Pan-cancer quantification of neoantigen-mediated immunoediting in cancer evolution

Tao Wu[1-3], Guangshuai Wang[1], Xuan Wang[1], Shixiang Wang[1], Xiangyu Zhao[1], Chenxu Wu[1], Wei Ning[1], Ziyu Tao[1], Fuxiang Chen[4] and Xue-Song Liu[1]

Affiliations of authors:

1 School of Life Science and Technology, ShanghaiTech University, Shanghai 201203, China;

2 Shanghai Institute of Biochemistry and Cell Biology, Chinese Academy of Sciences, Shanghai, China;

3 University of Chinese Academy of Sciences, Beijing, China;

4 Department of Clinical Immunology, Ninth People's Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, 200011, People's Republic of China.

Correspondence: Xue-Song Liu, School of Life Science and Technology, ShanghaiTech University, 230 Haike Road, Shanghai 201210, China. E-mail: liuxs@shanghaitech.edu.cn.

Key words: immunoediting, cancer biomarker, Neoantigen; immunoediting-elimination; immunoediting-escape; Negative selection; Tumor evolution;

Conflict of interest statement: The authors declare no potential conflicts of interests.

Abstract

Immunoediting, which includes three temporally distinct stages, termed elimination, equilibrium, and escape, has been proposed to explain the interactions between cancer cells and the immune system during the evolution of cancer. However the status of immunoediting in cancer remains unclear, and the existence of neoantigen depletion signal in untreated cancer has been debated. Here we developed a distribution pattern based method for quantifying neoantigen mediated negative selection in cancer evolution. Our method provides a robust and reliable quantification for immunoediting signal in an individual cancer patient. The prevalence of immunoediting signal in immunotherapy untreated cancer genome has been demonstrated with this method. Importantly, the elimination and escape stages of immunoediting can be quantified separately, tumor types with strong immunoediting-elimination tend to have weak immunoediting-escape signal, and vice versa. Quantified immunoediting-elimination signal predicts cancer immunotherapy clinical response. Immunoediting quantification provides an evolutional perspective for evaluating the antigenicity of neoantigen, and reveals a potential biomarker for cancer precision immunotherapy.

Introduction

During cancer evolution, some genome DNA alterations can be positively selected, such as driver mutations. Some genome alterations could posit a deleterious effect, and consequently are negatively selected or depleted during cancer evolution. Some (or the majority of) genome DNA alterations do not have driving or deleterious effects on cancer, and follow a neutral evolution pattern. The interactions between immune cells and tumor cells are reflected as immunoediting, which could mediate the negative selection of DNA alterations encoding high antigenicity (also known as neoantigenic mutations) (1,2). Positively selected genome alterations can be readily detected in the final mutation reservoir, however, the quantification of negative selection in cancer evolution is still a significant challenge (3). The status of neoantigen mediated negative selection in cancer evolution has been evaluated in several studies, and controversial results have been reported (4-7).

Convincing cases of adaptive molecular evolution have been identified through comparison of synonymous (silent; dS) and nonsynonymous (amino acid-changing; dN) substitution rates in protein-coding DNA sequences. dN/dS is the ratio between the rate of non-synonymous substitutions per non-synonymous site and the rate of synonymous substitutions per synonymous site. dN/dS method was originally developed to quantify the molecular evolution from sequencing data (8,9). Recently, dN/dS method has been applied in cancer evolution study (3). An important consideration in dN/dS analysis is the selection of negative control regions. For example, a recent study reported that neoantigen depletion signal is undetectable in the pan-cancer dataset (5), however the selection of negative control region is questionable (10). In addition, the percentage of depleted neoantigen could be tiny, and this prohibits the accurate detection of negative selection signal

through dN/dS, especially in an individual cancer patient. Population genetics based method has been used to identify the neutral pattern of cancer evolution, this method is based on the assumption that the variant allele frequency (VAF) within a tumor follows a characteristic power-law distribution in case of neutral evolution (11,12). The detection of neutral evolution with this method has been questioned (13,14), and its application in immune mediated negative selection has not been fully established. In together, till now, the existence and the degree of neoantigen mediated negative selection in human cancer remain unclear.

It is known that immunoediting elimination can lead to the down-regulation of cancer cell fraction (CCF) of antigenic mutation (15). To fully utilize the CCF distribution information, here we build a new distribution pattern based method for quantifying neoantigen mediated negative selection in an individual cancer patient. With this new analysis framework, we demonstrate the pan-cancer existence of neoantigen mediated negative selection signal. Shut down the expression of antigenic mutations can be one way for tumor cells to escape the surveillance of immune system, consequently the mRNA down-regulation status of antigenic mutations can be a surrogate of immune escape. Thus the elimination and escape phases of immunoediting can be quantified separately. In total, this study not only provides a novel method for quantifying negative selection in cancer evolution, but also reveals a potential biomarker for cancer immunotherapy clinical response prediction.

Materials and Methods

Pan-cancer clinical and molecular data

The normalized gene-level RNA-seq data (TPM, transcripts per million) for 31 TCGA cohorts were downloaded from Xena (https://xenabrowser.net/ , dataset

ID: tcga_RSEM_gene_tpm). Pre-compiled curated somatic mutations for TCGA cohorts were downloaded from Xena (dataset ID: GDC-PANCAN.mutect2_snv.tsv), and missense variants are selected for downstream analysis (16,17). ABSOLUTE-annotated MAF file which contains cancer cell fraction (CCF) information of mutations was downloaded from GDC PanCanAtlas publications (https://gdc.cancer.gov/about-data/publications/pancanatlas), and then we used liftover function from R package "rtracklayer" to convert the hg37 genome coordinates to hg38. Clinical data was obtained from GDC PanCanAtlas publications. HLA typing data was downloaded from Thorsson et.al study (18). The downloaded mutation data, HLA typing data and CCF values for TCGA samples have also been validated with in house algorithm. Immune cell infiltration data for all TCGA tumors was downloaded from ImmuneCellAI study (19), which estimates the abundance of 24 immune cells comprised of 18 T-cell subtypes and 6 other immune cells. Other immune cell infiltration data including CIBERSORT (abs mode), Quantiseq were obtained from the TIMER2.0 study (20). For TCGA tumors which do not have HLA typing data in the mutation data set (2404 samples), we downloaded raw bam files, and performed HLA typing as described below. Driver mutation data was downloaded from Bailey et al study (21). This study utilized three different categories of tools to identify driver mutations: (1) tools distinguishing benign versus pathogenic mutations based on sequence; (2) tools distinguishing driver versus passenger mutations based on sequence; and (3) tools identifying statistically significant three-dimensional clusters of missense mutations (21). We keep mutations identified by ≥2 approaches as the final high confident driver mutations, including 3437 unique mutations.

Somatic mutation calling

For the cancer immune checkpoint inhibitor therapy datasets, raw sequences

were aligned to the reference human genome (hg38) using Burrows–Wheeler Alignment (BWA) tool (22). Preprocessing followed the GATK4 best practices workflow, including duplicate removal, base quality score recalibration. Somatic mutations were identified on processed data using Mutect2 (23). BCFtools was used to filter genome variants that passed all quality control filters (24). The resulting VCF files were annotated by VEP and further converted to MAF files by vcf2maf.pl (https://github.com/mskcc/vcf2maf). The MAF file was loaded into R, analyzed and visualized by Maftools (25).

Gene expression analysis

For the cancer immune checkpoint inhibitor therapy datasets, paired-end RNA-seq data were processed using hisat2 (26) aligner on the basis of the hg38 human genome assembly with default parameters. Then the aligned SAM files were transformed to BAM files using samtools (27). Normalized RNA expression values (TPM) were calculated by TPMCalculator (28).

Cancer cell fraction (CCF) calculation

We followed the GATK4 copy number analysis pipeline to get copy number segment files. CCF information for each mutation was calculated based on segment files and somatic mutation MAF files using ABSOLUTE software (29). Briefly, read counts for each of the exome targets were collected from all samples and calculated the coverage by count reads that overlap intervals which were formed by padding the target regions. Each of the tumor samples was compared to a panel of normal (PoN) controls for normalization and denoising. The tool standardizes counts by the PoN median counts. The normalization process includes log2 transformation and normalizing the counts data to center around one. Then, the tool denoised the standardized copy ratios using the principal components of the PoN. These normalized coverage profiles were then segmented using Gaussian-kernel binary-segmentation algorithm, which were fed into ABSOLUTE algorithm to determine CCF.

HLA typing and neoantigen prediction

HLA genotyping was performed with Optitype (30), using default parameters. Mutect2 mutation files were first transformed into VCF format by maf2vcf tools, and we used NeoPredPipe to predict neoantigen (31). Single-nucleotide variants leading to a single amino acid change are the focus of this study. From the output results, if the IC50 of a novel peptide is less than 50, the binding level is SB (strong binder, rank is less than 0.5%), and the expression level (TPM) is greater than 1, then this peptide is labeled as neoantigen. A mutation is considered antigenic if there is at least one peptide in all possible 8, 9, 10-mer peptides derived from the mutated site predicted as neoantigen. To validate the major conclusion of our study, we also used additional MHCflurry method implemented in pVACseq to predict neoantigen (32).

Enrichment Score calculation

The Kolmogorov–Smirnov (K-S) statistic can be used to quantify the distance between two cumulative distributions. We constructed a K-S like statistic to quantify the difference between the distribution of the CCF (or mRNA expression) of antigenic mutations and non-antigenic mutations in each sample.

