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], Alexander Solovyov[8], Naiyer A. Rizvi[9], Taha Merghoub[7,10,11], Arnold J. Levine[1], Timothy A. Chan[2,3,4,7], Jedd D. Wolchok[7,10,11,12], Benjamin D. Greenbaum[8,*]

[1]The Simons Center for Systems Biology, Institute for Advanced Study, Princeton, NJ, USA.
[2]Departments of Radiation Oncology, [5]Surgery and [12]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.
[8]Tisch Cancer Institute, Departments of Medicine, Oncological Sciences, and Pathology, Icahn School of Medicine at Mount Sinai, New York, NY, USA.
[9]Department of Medicine, Columbia University Medical Center, New York, NY, USA
[10]Ludwig Collaborative and Swim Across America Laboratory, Memorial Sloan Kettering Cancer Center, New York, NY, USA.
[11]Melanoma and Immunotherapeutics Service, Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, NY, USA; Weill Cornell Medical College, Cornell University, 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]. How these underlying processes determine the success of immunotherapies has remained unclear. Here, we show that a fitness model for tumors based on immune interactions of neoantigens predicts response to immunotherapy. Two factors determine a neoantigen's fitness cost. First, the cost depends on its presentation by the major histocompatibility complex (MHC), estimated as a function of that neoantigen's relative MHC binding affinity. Second, it depends on T-cell recognition of a neoantigen, which is modeled as a non-linear function of its sequence similarity to known antigens. To describe the evolution of a heterogeneous tumor, we evaluate its fitness as a weighted average over 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].

Recent 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 disinhibiting the immune system[1]. However, only a minority of patients achieves a durable clinical benefit, suggesting there may be 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. An elevated number of mutations or neoantigens has been linked to improved response to checkpoint blockade therapy in multiple malignancies[4-6]. Hence, inferred neoantigen burden is a coarse-grained proxy for whether a tumor is likely to respond to therapy. Other implicated biomarkers of response include T-cell receptor (TCR) repertoire profiles[10], assays of checkpoint status[11,12], immune based microenvironment signatures[4,13], and tumor heterogeneity[14]. Despite high overall mutational load, a heterogeneous tumor may have immunogenic neoantigens present only in certain subclones. As a result, therapies targeting only a fraction of the tumor could disrupt clonal competitive balance and inadvertently stimulate growth of untargeted clones[16,17]. Moreover, mass spectrometry-based validation of neoantigens, already limited by sensitivity, does not sample all of the many relevant clones in heterogeneous tumors nor account for clonal variations across metastases[15]. A mathematical model using 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.

We propose a fitness model of immune interactions to describe the evolutionary dynamics of cancer cell populations under checkpoint-blockade immunotherapy (Fig. 1). Fitness models of this kind have been successful in capturing immune interactions for human influenza[7], HIV[8] and chronic viral infections[9], and we aim to introduce this approach to the study of immunotherapy. Checkpoint blockade exposes cancer cells to strong immune pressure on their neoantigens and thereby reduces their reproductive success. Our fitness model, which is detailed below, predicts the evolution of a cancer cell population under such selection pressure. Specifically, we compute $n(\tau)$, the predicted 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),

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

where F_{α} is the fitness and X_{α} is the initial frequency of clone α and τ is a characteristic evolutionary time scale (Methods). Our effective size estimates the number of cancer cells required to generate the observed population diversity and is not an estimate of the physical tumor size. Patients with less immunologically fit tumors will have more significant size reductions and, we assume, improved survival prognosis, which is it what we aim to predict. 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[18]. Unlike in previous approaches[14], here we learn the ancestral dependencies between clones. These determine the mutations and neoantigens that are 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 that determine immunogenicity of a neoantigen: an amplitude determined by MHC-presentation, A, and the probability of TCR-recognition, R (defined below). We call the product of these two factors, $A \times R$, the *cross-reactivity load* of the neoantigen. Next, we quantify total fitness for cancer cells in a tumor clone by aggregating over the fitness effects due to its neoantigens (Fig 1b, Methods). Specifically, we model the fitness of a given clone α by the cross-reactivity load of the immunodominant neoantigen,

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

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 dissociation constants for both mutant and wildtype peptides for a patient's HLA type[18] (SI).

We quantify the MHC-presentation factor for a neoantigen using the relative MHC affinity between the wildtype and the mutant peptide. This ratio, which was used to analyze computational neoantigen predictions[20], defines our amplitude A (Methods). We show that, unlike considering the mutant or wildtype affinity value alone, the ratio has consistent predictive value within our model (Extended Data Table 1). The interpretation of this model component is consistent with the competitive advantage gained by a neoantigen due to increased concentration, and a neoantigen being less likely to have immune tolerance due to presentation of its closest self-peptide (see discussion in Methods).

For TCR-recognition, we model cross-reactivity of neoantigens with positive, class I restricted T-cell antigens from the Immune Epitope Database[21] (IEDB). This approach does not assume preexisting immunity due to this set of epitopes. Rather, we posit that neoantigens predicted to be more cross-reactive with a member of this set are more "non-self" and, therefore, more likely to be immunogenic. As cross-reactivity is caused by physical binding of a TCR and a neoantigen, we use an established thermodynamic model to estimate this binding probability from sequence[22]. 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 the TCR recognizing epitope $\mathbf{e}$, proportionally to the alignment mismatch cost. Importantly, the probability a TCR binds a neoantigen is given by a nonlinear logistic dependence on sequence alignment score (Fig. 2). A similar nonlinear dependence on sequence similarity was previously used to estimate cross-immunity between influenza strains: strains with homologous epitope regions are likely to be antigenically similar[7]. Our model does not require full 9-amino acid identity of the neoantigen and epitope sequences for recognition. The total TCR-recognition probability, R, is defined as the probability that neoantigen $\mathbf{s}$ is recognized by at least one TCR corresponding to an IEDB epitope (Methods).

We apply the 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]. We assess our predictions with available patient survival data: total survival times of patients in the melanoma cohorts and progression free survival data on the lung cohort. Neoantigen amino-acid anchor positions, 2 and 9, are constrained due to their molecular function and display a hydrophobic bias, which is also reflected by non-informative MHC affinity amplitudes (Extended Data Fig. 1a). Hence, neoantigens with mutations in these positions are excluded from predictions with our model. Amino-acid diversity in remaining positions is unconstrained (Extended Data Fig. 1b)[23]. 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 the therapy. This is the time at which, following equation (1), samples are predicted to have a resolved heterogeneity, with their highest fitness clone dominating the evolutionary dynamics. We show that we are able to choose a consistent value of τ across datasets and that predictions are stable in a broad interval around it (Methods and Extended Data 2). Two model parameters are optimized: the midpoint and the steepness of the logistic binding function, which describes the probability of binding between neoantigens and epitope-specific TCRs (Methods). We maximize the survival log-rank test score to fit the binding curve parameters to the data on the largest dataset, Van Allen et al.[5] (103 metastatic patients). The parameter choice is confirmed to give high log-rank test scores also in the two smaller datasets from Snyder et al., and Rizvi et al., (64 and 34 patients respectively) (Fig. 2 and Extended Data Fig. 3). When using these logistic function parameters in all three datasets, the binding probability of 0.5 is obtained by alignments of average length of 6.55 amino acids; for almost certain binding of probability above 0.95 the average alignment length is 6.98 amino acids.

