# Evolutionary dynamics of neoantigens in growing tumors

Eszter Lakatos[1], Marc J. Williams[1], Ryan O. Schenck[2,3], William C. H. Cross[1], Jacob Househam[1], Luis Zapata[4], Benjamin Werner[4], Chandler Gatenbee[2], Mark Robertson-Tessi[2], Chris P. Barnes[5], Alexander R. A. Anderson[2], Andrea Sottoriva[4,*], Trevor A. Graham[1,*]

[1] Evolution and Cancer Laboratory, Centre for Genomics and Computational Biology, Barts Cancer Institute, School of Medicine and Dentistry, Queen Mary University of London, London, UK.
[2] Integrated Mathematical Oncology, Moffitt Cancer Center, Tampa, FL, USA.
[3] Wellcome Centre for Human Genetics, University of Oxford, Oxford, UK.
[4] Evolutionary Genomics & Modelling Lab, Centre for Evolution and Cancer, Institute of Cancer Research, London, UK.
[5] Department of Cell and Developmental Biology, University College London, London, UK.
* Correspondence: andrea.sottoriva@icr.ac.uk; t.graham@qmul.ac.uk

## ABSTRACT

**Cancers accumulate mutations that lead to neoantigens, novel peptides that elicit an immune response, and consequently undergo evolutionary selection. Here we establish how negative selection shapes the clonality of neoantigens in a growing cancer, by constructing a mathematical model of neoantigen evolution. The model predicts that, without immune escape, tumor neoantigens are either clonal or at low frequency, and hyper-mutated tumors can only establish following the evolution of immune escape. Moreover, the site frequency spectrum of somatic variants under negative selection appears more neutral as the strength of negative selection increases, consistent with classical neutral theory. These predictions are corroborated by the analysis of neoantigen frequencies and immune escape in exome and RNA sequencing data from 879 colon, stomach and endometrial cancers.**

## INTRODUCTION

Mutations accrue throughout tumor development and provide 'fuel for the fire' of cancer evolution. However, mutations can also hinder tumor evolution if they lead to an anti-tumor immune response, via the generation of *neoantigens*, novel peptides presented on the cell's surface and recognized as 'non-self' by cells of the adaptive immune system[1,2]. The immune system is a major determinant of tumor evolution, most starkly demonstrated by the prognostic value of immune-infiltration[3] and the success of immunotherapy[4,5].

The landscape of neoantigenic mutations is shaped by ecological and evolutionary interactions between a tumor and its microenvironment[1,6,7]. In the absence of an immune system, neoantigens accumulate as a 'side-effect' of mutation acquisition[8], and are expected to follow neutral evolutionary dynamics[9]. *Immuno-editing* refers to immune-cell killing of antigenic cells[1] and so represents a negative selective pressure[6]. Tumor cells can also experience positive selection upon the evolution of mechanisms to inhibit the immune system's ability to recognize or react to cancer-associated antigens. These are termed *immune escape* mechanisms[7,8,10]. Cancer evolution in response to immune control is a 'hallmark of cancer'[11] and it is well-recognized that the tumor-specific immune microenvironment shapes the neoantigenic repertoire found in tumors[12–14].

Therapies that (re)activate the immune response following escape have achieved exceptional success (reviewed in ref[15]), especially in cancers of high mutational load[16–18]. Neoantigen-profiling is predictive of treatment response[19] and long-term survival[20]. However, a significant number of patients do not respond to immunotherapy regardless of a high mutational load and the presence of molecular markers of immune escape[21], and there is a need to better predict the likelihood of treatment response.

The evolutionary dynamics of tumor development can be partially decoded from the pattern of intra-tumor genetic heterogeneity[22]. Positive and negative selection, respectively, cause the expansion and contraction of subclones. Consequently, the site frequency spectrum of mutations, as measured by variant allele frequencies (VAF)[9,23] from genome sequencing data, and cohort-wide mutation frequencies (e.g. dN/dS analysis) can be used to infer the evolutionary dynamics that shaped the mutational landscape[24–28]. Population genetics has long been concerned with the dynamics of negative selection in *constant population* sizes[29–33], which has been extended for expanding populations with *rare mutations*[34,35]. However, cancer evolution represents a distinct evolutionary regime because neoantigens are common, making negative selection pervasive, immune escape can diminish selection; and tumors are growing populations. Therefore, the dynamics resulting from negative selection acting on neoantigens in a growing tumor remain to be determined.

