 35–44.

26. Van Allen, E.M., Miao, D., Schilling, B., Shukla, S.A., Blank, C., Zimmer, L., Sucker, A., Hillen, U., Foppen, M.H.G., Goldinger, S.M., et al. (2015). Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. Science *350*, 207–211.

27. Kenter, G.G., Welters, M.J., Valentijn, A.R., Lowik, M.J., Berends-van der Meer, D.M., Vloon, A.P., Essahsah, F., Fathers, L.M., Offringa, R., Drijfhout, J.W., et al. (2009). Vaccination against HPV-16 oncoproteins for vulvar intraepithelial neoplasia. N Engl J Med *361*, 1838–1847.

28. Bollard, C.M., Gottschalk, S., Torrano, V., Diouf, O., Ku, S., Hazrat, Y., Carrum, G., Ramos, C., Fayad, L., Shpall, E.J., et al. (2014). Sustained complete responses in patients with lymphoma receiving autologous cytotoxic T lymphocytes targeting Epstein-Barr virus latent membrane proteins. J. Clin. Oncol. *32*, 798–808.

29. Koup, R.A., Safrit, J.T., Cao, Y., Andrews, C.A., McLeod, G., Borkowsky, W., Farthing, C., and Ho, D.D. (1994). Temporal association of cellular immune responses with the initial control of viremia in primary human immunodeficiency virus type 1 syndrome. J. Virol. *68*, 4650–4655. D - NLM: PMC236393 EDAT- 1994/07/01 MHDA- 1994/07/01 00:01 CRDT- 1994/07/01 00:00 PST - ppublish.

30. Kuroda, M.J., Schmitz, J.E., Charini, W.A., Nickerson, C.E., Lifton, M.A., Lord, C.I., Forman, M.A., and Letvin, N.L. (1999). Emergence of CTL coincides with clearance of virus during primary simian immunodeficiency virus infection in rhesus monkeys. J. Immunol. Baltim. Md 1950 *162*, 5127–5133.

31. Nielsen, M., and Andreatta, M. (2016). NetMHCpan-3.0; improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets. Genome Med *8*, 33.

32. Turajlic, S., Litchfield, K., Xu, H., Rosenthal, R., McGranahan, N., Reading, J.L., Wong, Y.N.S., Rowan, A., Kanu, N., Al Bakir, M., et al. (2017). Insertion-and-deletion-derived tumour-specific neoantigens and the immunogenic phenotype: a pan-cancer analysis. Lancet Oncol *18*, 1009–1021.

33. Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A., Behjati, S., Biankin, A.V., Bignell, G.R., Bolli, N., Borg, A., Borresen-Dale, A.L., et al. (2013). Signatures of mutational processes in human cancer. Nature *500*, 415–421.

34. Allen, T.M., Yu, X.G., Kalife, E.T., Reyor, L.L., Lichterfeld, M., John, M., Cheng, M., Allgaier, R.L., Mui, S., Frahm, N., et al. (2005). De novo generation of escape variant-specific CD8+ T-cell responses following cytotoxic T-lymphocyte escape in chronic human immunodeficiency virus type 1 infection. J Virol *79*, 12952–12960.

35. Kaslow, R.A., Carrington, M., Apple, R., Park, L., Munoz, A., Saah, A.J., Goedert, J.J., Winkler, C., O'Brien, S.J., Rinaldo, C., et al. (1996). Influence of combinations of human major histocompatibility complex genes on the course of HIV-1 infection. Nat Med *2*, 405–411.

36. Migueles, S.A., Sabbaghian, M.S., Shupert, W.L., Bettinotti, M.P., Marincola, F.M., Martino, L., Hallahan, C.W., Selig, S.M., Schwartz, D., Sullivan, J., et al. (2000). HLA B*5701 is highly associated with restriction of virus replication in a subgroup of HIV-infected long term nonprogressors. Proc Natl Acad Sci U A *97*, 2709–2714.

37. O'Connor, D.H., Mothe, B.R., Weinfurter, J.T., Fuenger, S., Rehrauer, W.M., Jing, P., Rudersdorf, R.R., Liebl, M.E., Krebs, K., Vasquez, J., et al. (2003). Major Histocompatibility Complex Class I Alleles Associated with Slow Simian Immunodeficiency Virus Disease Progression Bind Epitopes Recognized by Dominant Acute-Phase Cytotoxic-T-Lymphocyte Responses. J. Virol. *77*, 9029–9040.

38. Abelin, J.G., Keskin, D.B., Sarkizova, S., Hartigan, C.R., Zhang, W., Sidney, J., Stevens, J., Lane, W., Zhang, G.L., Eisenhaure, T.M., et al. (2017). Mass Spectrometry Profiling of HLA-Associated

Peptidomes in Mono-allelic Cells Enables More Accurate Epitope Prediction. Immunity *46*, 315–326.

39. Bassani-Sternberg, M., Pletscher-Frankild, S., Jensen, L.J., and Mann, M. (2015). Mass spectrometry of human leukocyte antigen class I peptidomes reveals strong effects of protein abundance and turnover on antigen presentation. Mol Cell Proteomics *14*, 658–673.

40. Juncker, A.S., Larsen, M.V., Weinhold, N., Nielsen, M., Brunak, S., and Lund, O. (2009). Systematic characterisation of cellular localisation and expression profiles of proteins containing MHC ligands. PLoS One *4*, e7448.