ES_{CCF} quantification in individual cancer patient

We equally divided the whole CCF range (0-1) into 100 intervals (in descending order) and assigned each interval a rank value (from 100 to 1). To make the heavier weights on two tails of the rank distribution, we further normalized the ranks (Eq. A):

$$R_i = \left| \frac{L}{2} - r \right| + 1 \qquad \text{(A)}$$

Where *Ri* is the normalized rank value, i is the interval index, L is the total number of intervals (here L=100), and r is the original rank value.

Then we counted the number of mutations lied in each interval (m (i)) and assigned a value a(i) to each interval depending on mutation counts (m (i)) and interval rank (R (i)) (Eq. B):

$$a_i = \frac{m_i \times R_i}{\sum m_i \times R_i} \qquad \text{(B)}$$

We can calculate the empirical cumulative distribution of random variable a (i) by walking from top to bottom (Fig 2) (Eq. C):

$$F(n) = \sum_{i}^{n} a_i \quad ,n = 1, 2, ..., 100 \qquad \text{(C)}$$

n is the total number of intervals, i is the interval index.

We then constructed two distributions for antigenic mutations (F_N(n)) and non-antigenic mutations (F_M(n)) of an individual sample, respectively. Then K-S like statistics can be obtained by taking distance (D(n)) of two distributions (Eq. D):

$$D(n) = F_N(n) - F_M(n) \qquad \text{(D)}$$

Similar to GSVA (33), the enrichment score (ES) was defined as (Eq. E):

$$ES = \left|D(n)^+\right| - \left|D(n)^-\right| = \max(0, D(n)) - \left|\min(0, D(n))\right| \qquad \text{(E)}$$

Where D(n)+ and D(n)−are the largest positive and negative random walk deviations from zero, respectively.

ES_{RNA} quantification in individual cancer patient

For a sample, using to denote mRNA expression (TPM). To reduce the influence of potential outliers, we first convert z to rank z', and normalize further to r = |P/2 - z'| + 1, making the ranks symmetric around 1 (P is the number of mutations in a sample), making the heavier weights on two tails of the rank distribution. Then we got two cumulative distributions, for mutations which are neoantigens (Eq. F):

$$D_{neo}(S,i) = \sum_{r_j \in S, j \le i} \frac{|r_j|}{\sum_{r_j \in S} |r_j|} \qquad \text{(F)}$$

For mutations which are not neoantigens (Eq. G):

$$D_{\text{not}neo}(S,i) = \sum_{r_j \notin S, j \le i} \frac{|r_j|}{\sum_{r_j \notin S} |r_j|} \qquad \text{(G)}$$

Where S is the set of mutations which are neoantigens, the size of the set is P_s, P is the number of mutations in a sample, r is normalized rank of mutations, i and j are mutation index. Then we constructed a K-S like statistics (Eq. H):

$$T = D_{neo}(S,i) - D_{notneo}(S,i) \qquad \text{(H)}$$

We transform the K-S like statistic into neoantigen enrichment score (ES) as the difference between the largest positive and negative distribution deviations from zero (Eq. I):

$$ES = \max(0,T) - |\min(0,T)| \qquad \text{(I)}$$

Estimation of significance level of ES.

We employed a permutation method to derive a null distribution to calculate p value of the ES (ES_{CCF} or ES_{RNA}). For each sample, the same number of mutations as neoantigens are randomly selected from the mutation list and the corresponding ES is calculated. This process is repeated 1000 times to get the ES null distribution. The p value is calculated from the positive or negative region of the empirical null distribution (Eq. J):

$$P = \begin{cases} \frac{1}{1000} \sum_{n=1}^{1000} I(ES_n \ge ES) \ , ES \ge 0 \\ \frac{1}{1000} \sum_{n=1}^{1000} I(ES_n < ES) \ , ES < 0 \end{cases} \qquad \text{(J)}$$

Where I is an indicator function.

Neutral simulation

For each sample, we permute the neoantigen labeling (ie. randomly select the same number of mutations as the actual neoantigen number in the selected sample and label them as neoantigenic mutations) and calculate ES value. For pan-cancer or cancer type dataset, we can obtain the same number of simulated samples and corresponding ES values, then calculate the median ES of these simulated samples. This process was repeated many times (usually 2000 times) to get the simulated distribution of median ES. The actual pan-cancer or cancer type median ES values are compared with this simulated ES distribution, and p values are then reported.

Immune escape analysis

We consider the following immune escape mechanisms: 1, suppress the transcription of genome alterations encoding high antigenicity (quantified as ES_{RNA}); 2, antigen presentation pathway gene alterations; 3, PD-L1 or CTLA-4 overexpression; 4, loss of heterozygosity (LOH) on the HLA locus (34). Antigen presentation pathway genes were selected based on the list of antigen processing and presentation machinery (APM) from the Gene Ontology Consortium (GO:0002474) (35). Gene level non-silent mutation file was downloaded from UCSC Xena. Immune checkpoint gene overexpression was assessed using RNA-seq data. Normal expression values (in transcripts per million mapped reads (TPM)) of PD-L1 and CTLA-4 were established from the TCGA based on RNA-seq expression of the two genes in normal samples. Checkpoint overexpression was called if either PD-L1 or CTLA-4 expression in the tumor was higher than the mean plus two standard deviations of normal expression. The HLA LOH status data was obtained from Li et al study (36). If at least one HLA allele is subject to loss by LOH, then the sample is labeled as HLA LOH.

Cancer immunotherapy datasets analysis

To investigate the predictive performance of the quantified immunoediting-elimination signal in immune checkpoint inhibitor (ICI) therapy clinical response prediction for individual patient, we searched for public ICI datasets with available raw WES data and RNA-seq data. Three melanoma ICI datasets have been identified for this study. The Hugo et al dataset was related to anti-PD-1 therapy in metastatic melanoma (37). This dataset has 38 samples with WES data, 27 were also analyzed by RNA sequencing (RNA-seq). The Riaz et al dataset was related to anti-PD-1therapy in metastatic melanoma, and it has 64 samples with WES data, 51 with RNA-seq (38). The Liu et al cohort includes melanoma patients treated with anti-PD1 antibody, it has 124 samples with WES data and 121 samples with RNA-seq (39). All three melanoma studies used a very similar definition for clinical endpoints. Clinical response for patients was defined by RECIST v1.1, responding tumors were derived from patients who have complete or partial responses (CR/PR) in response to anti-PD-1 therapy; non-responding tumors were derived from patients who had progressive disease or stable disease (PD/SD). We only chose pre- immunotherapy treatment samples for analysis. Mutation calling, neoantigen prediction, expression quantified, CCF calculation and ES calculation were performed as described above.

The performance of ES_{CCF} has been compared with 15 biomarkers reported to have significant association with immune checkpoint inhibitor (ICI) response (40), including tumor mutation burden (TMB), clonal TMB, indel mutation burden (41), burden of indels escaping nonsense mediated decay (NMD) (42), *SERPINB3* mutations, CD274 (PD-L1) expression, *CD38* expression, *CD8A* expression, *CXCL13* expression, *CXCL9* expression, T cell inflamed gene expression signature (43), IMPRES (44), CD8 T effector from the POPLAR trial (45), cytolytic score, and UV signature. TMB was calculated as the number of missense mutations per megabase; clonal TMB was measured as missense mutations which CCF exceed 0.9. Indel mutation burden was

calculated by the counts of frameshift indel mutation counts. The counts of indels having no overlap with the nucleotides of NMD score > 0.52 was considered as NMD-escape indel burden (40). T cell inflamed gene expression signature (Ayer score) was calculated as average expression (TPM) of 18 genes (*CD3D, IDO1, CIITA, CD3E, CCL5, GZMK, CD2, HLA-DRA, CXCL13, IL2RG, NKG7, HLA-E, CXCR6, LAG3, TAGAP, CXCL10, STAT1, GZMB*). IMPRES values was calculated based on expression of 15 gene pairs. For each gene pair gene_i/gene_j, using following formula to calculate gene pair value (Eq. K):

$$F_{i,j}(x) = \begin{cases} 1, \exp_i(x) < \exp_j(x) \\ 0, \text{otherwise} \end{cases} \qquad \text{(K)}$$

For each sample, we can get a gene pairs value vector of length 15, and The total number of '1's in this vector denotes the sample's IMPRES score (44). CD8 T effector signature from the POPLAR trial was calculated as the average expression (TPM) of 8 genes (*CD8A, GZMA, GZMB, IFNγ, EOMES, CXCL9, CXCL10*, and *TBX21*). Cytolytic activity score (CYT) was calculated as the geometric mean expression (TPM) of *GZMA* and *PRF1*.

Stochastic branching process model for cancer evolution and power analysis

The tumor evolution model constructed by Lakatos et al has been applied in this study (15). In this model, tumor evolution was initiated by a single transformed cell. At any simulation step, a cell is randomly selected and has three events that could happen: birth (divide to produce two offspring), death and waiting. For a birth event, new cells could acquire some new mutations (counts are sampled from Poisson distribution) and each mutation can become neoantigen as a specific probability. Under negative selection on neoantigen, the death rate of cells could increase from d_0 to d_i with neoantigen accumulation. Selection strength (s) of neoantigen mediated negative

selection can be calculated as (Eq. L):

$$1+s\times n=\frac{b-d_i}{b-d_0} \qquad \text{(L)}$$

n is the number of neoantigens in a cell, b is the birth rate (for simplicity, set b=1) In addition, every mutation has a probability (p_{esc}) to escape. Once a mutation is escaped, the death rate of the cell which contains this mutation decrease to basal death rate d_0. This simulation step continues until the population reaches a pre-defined size. Similar to the original study (15), the following parameters were applied: neoantigen probability p=0.1, birth rate b=0.1, basal death rate d_0=0.1, Poisson distribution parameter (mutation rate) μ=1, escape probability $p_{esc}=10^{-6}$, selection strength $-0.25\leq s\leq 0$, final population size popSize=10^5. At each selection strength, we run the simulation 100 times. The model was implemented with Julia (v1.3.1, revised from the original Julia code provided by Lakatos et.al).

