A neoantigen fitness model predicts tumor response to checkpoint blockade immunotherapy

Marta Łuksza[1,*], Nadeem Riaz[2,3], Vladimir Makarov[3,4], Vinod P. Balachandran[5,6,7], Matthew D. Hellmann[7,8,9], Alexander Solovyov[10], Naiyer A. Rizvi[11], Taha Merghoub[7,12,13], Arnold J. Levine[1], Timothy A. Chan[2,3,4,7], Jedd D. Wolchok[7,8,12,13], Benjamin D. Greenbaum[10,*]

[1]The Simons Center for Systems Biology, Institute for Advanced Study, Princeton, NJ, USA.
[2]Departments of Radiation Oncology, [5]Surgery and [8]Medicine, Memorial Sloan Kettering Cancer Center, New York, NY, USA.
[3]Immunogenomics and Precision Oncology Platform, Memorial Sloan Kettering Cancer Center, New York, NY, USA.
[4]Human Oncology and Pathogenesis Program, Memorial Sloan Kettering Cancer Center, New York, NY, USA.
[6]David M. Rubenstein Center for Pancreatic Cancer Research, Memorial Sloan Kettering Cancer Center, New York, NY, USA.
[7]Parker Institute for Cancer Immunotherapy, Memorial Sloan Kettering Cancer Center, New York, NY, USA.
[9]Department of Medicine, Weill Cornell Medical College, Cornell University, New York, NY, USA.
[10]Tisch Cancer Institute, Department of Medicine, Hematology and Medical Oncology, Departments of Oncological Sciences and Pathology, Icahn School of Medicine at Mount Sinai, New York, NY, USA.
[11]Department of Medicine, Columbia University Medical Center, New York, NY, USA
[12]Ludwig Collaborative and Swim Across America Laboratory, Memorial Sloan Kettering Cancer Center, New York, NY, USA.
[13]Melanoma and Immunotherapeutics Service, Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, NY, USA.

[*] Corresponding Authors:
Marta Łuksza, PhD
The Simons Center for Systems Biology
School of Natural Sciences
Institute for Advanced Study
Princeton, NJ 08540
Tel: (609) 734-8387
Fax: (609) 951-4438
E-mail: mluksza@ias.edu

Benjamin D. Greenbaum, PhD
Tisch Cancer Institute
Icahn School of Medicine at Mount Sinai
New York, NY 10029
Tel: (212) 824-8434
E-mail: benjamin.greenbaum@mssm.edu

Checkpoint blockade immunotherapies enable the host immune system to recognize and destroy tumor cells[1]. Their clinical activity has been correlated with activated T-cell recognition of neoantigens, which are tumor-specific, mutated peptides presented on the surface of cancer cells[2,3]. Here, we present a fitness model for tumors based on immune interactions of neoantigens that predicts response to immunotherapy. Two main factors determine neoantigen fitness: its likelihood of presentation by the major histocompatibility complex (MHC) and its subsequent T-cell recognition. We estimate these two components using a neoantigen's relative MHC binding affinity and a non-linear dependence on its sequence similarity to known antigens. To describe the evolution of a heterogeneous tumor, we evaluate its fitness as a weighted effect of dominant neoantigens in the tumor's subclones. Our model predicts survival in anti-CTLA-4 treated melanoma patients[4,5] and anti-PD-1 treated lung cancer patients[6]. Importantly, low-fitness neoantigens identified by our method may be leveraged for developing novel immunotherapies. By using an immune fitness model to study immunotherapy, we reveal broad evolutionary similarities between cancers and fast-evolving pathogens[7-9].

Although T-cell receptors are capable of recognizing and eliminating tumors, cancers evolve resistance mechanisms by display of checkpoint blockade molecules and disrupt the processes of immune recognition and attack. Clinical trials using immune checkpoint blocking antibodies, such as anti-cytotoxic T-lymphocyte-associated protein 4 (anti-CTLA-4), or anti-programmed cell death protein-1 (anti-PD-1), have improved overall survival in many malignancies by inhibiting these immune checkpoints[1]. Though only a minority of patients achieves durable clinical benefit, multiple studies have shown genetic determinants of response. De novo somatic mutations within coding regions can create *neoantigens* – novel protein epitopes specific to tumors, which MHC molecules present to the immune system and which may be recognized by T-cells as non-self[2,3]. An elevated number of mutations or neoantigens has been linked to improved response to checkpoint blockade therapy in multiple malignancies[4-6,10]. Hence, inferred neoantigen burden is a coarse-grained proxy for whether a tumor is likely to respond. Other implicated biomarkers of response include T-cell receptor (TCR) repertoire profiles[11], assays of checkpoint status[12,13], immune microenvironment landscape[4,14,15] and tumor heterogeneity[16]. Despite high overall mutational load, a heterogeneous tumor may have immunogenic neoantigens present only in certain subclones[16]. Therapies targeting a fraction of the tumor could disrupt clonal competitive balance and inadvertently stimulate growth of untargeted clones[17,18]. A mathematical model integrating genomic data has the advantage of broad consideration of neoantigen space. Worldwide efforts are being undertaken to model neoantigens and quantify neoantigen features from genomic data, and a predictive neoantigen-based model for immunotherapy response is therefore a highly sought-after goal, complementing mass spectrometry-based validation of neoantigens[19].

We propose a fitness model of tumor-immune interactions as a general mathematical framework to describe the evolutionary dynamics of cancer cell populations under checkpoint-blockade immunotherapy (Fig. 1). Analogous fitness models based on immune interactions have been successfully applied to human influenza[7], HIV[8] and chronic viral infections[9]. We aim to introduce this approach to the study of immunotherapy and provide an initial proof of concept regarding its potential utility. Checkpoint blockade exposes cancer cells to strong immune pressure on their neoantigens and thereby reduces their reproductive success. Our fitness model predicts the evolution of a cancer cell population under such pressure. We compute $n(\tau)$, the predicted future effective size of a cancer cell population in a tumor relative to its effective size at the start of therapy. This effective size is a weighted sum over tumor's genetic clones (Fig. 1a, Methods),

$$n(\tau) = \sum_{\alpha} X_{\alpha} \exp(F_{\alpha}\tau) \quad (1)$$

where F_α is the fitness, X_α is the initial frequency of clone α and τ is a characteristic evolutionary time scale when the prediction is evaluated. The effective size estimates the relative number of cancer cells required to generate the observed population diversity but, as tumors may also include other cell types, it is not to be interpreted as a direct measure of physical tumor size. Patients with less immunologically fit tumors will have more significant effective size reductions and, presumably, improved overall survival after checkpoint blockade therapy. To reconstruct the clonal tree structure of a tumor from exome sequencing data, we use a likelihood scheme based on the allele frequencies of its mutations[20]. Unlike in previous approaches[16], we learn the ancestral dependencies between clones, and these determine the mutations and neoantigens inherited by clones from their ancestors (Fig. 1a). Our fitness model assigns to subclones the same or lower fitness than their ancestral clones, depending on whether they acquired new dominant neoantigens.

Our approach quantifies two essential factors in determining immunogenicity of a neoantigen: an amplitude, A, determined by mutant and wildtype MHC-presentation, and an intrinsic TCR-recognition probability, R (both factors are defined below). We call the product of these two factors, $A \times R$, a neoantigen's *recognition potential*. We quantify total fitness for cancer cells in a clone by aggregating over the fitness effects due to immune recognition of its neoantigens (Fig 1b, Methods). Here, we model the fitness of a given clone α by the recognition potential of its dominant neoantigen,

$$F_\alpha = -\max_{i \in \text{Clone } \alpha} (A_i \times R_i) \quad (2)$$

where index i runs over all neoantigens in clone α (we discuss other choices for aggregating neoantigen fitness effects in Methods).

We utilize nonamer neoantigens inferred by a consistent identification pipeline with affinities, standing in for dissociation constants, for both mutant and wildtype peptides for a patient's HLA type[21] (SI), and define the amplitude A using the relative MHC affinity between the wildtype and the mutant peptide (Methods). Unlike considering solely mutant or wildtype affinities, the amplitude has consistent predictive value within our model (Extended Data Table 1). A simple interpretation of this observation is that the amplitude is related to the quantity of TCRs available to recognize the neoantigen. That is, a neoantigen needs to have low dissociation constant (i.e. high binding affinity) to be presented and to generate a TCR response. However, if the wildtype peptide also has a low dissociation constant, tolerance mechanisms could have removed wildtype peptide specific TCRs. Due to cross-reactivity, the quantity of mutant specific TCRs could be reduced (see discussion in Methods).