41. Yarchoan, M., Hopkins, A., and Jaffee, E.M. (2017). Tumor Mutational Burden and Response Rate to PD-1 Inhibition. N Engl J Med *377*, 2500–2501.

42. Huang, M.N., McPherson, J.R., Cutcutache, I., Teh, B.T., Tan, P., and Rozen, S.G. (2015). MSIseq: Software for Assessing Microsatellite Instability from Catalogs of Somatic Mutations. Sci Rep *5*, 13321.

43. Konishi, H., Mohseni, M., Tamaki, A., Garay, J.P., Croessmann, S., Karnan, S., Ota, A., Wong, H.Y., Konishi, Y., Karakas, B., et al. (2011). Mutation of a single allele of the cancer susceptibility gene BRCA1 leads to genomic instability in human breast epithelial cells. Proc Natl Acad Sci U A *108*, 17773–17778.

44. Nik-Zainal, S., Davies, H., Staaf, J., Ramakrishna, M., Glodzik, D., Zou, X., Martincorena, I., Alexandrov, L.B., Martin, S., Wedge, D.C., et al. (2016). Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature *534*, 47–54.

45. Venkitaraman, A.R. (2014). Cancer suppression by the chromosome custodians, BRCA1 and BRCA2. Science *343*, 1470–1475.

46. Schouten, P.C., Grigoriadis, A., Kuilman, T., Mirza, H., Watkins, J.A., Cooke, S.A., van Dyk, E., Severson, T.M., Rueda, O.M., Hoogstraat, M., et al. (2015). Robust BRCA1-like classification of copy number profiles of samples repeated across different datasets and platforms. Mol Oncol *9*, 1274–1286.

47. Strickland, K.C., Howitt, B.E., Shukla, S.A., Rodig, S., Ritterhouse, L.L., Liu, J.F., Garber, J.E., Chowdhury, D., Wu, C.J., D'Andrea, A.D., et al. (2016). Association and prognostic significance of BRCA1/2-mutation status with neoantigen load, number of tumor-infiltrating lymphocytes and expression of PD-1/PD-L1 in high grade serous ovarian cancer. Oncotarget *7*, 13587–13598.

48. Neelapu, S.S., Locke, F.L., Bartlett, N.L., Lekakis, L.J., Miklos, D.B., Jacobson, C.A., Braunschweig, I., Oluwole, O.O., Siddiqi, T., Lin, Y., et al. (2017). Axicabtagene Ciloleucel CAR T-Cell Therapy in Refractory Large B-Cell Lymphoma. N Engl J Med *377*, 2531–2544.

49. Chang, T.C., Carter, R.A., Li, Y., Li, Y., Wang, H., Edmonson, M.N., Chen, X., Arnold, P., Geiger, T.L., Wu, G., et al. (2017). The neoepitope landscape in pediatric cancers. Genome Med *9*, 78.

50. Capietto, A.-H., Hoshyar, R., and Delamarre, L. (2022). Sources of Cancer Neoantigens beyond Single-Nucleotide Variants. Int. J. Mol. Sci. *23*, 10131.

51. El-Bchiri, J., Guilloux, A., Dartigues, P., Loire, E., Mercier, D., Buhard, O., Sobhani, I., de la Grange, P., Auboeuf, D., Praz, F., et al. (2008). Nonsense-mediated mRNA decay impacts MSI-driven carcinogenesis and anti-tumor immunity in colorectal cancers. PLoS One *3*, e2583.

52. Pastor, F., Kolonias, D., Giangrande, P.H., and Gilboa, E. (2010). Induction of tumour immunity by targeted inhibition of nonsense-mediated mRNA decay. Nature *465*, 227–230.

53. McGranahan, N., Furness, A.J., Rosenthal, R., Ramskov, S., Lyngaa, R., Saini, S.K., Jamal-Hanjani, M., Wilson, G.A., Birkbak, N.J., Hiley, C.T., et al. (2016). Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. Science *351*, 1463–1469.

54. Landau, D.A., Carter, S.L., Stojanov, P., McKenna, A., Stevenson, K., Lawrence, M.S., Sougnez, C., Stewart, C., Sivachenko, A., Wang, L., et al. (2013). Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. Cell *152*, 714–726.

55. Anagnostou, V., Smith, K.N., Forde, P.M., Niknafs, N., Bhattacharya, R., White, J., Zhang, T., Adleff, V., Phallen, J., Wali, N., et al. (2017). Evolution of Neoantigen Landscape during Immune Checkpoint Blockade in Non-Small Cell Lung Cancer. Cancer Discov *7*, 264–276.

56. Cibulskis, K., Lawrence, M.S., Carter, S.L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E.S., and Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol *31*, 213–219.

57. Blank, C.U., Haanen, J.B., Ribas, A., and Schumacher, T.N. (2016). CANCER IMMUNOLOGY. The "cancer immunogram." Science *352*, 658–660.

58. Chen, D.S.S., and Mellman, I. (2013). Oncology meets immunology: the cancer-immunity cycle. Immunity *39*, 1–

59. Cornish, A., and Guda, C. (2015). A Comparison of Variant Calling Pipelines Using Genome in a Bottle as a Reference. Biomed Res Int *2015*, 456479.