Mutations harbored in at least 5 cells out of 10^5 were collected at the end of each simulation and the CCF was calculated. To account for imperfect sequencing measurements, CCF values were computed via a simulated sequencing step introducing noise to these frequencies with the indicated read depth. For a given read depth D, each frequency value f was substituted by a new frequency f' sampled from a binomial distribution with parameters D and f: f' ~ Binom(D,f)/D . We filtered for mutations with f' above 0 to discard mutations that are not picked up due to limited detection power. In addition to sequencing limitations, we also added different proportions of false positive neoantigen when evaluating the power of detecting negative selection: we randomly sampled nonantigenic mutations of simulated tumors (varied between 5 and 500% of the number of true neoantigen) that were falsely labeled as neoantigen. To calculate the power of derivation from neutral VAF distribution method (15), we used two side K-S test to detect the difference between the VAF distribution of all mutations and neoantigenic mutations and reported K-S

375 statistic and corresponding *P* value.

376

377 **Statistical analysis**

378 All statistical tests were performed using R statistical language. In all boxplots,
379 the center lines represent the median, low and upper box limits are the first
380 and third quartiles, respectively, and whiskers represent the values up to 1.5
381 times of the interquartile range. *P* values for comparisons between boxplots
382 were calculated by Wilcoxon rank sum test. Correlation and corresponding p
383 values were calculated by Pearson method using R function cor.test.
384 Kaplan-Meier survival analysis was performed using the R package "survival"
385 with log-rank test, and Cox-proportional hazard analysis was performed using
386 the R package "ezcox". The cutoff value of ES_{CCF} in Kaplan-Meier overall
387 survival analysis was determined by surv_cutpoint function of "survminer"
388 package. R function ks.test was used to perform two-sided K-S test.

389

390 **Software and data availability**

391 Custom code for quantifying immunoediting-elimination and
392 immunoediting-escape are available in
393 https://github.com/XSLiuLab/Immunoediting/tree/main . All code required to
394 reproduce the analysis outlined in this manuscript, and R markdown analysis
395 report are available in https://xsliulab.github.io/Immunoediting/.

396

397

398

399

400 **Results**

401 **Conceptual framework for the elimination and escape phases of cancer**
402 **immunoediting**

403 The interactions between cancer cells and immune cells are manifested as
404 immunoediting, which consists of three sequential phases: elimination,

equilibrium, and escape (1,2). In the elimination phase, tumor cells with genome alterations encoding high antigenicity are partially or completely eliminated by immune cells, and this leads to the down-regulation of the cancer cell fraction (CCF) of genome alterations encoding high antigenicity (**Fig. 1A**). In the escape phase, tumor cells escape the surveillance of immune system through multiple mechanisms, including the following: 1. Suppress the transcription or expression of genome alterations encoding high antigenicity; 2. Antigen presentation pathway down-regulation; 3. Up-regulate the expression of immune suppressive molecules, including PD-L1, CTLA-4, etc. (**Fig. 1A**).

The elimination phase of immunoediting will lead to the down-regulation of CCF of neoantigenic mutations, and consequently this CCF down-regulation status of neoantigenic mutations can reflect the strength of the elimination phase of immunoediting. The mRNA down-regulation status of neoantigenic mutations is a partial reflection of the strength of immunediting-escape phase. Here we use TCGA pan-cancer dataset to investigate this immunoediting signal. TCGA dataset includes 31 cancer types and 9511 samples with available WGS or WES data and mRNA expression profiling (RNA-seq) data, and neoantigenic genome alterations can be found in 9166 samples (Supplementary Fig. S1) (46). In the following section we build a distribution pattern based method to quantify the selection strength acting on the CCF or mRNA expression of neoantigenic mutation.

Method for quantifying neoantigen mediated negative selection

For each genome mutation, we have CCF and normalized mRNA expression (transcripts per million, TPM) information. The antigenicity value of genome mutation can be calculated as the possibility of the mutated peptide to be presented by HLA type I, and mutated peptides with predicted HLA I binding affinity (IC50) less than 50nM are labeled as neoantigens. A mutation was considered neoantigenic if there was at least one peptide derived from the

mutated sequence is predicted as neoantigen. The consequence of immunoediting-elimination phase will lead to an unbalanced distribution of the CCF of neoantigenic mutations (15), and this CCF distribution pattern of genome alterations encoding antigenicity can reflect the selection strength of immunoediting-elimination phase. This distribution enrichment status of CCF was calculated following a similar principle of gene set variation analysis (GSVA) or gene set enrichment analysis (GSEA), which was originally developed in the estimation of the variation of pathway activity over a sample population in an unsupervised manner (33,47).

The mutations in an individual sample or a cancer type as a whole, are ordered by CCF or mRNA expression (TPM) as a ranked list L. The mutations with antigenicity are defined as a set S. The goal of this analysis is to determine whether the members of S are randomly distributed throughout L or primarily found at the top or bottom. There are two key steps of this method (**Fig. 1B**):

1. Enrichment score (ES) calculation based on the distribution of neoantigen. We calculate an ES that reflects the degree to which a set S is overrepresented at the extremes (top or bottom) of the entire ranked list L. The score is calculated by walking down the list L, increasing a running-sum statistic when we encounter a mutation in S and decreasing it when we encounter mutations not in S. The ES is calculated based on the maximum deviations from zero during the random walk, it corresponds to a weighted Kolmogorov–Smirnov (K-S) like statistic (see details in the Methods). The CCF values of mutations are in the range of 0-1, and in TCGA dataset, the CCF values of mutations do not show normal distribution, and many mutations have CCF values equal to 1 (Supplementary Fig. S2). A fixed CCF rank from 1 to 100 has been constructed in the quantification of CCF distribution enrichment status of neoantigenic mutation (ES_{CCF}). CCF distributions of neoantigenic and non-neoantigenic mutations in TCGA cancer types are shown, an apparent

shift in the CCF distribution of neoantigenic mutations compared with non-neoantigenic mutations can be observed (Supplementary Fig. S3).

2. Estimation of the significance level of ES.

We estimate the statistical significance (nominal *P* value) of the ES by using a permutation test procedure, this procedure permute the neoantigen labels and recomputed the ES of each patient, and this generates a null distribution for the ES. The *P* value of the observed ES is then calculated according to this null distribution. For ES significance analysis in TCGA pan-cancer or individual cancer type cohort level, the observed median ES value of the test cohort is compared with the distribution of median ES values from 2000 simulations (**Fig. 1C**). The calculated *P* values are dependent on the mutation rank, and also the number of total mutations and the number of antigenic mutations. Minimum number of total mutations and antigenic mutations are required for confident quantification of ES values (Supplementary Fig. S4).

Tumor cells can evolve multiple strategies to escape the surveillance of immune system, and down-regulating the mRNA expression of neoantigenic mutation is one of these strategies (**Fig. 1A**). Similar to CCF values, the mRNA expression values of mutations are independent variables from antigenicity IC50 values. Similar strategy can be applied to quantify this mRNA expression down-regulation mediated immunoediting, and the resulting ES_{RNA} is a partial reflection of the strength of immunoediting-escape signal (**Fig. 1A** and **B**).

The existence of significant immunoediting signal

Previous studies have debated the existence of neoantigen depletion signals in cancer evolution. Van den Eynden J. *et al.* reported that neoantigen depletion signal is undetectable in TCGA pan-cancer dataset (5). However as pointed out in a preprint, their method for neoantigen depletion signal detection is problematic, as the actual neoantigens with antigenicity are not located in

their defined "HLA-binding regions" (10). We investigate the status of this immunoediting signal with the new method developed in this study using TCGA pan-cancer dataset. The antigenicity IC50 value is calculated based on the mutated DNA sequence and HLA status, and the CCF information is independently obtained from high-throughput sequencing. Mutation types do not influence the CCF values, and the distribution of antigenicity is not influenced by mutation types either. The antigenicity IC50 values are thus independent variables from CCF values, and this is different from the calculation of dN/dS, where the two variables dN, dS are interconnected and are both significantly influenced by mutation types (3,48).

Since the variables (CCF and HLA binding IC50 status) are independent, we use random simulation to generate a null distribution of ES_{CCF}. For TCGA pan-cancer or individual cancer type cohort, the median ES_{CCF} values are recorded after each simulation. The observed median ES_{CCF} values are compared with the simulated ES_{CCF} values. In TCGA pan-cancer cohort with at least 1 neoantigenic and 1 subclonal mutation (CCF<0.6) (n=5900), the observed ES_{CCF} is -0.017 (*P*=0.051) (**Fig. 2A** and **D**). In PAAD and LUAD, the observed ES_{CCF} values are significant lower compared with the random simulations, suggesting the existence of immunoediting-elimination signal (**Fig. 2A;** Supplementary Fig. S5). Since some neoantigenic mutations can be cancer drivers, which are known to undergo positive selection during the evolution of cancer. Neoantigens that happen to be cancer drivers are not undergoing immune based negative selection (Supplementary Fig. S6). In TCGA pan-cancer cohort, when samples with neoantigenic and driver mutations lying on the same gene are not included, the observed median ES_{CCF} is -0.023 (n=5295, P=0.0055) (**Fig. 2B** and **D**). Several cancer types including ACC, PAAD, UCEC, LUAD show significant low ES_{CCF} values (**Fig. 2B;** Supplementary Fig. S7). This data demonstrates the existence of immunoediting-elimination signal in TCGA dataset.

