HHS Public Access

Author manuscript

Nat Med. Author manuscript; available in PMC 2021 March 08.

Published in final edited form as:

Nat Med. 2019 November ; 25(11): 1715–1720. doi:10.1038/s41591-019-0639-4.

Evolutionary divergence of HLA class I genotype impacts efficacy of cancer immunotherapy

Diego Chowell[1,2,9], **Chirag Krishna**[3,9], **Federica Pierini**[4,9], **Vladimir Makarov**[1,2], **Naiyer A. Rizvi**[5], **Fengshen Kuo**[2], **Luc G. T. Morris**[2,6], **Nadeem Riaz**[2,7], **Tobias L. Lenz**[4,10,*], **Timothy A. Chan**[1,2,7,8,10,*]

[1]Human Oncology and Pathogenesis Program, Memorial Sloan Kettering Cancer Center, New York, NY, USA.

[2]Immunogenomics and Precision Oncology Platform, Memorial Sloan Kettering Cancer Center, New York, NY, USA.

[3]Computational and Systems Biology Program, Memorial Sloan Kettering Cancer Center, New York, NY, USA.

[4]Research Group for Evolutionary Immunogenomics, Max Planck Institute for Evolutionary Biology, Plön, Germany

[5]Department of Medicine, Columbia University Medical Center, New York, NY, USA.

[6]Department of Surgery, Memorial Sloan Kettering Cancer Center, New York, NY, USA.

[7]Department of Radiation Oncology, Memorial Sloan Kettering Cancer Center, New York, NY, USA.

[8]Weill Cornell School of Medicine, New York, NY, USA.

[9]These authors contributed equally: Diego Chowell, Chirag Krishna, Federica Pierini.

[10]These authors jointly supervised this work: Tobias L. Lenz, Timothy A. Chan.

Abstract

Functional diversity of the highly polymorphic human leukocyte antigen class I (*HLA*-I) genes underlies successful immunologic control of both infectious disease and cancer. The divergent allele advantage hypothesis dictates that an *HLA*-I genotype with two alleles with sequences that are more divergent enables presentation of more diverse immunopeptidomes[1–3]. However, the effect of sequence divergence between *HLA*-I alleles—a quantifiable measure of *HLA*-I evolution

[*]**Correspondence and requests for materials** should be addressed to T.L.L. or T.A.C. lenz@post.harvard.edu; chant@mskcc.org.

Author contributions

D.C., C.K., F.P., V.M., T.L.L., T.A.C., F.K., L.G.T.M. and N.A.R performed the data acquisition and analyses. D.C., C.K., F.P., T.L.L. and T.A.C. wrote the manuscript, with input from all authors.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41591-019-0639-4.

Supplementary information is available for this paper at https://doi.org/10.1038/s41591-019-0639-4.

Peer review information Saheli Sadanand was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

—on the efficacy of immune checkpoint inhibitor (ICI) treatment for cancer remains unknown. In the present study the germline HLA-I evolutionary divergence (HED) of patients with cancer treated with ICIs was determined by quantifying the physiochemical sequence divergence between *HLA*-I alleles of each patient's genotype. HED was a strong determinant of survival after treatment with ICIs. Even among patients fully heterozygous at *HLA*-I, patients with an HED in the upper quartile respond better to ICIs than patients with a low HED. Furthermore, HED strongly impacts the diversity of tumor, viral and self-immunopeptidomes and intratumoral T cell receptor donality. Similar to tumor mutation burden, HED is a fundamental metric of diversity at the major histocompatibility complex-peptide complex, which dictates ICI efficacy. The data link divergent *HLA* allele advantage to immunotherapy efficacy and unveil how ICI response relies on the evolved efficiency of *HLA*-mediated immunity.

Checkpoint blockade immunotherapies such as anti-PD-1, anti-PD-L1 and anti-CTLA-4 have revolutionized the treatment of advanced-stage cancers, but only a minority of patients respond. A critical determinant of ICI response is tumor mutational burden (TMB), a proxy for the number of tumor-derived neoantigens that can be presented on the cell surface by major histocompatibility complex (MHC) molecules and subsequently recognized by cytotoxic T cells[4–9]. Another genetic factor that determines ICI response is heterozygosity at the highly polymorphic *HLA*-I loci[10]. According to heterozygote advantage, originally observed in studies of infectious diseases, heterozygous *HLA*-I genotypes facilitate presentation of a more diverse set of tumor antigens to T cells[10–15].

Each individual's *HLA*-I genotype consists of a pair of alleles at each of the classic class I genes—*HLA-A, -B* and *-C*—and their polymorphism is concentrated within their peptide-binding domains[16,17]. The set of peptides bound by each MHC class I (MHC-I) molecule is collectively referred to as its immunopeptidome, and *HLA*-I alleles have different peptide-binding specificities with varying overlap according to the physiochemical sequence divergence between alleles[1,18,19]. The concomitant diversity of *HLA*-I genotypes and peptide-binding specificities yields marked interindividual variability in immunopeptidome diversity[1,19]. This variation may affect the ability of each individual's immune system to recognize tumor antigens, and thus may influence response to ICI. Furthermore, recent studies have shown that the *HLA* genotype shapes the landscape of oncogenic mutations observed in tumors, and that somatic loss of *HLA*-I is one possible mechanism by which tumors evade immune control[10,20–23].

Motivated by the divergent allele advantage proposed three decades ago[1,2], the present study hypothesizes that the effect of *HLA*-I heterozygosity on response to ICIs may be modulated by the amount of sequence divergence between the peptide-binding domains of patient *HLA*-I alleles. High sequence divergence between the alleles' peptide-binding domains strongly affects the combined peptide-binding properties of the corresponding MHC-I molecules[2,3,24,25]. Thus, heterozygous patients with more divergent alleles may present a broader set of peptides for T cell recognition than heterozygous patients with less divergent *HLA*-I alleles[2,3,25].

HED was first determined using *HLA*-I genotypes across multiple cohorts of patients with metastatic melanoma or non-small-cell lung cancer (NSCLC) treated with anti-CTLA-4 or

anti-PD-1/-PD-L1 (Fig. 1a, and see Supplementary Tables 1 and 2). For each patient, the HED was calculated at each of *HLA-A, HLA-B* and *HLA-C* by measuring the Grantham distance[3,26] between the peptide-binding domains of the two alleles. The Grantham distance is a classic metric that allows quantification of physiochemical differences between protein amino acid sequences, taking into account composition, polarity and volume. To explore the landscape of HEDs in the dataset of the present study, hierarchical clustering of HED per *HLA*-I locus was performed for all pairwise allele combinations. Hierarchical clustering of HEDs demonstrated distinct clusters of high and low divergence between alleles (Fig. 1b and Extended Data Fig. 1), consistent with known interrelationships of *HLA-A, HLA-B* and *HLA-C* loci[17,27]. *HLA-B* pairwise divergences were higher relative to *HLA-A* and *HLA-C* (Fig. 1c), consistent with prior reports that *HLA-B* is the oldest and most diverse of the three *HLA*-I loci[17,27]. Moreover, *HLA-C* alleles had the lowest pairwise divergences, in line with prior studies that *HLA-C* has evolved most recently[17,27,28] (Fig. 1c). Next, for each patient, the mean HED was calculated as the mean of the three pairwise divergences of *HLA-A, HLA-B* and *HLA-C,* assuming that each locus contributes equally to presentation of antigenic peptides. Mean HED distributions in patients from the cohorts in the present study were similar to those observed in The Cancer Genome Atlas (TCGA) (Fig. 1d,e). A prior comparison of the Grantham distance to other common metrics of sequence divergence showed that the Grantham distance best captured the functional properties of HLA-I molecules[3]. The Grantham distance is a well-recognized metric that has been applied to measure amino acid polymorphism in studies of comparative evolution, cancer, infectious disease and immunity[29–34]. Furthermore, in an analysis of *HLA*-I allele pairs and naturally eluted peptides derived from mass spectrometry and monoallelic cell lines[35], an association was detected between HED and peptidome diversity (Supplementary Fig. 1). Taken together, these data verify that the Grantham distance is a suitable measure of *HLA*-I polymorphism in the patient cohorts.

Next it was asked whether HED is associated with the response to ICIs. Patients were stratified by mean HED in a cohort of 100 patients with melanoma treated with anti-CTLA-4[8] (hereafter called cohort 1). Improved overall survival was observed after ICIs in patients with high mean HED, where high was defined as mean HED greater than or equal to the top quartile, and low was defined as mean HED less than the top quartile ($P = 0.0072$, hazard ratio (HR) = 0.47, 95% confidence interval (CI) = 0.26–0.82) (see Extended Data Fig. 2a). These results were similar across different metrics (that is, sum, median or geometric mean) used to combine pairwise divergences of *HLA-A, HLA-B* and *HLA-C* alleles (see Supplementary Table 3). It was also found that the effect of mean HED on survival was independent of TMB and other genomic and clinical variables, when these were included in a multivariable Cox regression model of survival (see Extended Data Fig. 2d). Finally, it was found that the effect of both high mean HED and high TMB on overall survival after ICIs was more pronounced than the effect of either alone, as reflected by the reduction in HR (commonly considered to be the effect size in survival analyses)[36,37] when considering both variables (see Extended Data Fig. 2a-c).

Prior studies of divergent allele advantage have suggested that the diversity of immunopeptidomes of fully heterozygous *HLA*-I genotypes varies with sequence divergence[1,3]. Therefore, it was hypothesized that, even among patients fully heterozygous

at *HLA*-I, response to ICIs may also vary with HED. Strikingly, it was found that high mean HED was associated with improved survival after ICIs in the 78 fully heterozygous patients from cohort 1 (ref. [8]) ($P = 0.0094$, HR = 0.43, 95% CI = 0.22–0.83) (Fig. 2a). In a second cohort of 76 fully heterozygous patients with NSCLC treated primarily with anti-PD-1 (refs. [7,10]), it was also found that high mean HED was associated with better overall survival ($P = 0.049$, HR = 0.32, 95% CI = 0.10–1.06) (Fig. 2b). The same was observed in an additional third cohort of 95 fully heterozygous patients with metastatic melanoma treated with anti-PD-1/-PD-L1 (refs. [10,38]) ($P = 0.025$) (Fig. 2c). In a combined analysis of all three cohorts, a negative relationship was noted between mean HED and HR, indicating that, in general, an increase in mean HED corresponds to improved overall survival (see Extended Data Fig. 3). Beyond survival, clinical response to ICIs was also associated with a high mean HED when considering all patients (*HLA*-I homozygotes or heterozygotes) (57.4% versus 32.0%, $P = 0.003$, odds ratio (OR) = 0.35) (Fig. 2d), or only fully heterozygous patients (55.6% versus 35.3%, $P = 0.03$, OR = 0.44) (Fig. 2e) across all cohorts.

To determine whether HED might simply reflect a general prognostic factor in cancer, the association of *HLA*-I heterozygosity or HED with overall survival was examined among patients with melanoma and NSCLC who did not receive ICI therapy, and no effect was observed (see Extended Data Figs. 4 and 5). This suggests that mean HED is predictive of response to ICIs, and may not be prognostic in the setting of patients with advanced cancer not treated with ICIs.

All cohorts from Fig. 2 were examined to investigate the combined effect of mean HED and TMB on response to ICIs. It was found that the effect of mean HED on improved survival after ICIs (Fig. 3a) was independent of other clinical variables in multivariable Cox's regression analysis (see Extended Data Fig. 6a), and that high HED did not co-occur with known mutations in genes that have been reported to impact response to ICIs (see Extended Data Fig. 7). Furthermore, it was found that the combined effect of high HED and high TMB on overall survival after ICIs was stronger than the effect of either alone, as evidenced by the reduction in the HR when stratifying patients by both variables[36,37] (Fig. 3a–c). This combined effect was also observed when analyzing only fully heterozygous patients (Fig. 3d–f, and see Extended Data Fig. 6b). Furthermore, the effect remained robust across a wide range of cut points for HED and TMB (Fig. 3g and see Extended Data Fig. 8a) used to stratify patients into groups for survival analysis. High HED at each of *HLA-A* and *HLA-B* was associated with improved survival after ICI administration, when considering all patients or only fully heterozygous patients (Fig. 3h). On multi- variable analysis, it was found that high HED at both *HLA-A* and *HLA-B* was independently associated with improved survival (see Extended Data Fig. 8b), suggesting that divergence at individual class I loci may differentially affect ICI efficacy. Moreover, the effect of high mean HED on improved overall survival after ICI was detected in an additional pan-cancer dataset of over 1,000 patients (see Extended Data Fig. 9).

Next it was hypothesized that high HED may be associated with increased diversity of the neopeptide repertoire presented by *HLA*-I. In an exploratory analysis limited to patients fully heterozygous at each locus, it was found that the number of candidate neopeptides bound by heterozygous genotypes correlates with mean HED (Fig. 4a). Moreover, mean

HED did not correlate with TMB (Fig. 4b), indicating that the diversity in *HLA*-I peptide-binding domains specifically reflects the diversity of the neopeptides bound to *HLA*-I molecules, rather than the diversity of all tumor mutations. Furthermore associations were detected between HED and diversity of the neopeptide repertoire at individual class I loci (see Extended Data Fig. 10a-c). Consistent with these results, HED was also correlated with the abundance of viral peptides derived from a number of pathogens (Fig. 4c, and see Extended Data Fig. 10d-f and Supplementary Table 4).