The predicted evolutionary dynamics of tumors separates long- and short-term survivors in our datasets (Fig. 3). Long-term survivors (patients with survival time longer than two years in the Van Allen et al. and Snyder et al. datasets, and one year of progression free survival in Rizvi et al. dataset) are predicted to have faster decreasing relative population sizes $n(\tau)$. Moreover, our fitness model results in highly significant separation of patients in survival analysis of all three datasets (Fig. 4). We use the median value of $n(\tau)$ to separate patients into high and low predicted response groups. Using the median as opposed to an optimized threshold[4,5,14] prevents overfitting and allows for robust validation. Log-rank test *p*-values are *p*=0.001 for the Van Allen et al., *p*=0.011 for Snyder et al., and *p*=7.8e-5 for Rizvi et al. For comparison, a model considering only total neoantigen burden is significant only for Rizvi et al. (*p*=0.007), when also using unsupervised median partitioning of patients. We also use an alternative neoantigen load model that accounts for clonal structure (Methods). Again, only the Rizvi et al. cohort has a significant patient survival separation (*p*=0.0009, Extended Data Table 1).

The success of our model strongly depends on the joint contribution of two fitness components, the MHC presentation amplitude and TCR recognition probability in equation (2). We deconstruct the model by removing each of the components one at a time (Fig 4, bottom panels and Extended Data Table 1). The MHC-only model, achieved by fixing $R_i = 1$, results in consistently worse segregation of patients (not significant in Snyder et al., decreased significance in Van Allen, et. al, and Rizvi, et. al, *p*=0.027 and *p*=0.004 respectively). The TCR-recognition-only model, achieved by fixing $A_i = 1$, does not result in a significant segregation in any cohort. It is important to assess the clonal structure of a tumor when trying to identify dominant neoantigens. We compare the performance of the full model to one assuming homogenous, single-clone tumor structure, with all neoantigens at tumor frequency = 1 (Methods). This model does not segregate patients significantly in Van Allen et al., and performs worse in Rizvi et al. (*p* =0.019). In Snyder et al., the homogenous structure model gives slightly better separation than the full model (*p*=0.008); however, the score difference between the two is within error bars of the original model.

In a broader context, our model suggests strong similarities in the evolution of cancers and fast-evolving pathogens. In both systems, immune interactions govern the dynamics of a genetically heterogeneous population; fitness models can predict these dynamics over limited periods, as recently shown for seasonal human influenza[7]. Yet there are important differences between the immune interactions of these systems. Influenza evolution is determined by antigenic similarity with previous strains in the same lineage. Cancer cells originate from normal cells and acquire mutations in a large set of proteins. Hence, their immune interactions are distributed in a larger antigenic space. The fitness effects of these interactions have a specific interpretation: they capture neoantigen “non-selfness”; that is, they formalize aspects of what makes a tumor immunologically different from its host[24]. Thus, our fitness model quantifies the presence of non-self peptides in cancers, which offers insight into adaptive immunity analogous to that for innate recognition of non-self nucleic acids[25].

Our approach can be naturally extended to other fitness effects, such as positive selection due to acquisition of driver mutations, the impact of other components in the tumor microenvironment or the hypothesized role of the microbiome[26,27,28]. Modeling evolutionary dynamics of a cancer cell population can also be useful in studies of acquired resistance to therapy, which is a more distant response effect. The proposed fitness model is based on biophysical interactions underlying the presentation of neoantigens and their immune cross-reactivity. Therefore, besides its predictive function, it may also inform the choice of therapeutic targets for tumor vaccine design. Moreover, this insight may be crucial for understanding when cross-reactivity with self-peptides may result in side effects[29,30].

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. Tumeh, P.C., et al. PD-1 blockade induces responses by inhibiting adaptive immune resistance. *Nature* **515**, 568–571 (2014).
11. 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).
12. Herbst, R.S., et al. Predictive correlates of response to the anti-PD-L1 antibody MPDL3280A in cancer patients. *Nature,* **515**, 563-567 (2014).
13. de Henau, O. et al. Overcoming resistance to checkpoint blockade therapy by targeting PI3Kγ in myeloid cells. *Nature* **539**, 443–447 (2016).
14. McGranahan, N. et al. Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. *Science* **351**, 1463-1469 (2016).
15. 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.
16. Fisher, A., Vazquez-Garcia, I., Mustonen V. The value of monitoring to control evolving populations. *Proc. Natl. Acad. Sci.* **112(4)**, 1007-1012 (2015)
17. Anagnostu, V. et al. Evolution of neoantigen landscape during immune checkpoint blockade in non-small cell lung cancer. Cancer Discov. ***in press*** (2016).
18. Deshwar, A. G. *et al.* PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biol.* **16**, 35 (2015).
19. Andreatta, M. & Nielsen, M□. Gapped sequence alignment using artificial neural networks: application to the MHC class I system. *Bioinformatics* **32**, 511-517 (2016).
20. Hundal, J. et al. pVAC-Seq: A genome-guided in silico approach to identifying tumor neoantigens. *Genome* Med. **8**, 1-11 (2016).
21. Vita, R. et al. The immune epitope database (IEDB) 3.0. *Nucleic Acids Res.* **43**, D405-D412 (2014).
22. Berg, O.G. and von Hippel, P.H. Selection of DNA binding sites by regulatory proteins: Statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol.* **193**, 723-743 (1987).
23. Lehmann, J., Libchaber, A., & Greenbaum, B.D. Fundamental amino acid mass distributions and entropy costs in proteomes. *J. Theor. Biol.* **410**, 119-124 (2016).
24. Old, L.J. & Boyse, E.A. Immunology of experimental tumors. *Ann. Rev. Med.* **15**, 167-186 (1964).
25. Tanne, A et al. Distinguishing the immunostimulatory properties of noncoding RNAs expressed in cancer cells. *Proc. Natl. Acad. Sci. USA* **112**, 5154-15159 (2015).
26. Vétizou, M. *et al.* Anticancer immunotherapy by CTLA-4 blockade relies on the gut microbiota. *Science* **350**, 1079–1084 (2015).
27. Zitvogel, L., Ayyoub, M., Routy, B. & Kroemer, G. Microbiome and Anticancer Immunosurveillance. *Cell* **165**, 276–287 (2016).
28. Dubin, K., et al. Intestinal microbiome analyses identify melanoma patients at risk for checkpoint-blockade-induced colitis. *Nat. Commun.*, **7,** 10391 (2016)

29. Johnson D.B. et al. Fulminant myocarditis with combination immune checkpoint blockade. *New Engl. J. Med.* **375**, 1749-1755 (2016).