Similarly random simulations were performed to evaluate the significance of the observed ES_{RNA} values. Compared with ES_{CCF}, the observed ES_{RNA} show much strongly significant difference when compared with the random simulated values. In TCGA pan-cancer dataset with at least 1 neoantigenic mutation and accompanied mRNA expression information (n=6974), the observed ES_{RNA} =-0.048 ($p<0.0005$) (**Fig. 2C** and **D**). In the majority of cancer types (including DLBC, CHOL, SARC, LUAD, PRAD, UCS, STAD, LGG, LUSC, HNSC, OV, UCEC, BRCA, LIHC, READ, TGCT, KIRP), a significant low ES_{RNA} values are observed (**Fig. 2C;** Supplementary Fig. S8). This study demonstrates that the immunoediting escape through down-regulating the expression of neoantigenic alteration is prevalent in human cancer (**Fig. 2C**). Furthermore, the immunoediting-escape signal is more prevalent than the immunoediting-elimination signal (**Fig. 2A-C**). This is in line with the fact that clinically detectable tumors need to have immune escape capacity, and the tumors with strong immunoediting-elimination signal may not have the chance to become clinically apparent lesions.

Interestingly we observed that in cancer types with strong immunoediting-elimination signal, a weak immunoediting-escape signal exist, and vice versa (**Fig. 2E** and **F**). ES_{RNA} signal is only a partial mechanism of cancer immune escape, known additional mechanisms include overexpression of immune checkpoint genes (for example PD-L1, CTLA-4), antigen presentation pathway gene alterations, and loss of heterozygosity on the HLA locus. Pan-cancer distributions of these immune escape mechanisms are shown, and different cancer types show different proportion of tumors with each specific immune escape mechanisms (Supplementary Fig. S9). When TCGA samples are divided into two parts based on the existence of known immune escape mechanisms, significant immune elimination signal (ES_{CCF}) can only be observed in patients without immune escape mechanisms

(Supplementary Fig. S10). ES_{CCF} values of patients without immune escape mechanisms are significantly lower than patients with immune escape mechanisms (Supplementary Fig. S11).

To further validate the immunoediting signal, we select different neoantigen prediction cutoff, which increase the percentage of antigenic mutations from 9% to 19%, and under this new situation, we still observe a significant immunoediting elimination signal (ES_{CCF}=-0.013, *P*=0.054), and also a more significant ES_{RNA} signal (ES_{RNA}=-0.056, *P*<0.0005) (Supplementary Fig. S12). When another neoantigen prediction tool MHCflurry is applied, a significant immunoediting elimination signal (ES_{CCF}=-0.017, *P*=0.015) and ES_{RNA} signal (ES_{RNA}=-0.029, *P*<0.0005) can also be observed (Supplementary Fig. S13). In pan-cancer or individual cancer type level, the immunoediting elimination and escape signal exist, however in majority of cancer patients, both of immunoediting-elimination and escape signals are weak or undetectable (Supplementary Fig. S14 and Supplementary Fig. S15). Sufficient sequencing depth is required for the detection of this immunoediting signal, and the required sequencing depth is not reached in many TCGA samples.

Neoantigen enrichment score and immune negative selection strength quantification

Recently, immune based negative selection has been simulated using a stochastic branching process model (15). The neoantigen mediated negative selection strength (s) is an inherent feature of each patient. However method for accurately quantifying this immune based negative selection strength is still lacking. Here we investigate the connections between neoantigen enrichment score (ES_{CCF}) and immune negative selection strength *s* using a stochastic branching cancer evolution model as previously described (15). For each fixed selection strength s, the resulting ES_{CCF} was calculated (**Fig. 3A** and **B**). ES_{CCF} show near linear correlation with s values (**Fig. 3A**). This analysis suggests

that the quantified ES_{CCF} can be used to infer the immune selection strength in patient. The median ES_{CCF} in TCGA datasets is -0.023, suggesting a median immune negative selection strength *s*=-0.08 (**Fig. 3A**).

Proportional neoantigen burden measures the percentage of neoantigenic mutations in individual sample or cancer types. Proportional neoantigen burden was originally designed to compare the immune negative selection strength between two or more samples (15). The baseline values of proportional neoantigen burden cannot be obtained, and consequently proportional neoantigen burden method could not be applied in quantifying the strength of immune negative selection in individual cancer patient, or in individual cancer type. Derivation from neutral VAF distribution (*1/f* dependence of the cumulative VAF distribution) has been suggested to reflect the selection status (11,15). However neutral VAF distribution method is not suitable in negative selection quantification due to strict requirement in sequencing depth and neoantigen prediction accuracy (**Fig. 3C-F**).

Pan-cancer features and correlations of immunoediting signal

Human cancer evolve over a long time interval, usually in decades. The immunoediting-elimination signal quantified in this study suggests the existence of an already happened neoantigen mediated tumor elimination process. While the quantified immune cell infiltration level represent the current immune response status. We calculated the immunoediting status in TCGA pan-cancer datasets (**Fig. 2A**). The unbalanced distribution of CCF in neoantigenic vs non-neoantigenic mutations quantified as ES_{CCF} could reflect the status of neoantigen mediated tumor elimination. In tumors with detectable immunoediting-elimination signal (ES_{CCF}<0, *P*<0.05), a slightly increased $CD8^+$ T plus natural killer (NK) cell infiltration status compared with the remaining samples were observed, and the difference does not reach statistical significance (*P*=0.2) (**Fig. 4A** and **B**). The immune cell infiltration status was

615 further evaluated using additional methods, such as CIBERSORT (49),
616 Quantiseq (50), similarly no significant difference can be observed in $CD8^+$ T
617 cell and regulatory T cell (Treg) cell levels, $CD8^+$/Treg ratio between tumors
618 with and without ES_{CCF} signal (Supplementary Fig. S16, S17, S18). This data
619 suggests that historically happened immunoediting-elimination process may
620 not be reflected in the current immune cell infiltration status.
621
622 The down-regulation of antigenic mutation encoded mRNA is a partial
623 reflection of immunoediting-escape phase (**Fig. 1A**). Pan-cancer status of this
624 ES_{RNA} is shown, and different cancer types have different median ES_{RNA}
625 scores (**Fig. 2C**). The immunoediting-escape signal quantified as ES_{RNA} also
626 does not show a statistically significant difference between samples with
627 detectable immunoediting-escape signal (ES_{RNA}<0, P<0.05) and the remaining
628 samples in $CD8^+$ T plus NK cell infiltration (**Fig. 4C**). Treg percentage appear
629 to be up-regulated in samples with detectable immunoediting-escape signal
630 (P=0.03) (**Fig. 4D**), while this up-regulated Treg signal can be reproduced with
631 Quantiseq analysis, but not CIBERSORT analysis (Supplementary Fig. S16,
632 S17, S18). These analyzes suggest that there are no strong and direct
633 connections between the immune cell infiltration status of present time point
634 and the immune escape signal that historically happened during the evolution
635 of tumor.
636
637 **Quantified immunoediting-elimination signal predicts the clinical**
638 **response of cancer immunotherapy**
639 Immunotherapy, represented by immune checkpoint inhibitors (ICI), is
640 transforming the treatment of cancer. However, only a small percentage of
641 patients show response to ICI, and effective biomarkers for ICI clinical
642 response prediction is still urgently needed (51). To investigate the predictive
643 performance of the quantified immunoediting-elimination signal (ES_{CCF}) in ICI
644 response prediction for individual patient, we searched for public ICI datasets

with raw WES data and RNA-seq data available, and three melanoma ICI datasets have been identified (37-39) (Supplementary Fig. S19).

We calculate the immunoediting-elimination signal (ES_{CCF}) for each patient. In univariate Cox proportional hazards regression analysis, quantified ES_{CCF} value is significantly associated with cancer patients' survival (p=0.03), and low ES_{CCF} value (suggest the presence of high immunoediting-elimination signal) is associated with improved ICI clinical response (Hazard ratio (HR)=3.74, 95%CI=1.11-12.6) (**Fig. 5A**). Patients are divided into three groups based on ES_{CCF} value, patients with the lowest ES_{CCF} values (indicate the presence of immunoediting-elimination signal) show the best survival after ICI (**Fig. 5B**). Fifteen additional biomarkers, including tumor mutational burden (TMB), clonal TMB, indel mutation burden, burden of indels escaping nonsense mediated decay (NMD), SERPINB3 mutation, PD-L1 expression, CD38 expression, CD8A expression, CXCL13 expression, CXCL9 expression, T cell inflamed gene signature, IMPRES, CD8 T effector, cytolytic score, UV signature mentioned in Litchfield et al study have been evaluated and compared with the ES_{CCF} (40). In these melanoma datasets, only ES_{CCF} and UV signature show significant HR (Supplementary Fig. S20).

Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Here we use logistic regression to compare the efficiency of ES_{CCF}, TMB and neoantigenic mutation count in predicting immunotherapy clinical response. Relationship between prognosis (patients with clinical response or without clinical response) and ES_{CCF}, TMB and neoantigenic mutation count was analyzed. The goodness of fit was performed by Hosmer–Lemeshow test (H-L test). The H-L test *P*-value of TMB is 0.051 (**Fig. 5C**, middle), close to 0.05, implicating the difference between prediction and expectation is close to significant. The H-L test *P*-value of ES_{CCF} is 0.771 (**Fig. 5C**, right), higher than the H-L test *P*-value of TMB and

neoantigen count. This study suggests that the quantified immunoediting-elimination signal can be biomarker for ICI clinical response prediction. ICI clinical responses are known to be influenced by variables like gender (52,53), When considering ICI type as a covariate, the HR of ES_{CCF}=3.75, *P*=0.03; When considering gender as a covariate, the HR of ES_{CCF}=3.96, *P*=0.09 (Supplementary Fig. S21). Based on this analysis, the effects of ES_{CCF} is not influence by ICI type and gender, however more samples are needed to further validated the clinical effects of ES_{CCF} in ICI clinical response prediction.

Discussion

Here we provide reliable evidence to demonstrate the pan-cancer existence of immunoediting signal. Importantly, the elimination and escape phases of immunoediting can be separately quantified with our method. Cancer types with strong immunoediting elimination signal usually have low immunoediting escape signal, and vice versa. Furthermore, the quantified immunoediting elimination signal predict cancer immunotherapy clinical response.

This study provides an initial method to reliably quantify immunoediting signal in individual cancer patient. To quantify the immunoediting signal for an individual patient, at least one neoantigenic mutation is required. The mechanisms employed by tumor cells to escape immune surveillance is very complex, and the shutdown of the expression of neoantigen mutation is only one of the mechanisms. In addition, the mRNA expression is the combination of both wild type and mutated alleles. Lack of ES_{RNA} signal does not mean that the immune escape signal does not exist in the specific cancer or cancer types. Usually neoantigens are believed to be able to mediate the negative selection of cancer cells, the possibilities that some mutations could encode peptides

suppressing the immune function cannot be ruled out. For example, neoantigens that bind the TCR of Tregs, could have potential immunosuppressive function (54).

The existence of neoantigen mediated negative selection status in untreated cancer has been debated (4-7). Existing methods for negative selection study include dN/dS and population genetics method. Rooney *et al*. use TCGA pan-cancer CDS as the control sequence to calculate the expected neoantigen number per non-silent mutation (Bpred/Npred), then the actual observed neoantigen number per non-silent mutation (Bobs/Nobs) are compared with Bpred/Npred (55). Since the pan-cancer CDS sequence has already been immune edited, the neoantigen depletion signal reported in this study is systematically underestimated. In addition, as pointed out by Van den Eynden *et al*. There is HLA typing mistake in this study (5). Van den Eynden J. *et al.* reported that neoantigen depletion signal is undetectable in untreated cancer (5). This study selects the “non HLA-binding regions” as the control, and compare the nonsynonymous vs synonymous mutation ratio (n/s) in “HLA-binding regions” vs “non HLA-binding regions”. They did not identify any difference in these two regions in regard to n/s using TCGA pan-cancer dataset. However their method is problematic, as the actual neoantigen with antigenicity are not located in their defined “HLA-binding regions” (10). Martincorena *et al*. performed a comprehensive gene level evolution selection study with dN/dS method, and reported significant neutral and positive selection, but not negative selection in cancer genome (3). Since antigenicity mutations occupy less than 5% of total mutations. In gene level, the selection on neoantigen mutations is overshadowed by other driving or neutral mutations. Neoantigen mediated negative selection not being observed in gene-level does not mean the absence of immune based neoantigen depletion. Zapata *et al*. investigated immune based negative selection with dN/dS method (4). However same problem exists in the selection of control DNA

735 sequence. Similar to Van den Eynden J. *et al.* CDS was divided into epitope
736 region and non-epitope region, dN/dS was compared in these two regions.
737 Since neoantigen with antigenicity are not necessarily located in the “epitope
738 region”, the results reported in this study is also questionable. Population
739 genetics model for neutral evolution has been proposed to detect neutral
740 evolution based on cumulative VAF distribution. For instance, a recent study
741 test the neutrality of cancer based on the VAF of mutations in a limited
742 subclonal frequency range (11). However, that test of neutrality has been
743 questioned, the frequency distribution of mutated alleles in a limited frequency
744 range is not an accurate statistic for detecting selection in cancer (13,14).
745 Furthermore the application of this population genetics method in neoantigen
746 mediated negative selection quantification in individual cancer patient has not
747 been established (15) (**Fig. 3D** and **F**).
748
749 The quantification of negative selection in cancer evolution has been a major
750 scientific challenge, the method developed here for neoantigen mediated
751 negative selection quantification could be instructive for the future design of
752 strategies for studying negative selection in cancer evolution. The existence of
753 neoantigen mediated negative selection has been demonstrated with our new
754 method. Importantly we observed a strong immunoediting-escape signal
755 reflected as the down-regulation of mRNA encoded by neoantigenic mutations.
756 The quantification of immunoediting provides an evolutionary perspective for
757 the design of neoantigen vaccine for cancer therapy. The immune based
758 negative selection is an inherent feature of a cancer patient, the quantified
759 immunoediting signal can be used in cancer precision stratification, including
760 the clinical response prediction for cancer immunotherapy.
761
762
763
764

Acknowledgments

We thank ShanghaiTech University high performance computing public service platform for computing services. We thank Haopeng Wang of ShanghaiTech for critical discussion. We thank Raymond Shuter for editing the text. We thank multi-omics facility, molecular cellular facility of ShanghaiTech University for technical help. This work was supported by the national natural science foundation of China (31771373), Shanghai science and technology commission (21ZR1442400), and startup funding from ShanghaiTech University.

Contributions

TW collected the data, developed the immunoediting quantification analysis method and performed the computational analysis. GW participated in data collection and preprocessing. XW participated in neoantigen prediction. SW help to build the method for immunoediting quantification. XZ, CW, WN, ZT, FC participated in critical project discussion. XSL conceptualized the idea, designed, supervised the study and wrote the manuscript.

Conflict of interest

The authors declare no competing interests.

Reference :

1. Schreiber RD, Old LJ, Smyth MJ. Cancer Immunoediting: Integrating Immunity's Roles in Cancer Suppression and Promotion. Science **2011**;331:1565-70

2. O'Donnell JS, Teng MWL, Smyth MJ. Cancer immunoediting and resistance to T cell-based immunotherapy. Nat Rev Clin Oncol **2019**;16:151-67
3. Martincorena I, Raine KM, Gerstung M, Dawson KJ, Haase K, Van Loo P, *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. Cell **2017**;171:1029-+
4. Zapata L, Pich O, Serrano L, Kondrashov FA, Ossowski S, Schaefer MH. Negative selection in tumor genome evolution acts on essential cellular functions and the immunopeptidome. Genome Biol **2018**;19
5. Van den Eynden J, Jimenez-Sanchez A, Miller ML, Larsson E. Lack of detectable neoantigen depletion signals in the untreated cancer genome. Nat Genet **2019**;51:1741-+
6. Marty R, Kaabinejadian S, Rossell D, Slifker MJ, van de Haar J, Engin HB, *et al.* MHC-I Genotype Restricts the Oncogenic Mutational Landscape. Cell **2017**;171:1272-+
7. Claeys A, Luijts T, Marchal K, Van den Eynden J. Low immunogenicity of common cancer hot spot mutations resulting in false immunogenic selection signals. Plos Genet **2021**;17
8. Goldman N, Yang ZH. Codon-Based Model of Nucleotide Substitution for Protein-Coding DNA-Sequences. Mol Biol Evol **1994**;11:725-36
9. Yang ZH, Bielawski JP. Statistical methods for detecting molecular adaptation. Trends Ecol Evol **2000**;15:496-503
10. Wang S, Wang X, Wu T, He Z, Li H, Sun X, *et al.* Revisiting neoantigen depletion signal in the untreated cancer genome. bioRxiv **2020**
11. Williams MJ, Werner B, Barnes CP, Graham TA, Sottoriva A. Identification of neutral tumor evolution across cancer types. Nat Genet **2016**;48:238-44
12. Williams MJ, Werner B, Heide T, Curtis C, Barnes CP, Sottoriva A, *et al.* Quantification of subclonal selection in cancer from bulk sequencing data. Nat Genet **2018**;50:895-+
13. Tarabichi M, Martincorena I, Gerstung M, Leroi AM, Markowetz F, Spellman PT, *et al.* Neutral tumor evolution? Nat Genet **2018**;50:1630-3
14. McDonald TO, Chakrabarti S, Michor F. Currently available bulk sequencing data do not necessarily support a model of neutral tumor evolution. Nat Genet **2018**;50:1620-3
15. Lakatos E, Williams MJ, Schenck RO, Cross WCH, Househam J, Zapata L, *et al.* Evolutionary dynamics of neoantigens in growing tumors. Nat Genet **2020**;52:1057-+
16. Wang S, Liu X. The UCSCXenaTools R package: a toolkit for accessing genomics data from UCSC Xena platform, from cancer multi-omics to single-cell RNA-seq. Journal of Open Source Software **2019**;4
17. Wang SX, Xiong Y, Zhao LF, Gu K, Li Y, Zhao F, *et al.* UCSCXenaShiny: an R/CRAN package for interactive analysis of UCSC Xena data. Bioinformatics **2022**;38:527-9
18. Thorsson V, Gibbs DL, Brown SD, Wolf D, Bortone DS, Yang THO, *et al.* The Immune Landscape of Cancer. Immunity **2018**;48:812-+
19. Miao YR, Zhang Q, Lei Q, Luo M, Xie GY, Wang HX, *et al.* ImmuCellAI: A Unique Method for Comprehensive T-Cell Subsets Abundance Prediction and its Application in Cancer Immunotherapy. Adv Sci **2020**;7
20. Li TW, Fu JX, Zeng ZX, Cohen D, Li J, Chen QM, *et al.* TIMER2.0 for analysis of tumor-infiltrating immune cells. Nucleic Acids Res **2020**;48:W509-W14
21. Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. Cell **2018**;173:371-+
22. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform.