Next it was hypothesized that HED may be associated with the diversity of the total human self immunopeptidome, of which a fraction may potentially generate neoepitopes. All unique peptides of length nine from the entire human proteome were computationally generated to enable a common reference self-proteome across all patients, and *HLA*-I binding predictions performed. It was found that HED was correlated with the diversity of the predicted self immunopeptidome (Fig. 4d, and Extended Data Fig. 10g-i). Then HED was determined in an independent cohort of 18 individuals for whom *HLA-A* and *HLA-B* genotypes and naturally eluted peptide data were available[39], and an association was observed between HED and self-immunopeptidome diversity (see Supplementary Fig. 2). An additional dataset of mass spectrometry-derived peptidomes from monoallelic cells was analyzed[35], which includes peptide data for 10 *HLA-A* and 6 *HLA-B* alleles. HEDs and the number of peptides bound by all possible pairs of *HLA-A* and *HLA-B* alleles were computed (n = 120), and a signficant negative correlation was found between HED and the overlap of peptides bound by both alleles of a given pair (see Supplementary Fig. 1a). These data indicate that the more divergent *HLA*-I alleles are, the more distinct the peptides they present. A similar negative correlation was also detected when considering *HLA-A* alleles alone (see Supplementary Fig. 1b), or *HLA-B* alleles alone (see Supplementary Fig. 1c). Furthermore, it was found that HED was positively correlated with the abundance of peptides bound to pairs of alleles at each individual locus (see Supplementary Fig. 1d,e). Altogether, these data suggest that increased sequence divergence of an HLA-I genotype is associated with increased diversity of self, tumor and viral immunopeptidomes.

Next it was investigated whether the association of high HED with a broader neopeptide repertoire would increase the probability of neoantigen recognition by tumor-infiltrating T cells, and subsequently influence T cell clonal expansion. Accordingly, in a subset of patients treated with ICI therapy for whom next-generation deep sequencing of TCR complementarity-determining regions (CDR3s) was available[40], a positive correlation was found between mean HED and clonality of TCR CDR3s (Fig. 4e). However, additional data will be required to validate this result. Importantly, as TCRs interact with self-peptides presented by each individual's *HLA*-I molecules during thymic selection, HED may affect the diversity of the TCR repertoire of T cells in peripheral blood. Although blood for TCR sequencing was not available from the patients analyzed in the present study, it is hoped that this hypothesis will be evaluated in the near future.

Taken together, these data show that HED—as measured by sequence divergence between alleles of a *HLA*-I genotype—is associated with response to checkpoint blockade immunotherapy in patients treated for cancer, and with the diversity of tumor, viral and human immunopeptidomes. Compared with TMB, which can be challenging to accurately

estimate due to tumor purity or clonal fraction, HED can be reliably inferred from normal tissue DNA sequencing. Furthermore, the results of the present study suggest that patients with both high TMB and high HED are most likely to benefit from ICIs. Importantly, HED and TMB are both genetic variables that affect anti-tumor immunity. Critically, HED is different from neoantigen burden, which represents only a subset of tumor peptides that can potentially be presented by a patient's MHC-I molecules. In addition, neoantigen burden estimates suffer from imperfect peptide-HLA-binding prediction algorithms. We propose that, unlike neoantigen burden, HED is a granular metric of functional HLA diversity and, together with TMB, determines the potential for T cell-mediated tumor control (Fig. 4f). Therefore, both TMB and HED should be considered in the design of future clinical trials. Further studies will investigate the effect of HED on tumor evolution and the host TCR repertoire.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of code and data availability are available at https://doi.org/10.1038/s41591-019-0639-4.

Methods

Description of patient cohorts.

Seven previously published cohorts of patients were used who had late-stage melanoma and NSCLC treated with anti-CTLA-4, or PD-1/PD-L1 blockade[6–8,10,38,40,41]. Ten patients from the Van Allen et al.[8] cohort were excluded, because they achieved long-term survival after anti-CTLA-4 treatment with early tumor progression[8]. The NSCLC data are from patients with metastatic disease treated mainly with anti-PD-1 monotherapy. They are from a prospective trial that has been reported previously[7] and from the New York-Presbyterian/Columbia University Medical Center[10]. From these NSCLC cohorts, for the analyses involving combination of HED and TMB (see Fig. 3), only patients with exome sequencing data were included, because mutation data were not available for 66 patients with NSCLC. For the analyses involving HED only (see Fig. 2), all NSCLC patients were included, because *HLA* types were available for all patients. All patients were treated under institutional review-approved prospective protocols. Clinical characteristics of patient cohorts are provided in the original studies. The Cancer Genome Atlas (TCGA exome data for the patients with melanoma (n = 446) and lung cancer (n = 473) were obtained from TCGA (http://cancergenome.nih.gov).

Overall survival and clinical response.

Overall survival was defined as the length of time from treatment start to time of event (survival or censor). Response data were available for some cohorts[7,8,10]; clinical benefit was defined as complete response (CR), partial response (PR) or stable disease (SD), as indicated in previous studies[7,8,10]. No clinical benefit was defined as progressive disease. All clinical data, including overall survival and clinical response data, were obtained from

the original studies. Clinical data for TCGA patients with melanoma and NSCLC were obtained through TCGA data portal.

HLA-I genotyping.

HLA-I genotyping was performed as described previously[10]. Briefly, high-resolution *HLA*-I genotyping from germline normal DNA exome sequencing data was performed directly or using a clinically validated *HLA* typing assay (LabCorp). Patient exome data or targeted gene panels were obtained and the well-validated tool Polysolver was used to identify *HLA*-I alleles with default parameter settings[42]. For the 66 patients with NSCLC and no available exome sequencing data, *HLA*-I typing was done at LabCorp. For quality assurance of *HLA*-I genotyping using MSK-IMPACT (CLIA-approved hybridization-capture-based assay) with melanoma samples from anti-PD1-treated patients, *HLA*-I typing by Polysolver was compared across 37 samples that were sequenced with MSK-IMPACT and whole exome. The MSK-IMPACT panel successfully captured *HLA-A*, *-B* and -C reads and validation had previously been performed[10]. The overall concordance of class I typing between the MSK-IMPACT samples and their matched whole-exome sequencing samples was 96%. To make sure that *HLA*-I genes have adequate coverage in MSK-IMPACT bam files, the bedtools multicov tool (http://bedtools.readthedocs.io/en/latest/content/tools/multicov.html) was also applied, which reports the count of alignments from multiple position-sorted and indexed BAM files that overlap with target intervals in a BED format. Only high-quality reads were counted and only samples with sufficient coverage were used. Patients were considered fully heterozygous at *HLA-I* if they have six different *HLA-I* alleles.

Calculation of patient HED.

HED was calculated as described in Pierini and Lenz[3]. Briefly, first the protein sequence of exons 2 and 3 of each allele of each patients *HLA*-I genotype was extracted; these sequences correspond to the peptide-binding domains. Protein sequences were obtained from the ImMunoGeneTics/HLA database[43], and exons coding for the variable peptide-binding domains were selected following the annotation obtained from the Ensembl database[44]. Divergences between allele sequences were calculated using the Grantham distance metric[26], as implemented in Pierini and Lenz[3]. The Grantham distance is a quantitative pairwise distance in which the physiochemical properties of amino acids, and hence the functional similarity between sequences, are considered[26]. Given a particular *HLA*-I locus with two alleles, the sequences of the peptide-binding domains of each allele are aligned[45], and the Grantham distance is calculated as the sum of amino acid differences (taking into account the biochemical composition, polarity and volume of each amino acid) along the sequences of the peptide-binding domains, following the formula by R. Grantham[26]:

$$\begin{aligned}\text{Grantham distance} &= \sum D_{ij} \\ &= \sum \left[\alpha\left(c_i - c_j\right)^2 + \beta\left(p_i - p_j\right)^2 + \gamma\left(v_i - v_j\right)^2\right]^{1/2}\end{aligned} \quad (1)$$

where i and j are the two homologous amino acids at a given position in the alignment and D is the Grantham distance between them, c, p and v represent composition, polarity and volume of the amino acids, respectively, and α, β and γ are constants; all values are taken from the original study[34]. The final Grantham distance is calculated by normalizing the

value from Eqn (1) by the length of the alignment between the peptide-binding domains of a particular *HLA*-I genotypes two alleles. A prior analysis of multiple common sequence divergence measures showed that the correlation of Grantham distance with the number of peptides bound by both alleles of a heterozygous genotype exceeded that of the other distance measures[3]. Patient mean HED was calculated as the mean of divergences at *HLA-A*, *HLA-B* and *HLA-C.*

Tumor mutational analysis pipeline.

For cohorts that received whole-exome sequencing, reads in FASTQ format were aligned to the reference human genome GRCh37 using the Burrows-Wheeler aligner (BWA v.0.7.10)[46]. Local realignment was performed using the Genome Analysis Toolkit (GATK v.3.7)[47]. Duplicate reads were removed using Picard v.2.13. To identify somatic single-nucleotide variants (SNVs), a validated pipeline was used that integrates mutation calls from four different mutation callers: MuTect 1.1.7, Strelka 1.0.15, SomaticSniper 1.0.4 and Varscan 2.4.3 (refs.[48–51]). SNVs with an alternative allele read count <4, total coverage <10 or corresponding normal coverage <7 reads were filtered out.

Computational identification of *HLA*-I-restricted neopeptides.

Each non-synonymous SNV was translated into a 17-mer peptide sequence, centered on the mutated amino acid. Adjacent SNVs were corrected using MAC[52]. Subsequently, the 17-mer was then used to create 9-mers via a sliding window approach for determination of *HLA*-I-binding predictions for neopeptides using NetMHCpan-4.0 (ref.[53]). All peptides with a rank <2% were considered for further analyses.

Computational identification of HLA-I-restricted peptides from the human proteome and viral antigens.

Peptides from the entire human proteome that binds to patient-specific *HLA*-I alleles were identified. The human peptidome was downloaded from Ensembl[44] (ftp://ftp.ensembl.org/pub/grch37/update/fasta/homo_sapiens/pep//Homo_sapiens.GRCh37.pep.all.fa). Only sequences annotated as gene_biotype:protein_coding and transcript_biotype:protein_coding were kept. Transcripts with identical sequences were de-duplicated. The resulting FASTA file was submitted to NetMHCpan 4.0 (ref.[53]) to determine *HLA*-I-binding predictions. All peptides from the human proteome with a rank <2% were considered for further analyses. For the correlation analyses in Supplementary Fig. 2, self-peptides were used that were identified via mass spectrometry and *HLA*-I genotypes from Pearson et al.[39]. For the correlation analyses in Supplementary Fig. 1, naturally eluted self-peptides were used that were derived from mass spectrometry and mono allelic cell lines from Abelin et al.[35]. All correlation analyses were limited to peptides of length 9. In addition, predicted viral peptides were generated from a number of antigens (see Supplementary Table 4).

TCR β-chain sequencing and analysis.

Next-generation sequencing of TCR β-chain CDR3s (TCR sequencing) (Adaptive Biotechnologies)[54,55] was used from a subset of tumor samples collected pre-therapy from

responders (CR/PR/SD) in the Riaz et al. cohort[40]. Subsequently the clonality of the TCR CDR3 repertoire, defined as the complement of evenness (that is, 1 – evenness), was calculated. Evenness is defined as the observed Shannon entropy *(H)* divided by the maximum possible *H*, given the number of unique elements in a population. Correlation analyses were performed using Pearson's *r*.

Genomic oncoprint.

The oncoprint displays mutated genes that have been reported to impact response to ICIs. The genes in the *IFNG* gene cluster on 9p are: *IFNA1, IFNA10*, *IFNA13*, *IFNA14*, *IFNA16*, *IFNA17*, *IFNA2, IFNA21*, *IFNA22P*, *IFNA4, IFNA*, *IFNA6, IFNA7*, *IFNA8*, *IFNB1, IFNE*, *IFNW1.* The Loss events were identified in the following manner: (1) rounded FACETS ploidy value to the nearest whole number; (2) used rounded ploidy value to correct total copy number (tcn.em) with: Corrected_TCN = tcn.em – rounded_ploidy; (3) if Corrected_TCN <= – 1, then marked as a "Loss" event. Note that this computation was performed for each FACETS[56] segment on chromosome 9 and was assigned to individual genes with coordinates within the FACETS segment. Homozygous loss events were identified if tcn.em = 0 (ploidy-corrected TCN was not used). All losses were manually verified. For assessing loss of heterozygosity of HLA-I, copy number variation analysis was performed using FACETS 0.5.6 to determine allele-specific copy number[56]. Segments within the chromosome 6p locus were identified containing the *HLA-A*, *HLA-B* and *HLA-C* loci. Loss of heterozygosity was defined as a minor allele copy number estimate of 0 for any of the *HLA*-I loci using the expectation-maximization model[56].