30. 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

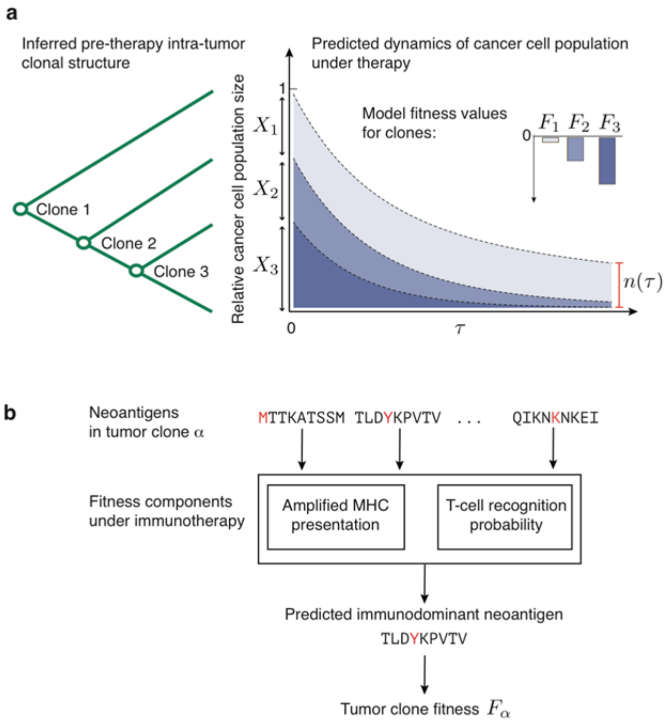

a
Inferred pre-therapy intra-tumor clonal structure
Clone 1
Clone 2
Clone 3
Predicted dynamics of cancer cell population under therapy
Relative cancer cell population size
1
X_1
X_2
X_3
Model fitness values for clones:
0
F_1 F_2 F_3
$n(\tau)$
0
τ
b
Neoantigens in tumor clone α
MTTKATSSM TLDYKPVTV ... QIKNKNKEI
Fitness components under immunotherapy
Amplified MHC presentation
T-cell recognition probability
Predicted immunodominant neoantigen
TLDYKPVTV
Tumor clone fitness F_α

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.

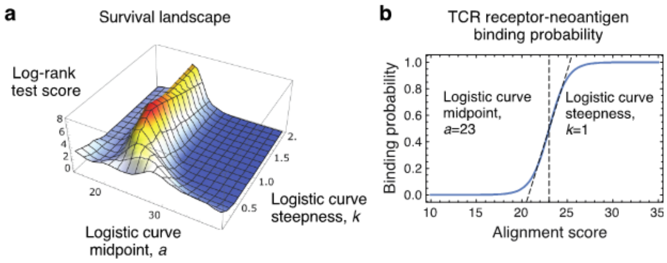


Figure 2 | Survival landscape as a function of TCR binding model. **a**, The landscape is a contour plot of log-rank test scores in survival analysis with patient data split by median relative population size (equation (1)). The locally smoothed landscape is plotted for the Van Allen et al. dataset as a function of the model parameters for the logistic curve midpoint (a) and steepness (k) (Methods). **b**, Logistic binding curve at inferred midpoint and steepness parameters used across all three datasets from parameters in Van Allen et al. The curve represents the binding probability of a neoantigen and a T-cell receptor specific to a given IEDB epitope as a function of alignment score between the neoantigen's peptide sequence and the epitope.

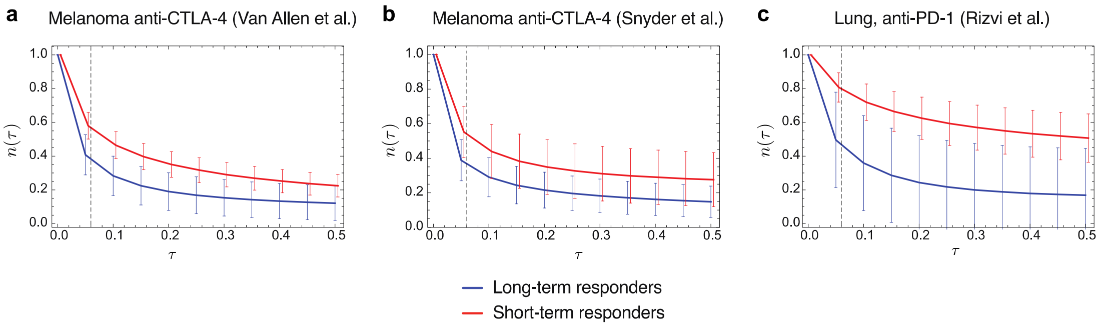


Figure 3 | Evolutionary dynamics predictions in patient cohorts. a, Relative population size predictions for long-term and short tem survivors across the **a,** Van Allen et al.; **b,** Snyder et al.; and **c,** Rizvi et al. cohorts. Long-term survivors are defined in the text. Error bars are 95% confidence intervals around the population average. The dashed line indicates the consistent choice of $\tau = 0.06$ used across all three datasets for patient survival predictions (Methods and Extended Data Figure 3).