We model the probability of TCR-recognition of a neoantigen based on the strength of its alignments to positively recognized, class-I restricted T-cell

antigens from the Immune Epitope Database[22] (IEDB). This approach does not assume preexisting host immunity due to this set of epitopes. Rather, we posit that the high scoring neoantigens are more "non-self", their distribution possibly reflecting intrinsic biases in TCR generation probabilities[23]. Therefore, these neoantigens are also more likely to be immunogenic as TCRs have the ability to recognize large classes of homologous peptides via cross reactivity[24]. We use a thermodynamic model to estimate this recognition probability (Methods): for a neoantigen with peptide sequence $\mathbf{s}$ and IEDB epitope with sequence $\mathbf{e}$, the alignment score between $\mathbf{s}$ and $\mathbf{e}$ is used as a proxy for the binding energy between this neoantigen and a TCR specific to epitope $\mathbf{e}$. Under this assumption, each mutation that changes a residue in $\mathbf{e}$ into a corresponding residue in $\mathbf{s}$ in their alignment will increase the binding energy between $\mathbf{s}$ and a TCR recognizing epitope $\mathbf{e}$, proportionally to the alignment mismatch cost. Importantly, the probability that a neoantigen is bound by a TCR is given by a nonlinear logistic dependence on sequence alignment scores to all IEDB epitopes $\mathbf{e}$ (Methods). Approximately 72% of the mutant neoantigen peptides are classified as TCR-recognizable according to this criterion ($R > 0.5$, Extended Data Fig. 1) – those remaining are penalized in our model in equation (2).

We apply this model to three datasets: two melanoma patient cohorts treated with anti-CTLA-4[4,5], and one lung tumor cohort treated with anti-PD-1[6]. Our model's efficacy is assessed by its ability to predict overall survival of patients from the time of beginning immunotherapy. Neoantigen amino-acid anchor positions 2 and 9, for the majority of HLA types, are constrained due to their molecular function and display a hydrophobic bias, as reflected by decreased amino-acid diversity at these positions[25] (Extended Data Fig. 2). We observe computational predictions of MHC affinities for wildtype peptides with non-hydrophobic anchor residues lead to non-informative amplitudes. Hence, neoantigens mutated on positions 2 and 9 where the wildtype peptide residue is non-hydrophobic are excluded from our model. Parameter τ, a characteristic evolutionary time scale for a patient cohort, is a finite value at which we expect cancer populations from responding tumors to have been affected by therapy. This is the time at which, following equation (1), tumor clones with neoantigens of amplitudes larger than $1/\tau$ will have been suppressed. The model has two other free parameters: the midpoint and steepness defining R (Methods). For each cohort we infer the parameters by maximizing the survival log-rank test score on independent training data. In the survival analyses, we use the median value of $n(\tau)$ to separate patients into high and low fitness groups.

We use the Snyder melanoma cohort with 64 patients to train parameters for the Van Allen melanoma cohort with 103 metastatic patients and vice versa use Van Allen cohort to train parameters for the Snyder cohort; we use the total score of both melanoma cohorts to train parameters for the smaller lung cancer cohort from Rizvi et al. with 34 patients (Methods). For each cohort we obtain significant stratification of patients: log-rank test *p*-values are p=0.0049 for the Van Allen et al., p=0.0026 for Snyder et al., and p=0.0062 for Rizvi et al. (Extended Data

Table 1). The parameters we obtain with this procedure are consistent between the datasets and are mutually included within each other's error bars (Extended Data Table 1, Methods). Based on this result, we further perform a joint optimization of the cumulative log-rank test score of the three cohorts, obtaining a single set of parameters with predictions highly stable around these values (Fig 2). The alignment threshold parameter of the binding function is consistently set to 26 (Extended Data Table 1), which in our datasets is obtained by alignments of average length of 6.8 amino-acids, just above the length of peptide motifs one would expect the TCR repertoire to discriminate (SI). The slope parameter of the binding function is set to 4.9 defining a strongly nonlinear dependence on the alignment score, with the probability of binding dropping below 0.01 for alignment score 25 and reaching probability above 0.99 at alignment score 27 (Extended Data Fig. 1a). The τ parameter is set to 0.09, meaning that clones with amplitudes larger than 11.1 are, on average, suppressed at prediction time. At these consistent parameters, separation of patients does not change for Van Allen and Rizvi (log-rank score increases by less than 1 unit, p=0.004 for Van Allen at al. and p=0.0062 for Rizvi et al.), and it improves to p=0.00026 for Snyder et al. (Fig. 3). Finally, the predicted evolutionary dynamics of tumors clearly separate therapy responders and non-responders, using patient classifications defined in the original studies[4,5,6]. In all datasets responders are predicted to have significantly faster decreasing relative population sizes $n(\tau)$, across a broad interval τ values (Fig. 4). The performance of the model deteriorates when we disrupt the biological relevance of the input data. When using the IEDB epitopes that are not supported by positive T-cell assays, the model looses the predictive ability in both melanoma cohorts (Methods, Extended Data Table 1 and Extended Data Fig. 3). Similarly, the model generally does not give significant patient separations when we perform the analysis with neoantigens derived with randomized HLA types of patients (Extended Data Fig. 4, SI).

The success of our model strongly depends on the joint contribution of two fitness components, A and R, in equation (2). We deconstruct the model by removing each of the components one at a time and repeat the same training and validation procedure as for the full model (Fig 3, bottom panels and Extended Data Table 1). In all datasets, such partial models have lower log-rank scores than the full model and neither the A-only nor the R-only model result in significant segregation for any cohort. An alternative model, which assigns a uniform fitness cost to each neoantigen (hence the total clone fitness reflects the neoantigen load of the clone, Methods), does not separate patients in either cohort (Fig 3, Extended Data Table 1). It is important to assess the clonal structure of a tumor when trying to identify dominant neoantigens. We compare the performance of our model to one assuming homogenous structure of tumors (Methods). The homogenous model performs worse in all datasets, and does not show the predictive consistency of the heterogenous model. For other fitness models considered here the homogeneous structure versions occasionally outperform their heterogeneous counterparts (Fig 3, Extended Data Table 1),

248 though with either marginal or no statistical significance. We present additional
249 model decompositions in Methods and in Extended Data Table 1. Our model is
250 the only one to significantly segregate patients across all three datasets and it is
251 predictive independent of other clinical correlates (Proportional Hazard model,
252 Extended Data Table 2).
253
254 The presented framework allows for straightforward incorporation of information
255 about the tumor's microenvironment. For the cohort from Van Allen et al., gene
256 expression data is available on 40 patients and local cytolytic activity is
257 significantly associated with benefit (p=0.04, Methods), as also observed in the
258 original study by Van Allen et al[5]. As a proof of principle, we incorporated
259 cytolytic score[26] as an amplitude multiplying the T-cell recognition probability. Its
260 inclusion improves predictions on these 40 patients, as assessed with survival
261 analysis, (p=0.043 and p=0.0025 respectively, Extended Data Fig. 5).
262
263 Immune interactions govern the evolutionary dynamics of cancers under
264 checkpoint blockade immunotherapy and many fast-evolving pathogens; fitness
265 models can predict these dynamics over limited periods, as recently shown for
266 seasonal human influenza[7]. Yet while influenza evolution is determined by
267 antigenic similarity with previous strains in the same lineage, cancer cells
268 originate from normal cells and acquire somatic mutations in a large set of
269 proteins. Hence, their immune interactions are distributed in a larger antigenic
270 space. The fitness effects of these interactions have a specific interpretation:
271 they capture neoantigen "non-selfness". That is, our model provides a structure
272 to formalize what makes a tumor immunologically different from its host,
273 analogous to that for innate recognition of non-self nucleic acids[27].

274 The approach can be naturally extended to other fitness effects, such as positive
275 selection due to acquisition of driver mutations, the impact of additional
276 components in the microenvironment or the hypothesized role of the
277 microbiome[28,29]. Further advances in predicting proteosomal processing[30] and
278 stability[31] of neoantigen-MHC binding could improve predictions. Our
279 evolutionary framework should be useful in studies of acquired resistance to
280 therapy and may be crucial for understanding when cross-reactivity with self-
281 peptides may result in side effects[32,33]. As our fitness model is based on
282 biophysical interactions underlying presentation and recognition of neoantigens,
283 it may also inform the choice of therapeutic targets for tumor vaccine design.
284