60. Hwang, S., Kim, E., Lee, I., and Marcotte, E.M. (2015). Systematic comparison of variant calling pipelines using gold standard personal exome variants. Sci Rep *5*, 17875.

61. Liu, S., Tsai, W.H., Ding, Y., Chen, R., Fang, Z., Huo, Z., Kim, S., Ma, T., Chang, T.Y., Priedigkeit, N.M., et al. (2016). Comprehensive evaluation of fusion transcript detection algorithms and a meta-caller to combine top performing methods in paired-end RNA-seq data. Nucleic Acids Res *44*, e47.

62. Fortier, M.H., Caron, E., Hardy, M.P., Voisin, G., Lemieux, S., Perreault, C., and Thibault, P. (2008). The MHC class I peptide repertoire is molded by the transcriptome. J Exp Med *205*, 595–6

63. Garcia Alvarez, H.M., Koşaloğlu-Yalçın, Z., Peters, B., and Nielsen, M. (2022). The role of antigen expression in shaping the repertoire of HLA presented ligands. iScience *25*, 104975.

64. Rusan, M., Li, Y.Y., and Hammerman, P.S. (2015). Genomic landscape of human papillomavirus-associated cancers. Clin Cancer Res *21*, 2009–2019.

65. Fritsch, E.F., Rajasagi, M., Ott, P.A., Brusic, V., Hacohen, N., and Wu, C.J. (2014). HLA-binding properties of tumor neoepitopes in humans. Cancer Immunol Res *2*, 522–529.

66. Yoshihara, K., Wang, Q., Torres-Garcia, W., Zheng, S., Vegesna, R., Kim, H., and Verhaak, R.G. (2015). The landscape and therapeutic relevance of cancer-associated transcript fusions. Oncogene *34*, 4845–4854. 10.1038/onc.2014.406.

67. Torres-Garcia, W., Zheng, S., Sivachenko, A., Vegesna, R., Wang, Q., Yao, R., Berger, M.F., Weinstein, J.N., Getz, G., and Verhaak, R.G. (2014). PRADA: pipeline for RNA sequencing data analysis. Bioinformatics *30*, 2224–2226. 10.1093/bioinformatics/btu169.

Methods

Datasets

HIV validation, TCGA mutation calls and RNA expression was acquired and processed as described in **Chapter 4**.

ICGC data

CLLE-ES (chronic lymphocytic leukemia) and MALY-DE (malignant lymphoma) projects were downloaded from ICGC release 20. Somatic mutations were annotated with RNA expression levels from sample-matched expression data, matched on Ensembl gene identifiers.

CGCI data

Diffuse large B-cell lymphoma (DLBCL) RNA sequencing data was obtained from the NHL-DLBCL project, provided by the Cancer Genome Characterization Initiative.

CoMMpass data

Multiple myeloma DNA and RNA sequencing data was obtained from the CoMMpass website (`https://research.themmrf.org/`), provided by the Multiple Myeloma Research Foundation, in January 2016. Somatic variants were annotated with sample-matched gene expression levels from the provided gene expression matrix, matched on Ensembl gene identifiers.

Genotype-Tissue Expression Project (GTEx) data

Healthy spleen and EBV-transformed lymphocyte datasets were obtained from the GTEx Portal (dbGaP project phs000424) and used to quantify background and positive control EBV transcript expression levels. The raw data were processed as described under 'Viral gene expression'.

B-cell lymphoma idiotype data

Idiotype sequences obtained from Diffuse Large B-cell lymphoma (DLBCL) and Follicular lymphoma (FL) patients were provided by the Veelken lab at the Leiden University Medical Center, The Netherlands. In selected DLBCL cases where two heavy or light chains were identified, both chains were included. The FL dataset was restricted to heavy chain sequencing data.

Gene fusion calls

Gene fusion calls for the majority of TCGA donors were provided by the Verhaak lab at The University of Texas MD Anderson Cancer Center, USA[1], identified using the structural variant caller PRADA[2]. Of note, PRADA only detects sense-sense fusions, a limitation that to the best of our knowledge applies to all fusion gene callers. To determine the amino acid sequence of gene fusions, the canonical nucleotide sequences of both fusion partners were retrieved from Ensembl release 64 as this version of Ensembl was used for the fusion variants in the PRADA analysis. Genomic breakpoint coordinates were converted to cDNA coordinates and mRNA features (e.g. UTRs) at the breakpoint locations were annotated. Fusions that were deemed unlikely to result in a translatable amino acid sequence (e.g. UTR-UTR fusions) were discarded. The nucleotide sequences of fusion partners in the remaining fusion events were joined and translated into amino acid sequences. As out-of-frame fusions result in the loss of the canonical stop codon, the nucleotide sequence after the breakpoint was scanned until the first occurrence of an in-frame stop codon and the resulting sequence was translated to protein. The resulting amino acid sequence was subsequently used in the neo-antigen prediction pipeline described below. As the PRADA-analysis was performed on RNASeq data, NMD-targeting of detected fusion transcripts can at best have been incomplete, but we cannot exclude partial degradation of these fusion variant calls. For this reason, we did not apply the NMD-filter to fusion variant calls.

Genome- and patient-level annotation

MMR status of tumor samples, antigen presentation capability and sensitivity to T-cell attack, PAM50 subtyping of breast cancer samples, gene essentiality were performed as described in **Chapter 4**.