Bioinformatics **2010**;26:589-95

23. Benjamin D, Sato T, Cibulskis K, Getz G, Stewart C, Lichtenstein L. Calling Somatic SNVs and Indels with Mutect2. bioRxiv **2019**
24. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, *et al.* Twelve years of SAMtools and BCFtools. Gigascience **2021**;10
25. Mayakonda A, Lin DC, Assenov Y, Plass C, Koeffler HP. Maftools: efficient and comprehensive analysis of somatic variants in cancer. Genome Res **2018**;28:1747-56
26. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol **2019**;37:907-+
27. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, *et al.* The Sequence Alignment/Map format and SAMtools. Bioinformatics **2009**;25:2078-9
28. Alvarez RV, Pongor LS, Marino-Ramirez L, Landsman D. TPMCalculator: one-step software to quantify mRNA abundance of genomic features. Bioinformatics **2019**;35:1960-2
29. Carter SL, Cibulskis K, Helman E, McKenna A, Shen H, Zack T, *et al.* Absolute quantification of somatic DNA alterations in human cancer. Nat Biotechnol **2012**;30:413-+
30. Szolek A, Schubert B, Mohr C, Sturm M, Feldhahn M, Kohlbacher O. OptiType: precision HLA typing from next-generation sequencing data. Bioinformatics **2014**;30:3310-6
31. Schenck RO, Lakatos E, Gatenbee C, Graham TA, Anderson ARA. NeoPredPipe: high-throughput neoantigen prediction and recognition potential pipeline. Bmc Bioinformatics **2019**;20
32. Hundal J, Carreno BM, Petti AA, Linette GP, Griffith OL, Mardis ER, *et al.* pVAC-Seq: A genome-guided in silico approach to identifying tumor neoantigens. Genome Med **2016**;8
33. Hanzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-Seq data. Bmc Bioinformatics **2013**;14
34. McGranahan N, Rosenthal R, Hiley CT, Rowan AJ, Watkins TBK, Wilson GA, *et al.* Allele-Specific HLA Loss and Immune Escape in Lung Cancer Evolution. Cell **2017**;171:1259-+
35. Wang SX, He ZK, Wang X, Li HM, Liu XS. Antigen presentation and tumor immunogenicity in cancer immunotherapy response prediction. Elife **2019**;8
36. Li XY, Zhou C, Chen K, Huang BD, Liu Q, Ye H. Benchmarking HLA genotyping and clarifying HLA impact on survival in tumor immunotherapy. Mol Oncol **2021**;15:1764-82
37. Hugo W, Zaretsky JM, Sun L, Song CY, Moreno BH, Hu-Lieskovan S, *et al.* Genomic and Transcriptomic Features of Response to Anti-PD-1 Therapy in Metastatic Melanoma. Cell **2016**;165:35-44
38. Riaz N, Havel JJ, Makarov V, Desrichard A, Urba WJ, Sims JS, *et al.* Tumor and Microenvironment Evolution during Immunotherapy with Nivolumab. Cell **2017**;171:934-+
39. Liu D, Schilling B, Liu D, Sucker A, Livingstone E, Jerby-Amon L, *et al.* Integrative molecular and clinical modeling of clinical outcomes to PD1 blockade in patients with metastatic melanoma. Nat Med **2019**;25:1916-+
40. Litchfield K, Reading JL, Puttick C, Thakkar K, Abbosh C, Bentham R, *et al.* Meta-analysis of tumor- and T cell-intrinsic mechanisms of sensitization to checkpoint inhibition. Cell **2021**;184:596-+
41. Turajlic S, Litchfield K, Xu H, Rosenthal R, McGranahan N, Reading JL, *et al.* Insertion-and-deletion-derived tumour-specific neoantigens and the immunogenic phenotype: a pan-cancer analysis. Lancet Oncol **2017**;18:1009-21

42. Lindeboom RGH, Vermeulen M, Lehner B, Supek F. The impact of nonsense-mediated mRNA decay on genetic disease, gene editing and cancer immunotherapy. Nat Genet **2019**;51:1645-+

43. Ayers M, Lunceford J, Nebozhyn M, Murphy E, Loboda A, Kaufman DR, *et al.* IFN-gamma-related mRNA profile predicts clinical response to PD-1 blockade. J Clin Invest **2017**;127:2930-40

44. Auslander N, Zhang G, Lee JS, Frederick DT, Miao BC, Moll T, *et al.* Robust prediction of response to immune checkpoint blockade therapy in metastatic melanoma. Nat Med **2018**;24:1545-+

45. Fehrenbacher L, Spira A, Ballinger M, Kowanetz M, Vansteenkiste J, Mazieres J, *et al.* Atezolizumab versus docetaxel for patients with previously treated non-small-cell lung cancer (POPLAR): a multicentre, open-label, phase 2 randomised controlled trial. Lancet **2016**;387:1837-46

46. Hoadley KA, Yau C, Hinoue T, Wolf DM, Lazar AJ, Drill E, *et al.* Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. Cell **2018**;173:291-+

47. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. P Natl Acad Sci USA **2005**;102:15545-50

48. Van den Eynden J, Larsson E. Mutational Signatures Are Critical for Proper Estimation of Purifying Selection Pressures in Cancer Somatic Mutation Data When Using the dN/dS Metric. Front Genet **2017**;8

49. Newman AM, Liu CL, Green MR, Gentles AJ, Feng WG, Xu Y, *et al.* Robust enumeration of cell subsets from tissue expression profiles. Nat Methods **2015**;12:453-+

50. Finotello F, Mayer C, Plattner C, Laschober G, Rieder D, Hackl H, *et al.* Molecular and pharmacological modulators of the tumor immune contexture revealed by deconvolution of RNA-seq data. Genome Med **2019**;11

51. Nishino M, Ramaiya NH, Hatabu H, Hodi FS. Monitoring immune-checkpoint blockade: response evaluation and biomarker development. Nat Rev Clin Oncol **2017**;14:655-68

52. Wang SX, Zhang J, He ZK, Wu K, Liu XS. The predictive power of tumor mutational burden in lung cancer immunotherapy response is influenced by patients' sex. Int J Cancer **2019**;145:2840-9

53. Wang SX, Cowley LA, Liu XS. Sex Differences in Cancer Immunotherapy Efficacy, Biomarkers, and Therapeutic Strategy. Molecules **2019**;24

54. Ahmadzadeh M, Pasetto A, Jia L, Deniger DC, Stevanovic S, Robbins PF, *et al.* Tumor-infiltrating human CD4(+) regulatory T cells display a distinct TCR repertoire and exhibit tumor and neoantigen reactivity. Sci Immunol **2019**;4

55. Rooney MS, Shukla SA, Wu CJ, Getz G, Hacohen N. Molecular and Genetic Properties of Tumors Associated with Local Immune Cytolytic Activity. Cell **2015**;160:48-61

Figure legend

Figure 1. Conceptual framework for the quantification of elimination and escape phases of immunoediting. **A**, Phases of immunnoediting and the manifestations of the elimination and escape phases of cancer immunoediting. **B**, Detailed steps for CCF down-regulation based immunoediting-elimination (ES_{CCF}) quantification. 1. Equally divide the whole CCF range (0-1) into 100 intervals and calculate the distribution of CCF of neoantigenic mutations and non-neoantigenic mutations in these intervals; 2. Construct the Kolmogorov–Smirnov (K-S) statistics based on difference between the two distributions; 3. Calculate the enrichment score (ES_{CCF}). **C**, Detailed steps for mRNA down-regulation based immunoediting-escape (ES_{RNA}) quantification. 1. Rank mutations by corresponding mRNA expression and calculate the distribution of mRNA expression of neoantigenic mutations and non-neoantigenic mutations; 2. Construct the K-S statistics based on difference between the two distributions; 3. Calculate the enrichment score (ES_{RNA}). **D**, Random simulation to obtain the null distribution of ES. For each sample, we permute the mutation labeling (ie. randomly select the same number of mutations as the observed number in the sample, and label them as neoantigenic mutations) and calculate ES value, the processes are repeated for 2000 times, and the actual ES values are compared with the simulated values.