References

1. Topalian, S.L. et al. Immune checkpoint blockade: a common denominator approach to cancer therapy. *Cancer Cell* **27**, 450-61 (2015).
2. Schumacher, T.N. & Schreiber, R.D. Neoantigens in cancer immunotherapy. *Science* **348**, 69-74 (2015).
3. Gubin, M.M., Artyomov, M.N., Mardis, E.R. & Schreiber, R.D. Tumor neoantigens: building a framework for personalized cancer immunotherapy. *J. Clin. Invest.* **125**, 3413-3421 (2015).
4. Snyder, A. et al. Genetic Basis for Clinical Response to CTLA-4 Blockade in Melanoma. *N. Engl. J. Med.* **371**, 2189-2199 (2014).
5. Van Allen, E.M. et al. Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. *Science* **350**, 207-211 (2015)
6. Rizvi, N.A. et al. Mutational landscape determines sensitivity to PD-1 blockade in non–small cell lung cancer. *Science* **348**, 124-128 (2015).
7. Łuksza, M. & Lässig, M. Predictive fitness model for influenza. *Nature* **507**, 57-61 (2014).
8. Wang, S. et al. (2015) Manipulating the selection forces during affinity maturation to generate cross-reactive HIV antibodies. *Cell* **160**, 785–797 (2015).
9. Nourmohammad, A., Otwinowski, J. & Plotkin, J.B. Host-pathogen coevolution and the emergence of broadly neutralizing antibodies in chronic infections. *PLoS Genet* **12**, e1006171 (2016).
10. Le, D.T. et al. Mismatch-repair deficiency predicts response of solid tumors to PD-1 blockade. *Science* **eaan6733**, (2017).
11. Tumeh, P.C. et al. PD-1 blockade induces responses by inhibiting adaptive immune resistance. *Nature* **515**, 568–571 (2014).
12. Topalian, S.L. et al. Safety, activity, and immune correlates of anti-PD-1 antibody in cancer. *N. Engl. J. Med*. **366**, 2443–2454 (2012).
13. Herbst, R.S. et al. Predictive correlates of response to the anti-PD-L1 antibody MPDL3280A in cancer patients. *Nature* **515**, 563-567 (2014).
14. de Henau, O. et al. Overcoming resistance to checkpoint blockade therapy by targeting PI3Kγ in myeloid cells. *Nature* **539**, 443–447 (2016).
15. Ayers, M. et al. IFN-γ–related mRNA profile predicts clinical response to PD-1 blockade. *J. Clin. Invest.* In press (2017).
16. McGranahan, N. et al. Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. *Science* **351**, 1463-1469 (2016).
17. Gerlinger M, & Swanton C. How Darwinian models inform therapeutic failure initiated by clonal heterogeneity in cancer medicine. *Br. J. Cancer* **103,** 1139–1143 (2010).
18. Anagnostu, V. et al. Evolution of neoantigen landscape during immune checkpoint blockade in non-small cell lung cancer. *Cancer Discov*. **7**, 264-276 (2016).
19. Purcell, A.W., McCluskey, J. & Rossjohn, J. More than one reason to rethink the use of peptides in vaccine design. *Nature Rev. Drug Discov.* **6**, 404-414 (2017).
20. Deshwar, A. G. et al. PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biol.* **16**, 35 (2015).
21. Andreatta, M. & Nielsen, M□. Gapped sequence alignment using artificial neural networks: application to the MHC class I system. *Bioinformatics* **32**, 511-517 (2016).
22. Vita, R. et al. The immune epitope database (IEDB) 3.0. *Nucleic Acids Res.* **43**, D405-D412 (2014).
23. Murugan, A., Mora, T., Walczak, A.M. & Callan, C.G., 2012. Statistical inference of the generation probability of T-cell receptors from sequence repertoires. *Proc. Natl. Acad. Sci.* **109**, 16161-16166 (2012).
24. Birnbaum, M.E. et al. Deconstructing the peptide-MHC specificity of T cell recognition.*Cell* **157**, 1073-1087 (2014).
25. Lehmann, J., Libchaber, A. & Greenbaum, B.D. Fundamental amino acid mass distributions and entropy costs in proteomes. *J. Theor. Biol.* **410**, 119-124 (2016).
26. Rooney, M. S., Shukla, S. A., Wu, C. J., Getz, G. & Hacohen, N. Molecular and genetic properties of tumors associated with local immune cytolytic activity. *Cell* **160**, 48-61 (2015).
27. Tanne, A. et al. Distinguishing the immunostimulatory properties of noncoding RNAs expressed in cancer cells. *Proc. Natl. Acad. Sci. USA* **112**, 5154-15159 (2015).
28. Vétizou, M. et al. Anticancer immunotherapy by CTLA-4 blockade relies on the gut microbiota. *Science* **350**, 1079–1084 (2015).
29. Dubin, K. et al. Intestinal microbiome analyses identify melanoma patients at risk for checkpoint-blockade-induced colitis. *Nat. Commun*. **7,** 10391 (2016).
30. Abelin, J.G. et al. Mass spectrometry profiling of hla-associated peptidomes in mono-allelic cells enables more accurate epitope prediction. *Immunity* 46, 315-326 (2017).
31. Strønen, E. et al. Targeting of cancer neoantigens with donor-derived T cell receptor repertoires." *Science* **352**, 1337-1341 (2016).
32. Johnson, D.B. et al. Fulminant myocarditis with combination immune checkpoint blockade. *New Engl. J. Med.* **375**, 1749-1755 (2016).
33. Hofmann, L. et al. Cutaneous, gastrointestinal, hepatic, endocrine, and renal side-effects of anti-PD-1 therapy. *European J. Cancer* **60**, 190-209 (2016).

Figures

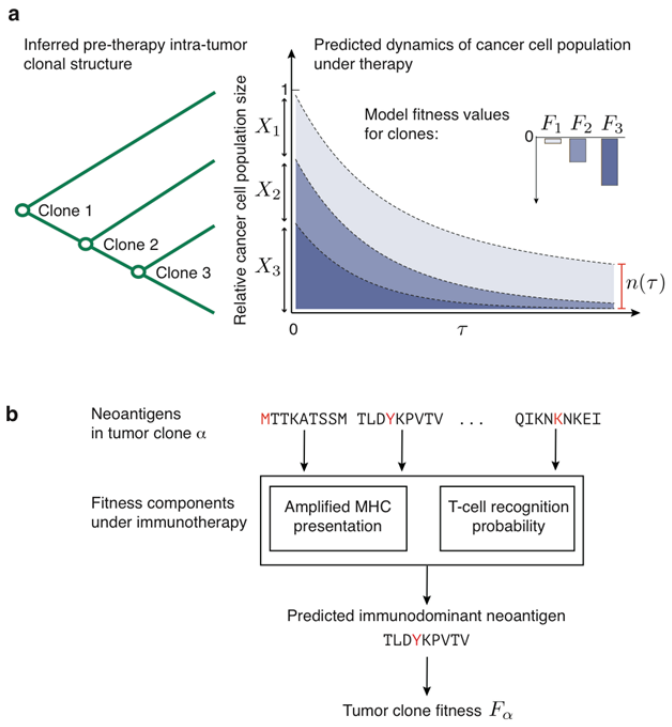


Figure 1 | Evolutionary tumor dynamics under strong immune selection and a neoantigen fitness model based on immune interactions. a, Clones are inferred from a tumor's phylogentic tree. We predict $n(\tau)$, the future effective size of the cancer cell population, relative to its size at the start of therapy (equation (1)) by evolving clones forward under the fitness model over a fixed time-scale, τ. Application of therapy can decrease fitness of tumor clones depending on their neoantigens. Tumors with strongly negative fitness have a greater loss of population size than more fit tumors. **b**, Our fitness model accounts for the presence of dominant neoantigens within a clone, α, by modeling the presentation and recognition of inferred neoantigens and assigning a fitness to a clone, F_α.

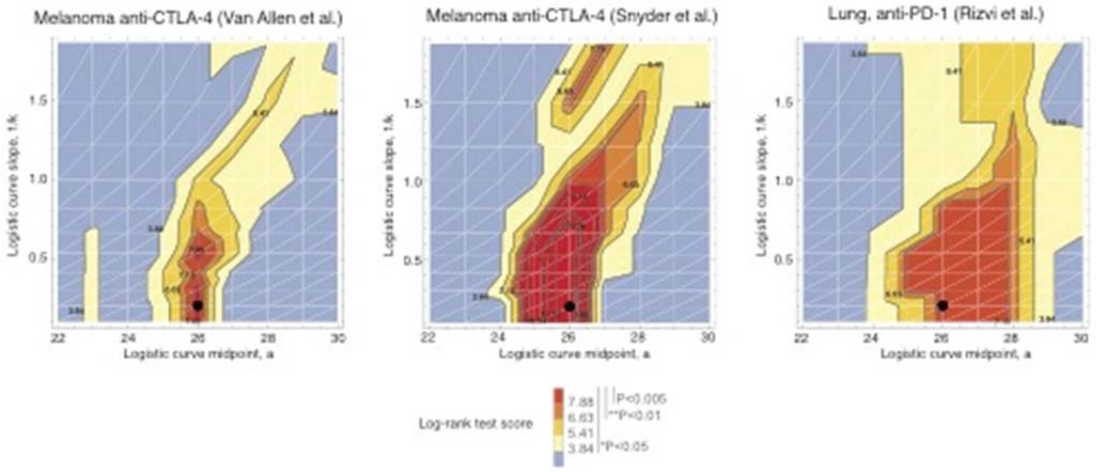


Figure 2 | Survival landscape as a function of the TCR binding model parameters. The landscape of log-rank test scores is shown for the consistent choice of $\tau = 0.09$, as the function of the parameters of the TCR binding model (a and $1/k$), colors represent the significance level of the long-rank test. Similar regions of high log-rank scores exist for the full model across all three datasets. The point corresponding to consistent parameters ($a = 26$ and $k = 4.9$) is marked by a black dot in each plot.