Peptide-level annotation

Similarity to self-repertoire of peptides was performed as described in **Chapter 4**.

Somatic variant annotation

Variant effect prediction, variant oncogenicity and essentiality, variant cellularity, variant expression, neoantigen prediction, validation of the neoantigen prediction pipeline were performed as described in **Chapter 4**.

Variant annotation with the likelihood of having been caused by any mutational process

Methods to extract the mutational signature of mutational processes compute the total burden caused by a mutational process and the mutational signatures of these processes for a large collection of genomes. They do not compute probabilities for single mutations to have been originated by a particular process, as one would require to compare the neo-antigen yielding capabilities of different mutational processes. We devised a methodology to do this, starting from mutational profiles and patient specific signature loadings.

We want to compute the likelihood of the n-th signature to have caused a somatic variant v of the ith type/letter in the alphabet of all 192 point mutations with the analyzed strand selected such that the reference allele is a pyrimidine and flanked by all possible nucleotide combinations (e.g. 'ApC', denoting a C to T mutation for which the 5' and 3' flanking variants are adenosine and guanine, respectively, employing the notation in Alexandrov et al. (2013)[33]). This measure should factor in the relative contribution of signature n, f^n, to the genome g in which the variant was observed. The absolute amount of mutational signature explained mutations of type i in genome g contributed by the the n-th signature, $N^i_{n,g}$, equals the forward probability of the n-the signature to cause a variant of the i-th letter multiplied by the presence of the n-th signature in the genome g: $N^i_{n,g} = p^i_{n,g}\, e^n_g$ where e^n_g represents the total contribution of signature n to genome g. From this, the probability for mutations of type i to have been caused by the n-th signature in that same genome g must be $\frac{N^i_{n,g}}{\sum_{n'} N^i_{n',g}}$. Somatic variants were assigned to the different mutational processes proportional to the likelihood of having been caused by these probabilities, after which somatic variant and neo-antigen loads could be tallied cognizant of partial class membership as described in **Chapter 4** 'Somatic variant and neo-epitope load tallying'.

Neoantigen load prediction

Computing the total foreign antigen load/foreignness

Foreign antigen load was computed as the sum of the neo-antigen load and viral antigen load (where applicable) on a sample-specific basis. For DLBCL and FL, idiotype-derived antigen loads were added to this by adding the medians of idiotype-derived foreign antigen loads found in non-matched DLBCL and FL patients, respectively.

Viral contribution to foreign antigen load

We noticed substantial differences in the expression levels of EBV genes and human genes in EBV^+ stomach cancer. In order not to overestimate the viral contribution to the foreign antigen load, we filtered out genes that had lower expression values than the first quartile of the human distribution on a per sample basis. Application of this rule had a relatively minor impact on the antigen contributions of HPV genes in HPV^+ cervical and head and neck tumors, with all viral genes being expressed at a higher level than the first quartile of the human protein-encoding transcriptome in the majority of tumors (68% and 59%, respectively) were unaffected due to all viral genes being expressed to a higher extent than the first quartile of human genes. Affected tumors lost an average of 42% (95% CI: [35%, 48%]) and 48% (95% CI: [36%, 61%]) of viral antigens due to low expression, respectively (Figure S2.3B). In contrast, all EBV^+ stomach tumors were strongly affected, as 64% (95% CI: [58%, 70%]) of viral antigens derive from genes that are expressed at a lower level than the first quartile of human protein-encoding transcriptome (Figure S2.3C).

Computation of heterogeneity aware neo-antigen loads

We computed two kinds of heterogeneity aware foreign antigen loads. First, the 'clonals only' measure, in which only neo-antigens derived from clonal mutations are included. Second, the 'cellularity-weighted' measure, in which neo-antigens are summed, weighted by the maximum likelihood cellularity estimates of their corresponding mutations. For both of these measures, we assumed the contribution of viral genes to be shared by all cells in HPV^+ and EBV^+ cancers due to their role in cellular transformation. In addition, idiotypic antigen contributions were weighted as one.

Statistical analysis

Antigen prediction quality assessment, somatic variant and neo-epitope load tallying was performed in as described in **Chapter 4**.

Neoantigen yield rate computation

To compare the neoantigen generating propensities of different somatic variant classes, we analyze their yield rates, defined for a class c as the number of neoantigens resulting from c divided by the mutation load contributed by c. When comparing the yield rates of different mutation classes c, they are computed for the aggregate of variants and neoantigens from multiple donors in a group of genomes G:

$$r_{c,G} \square \frac{\sum_{g \in G} N^{p}_{g,c}}{\sum_{g \in G} N^{v}_{g,c}}$$

which can be interpreted as the probability for a mutation of class c to yield a neoantigen. Invoking the binomial properties of the r for values of r in the range $[0, 1]$, we compute the 95% confidence interval of this estimate as $r_{c,G} \pm 1.96 \sqrt{\frac{1}{N^{n}_{g,c}} r_{c,G} \left(1 - r_{c,G}\right)}$.