Here we use stochastic modelling to study how the clonal structure and immunological phenotype of growing tumors is shaped by negative selection in response to neoantigenic mutations. We establish the dynamics expected under different selective environments and tumor mutator phenotypes. We characterize the emerging VAF distribution under pervasive negative selection, and determine the power to identify negative selection from genomics data. We compare our modelling predictions with whole-exome sequencing and RNA sequencing data from human cancers of the colon, stomach and endometrium.

## RESULTS

### Modelling predicts antigen-hot and antigen-cold tumors

We created a mathematical model of neoantigen evolution during tumor growth, based on a stochastic branching process (Fig. 1a and Methods). At each step, tumor cells of lineage *i* produced two surviving offspring at birth rate *b=1* per unit time and offspring accumulated mutations at rate $\mu$, which had antigenicity drawn from a pre-specified distribution. Cells died with death rate determined by the strength of negative selection, *s*, against the cumulative antigenicity of neoantigens in the lineage. *s* can be interpreted as the effectiveness of immune predation against an antigen: *s*=0 indicates no selection pressure (neutral evolution), and *s*<<0 strong negative selection (following ref[34]). Tumor growth was simulated until the tumor reached a predefined population size (approximating a clinically detectable size) or until a sufficiently long time elapsed without the tumor reaching detectable size.

We first examined the temporal neoantigen burden in simulated tumors. We defined the 'antigen score' of a tumor as the proportion of tumor cells carrying cumulative antigenicity $\geq T_c$. Tumors simulated with identical parameters separated into two distinct groups due to the stochasticity of neoantigen accrual: 'antigen-hot' and 'antigen-cold'. Antigen-hot tumors had an antigen score close to 1, corresponding to every tumor cell in the population being highly antigenic, whereas in antigen-cold tumors the majority of cells lacked immunogenic mutations (Fig. 1b-c). The proportion of antigen-hot tumors depended on the negative selection strength (Extended Data Fig. 1a): increased negative selection for neoantigens decreased the probability of observing antigen-hot tumors. In antigen-cold tumors, the proportion of neoantigen-carrying cells also decreased inversely with the strength of negative selection.

In the simulations, the antigenicity of newly accrued neoantigens was sampled from a 'prior' pre-

specified distribution. Regardless of the shape of the prior distribution, surviving lineages always showed enrichment for low-antigenicity alterations with an exponential-like distribution of final antigenicity values (Fig. 1d and Extended Data Fig. 1b).

We next simulated hyper-mutated tumors that generated a high number of mutations per cell division, causing lineages to rapidly accrue antigenicity. Consequently, most lineages rapidly became neoantigen-hot and were eradicated by negative selection (Fig. 1e). In rare tumors that survived to detectable size, high-frequency neoantigens were absent (Extended Data Fig. 2a-b and Supplementary Note).

Overall, we observed that negative selection prevented subclonal neoantigens rising to high frequency in a tumor, and this effect was exacerbated at higher mutation rates.

We compared the dynamics observed in our model to the dynamics of neoantigen accrual in a constant population size (Supplementary Note). Models of negative selection with constant population size[29–32] can lead to a broad range of evolutionary dynamics as the mutation rate and strength of negative selection are varied. In contrast, here we observed that allowing the population size to vary led to broadly consistent dynamics across the parameters space (Extended Data Fig. 2). We considered three scenarios: (i) High $s$, low $\mu$. When negative selection was strong and mutations rare, selection operated efficiently in a constant population rendering it devoid of neoantigenic mutations, but was attenuated in a variable-sized population due to population expansion decreasing the efficiency of selection, as previously reported for positive selection[23]. (ii) Low $s$, high $\mu$. Due to weak selection, only lineages with multiple mutations experienced non-negligible selection. As in the previous case, population growth attenuated the influence of selection relative to the constant-sized population model. (iii) High $s$, high $\mu$. In constant size populations, the population could not go extinct, and dynamics were determined by the relative strength of negative selection between lineages all accruing

neoantigenic mutations. The additive effect of any single mutation on fitness was proportionally diminished as mutation burden increased due to a Muller's Ratchet-like effect[33], leading to weakly selected dynamics. In a variable-sized population the dynamics were markedly different: populations where all lineages were strongly negatively selected went extinct, and surviving populations consisted of the 'lucky' lineages that had not accrued neoantigens (Extended Data Fig. 2a,d). These extinction-driven dynamics persisted in the growing population even in the special case of extremely high $\mu$ and low s, while the constant population became effectively neutral.