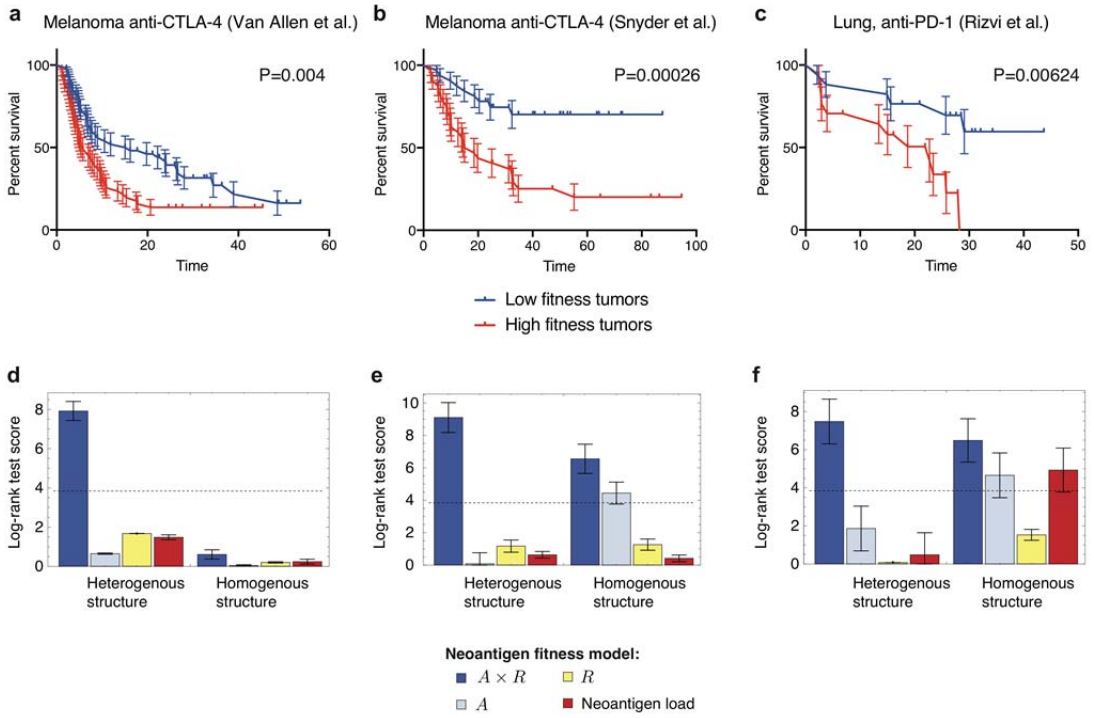


Figure 3 | **Neoantigen fitness model is predictive of patient survival after checkpoint blockade immunotherapy. a-c**, Tumor fitness is calculated across two melanoma patient datasets treated with anti-CTLA-4 antibodies[4,5] and one dataset of lung patients treated with anti-PD-1 antibodies[6] (see main text). Kaplan-Meier curves of overall survival are displayed for each cohort, with samples split by the median value of their tumor's relative population size $n(\tau)$ defined in equation (1). Error bars represent the standard error due to sample size. **d-f,** For comparison we show the log-rank test statistic for our full model and for models which account for removal of one feature of our model (bottom panels, higher score values indicate better patient segregation). We compare their values with a tumors' neoantigen load, which is the total number of neoantigens found in a sequenced tumor clone (red). All models are computed both over a tumor's clonal structure (heterogenous, left) and without taking heterogeneity into account (homogenous, right). All model scores are presented for parameters obtained on independent training data (Methods). The error bars are the standard deviation of the log-rank test score acquired from the survival analysis with one sample removed from the cohort at a time. Dashed lines on the bottom panels marks the score value corresponding to the log-rank test significance threshold of 5%; the full $A \times R$ model with heterogenous structure is the only model that gives scores that substantially exceed the threshold in all datasets, using a consistent set of parameters between all datasets (Extended Data Table 1).

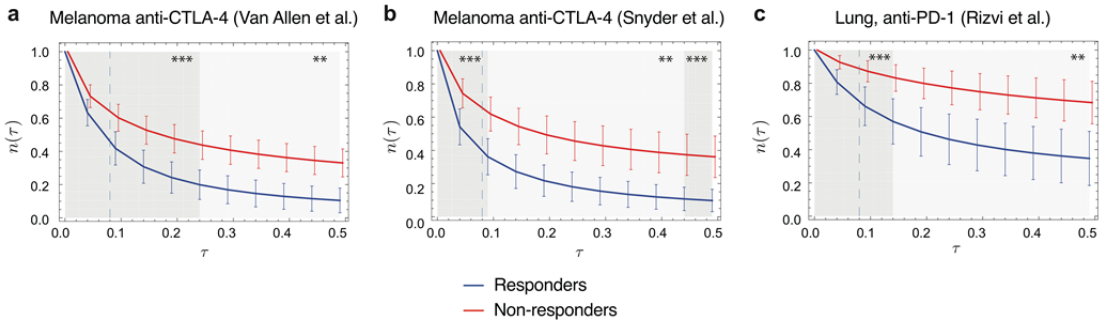
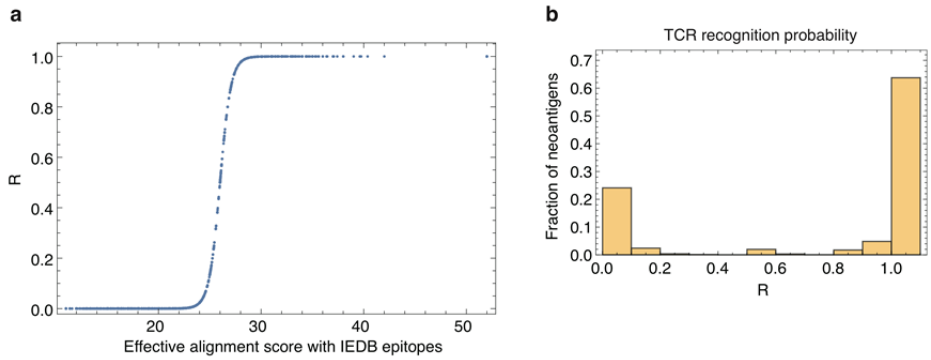
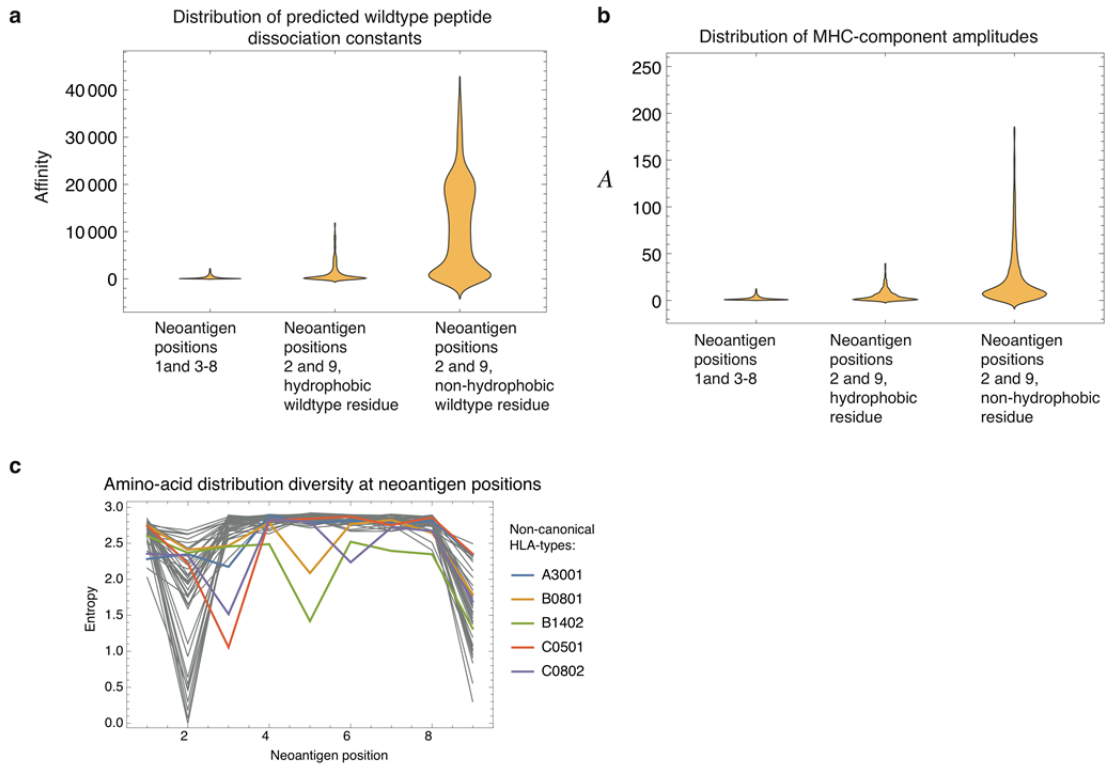


Figure 4 | **Evolutionary dynamics predictions in patient cohorts. a,** Relative population size predictions for responders and non-responders at consistent parameters across the **a,** Van Allen et al.; **b,** Snyder et al.; and **c,** Rizvi et al. cohorts. Responder and nonresponder patient classifications were defined by those studies. Error bars are 95% confidence intervals around the population average. The dashed line indicates the consistent choice of $\tau = 0.09$ used across all three datasets for patient survival predictions (Methods and Extended Data Fig. 3). The shading of the background represents the significance of separation of the two groups as computed with Kolmogorov-Smirnov test (**<0.01, ***<0.001).