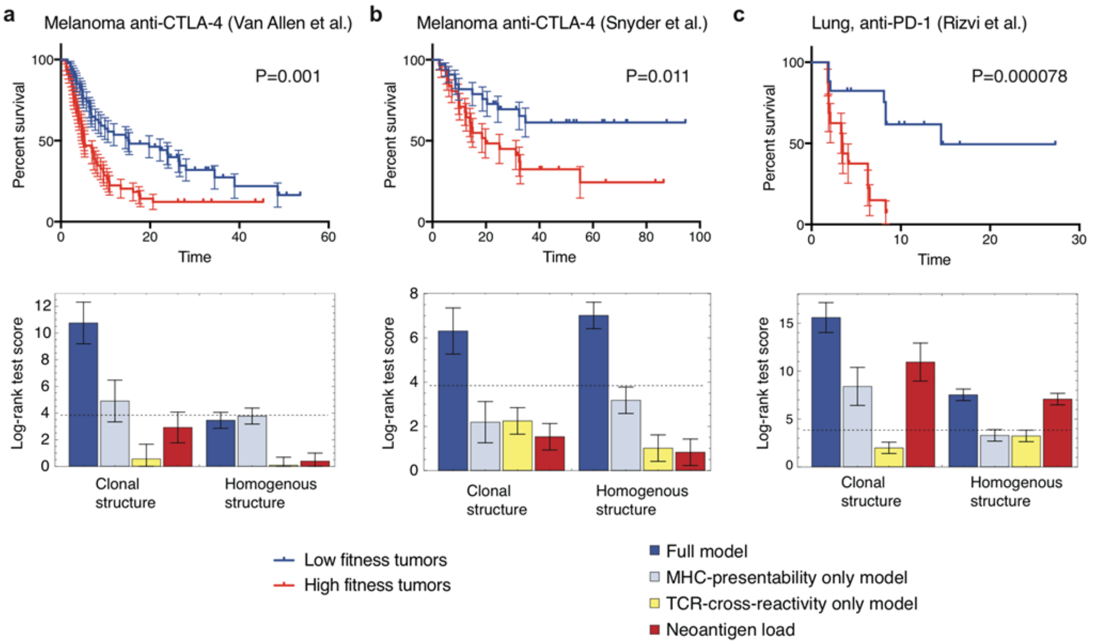
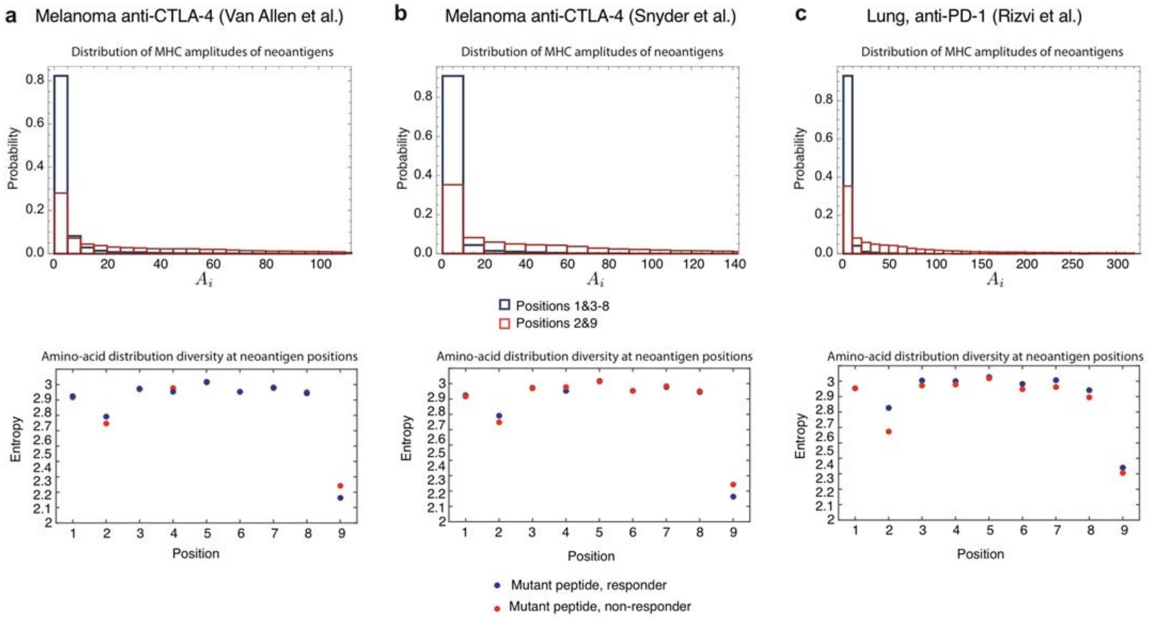
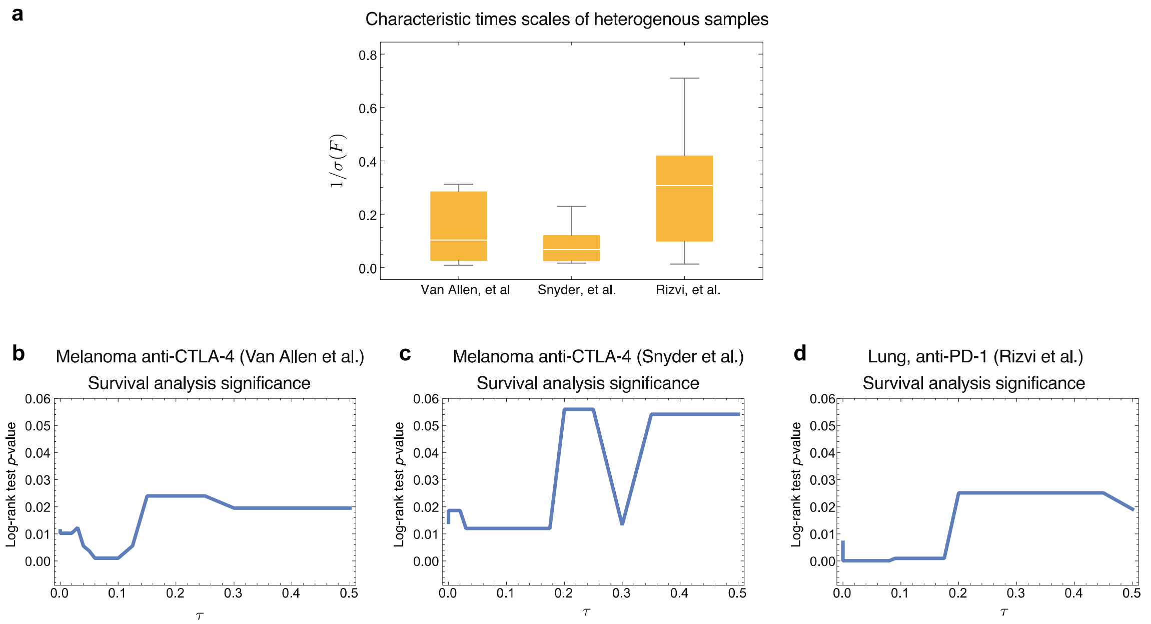


Figure 4 | Neoantigen fitness model is predictive of patient survival after checkpoint blockade immunotherapy. a,**b**, Kaplan-Meier survival curves are calculated across two melanoma patient datasets with patient survival data, which were treated with anti-CTLA-4 antibodies[4,5] and **c,** one dataset of lung patients with progression free survival data, which were treated with anti-PD-1 antibodies[6]. The samples are split in an unsupervised manner by the median value of their tumor's relative population size $n(\tau)$ defined in equation (1); the error bars represent the standard error. For comparison we show the log-rank test score for models, which account for removal of one feature of our model (bottom panels, higher score values indicate better patient segregation): an MHC-presentability only model (light blue) and a TCR-recognition only model (yellow). We compare their values with a tumors' neoantigen burden (red). All models are computed both over a tumor's clonal structure (clonal, left) and without taking heterogeneity into account (homogenous, right). Dashed lines on the bottom panels marks the score value corresponding to the significance threshold of 5%, scores above that threshold raise significant patient segregation. The error bars are the standard deviation of log-rank test score acquired from the survival analysis with one sample removed from the cohort at a time.