Figure 2. Pan-cancer distributions and features of the quantified immunoediting signals (ES_{CCF} and ES_{RNA}). **A**, Distribution of ES_{CCF} in pan-cancer (left) and in specific cancer type (right). The p values are calculated from simulated median ES distributions. ns: p > 0.05, *: P <= 0.05, **: P <= 0.01, ***: P <= 0.001, ****: P <= 0.0001. **B**, Distribution of ES_{CCF} in pan-cancer (left) and in specific cancer type (right), after removing samples with neoantigenic and driver mutations located in the same gene. The p values are calculated from simulated median ES distributions. **C**, Distribution of ES_{RNA} in pan-cancer (left) and in specific cancer type (right). The p values are

957 calculated from simulated median ES distributions. **D**, From left to right,
958 simulated median ES distribution and the observed median ES for Fig 3a, 3b
959 and 3c respectively. **E**, Correlation between median ES_{RNA} and ES_{CCF} of TCGA
960 cancer types. Pearson correlation coefficient and p value are shown. **F**,
961 Correlation between the percent of escape samples ($ES_{RNA} < 0$ and $P < 0.05$)
962 and median ES_{CCF} in TCGA cancer types. Pearson correlation coefficient and
963 p value are shown.
964
965 **Figure 3.** Immunoediting-elimination signal (ES_{CCF}) and neoantigen-mediated
966 negative selection strength quantification. **A**, ES_{CCF} as a function of
967 neoantigen-mediated negative selection strength s, computed from n=100
968 tumors, with simulated read depth of 200×for each indicated selection strength
969 s. The observed median ES_{CCF} of TCGA samples is indicated with a horizontal
970 dashed line. **B**, The proportion of 100 simulated tumors with significant ES_{CCF}
971 (FDR corrected p value less than 0.1) in each selection strength s. **C**,
972 Derivation from neutral VAF distribution quantification as a function of negative
973 selection strength s, computed from individual tumor of a simulation cohort
974 (n=100), with a simulated read depth of 200×. **D**, Proportion of 100 simulated
975 tumors with significant signal (FDR corrected p value less than 0.1) quantified
976 using derivation from neutral VAF distribution method under each negative
977 selection strength s. Of note, no tumors show significant signal under the same
978 simulated conditions as the data show in Fig. 3B. **E**, Power to detect negative
979 selection as a function of sequencing read depth (x axis) and false neoantigen
980 rate (y axis) using the enrichment score method developed in this study. Power
981 is the proportion of 100 simulated tumors with significant negative ES value
982 (FDR corrected *P* value less than 0.1). **F**, Power to detect negative selection
983 as a function of sequencing read depth (x axis) and false neoantigen rate (y
984 axis) using derivation from neutral VAF distribution method. Power is the
985 proportion of 100 simulated tumors with significant difference (two-sided K-S

test, FDR corrected *P* value less than 0.1) between the distribution of all mutations and neoantigenic mutations.

Figure 4. Immunoediting-elimination and escape signals and tumor immune cell infiltration status. **A** and **B**, Comparisons between TCGA cancer patients with detectable Immunoediting-elimination signal (ES_{CCF}<0, p<0.05) and the remaining patients in $CD8^+$ T plus natural killer (NK) cell (**A**) and Treg cell (**B**) infiltration status. **C** and **D**, Comparisons between TCGA cancer patients with detectable Immunoediting-elimination signal (ES_{RNA}<0, p<0.05) and the remaining patients in $CD8^+$ T plus NK cell (**C**) and Treg cell (**D**) infiltration status. Wilcoxon rank sum test *P* value is shown.

Figure 5. Quantified immunoediting-elimination signal (ES_{CCF}) predicts cancer immunotherapy clinical response. **A**, Univariate Cox regression analysis was performed to estimate the hazard ratio (HR) associated with ES_{CCF} values. The length of horizontal line represents the 95% confidence interval (CI) and the vertical dashed line represents HR = 1. **B**, Kaplan-Meier overall survival curves show the comparison between different groups stratified by ES_{CCF} value. Samples with ES_{CCF} values higher than the cutoff (-0.222, determined by surv_cutpoint function of “survminer” package) were classified as “high” group, and samples with ES_{CCF} value less than the cutoff were classified as “low” group. The remaining samples (without neoantigen or minimum CCF is higher than 0.6) were classified as “other” group. **C**, The goodness-of-fit is performed by Hosmer-Lemeshow test.