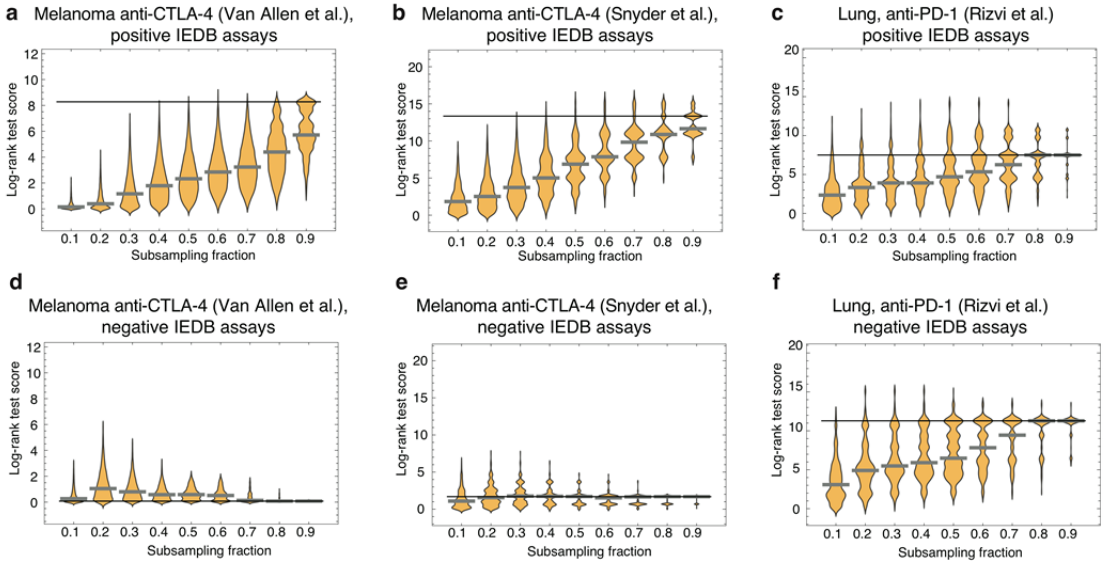
Extended Data Figures



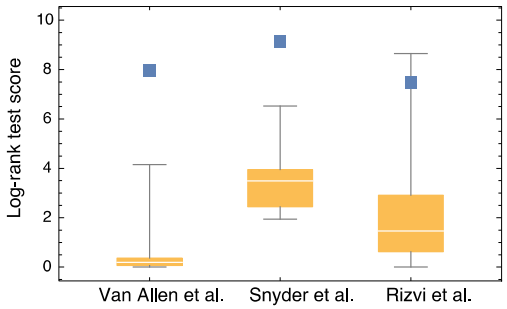
Extended Data Figure 1 | **Alignments to IEDB epitopes. a,** The TCR recognition probability for a neoantigen is a sigmoidal function of the neoantigen's alignment scores with IEDB epitopes, here shown as evaluated for the set of neoantigens from Van Allen et al. cohort patients, using the set of consistent parameters. **b,** The fraction of neoantigens with a given value of R.



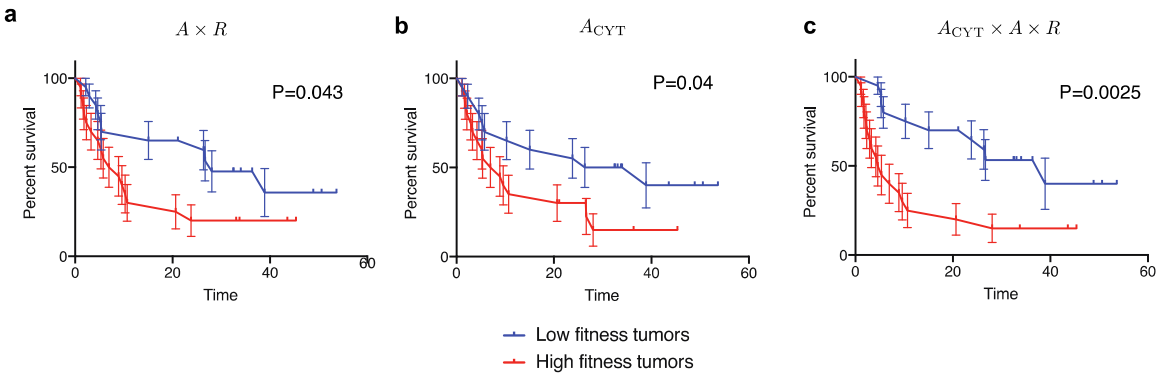
Extended Data Figure 2 | **Positions 2 and 9 have a subset of neoantigens with less predictive value. a,** Neoantigens coming from mutations at position 2 or 9 tend to have larger predicted affinities for their wildtype peptides. In particular this is magnified if the corresponding wildtype residue is non-hydrophobic. **b,** The observed biases at these positions are reflected in a wider distribution of amplitudes for wildtype peptides with non-hydrophobic residues at positions 2 and 9. **c,** Shannon entropy of amino acid diversity at 9 positions in neoantigen sequence, shown for all distinct HLA-types. The entropies are computed based on all neoantigens across all three datasets. Positions 2 and 9 have Shannon entropy lower than other residues. Other residue sites have the same entropy as the overall proteome[25] and are therefore unconstrained. Five HLA with non-canonical entropy profiles are singled out in the plot. These HLA-types contributed only 5 informative neoantigens with mutations on non-canonical anchor positions across all datasets and they are not treated differentially in our model.



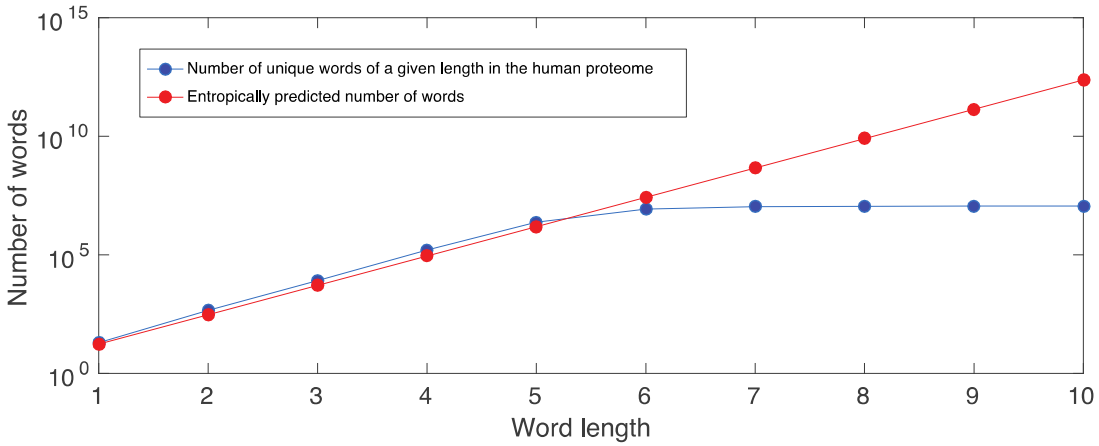
Extended Data Figure 3 | Effect of IEDB database sequence content on predictive power of neoantigen fitness model. Predictions were performed using subsampled IEDB epitope sequences, with subsampling rate varying between 0.1 and 0.9. For each subsampling rate 10000 iterations were performed to obtain a distribution of log-rank test scores. Solid black lines mark the log-rank test score of the prediction on the full set of epitope sequences, and gray thick lines mark the median scores on subsampled data. **a-c,** Subsampling of the original set of IEDB sequences, which are supported by positive T-cell assays, shows that the quality of predictions decreases with subsampling rate. The prediction quality is more robust in case of the Snyder et al. and Rizvi et al. datasets. **d-f,** Analogous subsampling procedure was repeated on IEDB sequences, which are not supported by positive T-cell assays. For the Van Allen et al. and Snyder et al. datasets the model performance is substantially lowered.