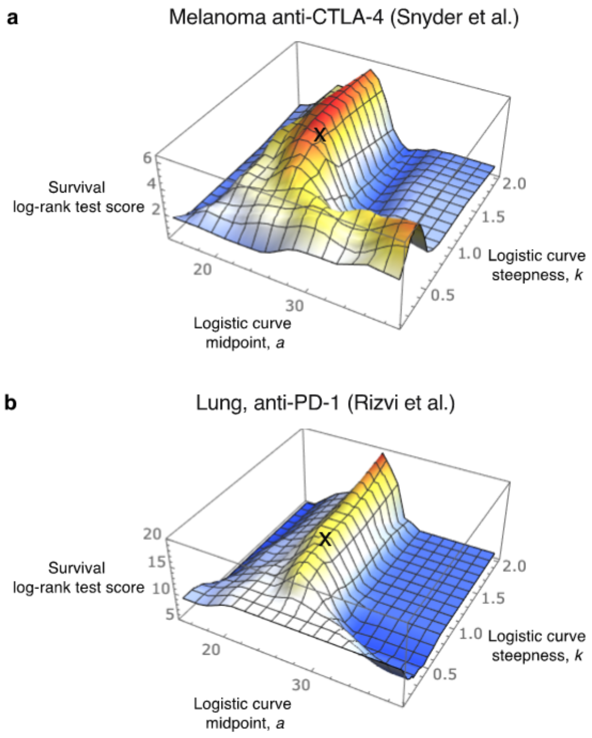
Extended Data Figures



Extended Data Figure 1 | Positions 2 and 9 in neoantigens are of less predictive value. a, Neoantigens with mutations at anchor residues at position 2 and 9 have highly diverging amplitude values and are of less overall predictive value than neoantigens at other positions. **b,** Patients classified in studies as responders are marked in blue and non-responders are marked in red. Positions 2 and 9 are highly constrained by a bias to be hydrophobic. Their Shannon entropy is lower than that of other residues, across all three datasets regardless of classification of their neoantigens in those datasets. Other residue sites have the same entropy as the overall proteome[22] and are therefore unconstrained.



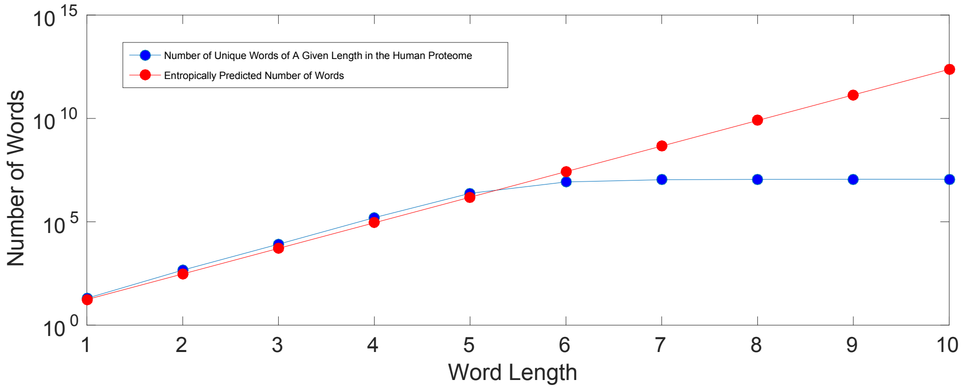
Extended Data Figure 2 | Consistency of evolutionary time-scale across datasets. a, The distribution of characteristic times scales of samples with clonal fitness heterogeneity (Methods) for the three patient cohorts. These distributions consistently define the interval for relevant time scales of τ, in all datasets we subsequently investigate $\tau \in [0{,}0.5]$. **b-d,** Significance of survival analysis reported as the result of the log-rank test on the three datasets with sample split at a median value $n(\tau)$ plotted as a function of τ. The chosen value of parameter τ =0.06 and a broad surrounding interval gives highly significant sample segregation in each of the datasets.



a Melanoma anti-CTLA-4 (Snyder et al.)

b Lung, anti-PD-1 (Rizvi et al.)

Extended Data Figure 3 | Survival landscape for Snyder et al., and Rizvi et al., cohorts. The survival landscape is defined by the log-rank test score as a function of the model parameters for the logistic curve shape, i.e the midpoint (a) and steepness (k) (Methods). The locally smoothed landscape is plotted for the **a,** Snyder et al., and **b,** Rizvi et al., datasets. An X marks the optimal parameters from Van Allen et al., a =23 and k =1 (cf. Fig. 2), which are used to derive survival curves for these two datasets and are at high score regions of the landscapes.



Extended Data Figure 4 | Word usage in the proteome is exhausted between 5 and 6 letter words. Given the entropy of the genome from Ref. 23, 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 MHC-presentability only model (Methods, equation (13)) a nd a TCR-recognition only model (Methods, equation (14)); absolute MHC-amplitude model and absolute wildtype MHC-amplitude model (Methods, equations (15) and (16) respectively); simple neoantigen load model and mutational load model (Methods, equations (17) and (18)); and finally an additive neoantigen fitness model (Methods, equation (19)), which summates fitness contributions of neoantigens in a clone as opposed to maximizing them as in our original model. **(2)** Above models evaluated without accounting for clonal structure structure of tumors. For each model we report the following parameters (if applicable): the aggregating function for neoantigen effects within a clone (MAX or SUM replacing Ag in equation (11)), the value of parameter τ used in predictions, the parameters of the logistic function a and k (Methods, equation (7)). Finally, we report the predictive value of the models as the log-rank test *p*-value and the corresponding log-rank test score. The comparison is shown on all three immunotherapy datasets.

Acknowledgments

We thank Curt Callan, Simona Cocco, Yuval Elhanati, Michael Lässig, 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, David T. Ting, Matthew Hellman and Alexandra Snyder-Charen for their critical reading of the manuscript. This work was supported by Stand Up to Cancer (M.Ł., A.S., J.D.W, B.D.G, T.A.C.), 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 Phil A. Sharp Innovation in Collaboration Award from Stand up to Cancer (B.D.G, J.D.W.), the Swim Across America, Ludwig Institute for Cancer Research, Parker Institute for Cancer Immunotherapy and the National Cancer Institute (NCI) K12 Paul Calabresi Career Development Award for Clinical Oncology K12CA184746-01A1 (V.P.B.). 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.

$$\frac{dN_\alpha}{d\tau} = F_\alpha N_\alpha \quad (3)$$

where N_α is the population size of clone α. 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 hence 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

$$N(\tau) = \sum\nolimits_\alpha N_\alpha(\tau) = \sum\nolimits_\alpha N_\alpha(0)\exp(F_\alpha \tau). \quad (4)$$

The absolute size $N(\tau)$ is meant as the effective population size, the number of cells estimated to have generated the observed clonal diversity; it is not to be understood as the physical tumor size. As our diagnostic of survival we use the relative effective population size $n(\tau) = N(\tau)/N(0)$, which compares the predicted evolved population size after a characteristic time scale of evolution τ (discussed below) to the initial pretreatment effective size $N(0)$. We denote the initial clone α frequency $X_\alpha = N_\alpha(0)/N(0)$, these frequencies are inferred from bulk exome reads from a tumor sample[17]. 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.

Clonal structure of a tumor and clone frequencies. Tumor clones are reconstructed using the PhyloWGS software package[18] (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 10 trees with the highest likelihood score.