Figure 1

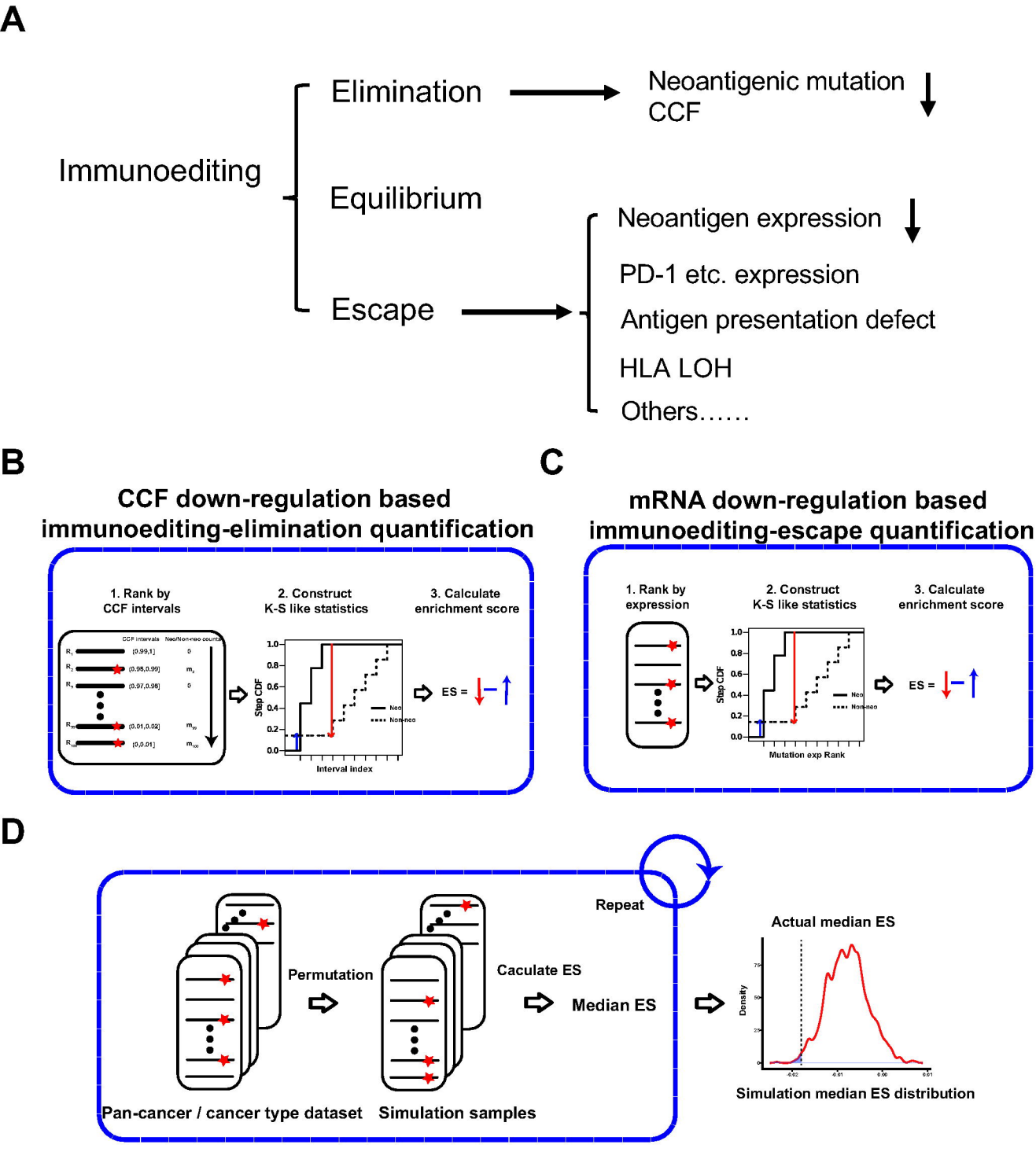

Figure 2

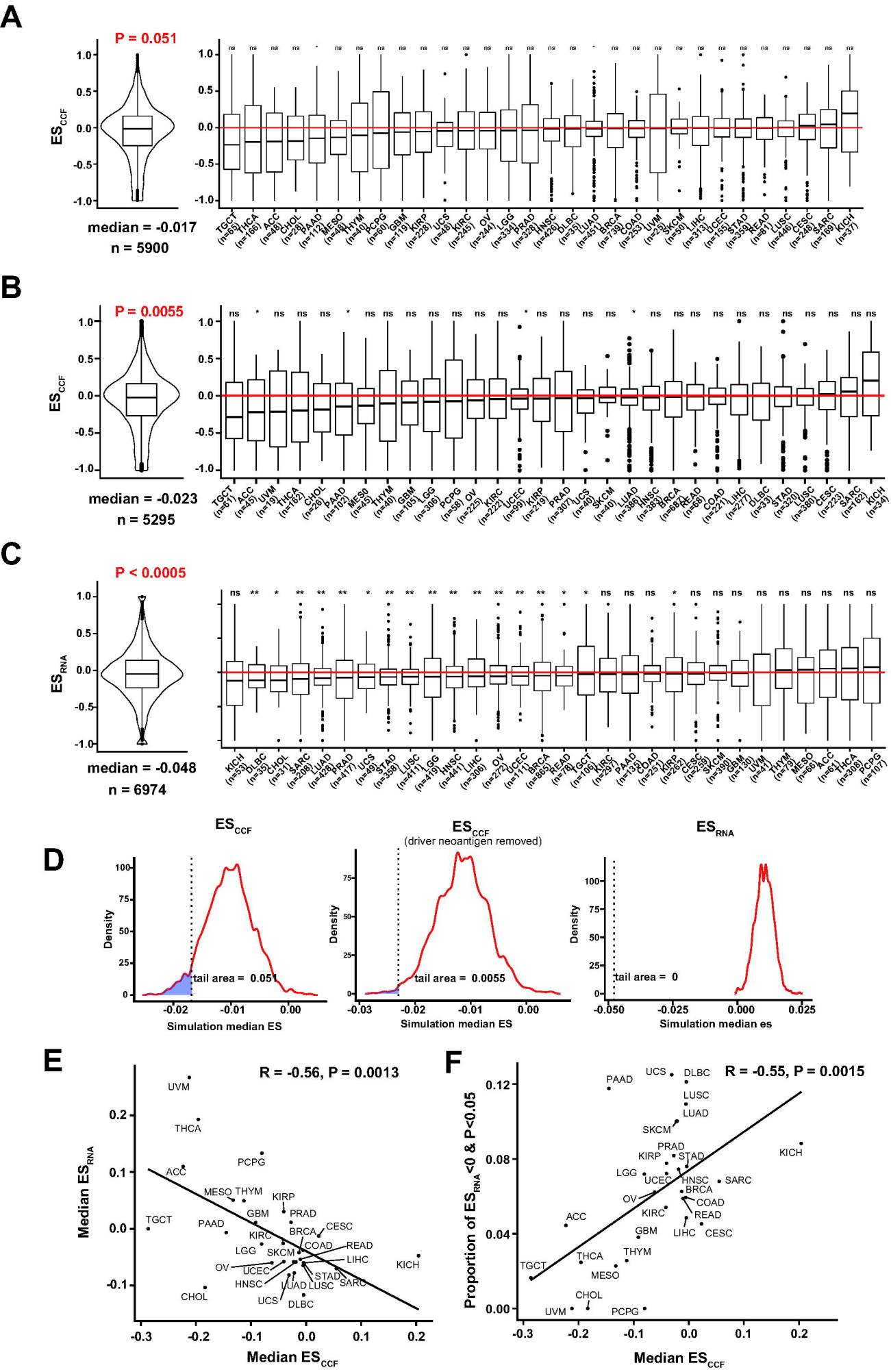

Figure 3

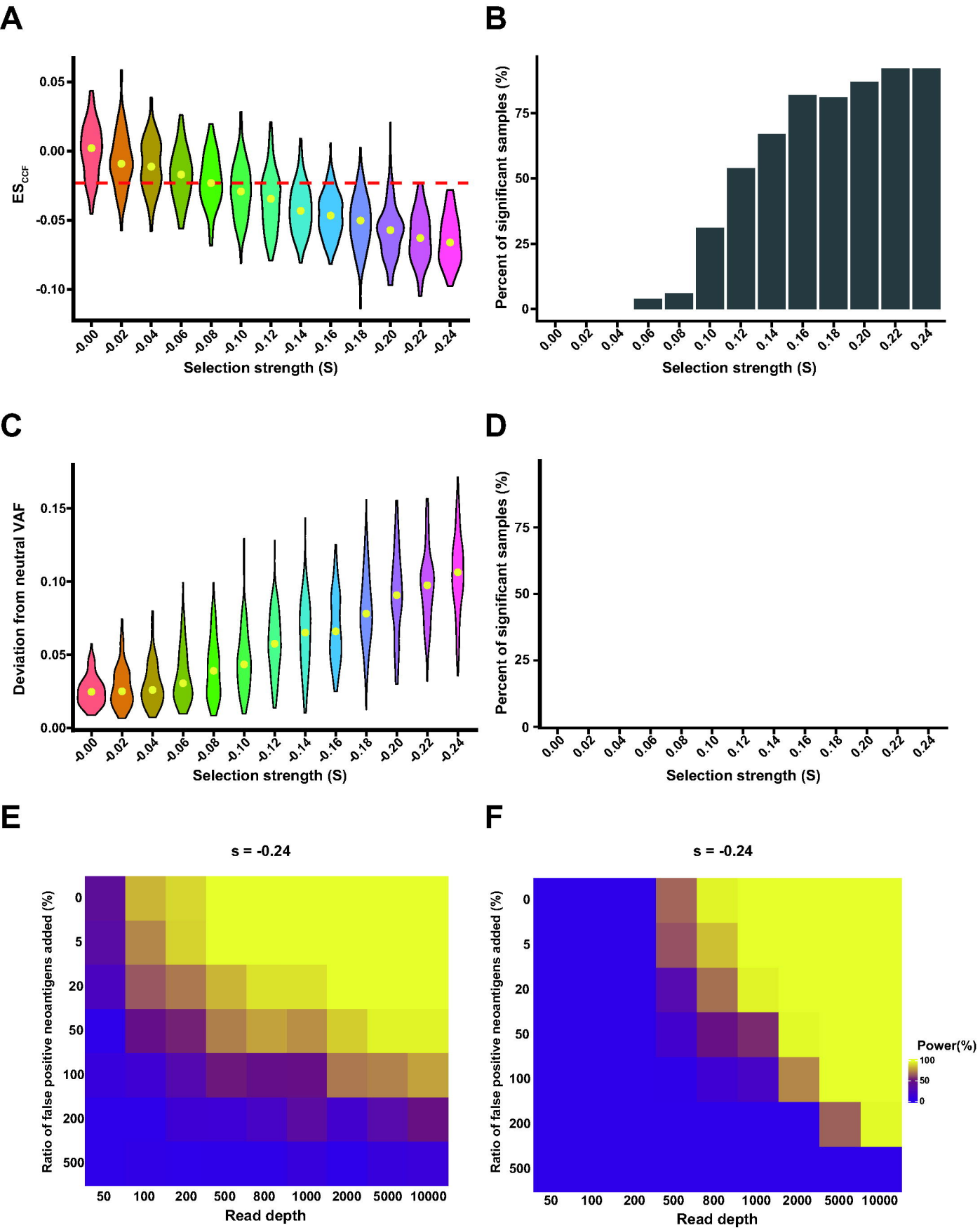

Figure 4

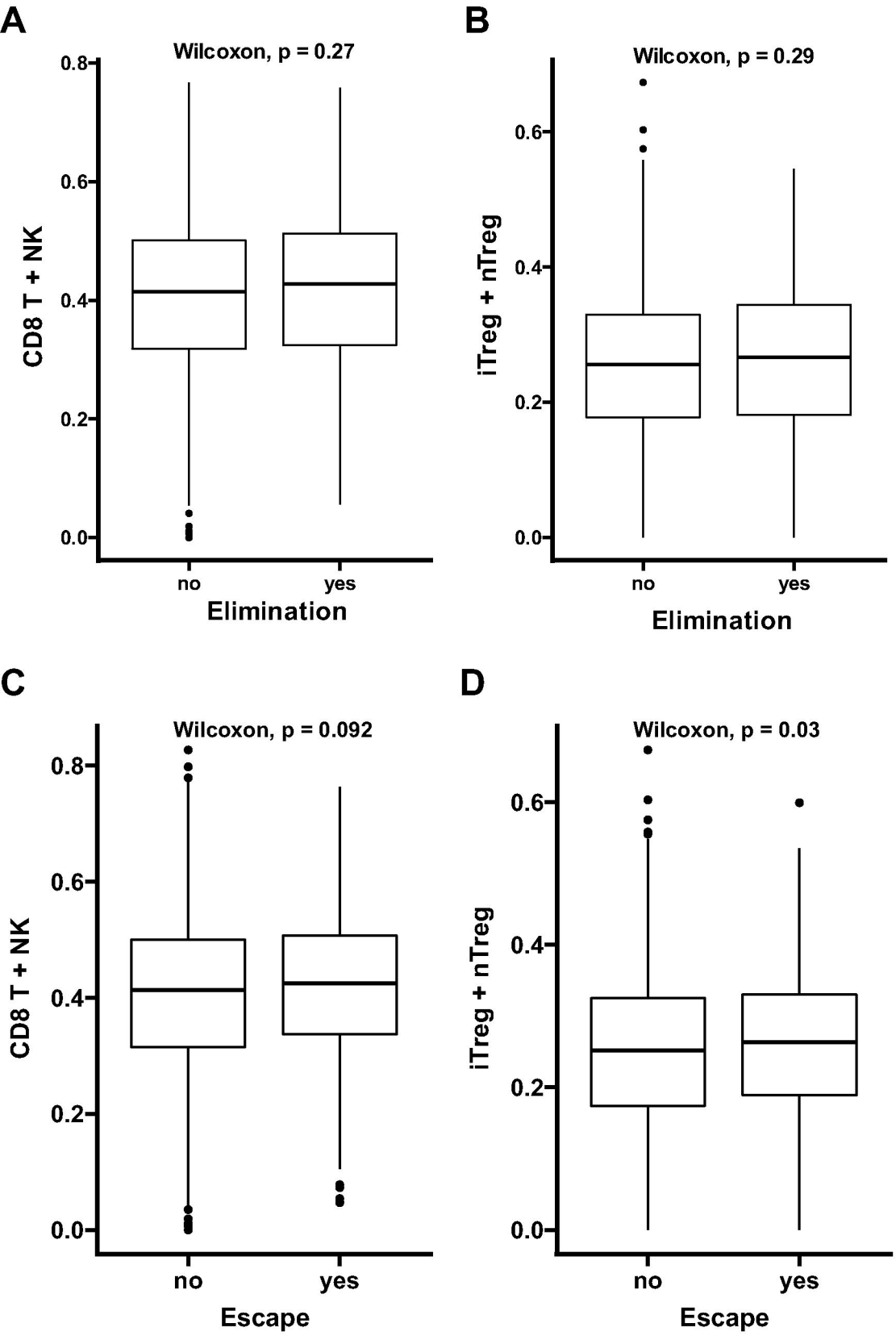

Figure 5