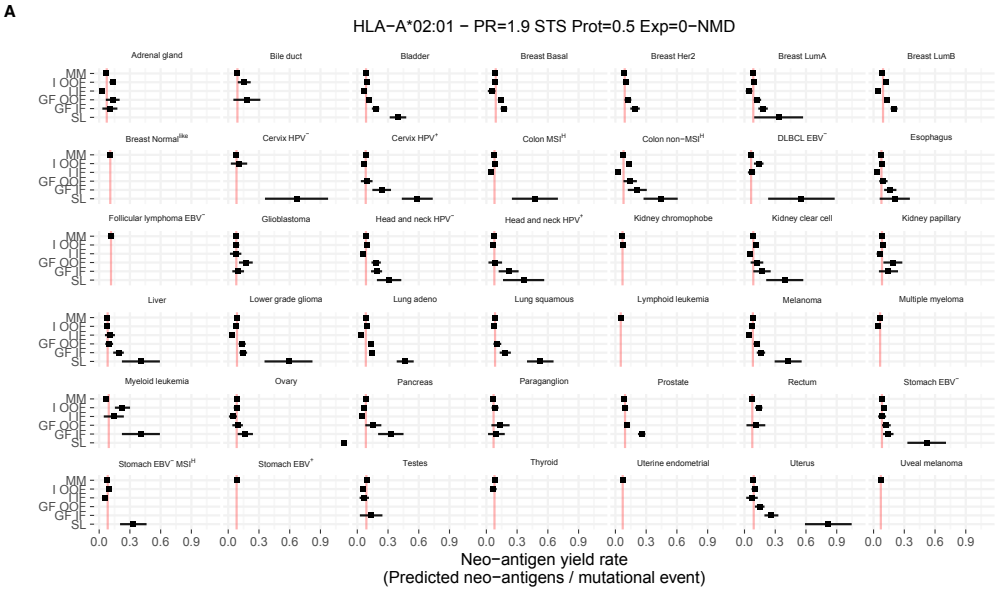
Data and software availability

Custom software was distributed over R and Perl packages as described in **Chapter 4**.

Supplemental Items

Table 1 can be retrieved from `https://doi.org/10.17632/mf39n7s2b9.1`.

Figure S2.1: Technical considerations in assessing foreignness
A Distance in base pairs between fusion gene breakpoints and focal DNA damage (SNVs and indels of max. 200 base pairs long).
B The pairwise Spearman correlation in predicted neoantigen loads between different settings of the neoantigen prediction pipeline across all included patients. Rows are identically ordered as columns. Gene expression filtering settings, rather than the choice of the HLA allele to use in neoantigen prediction, determine most of the variation. This indicates the choice of HLA allele is rather unimportant.


A
HLA−A*02:01 − PR=1.9 STS Prot=0.5 Exp=0−NMD
Adrenal gland
Bile duct
Bladder
Breast Basal
Breast Her2
Breast LumA
Breast LumB
Breast Normal^like
Cervix HPV^−
Cervix HPV^+
Colon MSI^H
Colon non−MSI^H
DLBCL EBV^−
Esophagus
Follicular lymphoma EBV^−
Glioblastoma
Head and neck HPV^−
Head and neck HPV^+
Kidney chromophobe
Kidney clear cell
Kidney papillary
Liver
Lower grade glioma
Lung adeno
Lung squamous
Lymphoid leukemia
Melanoma
Multiple myeloma
Myeloid leukemia
Ovary
Pancreas
Paraganglion
Prostate
Rectum
Stomach EBV^−
Stomach EBV^− MSI^H
Stomach EBV^+
Testes
Thyroid
Uterine endometrial
Uterus
Uveal melanoma
MM
I
OO
GF
SL
0.0 0.3 0.6 0.9
Neo−antigen yield rate
(Predicted neo−antigens / mutational event)