Peptide correlation analyses.

All HED-peptide correlation analyses were limited to patients heterozygous at each locus only. For the analyses of neo-, viral and self-peptides in Fig. 4, the *y* axes show the mean number of peptides bound uniquely to each allele for each of *HLA-A*, *-B* and *-C*. For the analyses of neo-, viral and self-peptides at individual loci in Extended Data Fig. 10, two patients had an *HLA-C* genotype (*C*03:03,C*03:04*) that bound 0 peptides. These patients were excluded from the plots for visualization purposes only. The correlations are significant regardless of whether these patients are included. The nonparametric Kendall's correlation was used as shown in Pierini and Lenz[3], because parameters were not normally distributed and ties could be detected in the data. For the analyses of neo- and viral peptides, one-sided *P* values—given the prior association of Grantham distance with diversity of nonself viral peptides shown by Pierini and Lenz[3]; it was hypothesized that a similar association would be observed for the neopeptide correlations. For the analyses of self-peptides from the human proteome, there was no prior hypothesis regarding the direction of the association between HED and peptide diversity; thus, two-sided *P* values were used.

Statistical analyses.

Comparisons of HED distributions across individual *HLA*-I loci were calculated using the Mann–Whitney test. Survival analyses were performed using the Kaplan–Meier estimator. For Figs. 2 and 3, all cutoffs for high germline HED and high TMB were determined using the top quartile, and for low HED and low TMB were defined as values less than the top quartile. For the pan-cancer analyses in Extended Data Fig. 9, cutoffs for high HED were

determined using the median or top quartile, and cutoffs for low HED were determined using all values less than the median or bottom quartile, respectively. For analyses in Fig. 3 combining cohorts with whole-exome and targeted panel sequencing, the TMB of the whole-exome cohorts was divided by 30 to normalize per megabase[38]. The survival analysis was performed in the Van Allen et al.[8] cohort (cohort 1) using the mean of divergences at *HLA-A, -B* and *-C* as well as the sum, median and geometric mean. Results were similar across all metrics used (see Supplementary Table 3). For the analysis in Extended Data Fig. 3, each mean HED value in the dataset was used as a cut point for high mean HED in survival analysis, and HRs were calculated from univariable Coxs regression. These HRs were plotted against all mean HED values. For the analysis in Fig. 3g, each value of mean HED in the data was used as a cut point for high HED, and the same was done for TMB. When combining mean HED and TMB, patients were in the high group if their mean HED and TMB were both greater than the cut points for mean HED and TMB, and in the low group if both variables were less than their respective cut points. For all multivariable analyses, *P* values and HRs were calculated using Coxs regression. For all survival analyses, *P* values were calculated using the log-rank test, and HRs using univariable Coxs regression. For the analyses of clinical response data, *P* values and ORs were calculated using Fishers exact test (two-sided). All survival and correlation analyses were performed in the R Statistical Computing Environment v.3.5.0 (http://www.r-project.org)

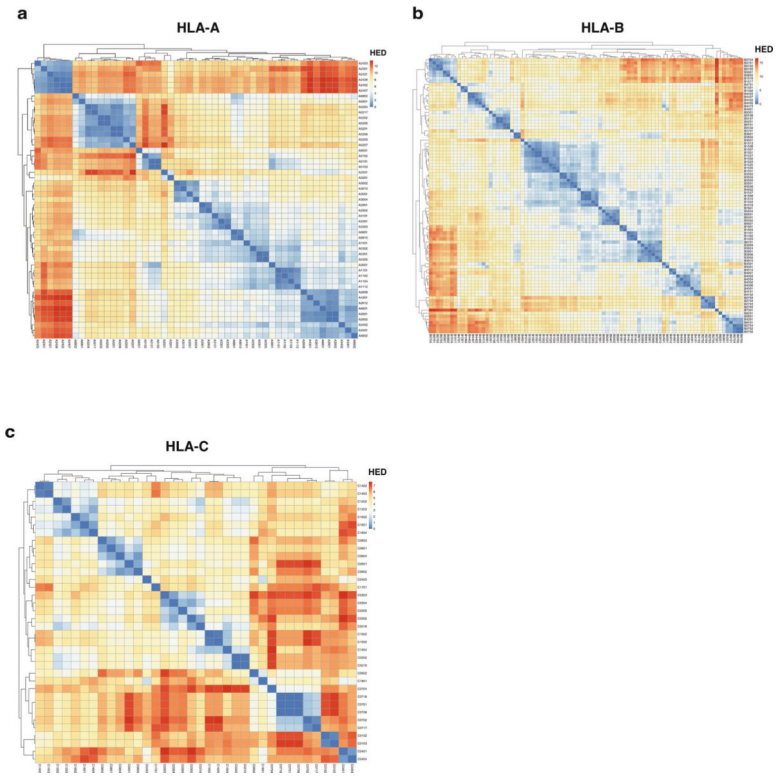
Reporting Summary.

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

Data availability

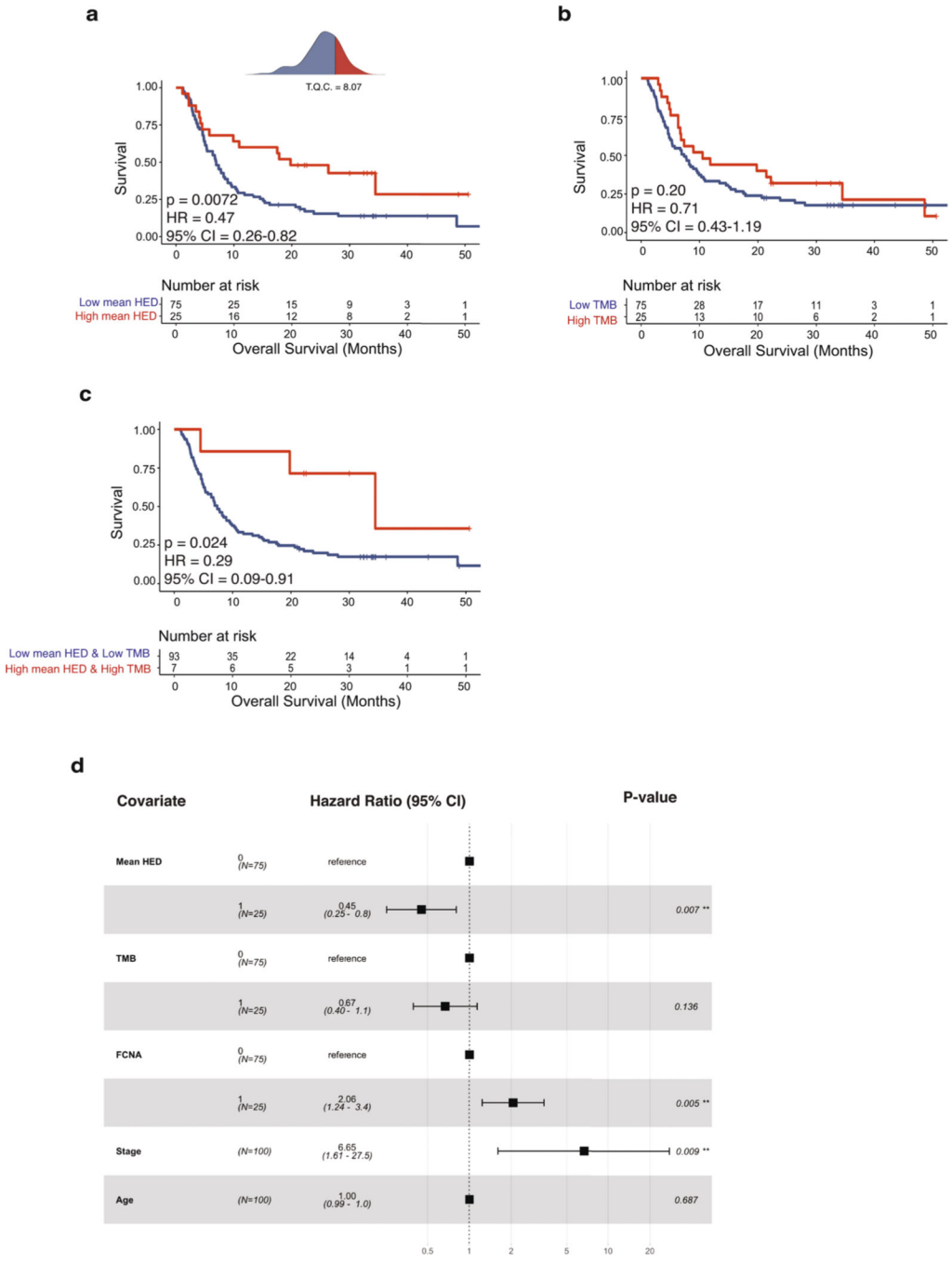
The data from prior studies are available at the following accession numbers: dbGaP, phs001041.vl.pl (Snyder et al.[6]); dbGaP, phs000452.v2.pl (Van Allen et al.[8]); SRA, SRP067938 (Hugo et al.[41]) and SRP090294 (Hugo et al.[41]); dbGaP, phs000980.vl.pl (Rizvi et al.[7]); SRA, SRP095809 and BioProject, PRJNA359359 (Riaz et al.[7]); SRA, PRJNA419415 (Chowell et al.[10]), PRJNA419422 (Chowell et al.[10]) and PRJNA419530 (Chowell et al.[10]); cBioPortal for Cancer Genomics, http://cbioportal.org/msk-impact.

Extended Data



Extended Data Fig. 1 |. Hierarchical clustering of HLA-I evolutionary divergences at individual HLA class I loci.

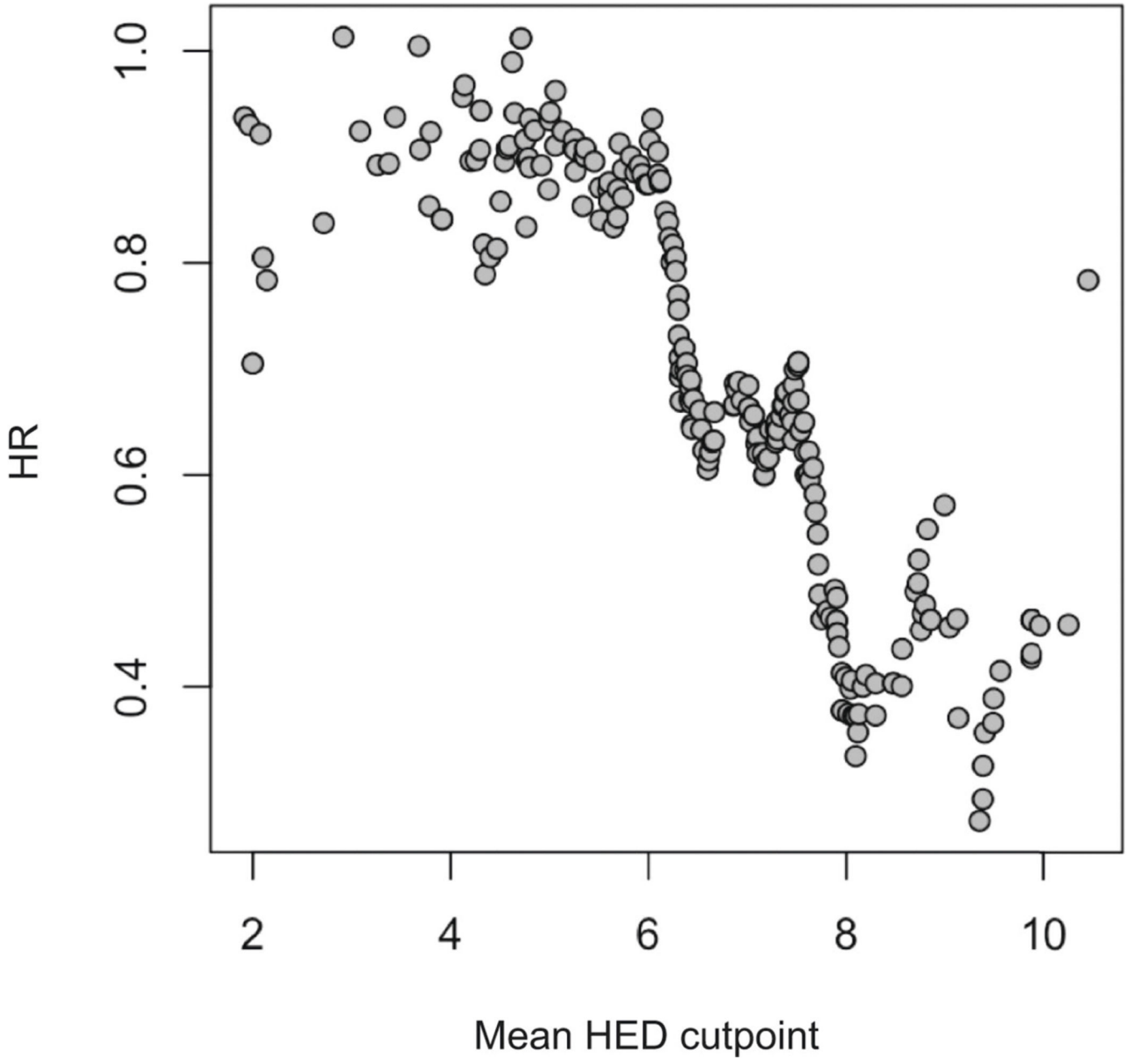
a, Hierarchical clustering of HED at *HLA-A* using all *HLA-A* alleles from all patient cohorts. **b,** Hierarchical clustering of HED at *HLA-B* using all *HLA-B* alleles. **c,** Hierarchical clustering of HED at *HLA-C* using all *HLA-C* alleles. Heat maps shows z-score normalized HED across all alleles. Color gradient of blue to red indicates low HED between allele pairs to high HED between allele pairs, respectively.