Extended Data Figure 4 | **Reshuffling patient HLA-types reduces predictive power of the neoantigen fitness model.** In each cohort, we performed 10 iterations of reshuffling patient HLA-types, followed by computational neoantigen prediction, fitness model calculation and survival analysis. We report the distribution of log-rank test scores over the iterations, boxes mark 75% confidence intervals, and whiskers mark the range of scores. The score values for the model on the original data are marked with blue squares.



Extended Data Figure 5 | **Inclusion of cytolytic score improves prediction quality. a,** Survival plot is shown for our model applied to Van Allen et al. on the 40 patient subset for which transcriptional data was available. **b,** An optimized model (Methods) for cytolytic score can significantly separate patients. **c,** Inclusion of cytolytic score in our model improves prediction on the 40 patient subset. In **a** and **c** we use consistent parameters trained on the three cohorts (Fig. 2); in **b** parameter τ is optimized.



Extended Data Figure 6 | Word usage in the proteome is exhausted between 5 and 6 letter words. Given the entropy of the genome from Ref. 25 we calculate the expected number of words of a given length in the proteome as a function of word length. We compare that to the number of unique words in the proteome of a given length. Between 5 and 6 letters the two curves diverge due to the finite size of the genome. By the time one reaches 9 letter nonamers (the length of a neoantigen) this divergence is of several orders of magnitude.

Extended Data Table 1 | **Ranking of fitness models.** We compare survival prediction of our full fitness model (Methods, equation (9)) with alternative models described in Methods: **(1)** models that eliminate one of the features of the full model, namely the A-only model (Methods, equation (12)) and the R-only model (Methods, equation (13)); models without the wildtype dissociation constant (Methods, equation (14)) and without the mutant dissociation constant (Methods, equation (15)); simple neoantigen load model (Methods, equation (17)); an additive neoantigen fitness model (Methods, equation (15)), which uniformly summates fitness contributions of neoantigens in a clone as opposed to maximizing them as in our original model. Additionally, we compare the model in which alignments to IEDB epitopes are evaluated only on position 3-8, a model that does not implement any filtering of neoantigens on position 2&9, and a model where the R component is evaluated on IEDB assays without positive validation. Finally, we test an alternative predictive criterion and instead of $n(\tau)$, we use the average fitness over tumor clones (Methods, equation (18)) to separate patients in the survival analysis. **(2)** Above models evaluated without accounting for clonal structure structure of tumors. For each model, if applicable, we report the parameters used for predictions, and error bars for these parameters (Methods). We also report the predictive power of the models as the log-rank test score, and the log-rank test *p*-value for models with p-values less than 0.05. The models with significant separation are highlighted: yellow for a model that is significant in a single cohort, orange for a model that is significant across two cohorts and red for a model that is significant in all three cohorts.

Extended Data Table 2 | **Multivariate analysis with a cox proportional hazards model.** A multivariate analysis with a cox proportional hazards model, was performed to adjust for important clinical covariates, while assessing for the predictive value of our $n(\tau)$ values. In the melanoma data sets, we controlled for stage, gender, and age. Stage IIIC and IVa are combined together, as both of these stages had limited number of patients in either cohort. Stage IIIc/IVa serve as the reference in the table. In both the Snyder and Van Allen cohorts, $n(\tau)$ predictions are independently associated with overall survival after anti-CTLA4 therapy. In the lung cancer cohort, all patients are Stage IV, so we correct for age, gender, and number of pack years smoked, and continued to find that $n(\tau)$, predictions are independently associated with overall survival after anti-PD1 therapy.

Acknowledgments

We thank Nina Bhardwaj, Curt Callan, Simona Cocco, Yuval Elhanati, Dmitry Krotov, Steven Leach, Stanislas Leibler, Albert Libchaber, Remi Monasson, Armita Nourmohammad, Vladimir Roudko, Zachary Sethna, Alexandra Snyder-Charen, Petr Sulc, and the members of Chan, Greenbaum, and Wolchok laboratories for many helpful discussions. We thank Michael Lässig for important suggestions about the biophysical model and for critical reading of the manuscript. We thank Alexandra Snyder-Charen, and David T. Ting, and for their critical reading of the manuscript. Research was supported by a Stand Up to Cancer-American Cancer Society Lung Cancer Dream Team Translational Research Grant (SU2C-AACR-DT17-15) (M.D.H., T.M., J.D.W.), a Stand Up to Cancer-National Science Foundation-Lustgarten Foundation Convergence Dream Team Grant (M.Ł., A.S., J.D.W, B.D.G, T.A.C.), a Phil A. Sharp Innovation in Collaboration Award from Stand up to Cancer (B.D.G, J.D.W.), NCI-NIH grant P01CA087497 (M.Ł.), the STARR Cancer Consortium (T.A.C.), the Pershing Square Sohn Cancer Research Alliance (T.A.C.), the National Institutes of Health (NIH) R01 CA205426 (N.A.R, T.A.C.), the V Foundation (V.P.B., A.S., J.D.W, B.D.G), the Lustgarten Foundation (A.S., J.D.W., B.D.G.), the National Science Foundation (NSF) 1545935 (B.D.G, J.D.W), the Swim Across America, Ludwig Institute for Cancer Research, Parker Institute for Cancer Immunotherapy, the National Cancer Institute (NCI) K12 Paul Calabresi Career Development Award for Clinical Oncology K12CA184746-01A1 (V.P.B.). Stand Up to Cancer is a program of the Entertainment Industry Foundation. The work was also supported in part by the MSKCC Core Grant (P30 CA008748).

Author Contributions

M.Ł. and B.D.G. designed the mathematical model and wrote the manuscript with critical comments from all the authors. M.Ł., N.R., V.M., V.P.B., A.S., N.A.R., T.M., A.J.L., T.A.C., J.D.W., and B.D.G contributed to data acquisition and analysis. M.Ł., T.A.C., J.D.W., and B.D.G. contributed to study conception and design. M.Ł., N.R., V.M., V.P.B., A.S., N.A.R., T.M., A.J.L., T.A.C., J.D.W., and B.D.G. interpreted the data and provided a critical reading of the manuscript.

Methods

1. Evolutionary dynamics of a cancer cell population in a tumor

The fitness of a cancer cell in a genetic clone α is its expected replication rate, i.e.

$$\frac{dN_\alpha}{d\tau} = F_\alpha N_\alpha \tag{3}$$

where N_α is the population size of clone α and F_α is that clone's fitness. Checkpoint-blockade immunotherapy introduces a strong selection challenge, which we anticipate overshadows pre-therapy fitness effects in a productive response. For a given clone α the dynamics of its absolute size are therefore given by $N_\alpha(\tau) = N_\alpha(0)\exp(F_\alpha\tau)$, and the total cancer cell population size is computed as a sum over its clones

$$N(\tau) = \sum_\alpha N_\alpha(\tau) = \sum_\alpha N_\alpha(0)\,\exp(F_\alpha\tau). \tag{4}$$

The absolute size $N(\tau)$ is an effective population size, the number of cells estimated to have generated the observed clonal diversity.

As our measure of survival, we use the evolved relative effective population size $n(\tau) = N(\tau)/N(0)$, which compares the predicted future population size after a characteristic dimensionless time scale of evolution τ to the initial pretreatment effective size $N(0)$, the assumption being that successful responders to therapy will have their future effective cancer cell population size more strongly suppressed. We denote the initial frequency of clone α as $X_\alpha = N_\alpha(0)/N(0)$, these frequencies are inferred from bulk exome reads from a tumor sample, as described in the Supplementary Information. Hence, to compute $n(\tau)$ we only require estimates of the initial frequencies and fitness values for each clone, as shown in equation (1); the absolute population size estimates are not needed. We model the hypothesis that due to the unleashing of a T cell mediated immune response by checkpoint-blockade immunotherapy, the deleterious effects due to recognition of neoantigens are a dominant fitness effect, and tumors with the greatest degree of selective immune challenge are better responders to therapy.

Clonal structure of a tumor and clone frequencies. Tumor clones are reconstructed using the PhyloWGS software package[20] (SI). The trees estimate the nested clonal structure of the tumor and the frequency of each clone, X_α. The differences between the high scoring trees are marginal on our data, concerning only peripheral clones and small differences in frequency estimates. We compute the predicted relative size of a cancer population, $n(\tau)$, as an averaged prediction over the 5 trees with the highest likelihood score, weighting their contribution proportionally to their likelihood.

2. Fitness model based on neoantigen recognition potential

Neoantigen recognition based fitness cost for a tumor clone. Our model associates each neoantigen with a fitness cost, which we term the *recognition potential* of a neoantigen. The recognition potential of a neoantigen is the likelihood it is productively recognized by the TCR repertoire. It is defined by two components. The first is the amplitude A, which is given by the relative probability that a neoantigen will be presented on class I MHC and the relative probability that its wildtype counterpart will not be presented. The second one is the probability R that a presented neoantigen will be recognized by the TCR repertoire. For a given neoantigen their product defines its recognition potential, $A \times R$. Both components are described in detail in the following paragraphs in this section.