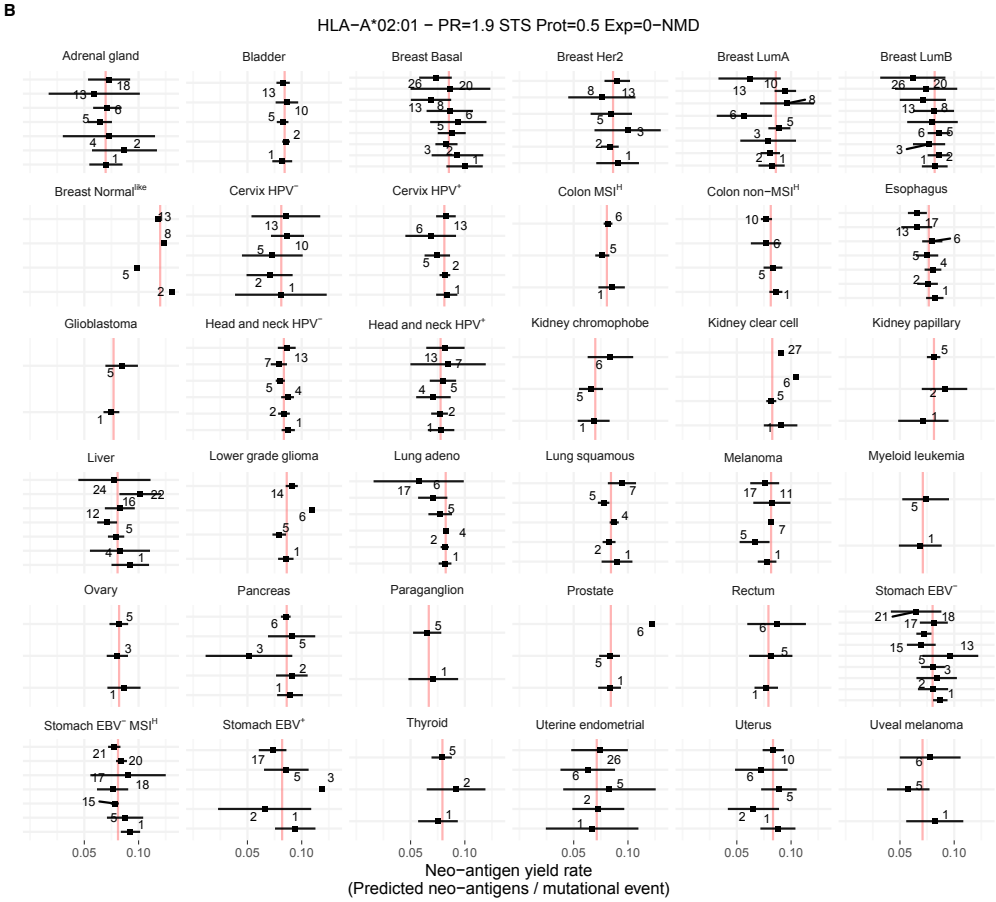

B
HLA−A*02:01 − PR=1.9 STS Prot=0.5 Exp=0−NMD
Adrenal gland
Bladder
Breast Basal
Breast Her2
Breast LumA
Breast LumB
Breast Normal^like
Cervix HPV^−
Cervix HPV^+
Colon MSI^H
Colon non−MSI^H
Esophagus
Glioblastoma
Head and neck HPV^−
Head and neck HPV^+
Kidney chromophobe
Kidney clear cell
Kidney papillary
Liver
Lower grade glioma
Lung adeno
Lung squamous
Melanoma
Myeloid leukemia
Ovary
Pancreas
Paraganglion
Prostate
Rectum
Stomach EBV^−
Stomach EBV^− MSI^H
Stomach EBV^+
Thyroid
Uterine endometrial
Uterus
Uveal melanoma
0.05 0.10
Neo−antigen yield rate
(Predicted neo−antigens / mutational event)

Figure S2.2: Relationship between mutational processes and foreign antigen loads across cancers.
A HLA-A*02:01 neo-antigen yield rates of different mutational events are highly similar across tumor types. Tumor-specific average yield rates are indicated by red lines.
B As in **A**, Comparing neo-antigen yield rates of the indicated mutational signatures within tumor types. No signatures are observed to be recurrently deviant from the other mutational signatures with respect to neo-antigen yield rates.