Extended Data Fig. 2 |. Mean HLA-I evolutionary divergence is associated with improved benefit from immune checkpoint inhibitors.

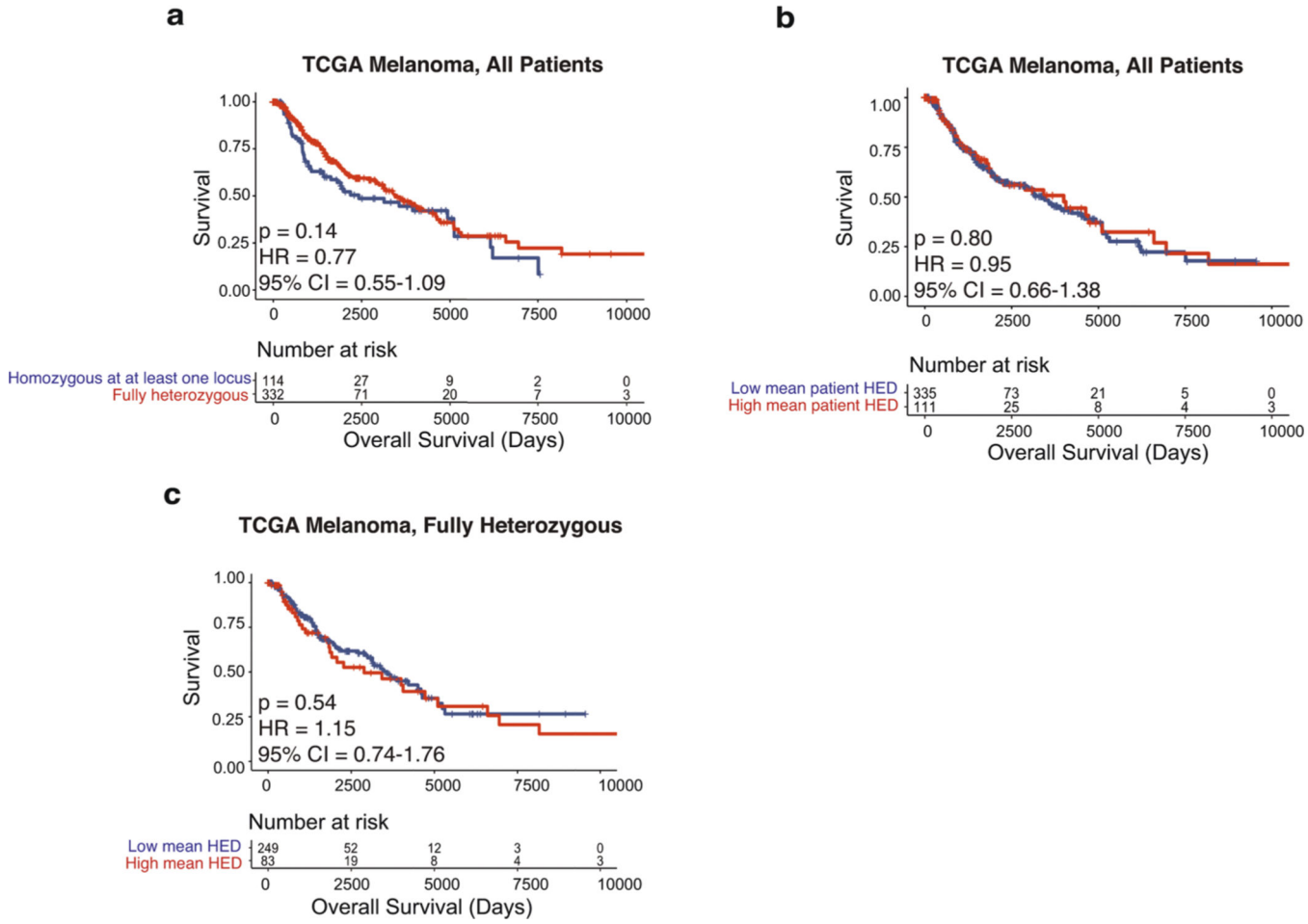
a, Association of high mean HED (red) with improved efficacy of anti-CTLA-4 treatment in a cohort of patients with metastatic melanoma; $P = 0.0072$; two-sided log-rank test. Density plots indicate the distribution of mean HED and cutoff used in the survival curves. T.Q.C. = top quartile cutoff, HR = hazard ratio, CI = confidence interval. **b,** Association of high (top quartile) tumor mutational burden (TMB) with overall survival after anti-CTLA-4 treatment; $P = 0.20$; two-sided log-rank test. **c,** Association of high mean HED and high TMB (red)

with improved overall survival after anti-CTLA4 treatment; $P = 0.024$; two-sided log-rank test. **d,** Multivariable Cox proportional-hazards model including mean HED and other clinical variables. Data show independent effect of mean HED associated with improved survival after anti-CTLA-4. HED, TMB, and fraction of copy number alterations (FCNA) are dichotomized into high (1) and low (0) groups based on the top quartile for each variable. *P* values calculated using two-sided log-rank test. Horizontal lines represent the 95% confidence interval.



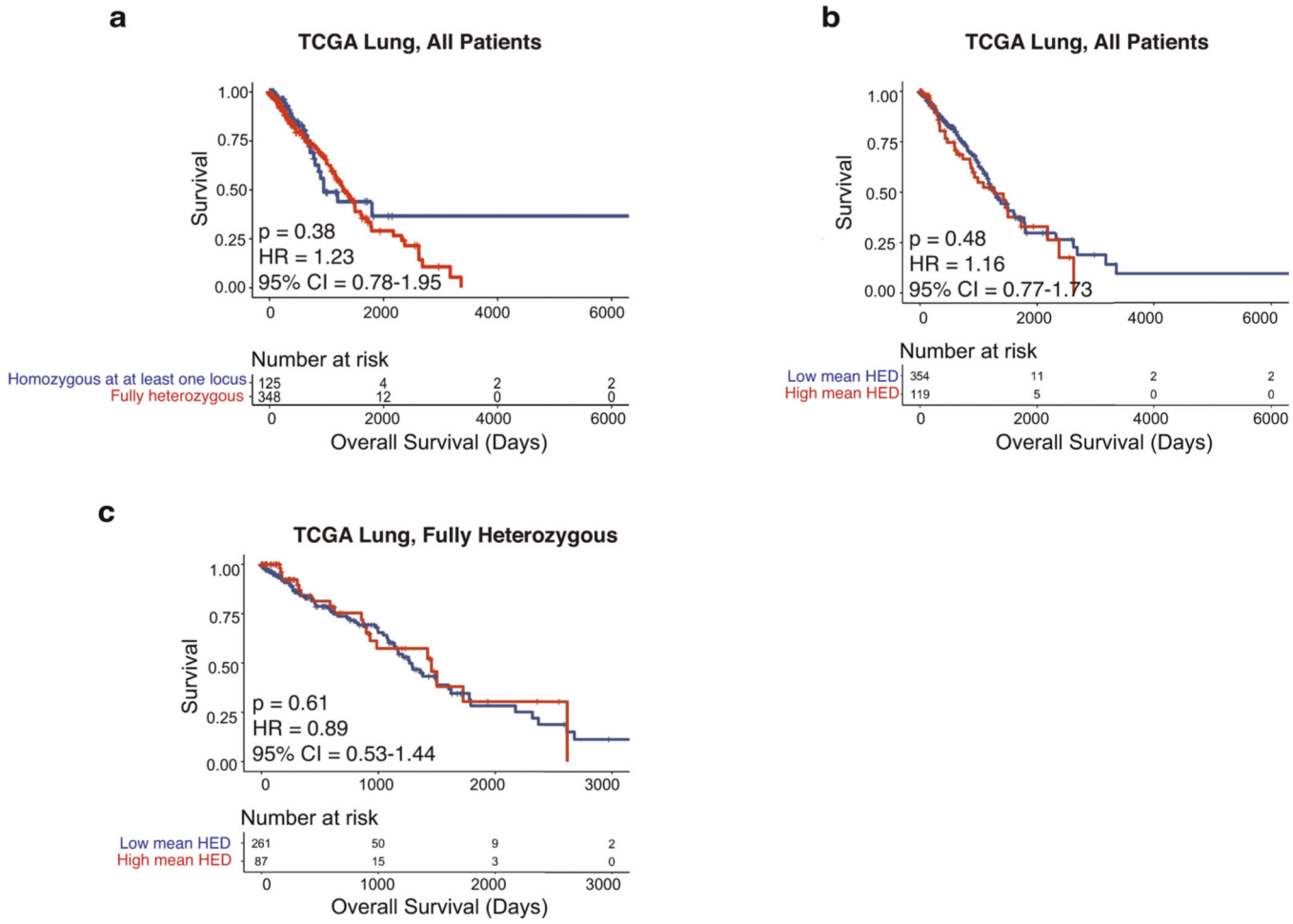
Extended Data Fig. 3 |. Effect of mean HLA-I evolutionary divergence on hazard ratio from survival across all possible cutpoints.

Cutpoint analysis showing the relationship between mean HED and hazard ratio. Data show a negative relationship between mean HED and hazard ratio across all possible cutpoints for mean HED, indicating improved overall survival as mean HED increases.



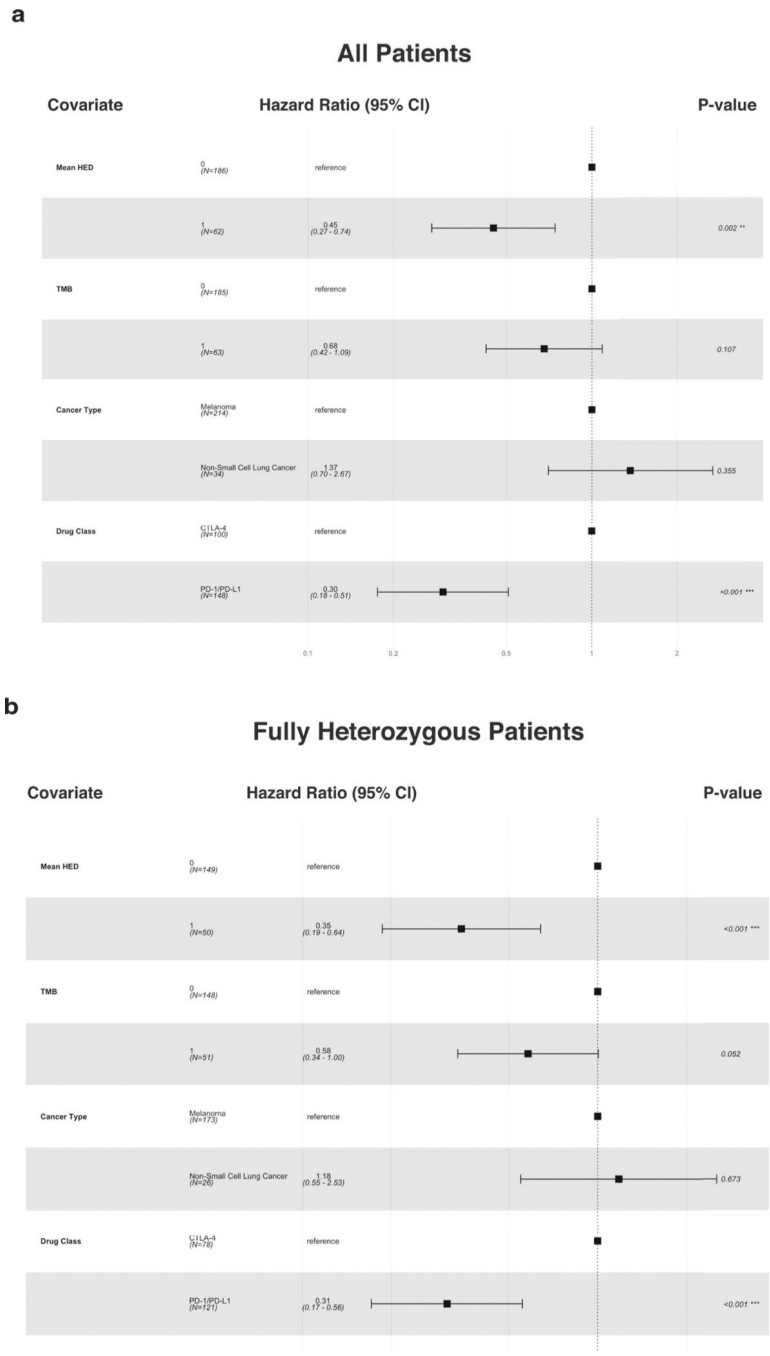
Extended Data Fig. 4 |. Neither HLA-I heterozygosity nor HLA-I evolutionary divergence is associated with prognosis in TCGA melanoma patients.

a, Full heterozygosity at *HLA*-I (red) is not associated with prognosis in TCGA melanoma patients; $P = 0.14$, two-sided log-rank test. **b,** High patient mean HED (red) is not associated with prognosis in TCGA melanoma patients; $P = 0.80$, two-sided log-rank test. **c,** High mean HED (red) is not associated with prognosis in TCGA melanoma patients fully heterozygous at *HLA*-I; $P = 0.54$; two-sided log-rank test.