2. Fitness model

MHC-amplitude. The amplitude due to the dissociation constant between a neoantigen and its wildtype peptide is defined as

$$A = K_D^{WT}/K_D^{MT}. \quad (5)$$

The dissociation constants are inferred for each peptide sequence and patient HLA type[19]; all mutant peptide sequences considered as neaontigens meet the standard cutoff, ${K_D}^{MT} < 500\,\mathrm{nM}$ (SI). The amplitude in this form has a high predictive value for patient survival predictions (discussed in section 4., demonstrated in Fig. 4 and Extended Table 1), consistently over the three patient cohorts, which is not the case of either the mutant or wildtype dissociation constants on their own.

We offer two interpretations of why this amplitude is relevant, which are not mutually exclusive of one another. The first is that the amplitude can be thought of as an approximate form derived with the use of simple equilibrium kinetics, where the concentration of peptide bound to MHC is given by their individual concentrations and inferred binding constant K_D, derived from NetMHC[19]. The underlying dependencies are

$$A = \frac{[\mathrm{MHC{:}\,neoantigen}]^{MT}}{[\mathrm{MHC{:}\,neoantigen}]^{WT}} = \frac{K_D^{WT}[\mathrm{MHC}]^{MT}[\mathrm{neoantigen}]^{MT}}{K_D^{MT}[\mathrm{MHC}]^{WT}[\mathrm{neoantigen}]^{WT}}, \tag{6}$$

where $[\mathrm{MHC{:}\,neoantigen}]^{MT}$ is the concentration of the mutant form of the neoantigen to MHC, with the WT superscript representing the same quantity for the wild-type peptide. This interpretation assumes the above quantity is dominated by the ratio of dissociation constants, which derives the formula for A in equation (5). In this sense, the amplitude reflects the relative concentration of mutant to wildtype peptide and therefore the likelihood that the mutant peptide would be presented versus its wildtype peptide. As such it may reflect the competitive advantage a neoantigen has acquired in terms of presentation through mutation, as posited in other in silico analyses[31].

The second interpretation is that the amplitude reflects the likelihood a neoantigen is similar to a peptide that has undergone immune tolerance. As we exclude neoantigens with mutations on positions 2 and 9, a high value of amplitude means the wildtype peptide is also likely to have hydrophobic residues at the anchor position and hence can be presented by the MHC. Since neoantigens differ from their wildtype peptides by a single mutation, and given the uniqueness of nonamer sequences in the proteome (Extended Data Fig. 4), the self-nonamer in the genome with the greatest similarity to a neoantigen is likely to be its wildtype peptide. We verified that this is the case for 92% of all neoantigens, with the remainder largely emanating from gene families with many paralogs (SI). Therefore a high amplitude usually stands for the self peptide most similar to a neoantigen not being likely to have been abundantly presented by the MHC. Following this reasoning, the mutant peptide with high affinity is likely to be novel to T-Cells as its immunogenicity is not mitigated by a homologous self-peptide.

TCR-recognition. We model R, the cross-reactivity of a neoantigen with a TCR-pool defined as the probability that a neoantigen cross-reacts with at least one TCR corresponding to a known immunogenic epitope. We profile *in silico* the cross-reactivity of neoantigen with a set of epitopes given by the Immune Epitope

Database and Analysis Resource[21] (IEDB). We restrict ourselves to IEDB epitopes that are positively recognized by T-cells after class I MHC presentation. We hypothesize that a neoantigen that is predicted to cross-react with a TCR from this pool of immunogenic epitopes is a neoantigen more likely to be immunogenic itself.

The probability that a TCR for a given epitope binds a given neoantigen is defined by a simple two-state thermodynamic model with logistic shape. In this model we use sequence alignment as a proxy for binding energy[22]. 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[32]. For an alignment score, $|\mathbf{s},\mathbf{e}|$, we compute the binding probability as

$$\mathrm{Pr}_{\mathrm{binding}}(\mathbf{s},\mathbf{e}) = \frac{1}{1+\ e^{-k(|\mathbf{s},\mathbf{e}|-a)}}, \tag{7}$$

where a represents the horizontal displacement of the binding curve and k sets the steepness of the curve at a. These are two free parameters to be fit in our model (see below). The parameters that we use in predictions are a=23 and k=1; these parameters give binding probability $\mathrm{Pr}_{\mathrm{binding}}(\mathbf{s},\mathbf{e})$ =0.5 at alignment score $|\mathbf{s},\mathbf{e}|$=23; the probability drops to below 0.05 at $|\mathbf{s},\mathbf{e}|$=20 and reaches value of above 0.95 at $|\mathbf{s},\mathbf{e}|$=26 (Fig. 2b). The corresponding alignment score span of 6 is close to the average identity match score in the BLOSUM62 matrix (5.64). The average alignment length corresponding to score 26 is 6.98 amino acids in our datasets and it is 6.55 for binding probability 0.5. The logistic function is therefore a strongly nonlinear function of the alignment score, where a mismatch on 1-2 positions can decide about lack of binding between the neoantigen and the epitope specific TCR.

For a given neoantigen $\mathbf{s}$ we calculate the probability it is recognized by a TCR within a repertoire as the probability it cross-reacts with at least one IEDB epitope:

$$R\ =\ 1-\prod_{\mathbf{e}\in\mathrm{IEDB}}[1-\ \mathrm{Pr}_{\mathrm{binding}}(\mathbf{s},\mathbf{e})] \tag{8}$$

Neoantigen-based fitness cost for a tumor clone. Our model associates each neoantigen with a fitness cost, the *cross-reactivity load*, defined as the product of the MHC-amplitude in equation (5) and TCR-recognition probability in equation (8), $A \times R$.

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 running over neoantigens in the clone. Therefore, the full form of the predicted relative cancer cell population size is given by

$$n(\tau) = \sum_{\alpha} X_{\alpha} \exp[-\max_{i\in \text{Clone } \alpha} (A_i \times R_i)\, \tau]. \tag{9}$$

One could use a more general model for fitness model of a clone,

$$F_{\alpha} = -\operatorname*{Ag}_{i\in \text{Clone } \alpha} (A_i \times R_i) \tag{10}$$

and use different function Ag to aggregate over cross-reactivity fitness effects of neoantigens within a clone, such as a summation over all neoantigens (Extended Data Table 1), summation over a fixed set, or other nonlinear dependency. Taking the best score within a clone is consistent with the notions of heterologous immunity and immunodominance – that a small set of antigens drive the immune response, whereas summing over neoantigens would imply a more uniform distribution of contributions.

3. Model parameters

Logistic binding function parameter optimization. To choose model parameters a and k in equation (7) we investigate the log-rank-test scores of patient segregation as a function of these parameters. The survival analysis is performed by splitting patient cohort into *high* and *low fitness* groups by the median cohort value of $n(\tau)$, the predicted relative cancer cell population size at a characteristic time τ (we discuss the choice of τ below). The survival score landscapes (Fig. 2 and Extended Data Fig. 3) appear to be consistent between the datasets, with an optimal value of parameter a around 23 and parameter k living on a trivial axis above value 1, suggesting strong nonlinear fitness dependence on the sequence alignment score. We choose parameters that optimize the log-rank-test score in the largest dataset in our study, the melanoma anti-CTLA4 cohort from Van Allen, et al[5].