To assess the total fitness effect for a clone α with multiple neoantigens, we aggregate the individual neoantigen fitness effects as $F_\alpha = -\max_{i \in \text{Clone }\alpha}(A_i \times R_i)$, where i is an index iterating over neoantigens in the clone. Therefore, the full form of the predicted relative cancer cell population size is given by

$$n(\tau) = \sum_\alpha X_\alpha \exp[-\max_{i \in \text{Clone }\alpha}(A_i \times R_i)\,\tau]\,. \tag{5}$$

One could use a more general model for aggregating neoantigen fitness effects within a clone,

$$n(\tau, \beta) = \sum_\alpha X_\alpha \exp\left[\sum_{i \in \text{Clone }\alpha} \frac{\exp(-\beta f_i)}{Z(\beta)}\, f_i\, \tau\right], \tag{6}$$

where $f_i = -A_i \times R_i$ and $Z(\beta) = \sum_{i \in \text{Clone }\alpha} \exp(-\beta f_i)$. In addition to equation (5), which corresponds to the limit $\beta \to \infty$, we show the case where $\beta = 0$ (uniform summation over all neoantigens, Extended Data Table 1). In that sense equation (6) represents a general mathematical framework for weighting neoantigen contributions with weights reflecting the probability of their productive recognition. The choice of β could be informed by additional data sources or defined in a clone specific manner, and it would then become an additional model parameter (or parameters). Taking the highest score within a clone as in equation (5) is consistent with notions of immunodominance – that a relatively small set of antigens drive the immune response.

MHC-amplitude. The amplitude, A, is the ratio of the relative probability that a neoantigen is bound on class I MHC times the relative probability that a neoantigen's wildtype counterpart is not bound. The amplitude is defined as $A = (P_U^{WT}/P_B^{WT}) \times (P_B^{MT}/P_U^{MT})$, where P_B^{MT} is binding probability of a neoantigen, P_B^{WT} is the binding probability of its wildtype counterpart, and $P_U^{WT} = 1 - P_B^{WT}$ and

$P_U^{MT} = 1 - P_B^{MT}$. As a result, the amplitude rewards cases where the discrimination energy between a mutant and wildtype peptide by the same class I MHC molecule (i.e. the same HLA allele) is large[34], while the mutant binding energy is also low. The τ parameter effectively sets this energy scale for dominant neoantigens in a clone when $R = 1$. Assuming similar concentrations for mutant and wildtype peptides, the amplitude is the ratio of wildtype to mutant dissociation constants,

$$A = K_d^{WT}/K_d^{MT}. \quad (6)$$

Negative thymic selection on TCRs is not absolute, but rather "prunes" the repertoire recognizing the self proteome[35,36]. We therefore use A as a proxy for the availability of TCRs in the repertoire to recognize a neoantigen. Neoantigens differ from their wildtype peptides by only a single mutation. Given the uniqueness of nonamer sequences in the self-proteome due to finite genome size (Extended Data Fig. 6) it is highly improbable that the mutant peptide would have another 8-mer match in the human proteome, so we only account for the comparison with the respective wildtype peptides. We verified that the above is the case for 92% of all neoantigens, with the remainder largely emanating from gene families with many paralogs (SI). The amplitude can be interpreted as a multiplicity of receptors available to cross-reactively recognize a neoantigen.

The MHC-binding probabilities are derived from the dissociation constants, which are themselves inferred from computationally predicted binding affinities, as justified below. Affinities are inferred for each peptide sequence and patient HLA type[21]; all mutant peptide sequences considered as neoantigens meet a standard 500 nM cutoff for their affinities (SI). The software, which predicts affinities occasionally predicts affinities with very high values where accuracy may be limited, and creating small denominators that can inflate the amplitude. This is a possibility in cancers such as melanoma and lung, where a high mutational burden inflates the probability of such events. As a remedy, a pseudocount, ε, is introduced so that, for both mutant and wildtype peptides $P_U/P_B \to (P_U + \varepsilon)/(P_B + \varepsilon)$. In this case the new dissociation constant divided by peptide concentration becomes

$$\frac{K_d/[L] + \varepsilon(1 + K_d/[L])}{1 + \varepsilon(1 + K_d/[L])} \approx \frac{K_d/[L]}{1 + \varepsilon K_d/[L]} \quad (7)$$

for small ε, where K_d was the original dissociation constant and $[L]$ is the peptide concentration. Consequently $1/\varepsilon$ sets a scale at which dissociation constants are not reliable for large K_d at a given concentration. To fix these scales, we note that assays to determine dissociation constants for peptide-MHC binding are typically performed at 0.1-1 nM where the ligand concentration is typically small compared to the dissociation constant[37]. In this regime, affinities can be interpreted as dissociation constants and 3687 nM is the outer range of predictability for the

assays upon which NetMHC is trained at no more that unit peptide concentrations. $\varepsilon/[L]$ is therefore chosen to be 0.0003≈1/3687 across datasets.

As the affinity is always less than 500 nM for the mutant peptide this correction is only relevant for the wildtype peptides. The corrected amplitude then becomes

$$A \approx \frac{K_d^{WT}}{K_d^{MT}} \cdot \frac{1}{1 + (\varepsilon/[L]) \cdot K_d^{WT}}. \quad (8)$$

The amplitude in this form, combined with the TCR-recognition term discussed below, has a high predictive value for patient survival predictions (Fig. 3), consistently over the three patient cohorts, which is not the case of either the mutant or wildtype dissociation constants on their own (Extended Data Table 1).

TCR-recognition. We model R, the probability that a neoantigen will be recognized by the T-cell receptor repertoire by alignment with a set of epitopes given by the Immune Epitope Database and Analysis Resource[22] (IEDB, described in the Supplementary Information). We restrict ourselves to IEDB epitopes that are positively recognized by T-cells after class I MHC presentation. In this approach, we assume that a neoantigen that predicted to cross-react with a TCR from this pool of immunogenic epitopes is a neoantigen more likely to be immunogenic itself, as members of the T-cell receptor repertoire both recognize a high number of presented antigens[38,39] and have intrinsic biases in their generation probabilities[23].

We use a multistate thermodynamic model to define R. In this model, we treat sequence similarities as a proxy for binding energies. To assess sequence similarity between a neoantigen with peptide sequence $\mathbf{s}$ and an IEDB epitope $\mathbf{e}$, we compute a gapless alignment between the two sequences with a BLOSUM62 amino-acid similarity matrix[40] and we denote their alignment scores as $|\mathbf{s}, \mathbf{e}|$, Given these sequence similarities, for a given neoantigen with peptide sequence $\mathbf{s}$, we compute the probability that it is bound by a TCR specific to some epitope $\mathbf{e}$ from the IEDB pool as

$$R = Z(k)^{-1} \sum_{\mathbf{e} \in \mathrm{IEDB}} \exp[-k(a - |\mathbf{s}, \mathbf{e}|)], \quad (9)$$

where a represents the horizontal displacement of the binding curve, k sets the steepness of the curve at a, and

$$Z(k) = 1 + \sum_{\mathbf{e} \in \mathrm{IEDB}} \exp[-k(a - |\mathbf{s}, \mathbf{e}|)] \quad (10)$$

is the partition function over the unbound state and all the bound states. In the model, k functions as an inverse temperature and $a - |\mathbf{s}, \mathbf{e}|$ functions as a binding

energy. These parameters define the shape of the sigmoid function (Extended Data Fig. 1) and, along with the characteristic time scale τ, are free parameters to be fit in our model (see below).

The parameters which give consistently informative predictions across all three datasets are $a = 26$ and $k = 4.9$. The logistic function is therefore a strongly nonlinear function of the effective alignment score, $\log(\sum_{\mathbf{e}\in \mathrm{IEDB}} \exp[-k(a - |\mathbf{s}, \mathbf{e}|)])$. The average alignment length corresponding to score 26 is 6.8 for neoantigens in our datasets, but the effective alignment score is occasionally increased by multiple contributions of shorter alignments. Under the interpretation where, for a sufficiently presented neoantigen, A represents the multiplicity of available TCRs and R represents an intrinsic probability of recognition, $A \times R$ represents the effective size of the overall TCR response. We present it as a core quantity that can be modulated by additional environmental factors such as the T-cell infiltration (discussed below).