Extended Data Fig. 5 |. Neither HLA-I heterozygosity nor HLA-I evolutionary divergence is associated with prognosis in TCGA lung cancer patients.

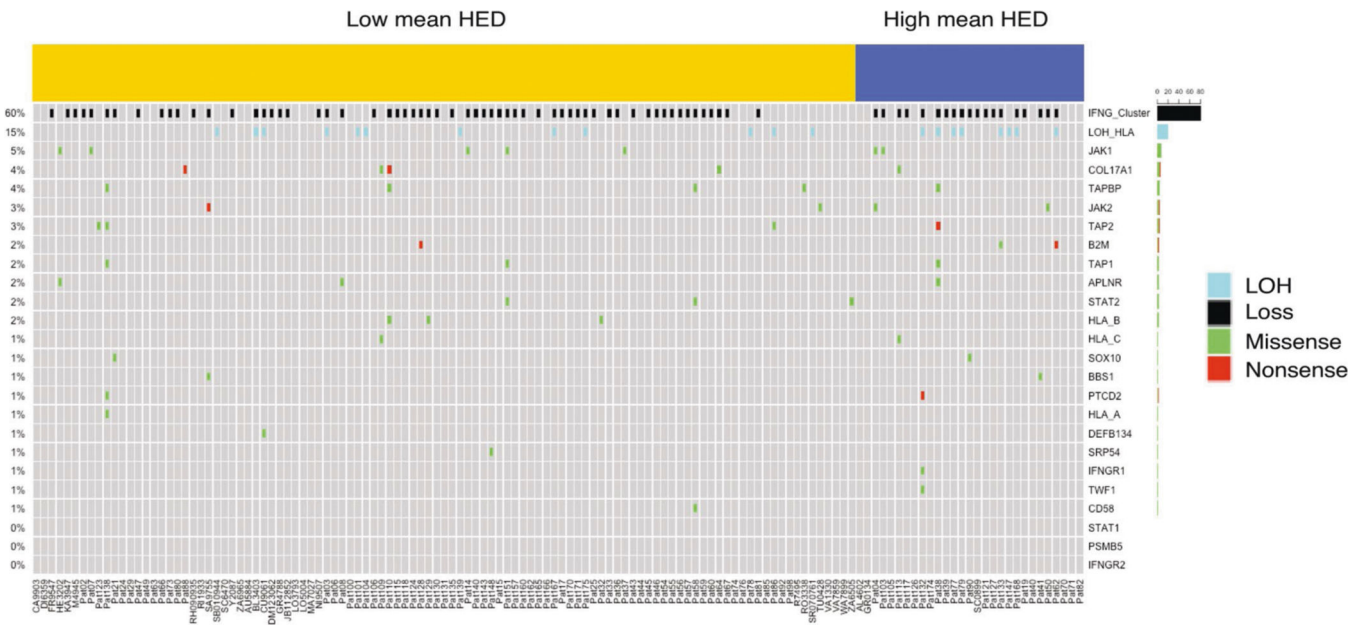
a, Full heterozygosity at *HLA*-I (red) is not associated with prognosis in TCGA lung cancer patients; *P*= 0.38, two-sided log-rank test. **b,** High mean HED is not associated with prognosis in patients from Extended Data Fig. 4a; *P* = 0.48, log-rank test. **c,** High mean HED (red) is not associated with prognosis in TCGA lung cancer patients fully heterozygous at *HLA*-I; *P* = 0.51, two-sided log-rank test



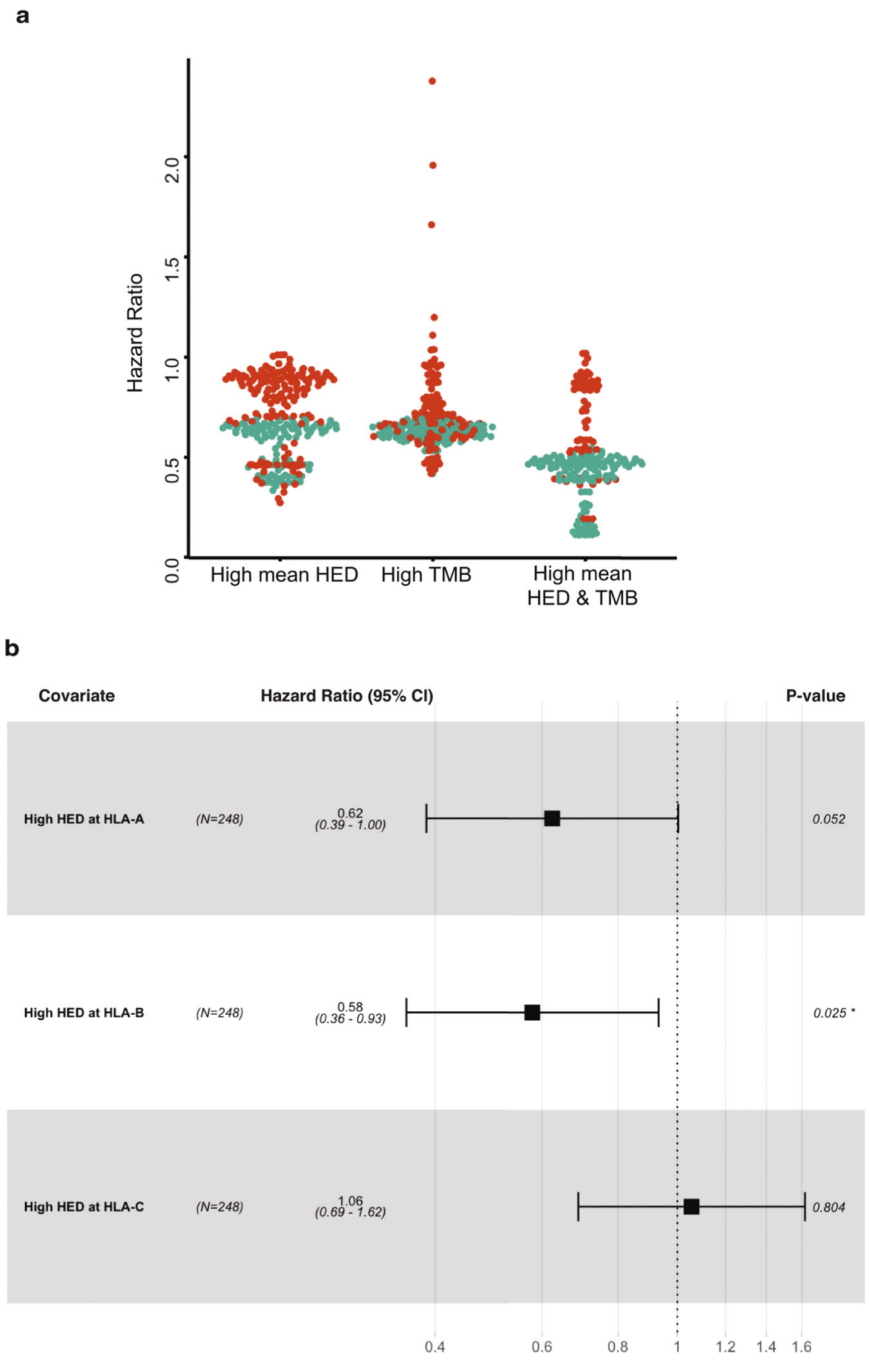
Extended Data Fig. 6 |. The effects of mean HLA-I evolutionary divergence and tumor mutational burden are independent of cancer type and drug class.

a, Multivariable Cox proportional-hazards model including mean HED and other clinical variables using all patients. Data show independent effect of mean HED in predicting response to ICI. Drug class $P = 8.14e\text{-}06$. *P* values calculated using two-sided log-rank test. Horizontal lines indicate 95% confidence interval. **b,** Multivariable Cox proportional-hazards model including mean HED and other clinical variables using patients fully heterozygous at *HLA*-I. Data show independent effect of mean HED associated with improved survival after ICI therapy. Mean HED $P = 7.25e\text{-}04$; Drug Class $P = 9.32e\text{-}05$.

HED and TMB are dichotomized into high (1) and low (0) groups using the top quartile for each variable. *P* values calculated using two-sided log-rank test. Horizontal lines represent the 95% confidence interval.



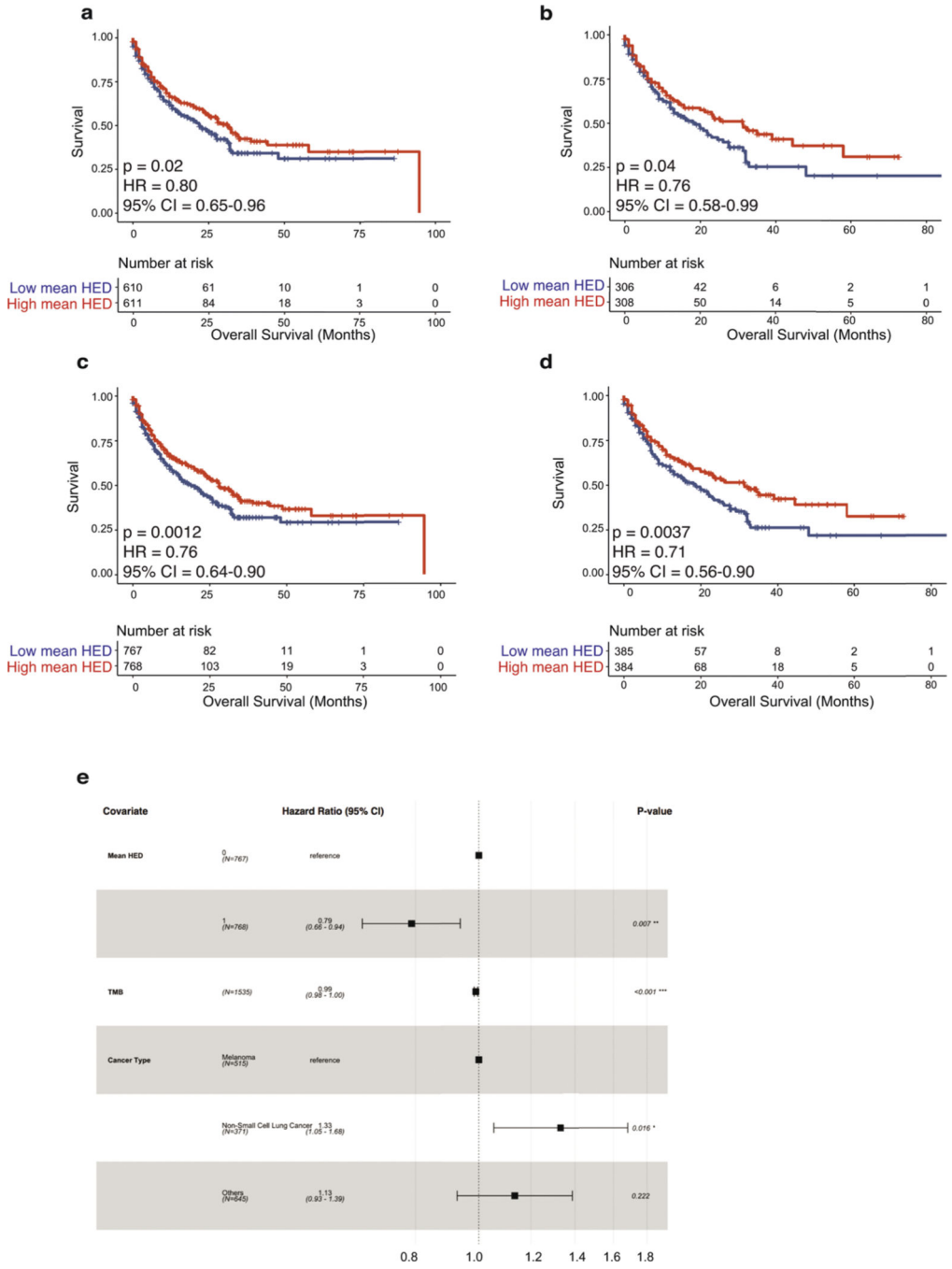
Extended Data Fig. 7 |. Oncoprint showing mutations in genes in our patient cohorts. Data show no difference in proportion of patients with mutations in the presented genes between patients with high mean HLA-I evolutionary divergence (HED) and low mean HED. LOH = loss of heterozygosity at *HLA*-I.