Characteristic time scale parameter estimation. In the survival analysis the samples are split by the median cohort value $n(\tau)$ at a specified time scale τ. Intuitively, this time should be set to a finite value at which the tumors are expected to have responded to therapy. At this value of τ the clonal heterogeneity of tumors is supposed to have decreased, with the highest fitness clone dominating in the population. For one tumor this time scale is inversely proportional to the standard deviation of intra-tumor fitness (i.e. of the order of $1/\sigma(F)$), where

$$\sigma^2(F) = \sum_{\alpha} X_{\alpha}\, F_{\alpha}^2 - \left(\sum_{\alpha} X_{\alpha} F_{\alpha}\right)^2. \tag{11}$$

In each cohort we determined the interval of characteristic times of heterogenous samples (Extended Data Fig. 2a) and we tested the dependence of prediction power on τ by performing log-rank test (Extended Fig. 2b-d). The optimal values of τ in each cohort belong to a relatively wide interval. The consistent broadness of these intervals suggests low sensitivity of predictive power on τ. Moreover, the parameter intervals giving highly significant patient segregation are also consistent between the cohorts. We choose $\tau = 0.06$ for our predictions in all datasets. As τ is an inverse fitness it also defines a typical maximum cross-reactivity load in a clone beyond which one would expect to have a clone that responded to therapy. For instance, at $\tau = 0.06$ this typical fitness value would be about 16.67. This would indicate that a neoantigen with a TCR recognition probability $R = 1$ would on average lead to a productive response if the ratio of its dissociation constants would be greater 16.67. Well beyond that value amplitudes would essentially carry the same predictive value.

Heterogenous samples were selected with criterion $e^{H_F} \geq 2,$ where H_F is clonal fitness entropy defined as

$$H_F = -\sum_{\beta} Y_\beta \log Y_\beta, \tag{12}$$

where the frequencies of clones with the same fitness are added together and denoted as Y_β. The index β then refers to all clones with a given fitness.

4. Alternative fitness models

We compare our full model in equation (9) to the following alternative models (Extended
Data Table 1):

1. **Heterogenous structure models**

   a. **MHC-presentability only model:**
   In this model the recognition factor is ignored and fitness is assumed to be determined only by MHC-amplitude of neoantigens. The defining equation is given by

$$n(\tau) = \sum_{\alpha} X_\alpha \, \text{exp}[-\max_{i \in \text{Clone } \alpha} A_i \, \tau]. \tag{13}$$

   b. **TCR-recognition only model:**
   Conversely, in this model the MHC-presentation factor is ignored and fitness is assumed to be determined only by TCR-recognition of neoantigens. The defining equation is given by

$$n(\tau) = \sum_{\alpha} X_\alpha \, \text{exp}[-\max_{i \in \text{Clone } \alpha} R_i \, \tau]. \tag{14}$$

c. Absolute MHC-amplitude model

In this model the likelihood of MHC presentation for a neoantigen is inversely correlated with its inferred dissociation constant, $A^{\text{abs}} = 1/K_D^{MT}$(cf. equation (5)). The model is defined as

$$n(\tau) = \sum_{\alpha} X_\alpha \, \text{exp}[- \max_{i \in \text{Clone } \alpha} \left(A_i^{\text{abs}} \times R_i\right) \tau]. \tag{15}$$

d. Absolute MHC-amplitude model

In this model the likelihood of MHC presentation for a neoantigen is inversely correlated with its inferred dissociation constant, $A^{\text{abs,WT}} = K_D^{WT}$(cf. equation (5)). The model is defined as

$$n(\tau) = \sum_{\alpha} X_\alpha \, \text{exp}[- \max_{i \in \text{Clone } \alpha} \left(A_i^{\text{abs,WT}} \times R_i\right) \tau]. \tag{16}$$

e. Neoantigen load model

This model assigns uniform fitness cost to each neoantigen. For L_α, the number of neoantigens in clone α, this model is defined by

$$n(\tau) = \sum_{\alpha} X_\alpha \, \exp[- L_\alpha \tau]. \tag{17}$$

We do not exclude neoantigens with mutations on positions 2 and 9 in the neoantigen load model.

f. Mutational load model

This model assigns uniform fitness cost to each somatic mutations. For, L_α^M, the number of somatic mutations (with respect to a normal cell) in clone α, this model is defined by

$$n(\tau) = \sum_{\alpha} X_\alpha \, \exp[- L_\alpha^M \tau]. \tag{18}$$

g. Additive neoantigen fitness model

This model implements an additive neoantigen aggregating function, namely

$$n(\tau) = \sum_{\alpha} X_\alpha \, \exp\left[-\left(\sum\nolimits_{i \in \text{Clone } \alpha} A_i \times R_i\right) \tau\right]. \tag{19}$$

2. Homogenous structure models

For each model defined in point (1) we can define its homogenous structure equivalent, which assumes tumor is strictly clonal with all neoantigens in the same clone at frequency 1.

We assess the predictive power of these models with a survival analysis, by separating patients by the median value of $n(\tau)$ in each patient cohort and computing the log-rank test for such segregation. For stringency of comparisons, we adjust the value of parameter τ in a supervised manner to optimize the performance of each alternative model (Extended Data Table 1).

5. Data availability

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

References

31. Khalili, J.S., Hanson, R.W., & Szallasi, Z. In silico prediction of tumor antigens derived from functional missense mutations of the cancer gene census. *Oncoimmunology* **1**,1281-1289 (2012).
32. Henikoff, S. & Henikoff, J.G. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA* **89**, 10915–10919 (1992).

Supplementary Information

Computational identification of neoantigens

Neoantigens from the three datasets were inferred using a consistent pipeline established at Memorial Sloan Kettering Cancer Center. Raw sequence data reads were aligned to the reference human genome (hg19) using the Burrows-Wheeler Alignment tool. Base-quality score recalibration, and duplicate-read removal were performed, with exclusion of germline variants, annotation of mutations, and indels as previously described[4]. Local realignment and quality score recalibration were conducted using the Genome Analysis Toolkit (GATK) according to GATK best practices[33,34]. For sequence alignment and mutation identification, the FASTQ files were processed to remove any adapter sequences at the end of the reads using cutadapt (v1.6)[35]. The files were then mapped using the BWA mapper (bwa mem v0.7.12[36], the SAM files sorted, and read group tags added using the PICARD tools. After sorting in coordinate order, the BAM's were processed with PICARD MarkDuplicates. First realignment was carried out using the InDel realigner followed by base quality value recalibration with the Base-QRecalibrator.

A combination of 4 different mutation callers (Mutect 1.1.4, Somatic Sniper 1.0.4, Varscan 2.3.7, and Strelka 1.013) were used to identify single nucleotide variants (SNVs)[37-39]. As previously described, SNVs with an allele read count of less than 4 or with corresponding normal coverage of less than 7 reads were filtered out[40].