IEDB sequences. The predictive value of R depends on the input set of IEDB sequences. The set we used in our analysis contained 2552 unique epitopes. We tested how the predictions depend on the content and size of the dataset by performing iterative subsampling of IEDB sequences at frequencies varying from 10% to 90% of the total set size. We repeated the survival analysis and log-rank test score evaluation (Extended Data Figure 3). For all three datasets removal of sequences has on average a negative impact on their predictive power, which monotonically decreases with the subsampling rate. In the Van Allen et al. cohort the median performance was below significance threshold already at 70% subsampling and lower, and for Snyder et al. and Rizvi et al. at 20% and lower. To investigate the biological input associated with the set of curated IEDB sequences that we use, we also evaluated the R component using an alternative set of IEDB sequences, coming from T-cell assays that did not have a positive validation. This is a larger set of 4657 sequences. In the two melanoma datasets, the predictions have gotten worse, not giving significant separation of patients in the survival analysis. This effect was also not due to the different sequence set size – subsampling of sequences did not improve the outcome. While in the Rizvi et al. dataset the predictions were still significant, this significance was not supported by consistency between all three datasets which is observed on the canonical IEDB sequence set.

Inclusion of microenvironment and proteosomal processing in fitness model. The role of the microenvironment in the likelihood of productive T-cell recognition of tumor neoantigens can be incorporated in a natural manner into our modeling framework. We utilize the cytolytic score (CYT), the geometric mean of the transcript per kilobase million of perforin and granzyme[26]. We do so for the 40 patients from the Van Allen, et al, anti-CTLA4 melanoma dataset, which have matched genome and transcriptome sequencing and where CYT had shown predictive value. For this set we also derive the CD8 T-cell fraction using CIBERSORT[41]. The two values have a Pearson correlation coefficient of 0.938.

Given their encapsulation of similar information we used CYT as it had previously been show to give significant segregation of patient benefit[5]. The score provides an additional amplitude A_{CYT} and the recognition potential becomes $f = A_{CYT} \times A \times R$. Therefore, the cytolytic score amplifies the recognition potential by the degree of cytolytic activity. We attempted to include proteosomal processing into our model as an additional criterion, as evaluated with NetCHOP[42]. We tested this procedure on the Rizvi et al. cohort; however, the imposed stronger filtering of neoantigens lead to the loss of predictive power of the model.

3. Model parameters

Parameter selection. To choose model parameters a and k in equation (9) and the characteristic time τ at which the prediction is evaluated (equations (2) and (5)), we select the parameters that maximize log-rank-test scores of survival analysis on patient cohorts. The survival analysis is performed by splitting patient cohort by the median value of $n(\tau)$ into *high* and *low fitness* groups. For each cohort, we perform parameter training on independent data: we use the melanoma cohorts to train parameters for each other by using the maximal score of one to define parameters for the other, and we use both melanoma cohorts and maximization of their total log-rank test score to train parameters for the lung cohort. To infer consistent parameters between all datasets, we maximize the total log-rank test score over the three cohorts.

For a given training set we compute the optimal parameters $\widehat{\Theta} = [a, k, \tau]$, as an average $\widehat{\Theta} = \langle \Theta \rangle_w$ over a distribution $w(\Theta)$ defined by the log-rank test score landscape on this set

$$w(\Theta) = Z^{-1}(\lambda)\ \exp[\lambda(S_{\max} - S(\Theta)]\,, \tag{11}$$

where $Z(\lambda)$ is the probability distribution normalization constant, $S(\Theta)$ is the value of the log-rank test score with parameters Θ and $S_{\max}$ is the maximal score value obtained over all possible parameters. The weight parameter λ is chosen such that the total statistical weight of the suboptimal parameter region is less than 0.01, the suboptimal scores are those less than $\max(3.841\,, S_{\max} - 2)$ (where 3.841 is the score value corresponding to 5% significance level of the log-rank test score). Using a smooth local neighborhood of parameters around the optimal values prevents overfitting on a potentially rugged score landscape. For each individual parameter, the error bars reported in Extended Data Table 1 are computed as standard deviation using marginalized probability distribution $w(\Theta)$ for this parameter.

The survival score landscapes (Fig. 2, at τ=0.09) are consistent between the datasets. The optimal value of parameter a, the midpoint of the logistic binding function is around 26 and parameter k, the steepness of the logistic function lives on a trivial axis above value 4, suggesting strong nonlinear fitness dependence on the sequence alignment score.

4. Model selection

Fitness models. We compare our full model in equation (5) to alternative models. We perform simple model decompositions, where only one component is used

$$F_\alpha = -\max_{i\in \text{Clone }\alpha} A_i, \tag{12}$$

$$F_\alpha = -\max_{i\in \text{Clone }\alpha} R_i. \tag{13}$$

We also further decompose the amplitude $A = K_d^{WT} \times \frac{1}{K_d^{MT}}$ and test various variants of the model, with and without the R component,

$$F_\alpha = -\max_{i\in \text{Clone }\alpha} {K_d^{WT}}_i [\times R_i], \tag{14}$$

$$F_\alpha = -\max_{i\in \text{Clone }\alpha} \frac{1}{{K_d^{MT}}_i} [\times R_i]. \tag{15}$$

Further, we investigate how informative the alignments contributing to the R_i components are. We test a model where alignments are restricted to the 6 residues in-between anchor positions, on positions 3-8. We also demonstrate the loss of predictive power of a model that does not implement any filtering of neoantigens mutated on position 2&9 (see discussion in section 2 of Methods and Extended Data Fig. 2).

We can also reduce the problem of choosing the neoantigen aggregating function to that of model selection. Here we test a fitness model where the fitness is defined by the total effect of all neoantigens in the clone (which is the limit case of $\beta = 0$ in equation (6)),

$$F_\alpha = -\sum_{i\in \text{Clone }\alpha} A_i \times R_i. \tag{16}$$

Finally, we formulate a simple fitness model that associates a constant fitness cost with each neoantigen,

$$F_\alpha = -L_\alpha, \tag{17}$$

where L_α is the number of neoantigens in clone α, referred to as the neoantigen load of clone α.

Homogenous structure models. For each fitness model, we can define its homogenous structure equivalent, which assumes tumor is strictly clonal with all neoantigens in the same clone at frequency 1. In a homogenous model the population size is thus modeled by a simple exponential,

$$n(\tau) = \exp[F\tau], \tag{18}$$

where F is the fitness of the homogenous tumor. Since in this model tumors show a constant decay over time, the ranking of $n(\tau)$ values of patients is defined only by fitness, and does not depend on τ. Therefore, τ is not a free parameter in these models when optimizing log-rank test score in survival analysis.

Average fitness. We also investigate the average fitness of clones,

$$\langle F \rangle = \sum_{\alpha} X_{\alpha} F_{\alpha}, \tag{19}$$

as a predictive marker and an alternative to $n(\tau)$. The average fitness reflects short-term dynamics – how fast the population is decreasing in size at the beginning of therapy. This is a lower complexity model because it does not include parameter τ. However, this model is less robust to outliers – small clones with very low fitness can dominate the average fitness, while the evolutionary projection in $n(\tau)$ naturally removes such effects.

We assess the predictive power of all models with a survival analysis, by separating patients by the median value of $n(\tau)$ (or median value of the average fitness $\langle F \rangle$) in each patient cohort and computing the log-rank test for such segregation. The results of this comparison are reported in Extended Data Table 1. To assign error bars to fluctuations of the log-rank test score we perform a leave-one-out analysis. That is we repeat the survival analysis for each dataset after leaving out one sample in a cohort and compute standard deviation of the test statistic over all leave-one-out iterations. Our approach also assesses the degree to which scores are robust to outliers. We claim a model is *predictive* if it has highly significant scores in all datasets with the same consistent set of parameters. Only the full neoantigen model meets these criteria. The average model, which uses $\langle F \rangle$ as a segregating criterion, marginally meets these requirements, but with less predictive power.

5. Data availability

Mutation data, inferred neoantigen peptide data for each dataset, and IEDB sequences are submitted as supplementary data.

References

34. Stormo, G.D. Modeling the specificity of protein-DNA interactions. *Quantitative Biol*. **1**, 115-130 (2013).
35. Yu, W., et al. Clonal deletion prunes but does not eliminate self-specific αβ CD8+ T lymphocytes. *Immunity* **42**, 929-941 (2015).
36. Legoux, F.P., et al. CD4+ T cell tolerance to tissue-restricted self antigens is mediated by antigen-specific regulatory T cells rather than deletion. *Immunity* **43**, 896-908 (2015).

37. 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).
38. Mason, D. A very high level of crossreactivity is an essential feature of the T-cell receptor. *Immunology Today* **19**, 395–404 (1999).
39. Sewell, A.K. Why must T cells be cross-reactive? *Nature Rev. Immunol.* **12**, 669-677 (2012).
40. Henikoff, S. & Henikoff, J.G. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA* **89**, 10915–10919 (1992).
41. Newman, A.M., et al. Robust enumeration of cell subsets from tissue expression profiles. *Nature Methods* **12**, 453-457 (2015).
42. Nielsen, M., Lundegaard, C., Lund, O., & Kesmir, C. The role of the proteasome in generating cytotoxic T cell epitopes: Insights obtained from improved predictions of proteasomal cleavage. *Immunogenetics* **57**, 33-41 (2005).