Extended Data Fig. 8 |. Combined effect of HED and TMB on survival after ICI administration and multivariable analysis of HED at individual loci.

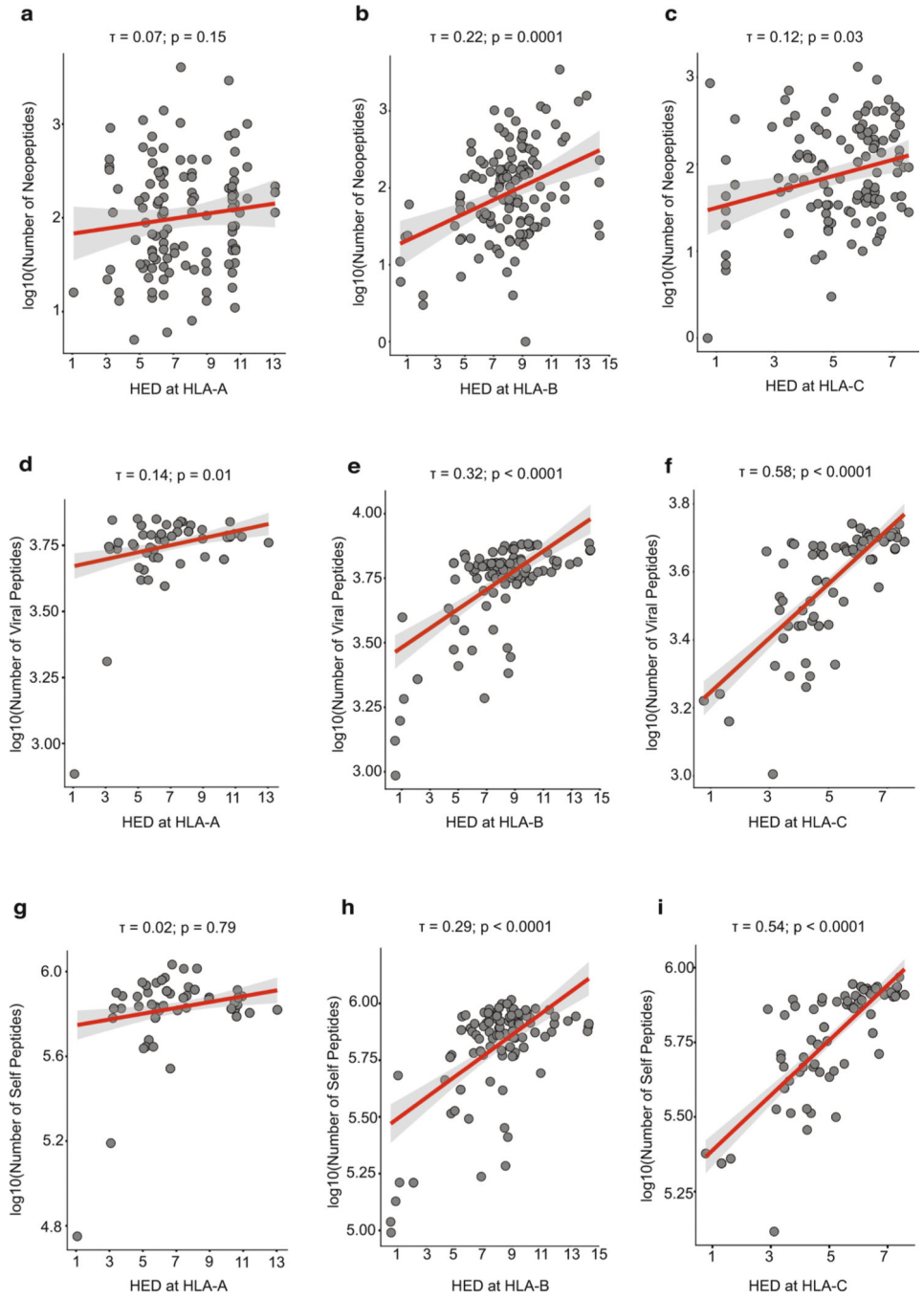
a, Cutpoint analysis showing the association of both high mean HED and high TMB with improved survival after ICI (same distributions as Fig. 3g; $n = 248$). Data show a reduction in hazard ratio when combining HED and TMB compared to either variable alone. Green indicates two-sided log-rank p-value < 0.05; red indicates non-significant log-rank p-value. **b,** Multivariable cox regression analysis demonstrating the effect of HED at individual loci on overall survival after ICI administration. Data indicate that high HED at *HLA-A* and

HLA-B are each associated with improved overall survival after ICI. *P* values calculated using two-sided log-rank test. Horizontal lines represent the 95% confidence interval.



Extended Data Fig. 9 |. Effect of mean HLA-I evolutionary divergence and tumor mutational burden on efficacy of immune checkpoint inhibitor treatment in an independent set of patients. **a,** Association of high mean HED (red) with improved overall survival after ICI in an independent pan-cancer dataset of patients described in Chowell *et al.* These patients do not overlap with those presented in Figs. 2 & 3. Cutoff was determined using the median mean HED across the cohort. *P* = 0.02; two-sided log-rank test. HR = hazard ratio, CI =

confidence interval. **b,** Association of high mean HED (red) with improved overall survival after ICI in an independent, pan-cancer dataset of patients described in Chowell *et al.* These patients do not overlap with those presented in Figs. 2 & 3. Patients in the red curve have mean HED greater than or equal to the top quartile; patients in the blue curve have mean HED less than or equal to the first quartile. $P = 0.04$; two-sided log-rank test **c,** Association of high mean HED (red) with improved overall survival after ICI in all patients described in Chowell *et al.* Cutoff was determined using the median mean HED across the cohort. $P = 0.0012$; two-sided log-rank test. **d,** Association of high mean HED (red) with improved overall survival after ICI in all patients described in Chowell *et al.* Patients in the red curve have mean HED greater than or equal to the top quartile; patients in the blue curve have mean HED less than or equal to the first quartile. $P = 0.0037$; two-sided log-rank test. **e,** Multivariable Cox proportional-hazards model including mean HED and other variables. Data show independent effect of mean HED on improved survival after ICI administration when adjusting for TMB and cancer type. Mean HED is dichotomized into high (1) and low (0) groups using the median; TMB is treated as a continuous variable. TMB $P = 0.0008$. P values calculated using two-sided log-rank test. Horizontal lines represent the 95% confidence interval.



Extended Data Fig. 10 |. Association of HLA-I evolutionary divergence at each class I locus with diversity of tumor and human immunopeptidomes.

a, Correlation of HED at *HLA-A* with number of unique neopeptides bound to *HLA-A* alleles of each patient genotype using all patients heterozygous at *HLA-A* from Fig. 2 (n = 118) for whom neopeptide data were available; P = 0.15; one-sided Kendall's rank correlation. **b,** Correlation of HED at *HLA-B* with number of unique neopeptides bound to *HLA-B* alleles of each patient genotype using patients heterozygous at *HLA-B* (n = 129); P = 0.001; one-sided Kendall's rank correlation **c,** Correlation of HED at *HLA-C* with number

of unique neopeptides bound to *HLA-C* alleles of each patient genotype using patients heterozygous at *HLA-C* ($n = 118$); $P = 0.03$; one-sided Kendall's rank correlation. **d,** Correlation of HED at *HLA-A* with number of unique viral peptides bound to *HLA-A* alleles of each patient genotype using patients heterozygous at HLA-A ($n = 118$); $P = 0.01$; one-sided Kendall's rank correlation. d, Correlation of HED at *HLA-B* with number of unique viral peptides bound to *HLA-B* alleles of each patient genotype using patients heterozygous at *HLA-B* ($n = 129$); $P < 0.0001$; one-sided Kendall's rank correlation. **f,** Correlation of HED at *HLA-C* with number of unique viral peptides bound to *HLA-C* alleles of each patient genotype using patients heterozygous at *HLA-C;* $P < 0.0001$. **g,** Correlation of HED at *HLA-A* with number of unique self peptides bound to *HLA-A* alleles of each patient genotype using patients heterozygous at *HLA-A* ($n = 118$); $P = 0.79$; two-sided Kendall's rank correlation. **h,** Correlation of HED at *HLA-B* with number of unique self peptides bound to *HLA-B* alleles of each patient genotype using patients heterozygous at *HLA-B* ($n = 129$); $P < 0.0001$; two-sided Kendall's rank correlation. **i,** Correlation of HED at *HLA-C* with number of unique self peptides bound to *HLA-C* alleles of each patient genotype using patients heterozygous at HLA-C ($n = 118$); $P < 0.0001$; two-sided Kendall's rank correlation. Red line indicates line of best linear fit.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank the Chan lab and members of the Immunogenomics and Precision Oncology Platform for advice and input. This work was supported in part by the National Institutes of Health (NIH) grant no. R35 CA232097 (to T.A.C.), NIH grant no. RO1 CA205426 (to T.A.C. and N.A.R.), the Paine Webber Chair (to T.A.C.), and the NIH/National Cancer Institute's Cancer Center support grant (no. P30 CA008748) and the Deutsche Forschungsgemeinschaft grant no. LE 2593/3-1 (to T.L.L.).

Competing interests

T.A.C. is a co-founder of Gritstone Oncology and holds equity. T.A.C. holds equity in An2H. T.A.C. acknowledges grant funding from Bristol-Myers Squibb, AstraZeneca, Illumina, Pfizer, An2H and Eisai. T.A.C. has served as an advisor for Bristol-Myers, MedImmune, Squibb, Illumina, Eisai, AstraZeneca and An2H. N.A.R. is a consultant/advisory board member for AstraZeneca, BMS, Roche, Merck, Novartis, Lilly and Pfizer. T.A.C. and N.A.R are cofounders of Gritstone Oncology. T.A.C., N.A.R, L.G.T.M. and D.C. hold ownership of intellectual property on using TMB to predict immunotherapy response, with pending patent, which has been licensed to PGDx.

References

1. Parham P & Ohta T Population biology of antigen presentation by MHC class I molecules. Science 272, 67–74 (1996). [PubMed: 8600539]
2. Wakeland EK et al. Ancestral polymorphisms of MHC class II genes: divergent allele advantage. Immunol Res. 9, 115–122 (1990). [PubMed: 2189934]
3. Pierini F & Lenz TL Divergent allele advantage at human MHC genes: signatures of past and ongoing selection. Mol. Biol. Evol 35, 2145–2158 (2018). [PubMed: 29893875]
4. McGranahan N et al. Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. Science 351, 1463–1469 (2016). [PubMed: 26940869]
5. Samstein RM et al. Tumor mutational load predicts survival after immunotherapy across multiple cancer types. Nat. Genet 51, 202–206 (2019). [PubMed: 30643254]

6. Snyder A et al. Genetic basis for clinical response to CTLA-4 blockade in melanoma. N. Engl J. Med 371, 2189–2199 (2014). [PubMed: 25409260]
7. Rizvi NA et al. Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. Science 348, 124–128 (2015). [PubMed: 25765070]
8. Van Allen EM et al. Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. Science 350, 207–211 (2015). [PubMed: 26359337]
9. Schumacher TN & Schreiber RD Neo antigens in cancer immunotherapy. Science 348, 69–74 (2015). [PubMed: 25838375]
10. Chowell D et al. Patient HLA class I genotype influences cancer response to checkpoint blockade immunotherapy. Science 359, 582–587 (2018). [PubMed: 29217585]
11. Carrington M et al. HLA and HIV-1: heterozygote advantage and B*35-Cw*04 disadvantage. Science 283, 1748–1752 (1999). [PubMed: 10073943]
12. Penn DJ, Damjanovich K & Potts WK MHC heterozygosity confers a selective advantage against multiple-strain infections. Proc. Natl Acad. Sci. USA 99, 11260–11264 (2002). [PubMed: 12177415]
13. Thursz MR, Thomas HC, Greenwood BM & Hill AV Heterozygote advantage for HLA class-II type in hepatitis B virus infection. Nat. Genet 17, 11–12 (1997). [PubMed: 9288086]
14. Doherty PC & Zinkernagel RM A biological role for the major histocompatibility antigens. Lancet i, 1406–1409 (1975).
15. Doherty PC & Zinkernagel RM Enhanced immunological surveillance in mice heterozygous at H-2 gene complex. Nature 256, 50–52 (1975). [PubMed: 1079575]
16. Hughes AL & Yeager M Natural selection at major histocompatibility complex loci of vertebrates. Annu. Rev. Genet 32, 415–435 (1998). [PubMed: 9928486]
17. Robinson J et al. Distinguishing functional polymorphism from random variation in the sequences of > 10,000 HLA-A, -B and -C alleles. PLoS Genet. 13, e1006862 (2017).
18. Gfeller D & Bassani-Sternberg M Predicting antigen presentation-what could we learn from a million peptides? Front. Immunol 9, 1716 (2018). [PubMed: 30090105]
19. Paul S et al. HLA class I alleles are associated with peptide-binding repertoires of different size, affinity, and immunogenicity. J. Immunol 191, 5831–5839 (2013). [PubMed: 24190657]
20. Marty R et al. MHC-I genotype restricts the oncogenic mutational landscape. Cell 171, 1272–1283.e15 (2017).
21. Marty R, Thompson WK, Salem RM, Zanetti M & Carter H Evolutionary pressure against MHC Class II binding cancer mutations. Cell 175, 416–428 e413 (2018).
22. McGranahan N et al. Allele-specific HLA loss and immune escape in lung cancer evolution. Cell 171, 1259–1271.e 11 (2017). [PubMed: 29107330]
23. Rosenthal R et al. Neo antigen-directed immune escape in lung cancer evolution. Nature 567, 479–485 (2019). [PubMed: 30894752]
24. Potts WK & Wakeland EK Evolution of diversity at the major histocompatibility complex. Trends Ecol. Evol 5, 181–187 (1990). [PubMed: 21232350]
25. Lenz TL Computational prediction of MHC II-antigen binding supports divergent allele advantage and explains trans-species polymorphism. Evolution 65, 2380–2390 (2011). [PubMed: 21790583]
26. Grantham R Amino acid difference formula to help explain protein evolution. Science 185, 862–864 (1974). [PubMed: 4843792]
27. McKenzie LM, Pecon-Slattery J, Carrington M & O'Brien SJ Taxonomic hierarchy of HLA class I allele sequences. Genes Immun. 1, 120–129 (1999). [PubMed: 11196658]
28. Buhler S, Nunes JM & Sanchez-Mazas A HLA class I molecular variation and peptide-binding properties suggest a model of joint divergent asymmetric selection. Immunogenetics 68, 401–416 (2016). [PubMed: 27233953]
29. Grueber CE, Wallis GP & Jamieson IG Episodic positive selection in the evolution of avian toll-like receptor innate immunity genes. PLoS ONE 9, e89632 (2014).
30. Subramanian S & Kumar S Evolutionary anatomies of positions and types of disease-associated and neutral amino acid mutations in the human genome. BMC Genom. 7, 306 (2006).