The assignment of a somatic mutation to a neoantigen was estimated using a previously described bioinformatics tool called NASeek[4]. Briefly, NASeek is a computational algorithm that first translates all mutations in exomes to strings of 17 amino acids, for both the wild type and mutated sequences, with the amino acid resulting from the mu-tation centrally situated. Secondly, it evaluates putative MHC Class I binding for both wild type and mutant nonamers using a sliding window method using NetMHC3.4[19] (http://www.cbs.dtu.dk/services/NetMHC-3.4/) for patient-specific HLA types, to gene-rate predicted binding affinities for both peptides. NASeek finally assesses for similarity between nonamers that predicted to be presented by patient-specific MHC Class I. All nonamers with binding scores (i.e. the inferred dissociation constants ${K_D}^{MT}$) below 500 nM are defined as neoantigens.

Clonal tree reconstruction with PhylowWGS

Tumor clones are reconstructed using the PhyloWGS software package[18]. The input data for the algorithm is extracted from exome sequencing data: (1) mutation reads obtained with the pipeline described above, and (2) allele-specific copy-number variant data, obtained with FACETS v0.5.0[41]. Briefly, the package clusters mutations into clones by the frequency of their reads and it infers possible nesting of clones (ancestral relations) between pairs of clones.

Intuitively, an ancestral clone needs to have higher frequency then its derived clone. From this information PhyloWGS reconstructs high likelihood tumor geneological trees.

Amino acid diversity

We define the amino acid diversity at i-th position in a neoantigen as e^{H_i}, where H_i is Shannon entropy[42] of amino acid usage at this position, i.e.

$$H_i = -\sum_{j=1}^{20} f(a_{ij}) \log(f(a_{ij})),$$

where $f(a_{ij})$ is frequency of the i-th position in all neoantigens in a group. Inferred neoantigens are nonamers, so i ranges in value from 1 to 9. The diversity of neoantigens at a given site were compared to the values found in the human proteome in Lehman, et al.[23].

To calculate the expected number of words in the proteome we utilize the frequency of amino acids from Lehman, et al. We compute the entropy associated with the frequency of amino acids in the human genome:

$$H(a) = -\sum_{j=1}^{20} f(a_j) \log(f(a_j)),$$

where $f(a_j)$ is the frequency of the j-th amino acid in the human genome. The expected number of words of length n is therefore $e^{nH(a)}$. This value is compared to the observed number of words of length n in the reference proteome for GRCh38.p7

Identification of closest nonamers in human proteome to neoantigens

We have mapped the WT and MT 9-mer peptides to all proteins in the current human reference genome (GRCh38.p7) with at least 8 out of 9 matches and no gaps (allowing only mismatches). For this we used LAST[43] (version 819) with the following parameters:

```
lastal -f BlastTab -j1 -r2 -q1 -e15 -y2 -m100000000 -l4 -L4 -P0
```

(9-mer mapping with at most one mismatch is guaranteed to have a matching 4-mer word).

One expects the mutated peptide to only map to the same location as the WT peptide, WT mapping exactly (9 matches) and MT mapping with one mismatch (8 matches). The expected case is that the WT peptide maps to the proteome exactly and the MT peptide maps to the proteome with one mismatch and only to the loci WT peptide maps to.

This rule can be violated in the following cases, sorted from the most to the least severe:

1. WT peptide does not map to the proteome exactly. Some possible reasons are: a difference in the reference assemblies used for mutation calling and peptide mapping, a germline mutation mistakenly identified as somatic, or a difference between the pa-tient genome and the reference genome used for alignments.
2. WT peptide maps to the proteome exactly (9 matches), MT peptide maps to the pro-teome exactly (9 matches) but to a different locus.
3. WT peptide maps to the proteome exactly, MT peptide maps to the proteome with one mismatch; however, MT peptide maps with one mismatch to the subjects WT does not map exactly.
4. WT peptide maps to the proteome exactly, MT peptide maps to the proteome with one mismatch; however, MT peptide maps with one mismatch to a different locus on the gene WT maps to.

We have examined each peptide for the worst possible scenario. We have gone from category 1 to 4 in the list. Category 1 indicates a difference in the reference genome. Categories 2-4 typically are due to mutations that occur in repetitive gene families with many paralogs. Once we identified that a peptide belongs to any category, we excluded it from further considerations. This way the numbers of peptides in each category add up to the total number of peptides. Below is a summary for the different datasets utilized in this study:

Van Allen, et al.[5]:
39373 total peptides, (1) 42 WT unmapped, leaving 39331
36783 expected peptides (93.42%), (2) 387 have 9 matches in MT, (3) 2076 have other alignments, (4) 85 have other alignments to the same subject.

Snyder, et al[4].:
29781 total peptides, (1) 35 WT unmapped, leaving 29746
27674 expected peptides (92.93%), (2) 361 have 9 matches in MT, (3) 1644 have other alignments, (4) 67 have other alignments to the same subject.

Rizvi, et al.[6]:
5581 total peptides, (1) 6 WT unmapped, leaving 5575
5125 expected peptides (91.83%), (2) 105 have 9 matches in MT, (3) 323 have other alignments, (4) 22 have other alignments to the same subject.

Additional supplementary files for each dataset are included as Supplementary Data:

mt-with-9.tsv – list of peptides from category 2 and the subjects each one aligns to .

peptides-with-extra-aln.tsv – peptides from group 3 and the subjects each one aligns to.

peptides-multimapping-same-subj.tsv – peptides from group 4 and their alignments including the start and end coordinates

Additional References

33. DePristo, M., et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genet.* **43**, 491-498 (2011).
34. Van der Auwera, G.A., et al. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Curr. Prot. in Bioinformatics* **43**, 11.10.1-11.10.33 (2013).
35. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10-12 (2011).
36. Li, H., & Durbin R. Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics* **25**, 1754-1760 (2009).
37. Wei, L. et al. MAC: identifying and correcting annotation for multi-nucleotide variations. *BMC Genomics* **16**, 569 (2015).
38. Snyder, A. & Chan, T.A. Immunogenic peptide discovery in cancer genomes. *Curr Opin Genet Dev* **30**, 7-16 (2015).
39. Nielsen, M. et al. Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. *Protein Sci* **12**, 1007-1017 (2003).
40. Riaz, N. et al. Recurrent SERPINB3 and SERPINB4 mutations in patients who respond to anti-CTLA4 immunotherapy. *Nat. Genet.* **48**, 1327-1329 (2016).
41. Shen, R. & Seshan, V.E. FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. *Nucleic Acids Res.* **44**, e131 (2016).
42. Shannon, C.E. A mathematical theory of communication. *Bell Sys. Tech. J.* **27**, 379-423 (1948).
43. Kiełbasa, S.M., Wan, R., Sato, K., Horton, P., & Frith, M.C. Adaptive seeds tame genomic sequence comparison. *Genome Res.* **21**, 487-493 (2011).