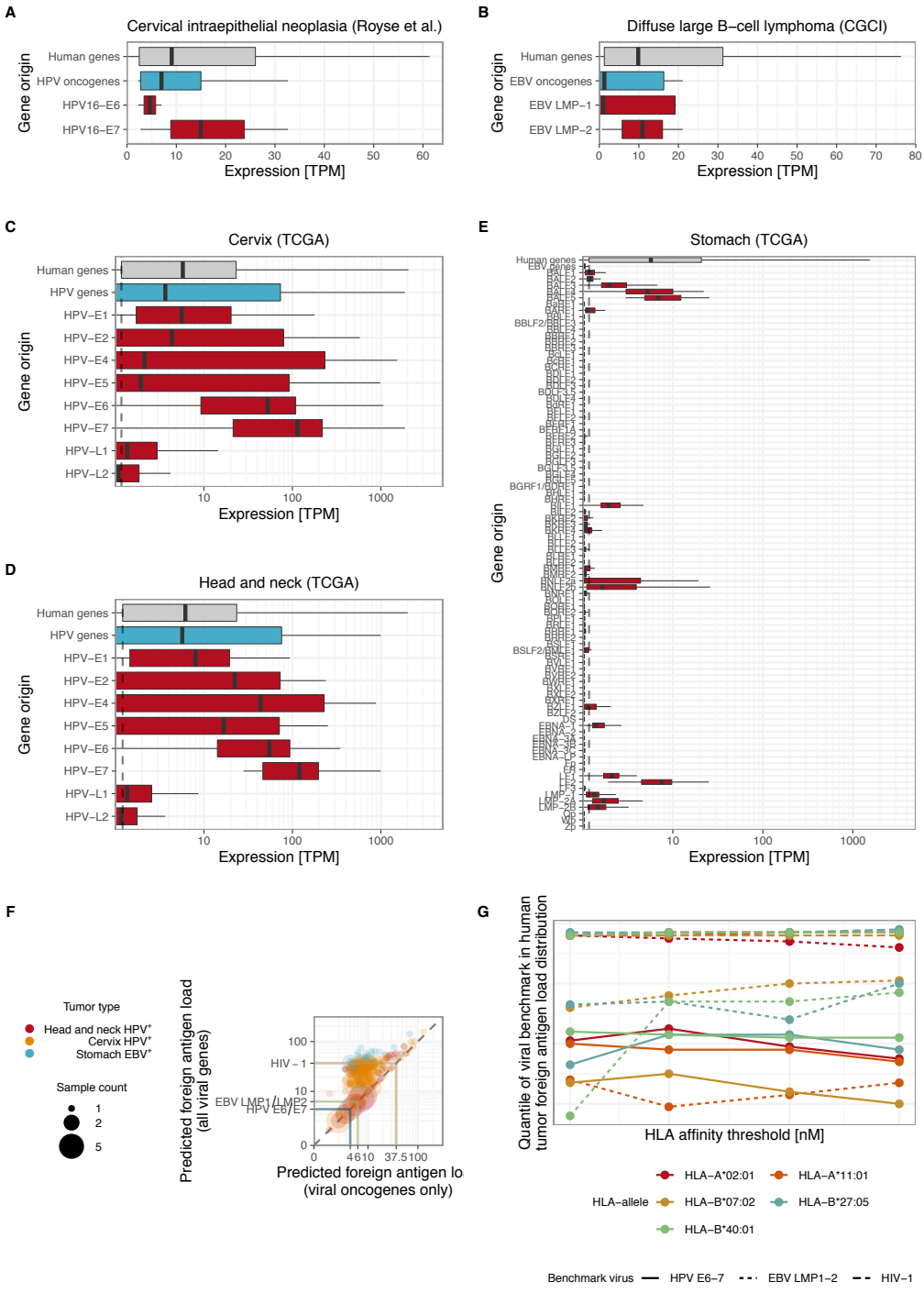


Figure S2.3: Validation of comparisons of viral antigens and antigens derived from somatic mutations

A (Left) Distributions of RNA expression levels of human genes and HPV-associated *E6* and *E7* oncogenes in cervical intraepithelial neoplasia. (Right) Same as left panel, but depicting expression of human genes and EBV-derived oncogenes in diffuse large B-cell lymphoma. Blue boxes represent expression levels of viral oncogenes.
B (Top) Distributions of RNA expression levels of human genes and HPV genes in cervical cancers. (Bottom) As in top panel, but for HPV^+ head and neck cancers. Blue boxes represent expression of all viral genes.
C RNA expression levels of human genes and EBV genes in EBV^+ stomach tumors. Dashed line indicates lower quartile expression level of human genes. The blue box represents expression across all EBV genes.
D Relationship between foreign antigen loads of virus-positive tumors when either only taking viral oncogenes (*E6* and *E7* for HPV, *LMP-1* and *LMP-2* for EBV) into account, or when taking all viral genes into account for which expression greater than the first quartile of the human genes is observed in a sample-specific manner. Dot size reflects the number of samples on a coordinate; color denotes tumor type.