31. Wain LV et al. Adaptation of HIV-1 to its human host. Mol. Biol. Evol 24, 1853–1860 (2007). [PubMed: 17545188]
32. International Wheat Genome Sequencing Consortium. A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. Science 345, 1251788 (2014).
33. Rentoft M et al. Heterozygous colon cancer-associated mutations of SAMHD1 have functional significance. Proc. Natl Acad. Sci. USA 113, 4723–4728 (2016). [PubMed: 27071091]
34. Sundaram L et al. Predicting the clinical impact of human mutation with deep neural networks. Nat. Genet 50, 1161–1170 (2018). [PubMed: 30038395]
35. Abelin JG et al. Mass spectrometry profiling of HLA-associated peptidomes in mono-allelic cells enables more accurate epitope prediction. Immunity 46, 315–326 (2017). [PubMed: 28228285]
36. Bradburn MJ, Clark TG, Love SB & Altman DG Survival analysis part II: multivariate data analysis-an introduction to concepts and methods. Br. J. Cancer 89, 431–436 (2003). [PubMed: 12888808]
37. Broström Gr Event history analysis with R. (CRC Press, Boca Raton, FL, 2012).
38. Zehir A et al. Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. Nat. Med 23, 703–713 (2017). [PubMed: 28481359]
39. Pearson H et al. MHC class I-associated peptides derive from selective regions of the human genome. J. Clin. Invest 126, 4690–4701 (2016). [PubMed: 27841757]
40. Riaz N et al. Tumor and microenvironment evolution during immunotherapy with nivolumab. Cell 171, 934–949.el6 (2017).

References

41. Hugo W et al. Genomic and transcriptomic features of response to anti-PD-1 therapy in metastatic melanoma. Cell 165, 35–44 (2016). [PubMed: 26997480]
42. Shukla SA et al. Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes. Nat. Biotechnol 33, 1152–1158 (2015). [PubMed: 26372948]
43. Robinson J et al. The IPD and IMGT/HLA database: allele variant databases. Nucleic Acids Res. 43, D423–D431 (2015). [PubMed: 25414341]
44. Zerbino DR et al. Ensembl 2018. Nucleic Acids Res. 46, D754–D761 (2018). [PubMed: 29155950]
45. Edgar RC MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797 (2004). [PubMed: 15034147]
46. Li H & Durbin R Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25, 1754–1760 (2009). [PubMed: 19451168]
47. McKenna A et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20, 1297–1303 **(**2010**).** [PubMed: 20644199] **().**
48. Cibulskis K et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat. Biotechnol 31, 213–219 (2013). [PubMed: 23396013]
49. Koboldt DC et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 22, 568–576 (2012). [PubMed: 22300766]
50. Larson DE et al. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. Bioinformatics 28, 311–317 (2012). [PubMed: 22155872]
51. Saunders CT et al. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics 28, 1811–1817 (2012). [PubMed: 22581179]
52. Wei L et al. MAC: identifying and correcting annotation for multi-nucleotide variations. BMC Genom. 16, 569 (2015).
53. Jurtz V et al. NetMHCpan-4.0: improved peptide-MHC Class I interaction predictions integrating eluted ligand and peptide binding affinity data. J. Immunol 199, 3360–3368 (2017). [PubMed: 28978689]
54. Carlson CS et al. Using synthetic templates to design an unbiased multiplex PCR assay. Nat. Commun 4, 2680 (2013). [PubMed: 24157944]

55. Robins HS et al. Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. Blood 114, 4099–4107 (2009). [PubMed: 19706884]

56. Shen RL & Seshan VE FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. Nucleic Acids Res. 44, el31 (2016).

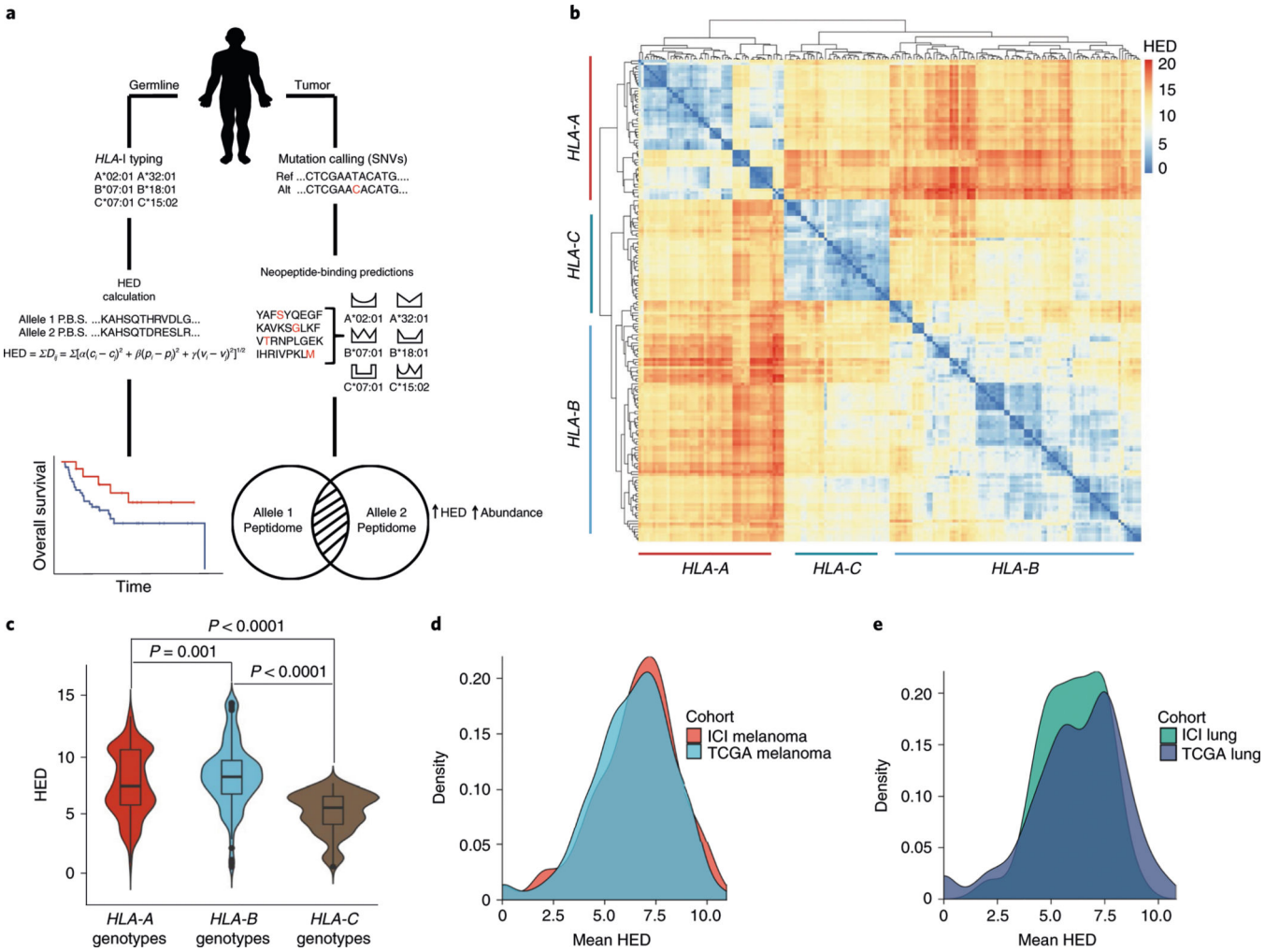


Fig. 1 |. Landscape of HEDs at *HLA-A, -B* and *-C*.

a, Schematic of experimental design. HEDs are calculated between peptide-binding domains using the Grantham distance and then used to stratify patients treated with ICI s. Predicted neopeptides are called using whole-exome sequencing from the patient's tumor, counted and correlated with HED. Predicted viral and self peptides were also correlated with HED. **b,** Hierarchical clustering of HED at *HLA-A, HLA-B* and *HLA-C* (*HLA*-I). The heatmap shows z score-normalized HED across all alleles in all patient cohorts. The color gradient of blue to red indicates low HED between allele pairs to high HED between allele pairs, respectively. **c,** Distributions of HED for each *HLA-A, HLA-B* and *HLA-C* heterozygous genotype. *HLA-A* (n = 279 patients; minimum= 1.08, median = 7.62, maximum = 13.20) versus *HLA-B* (n = 300 patients; minimum = 0.53, median = 8.10, maximum = 14.33) (P = 0.001); *HLA-A* versus *HLA-C* (n = 281 patients; minimum = 0.56, median = 5.60, maximum = 7.58; $P < 0.0001$); *HLA-B* versus *HLA-C* ($P < 0.0001$; two-sided Mann-Whitney test). **d,** Distribution of patient mean HED across all melanoma cohorts treated with ICIs (ICI melanoma) and TCGA (TCGA melanoma). **e,** Distribution of patient mean HED across all lung cancer cohorts treated with ICIs (ICI lung) and TCGA (TCGA lung).

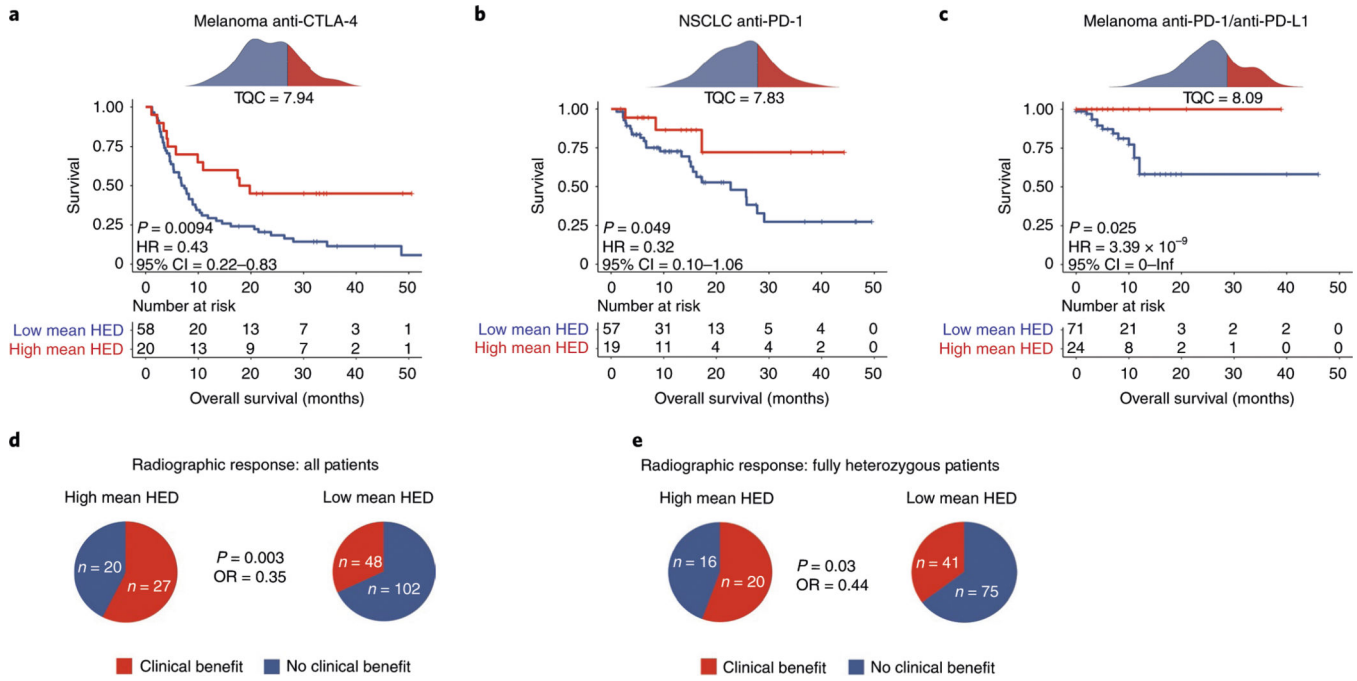


Fig. 2 |. High mean HED is associated with improved response and survival to ICIs.

a, Association of high mean HED (red) with improved survival after anti-CTLA-4 treatment in a cohort of metastatic melanoma patients fully heterozygous at HLA-I ($P = 0.0094$; two-sided log-rank test). Density plots indicate the distribution and cutoff for mean HED used in the survival curves. TQC, top quartile cutoff. **b,** Association of high mean HED (red) with improved survival after anti-PD-1 treatment in an independent cohort of patients with NSCLC fully heterozygous at HLA-I ($P = 0.049$; two-sided log-rank test). **c,** Association of high mean HED (red) with improved overall survival in an independent cohort of patients with melanoma fully heterozygous at HLA-I treated with anti-PDI ($P = 0.025$; two-sided log-rank test). **d,** Association of high patient mean HED with clinical response (red) to ICIs, including all patients (both homozygous and heterozygous at *HLA*-I) for whom clinical response data were available from **a-c** ($P = 0.003$; OR = 0.35; two-sided Fisher's exact test). Numbers on pie charts indicate number of patients deriving clinical or no clinical benefit. **e,** Association of high mean HED with clinical response (red) to ICIs, including only patients fully heterozygous at HLA-I for whom clinical response data were available from **a-c** ($P = 0.03$, OR = 0.44; two-sided Fisher's exact test). Numbers on pie charts indicate number of patients deriving clinical or no clinical benefit.

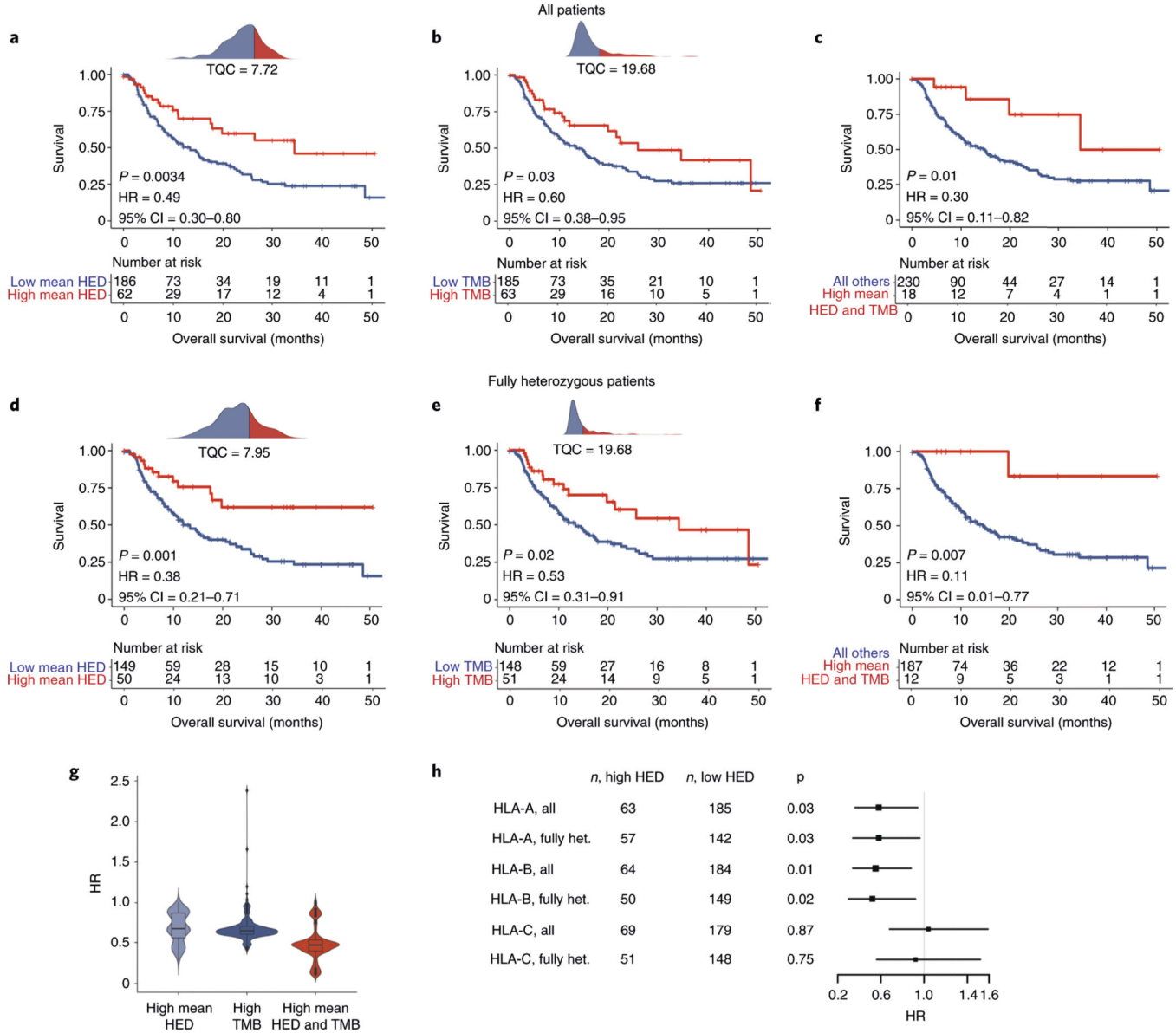


Fig. 3 |. Effect of high mean HED and high TMB on efficacy of ICI treatment.

a, Association of high mean HED (red) with improved overall survival after ICIs in all patients (*HLA*-I homozygous or heterozygous) from Fig. 2 for whom the TMB was available ($P = 0.0034$; two-sided log-rank test). Density plot indicates the distribution and cutoff for mean HED used in the survival curves. **b,** Association of high TMB (red) with improved overall survival after ICIs among all patients $P = 0.03$; two-sided log-rank test). The density plot indicates the distribution and cutoff for TMB used in the survival curves. **c,** Survival of patients with both high mean HED and high TMB (red) after ICI treatment among all patients ($P = 0.01$; two-sided log-rank test). **d,** Association of high mean HED (red) with improved overall survival after ICIs in patients fully heterozygous at *HLA*-I from Fig. 2 for whom TMB was available ($P = 0.001$; two-sided log-rank test). **e,** Association of high TMB with improved overall survival after ICIs among fully heterozygous patients ($P =$

0.02; two-sided log-rank test). **f,** Survival of patients with both high mean HED and high TMB after ICI treatment among fully heterozygous patients ($P = 0.007$; log-rank test). **g,** Cut-point analysis showing the association of both high mean HED and high TMB with improved survival after ICIs ($n = 248$; high mean HED: minimum = 0.27; median = 0.67; maximum= 1.01; high TMB: minimum = 0.42; median = 0.64; maximum = 2.38; high mean HED and TMB: minimum = 0.11; median = 0.47; maximum = 1.02). Data show a reduction in HR when combining HED and TMB compared with either variable alone. **h,** Univariable Cox regression analysis showing the association of high HED (top quartile) at individual *HLA*-I loci with improved survival after ICIs ('all', *HLA*-I homozygous or heterozygous; 'fully het.', fully heterozygous at *HLA*-I; n = number of patients). P values were calculated using a two-sided log-rank test. Horizontal lines represent 95% Cl.

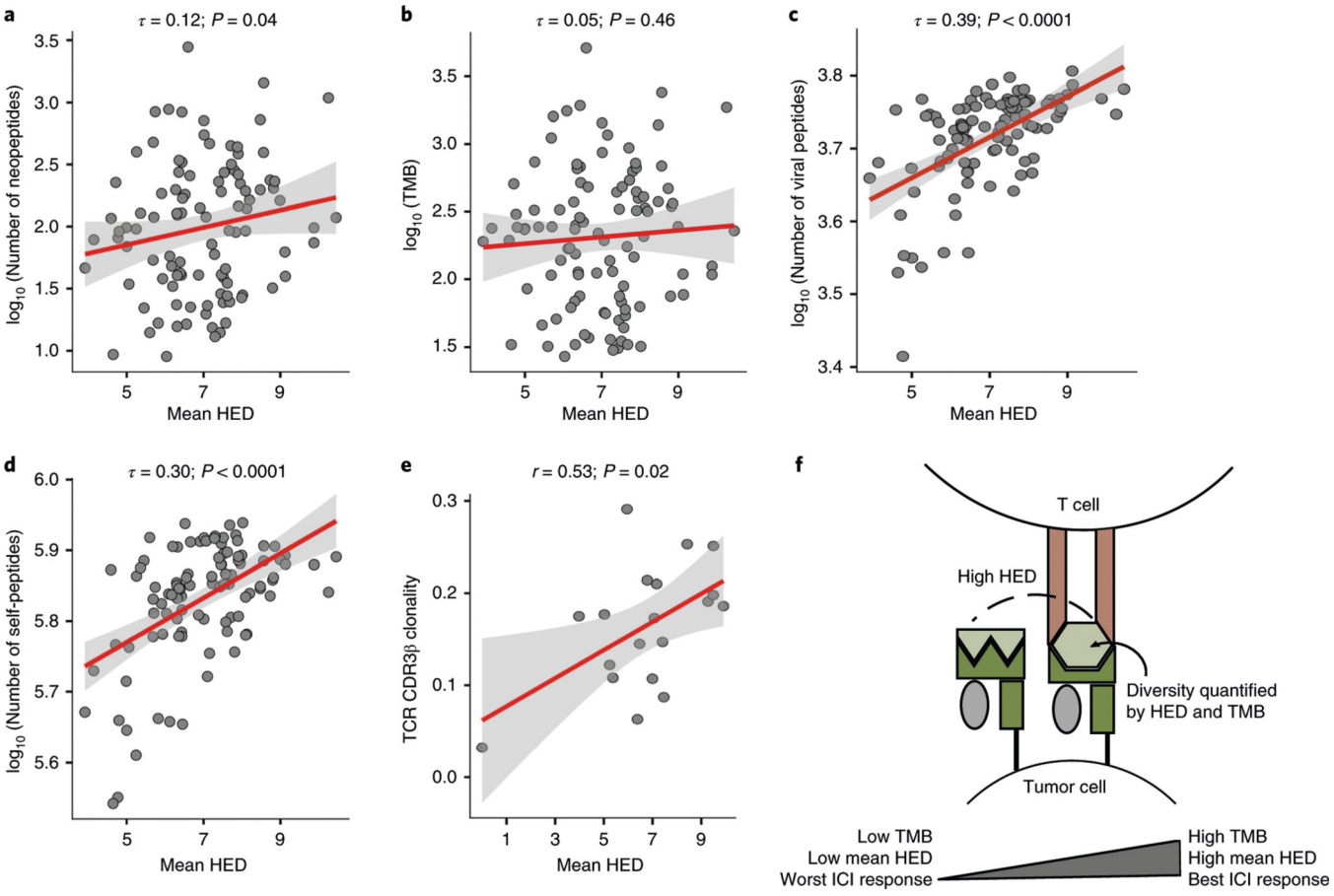


Fig. 4 |. Mean HED is positively correlated with diversity of the tumor, viral and human immunopeptidomes.

a, Correlation of mean HED with number of unique neopeptides bound to alleles of each patient genotype using all patients fully heterozygous at HLA-I from Fig. 2 for whom neopeptide data were available ($n = 103$; $P = 0.04$; one-sided Kendall's rank correlation). Each point represents a patient HLA-I genotype (*HLA-A, -B* and *-C*); the y axis depicts the mean number of neopeptides bound across *HLA-A, -B* and *-C* (see Methods). **b,** Correlation of mean HED with TMB ($n = 103$; $P = 0.46$; two-sided Kendall's rank correlation). **c,** Correlation of mean HED with number of unique viral peptides bound to alleles of each HLA-I genotype ($n = 103$; $P = 2.41 \times 10^{-9}$; one-sided Kendall's rank correlation). **d,** Correlation of mean HED with number of unique self-peptides from the human proteome bound to alleles of each HLA-I genotype ($n = 103$; $P = 6.46 \times 10^{-6}$; two-sided Kendall's rank correlation). The *y* axis depicts the mean number of self-peptides bound across *HLA-A, -B* and *-C*. **e,** Association of mean HED with intratumoral TCR CDR3β clonality ($n = 19$; $P = 0.02$; two-sided Pearson's correlation). The red line indicates the line of best linear fit. **f,** Schematic depicting the effects of HED and TMB on imunopeptidome diversity and response to ICIs. One representative *HLA*-I locus with high HED between the alleles is depicted.