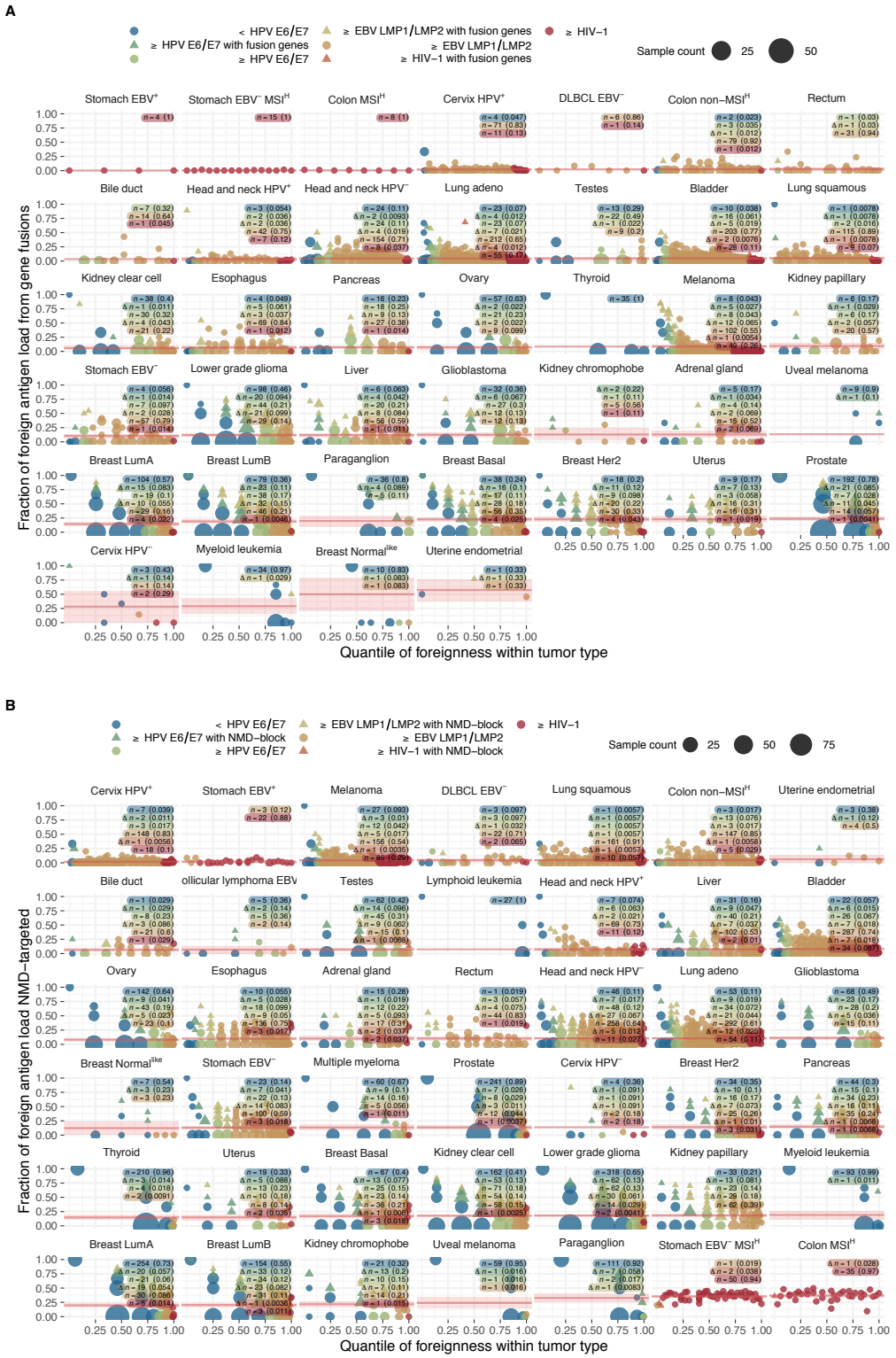

A
< HPV E6/E7
≥ HPV E6/E7 with fusion genes
≥ HPV E6/E7
≥ EBV LMP1/LMP2 with fusion genes
≥ EBV LMP1/LMP2
≥ HIV−1 with fusion genes
≥ HIV−1
Sample count 25 50
Fraction of foreign antigen load from gene fusions
Quantile of foreignness within tumor type
B
< HPV E6/E7
≥ HPV E6/E7 with NMD−block
≥ HPV E6/E7
≥ EBV LMP1/LMP2 with NMD−block
≥ EBV LMP1/LMP2
≥ HIV−1 with NMD−block
≥ HIV−1
Sample count 25 50 75
Fraction of foreign antigen load NMD−targeted
Quantile of foreignness within tumor type

Figure S2.4: Effect of tumor heterogeneity on foreign antigen loads.
A Relationship between the foreign antigen loads excluding fusion genes (quantile within tumor type, horizontal axis) and the fractional contribution of fusion-derived antigens to the neo-antigen load. Dot size reflects the number of samples on a coordinate. Only tumor samples for which the presence of gene fusion events was evaluated are incorporated. Samples are colored according to their foreignness compared to the three viral benchmarks. Tumors moving up to a higher level of foreignness, as compared to the viral benchmarks, when including fusion-derived antigens are denoted by triangles rather than circles. Legends indicate sample size per color group, in which black parenthesized numbers reflect the fraction of samples in each color group per tumor type. Mean fractional contributions and associated 95% confidence intervals of fusion genes to total foreignness are indicated by red lines and surrounding red transparent boxes, respectively. Rank quantiles shown on the horizontal axis are computed over the foreign antigen load excluding fusion genes, assigning tied samples identical maximal rank quantiles. Fusions genes are shown to infrequently contribute a sizeable fraction of foreign antigen load, especially for tumors that are relatively rich in focal mutation-derived foreign antigens.
B As in **A**, but showing the fractional contribution of NMD-targeted neo-antigens to the tumoral foreign antigen loads.

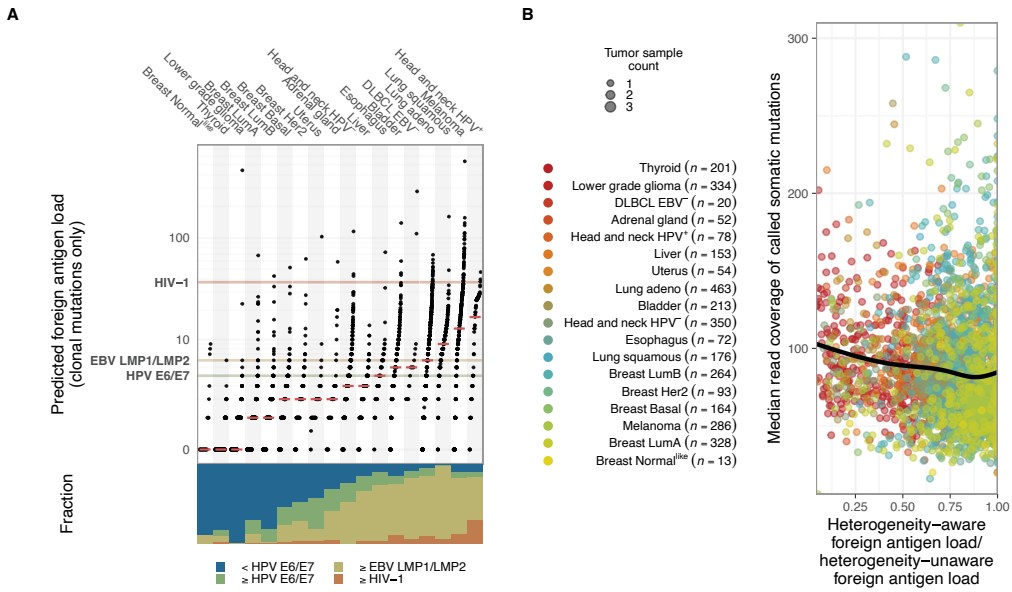


Figure S2.5: Effect of tumor heterogeneity on foreign antigen loads.
A Pan-cancer overview of foreignness in which neo-antigens contributed by focal DNA damage types are restricted to clonal mutations and combined with viral and idiotype antigen contributions, with the latter two antigen sources assumed to be clonal.
B Absence of relationship between sample DNA sequencing depth and the degree to which the foreign antigen load is reduced in heterogeneity aware estimates. Dot size reflects the number of samples at a coordinate. The black line represents a LOESS local smoothing regression.