**Immune escape leads to antigen-hot and antigen-warm tumors**

We next simulated *immune escape alterations* acquired by one cell that renders descendants less susceptible to immune predation[36,37]. Specifically, we set the death rate of immune escaped cells to the baseline non-immunogenic death rate irrespective of the cell's burden of antigenic mutations.

If the founder cell of the tumor contained an escape mutation (*clonal escape*), tumors with a continuum of antigenicity scores emerged (Fig. 1f). We termed these tumors 'antigen-warm' as they contained strong high-frequency and/or several subclonal neoantigens.

We then simulated tumors which could acquire immune escape at a random time (*probabilistic escape*) and evaluated the detectable neoantigen load in the emerging tumors (Methods). When the mutation rate was low, tumors that reached detectable size had rarely evolved immune escape, and the strength of negative selection imposed on growth was inversely correlated with the subclonal neoantigen burden observed in the final tumor (Fig. 1g). When the mutation rate was high, lineages rapidly accrued neoantigens and were driven to extinction by negative selection (Fig. 1e). Tumors only grew to detectable frequency if the founder lineage

stochastically acquired immune escape to 'rescue' them. Consequently, at high mutation rates, detectable tumors were exclusively immune escaped and had a high burden of high-frequency neoantigens (Fig. 1h).

Taken together, these results suggest that there is a non-linear relationship between the levels of immune surveillance in the microenvironment and the magnitude of immuno-editing seen in tumors of detectable size. Moving from low to moderate negative selection, the dynamics increasingly depart from strictly neutral dynamics as expected, and correspondingly the clonal and subclonal neoantigen burden is progressively decreased. At strong negative selection, detectable tumors are those that have stochastically accrued immune escape, and consequently show a high proportion of neoantigen-warm and –hot cases and evolve effectively neutrally. We also note that the mutation rate is a determinant of the strength of negative selection experienced by a lineage: at high mutation rates a lineage is likely to accrue multiple negatively selected variants and so experience stronger negative selection.

**Immune-infiltrated cancers are antigen-hot and escaped**

To compare model predictions to experimentally measured neoantigen landscapes, we analyzed neoantigens in 363 colorectal, 146 stomach and 370 endometrial cancers (CRC, STAD and UCEC, respectively) from The Cancer Genome Atlas (TCGA) (Fig. 2a). We focused on these cancer types because of the prevalence of mutator phenotypes, namely cancers with: polymerase-ε mutation (POLE – very high mutation rate), mismatch repair deficiency (MMR – high mutation rate, often responding well to immunotherapy[18,38]), and microsatellite stable tumors (MSS – lower mutation rate). Therefore, they provide a good model to explore the effect of different tumor-immune dynamics. TCGA samples filtered for high sequencing depth and purity were first HLA-typed *in silico*[39], and their neoantigens called and filtered[19] using the NeoPredPipe pipeline[40] (see Methods). We also evaluated T-cell infiltration from paired RNA-

168 seq data[41] as a measure analogous to negative selection strength *s* experienced by
169 neoantigens.

170 The vast majority of tumors (90%) had clonal neoantigens (Supplementary Table 1), and so
171 were defined as 'antigen-hot'. We observed that the mutation-antigenicity distribution of tumors
172 (see Methods) was enriched for low binding neoantigens irrespective of the level of T-cell
173 infiltrate, but still contained a tail of high-scoring neoantigens (Fig. 2b). Subclonal neoantigen
174 burden varied significantly between cancers: cancers with low or medium T-cell infiltration
175 (putative small or moderate *s*) had proportionally fewer subclonal neoantigens than high T-cell
176 infiltrate tumors (high *s*) (Fig. 2c), suggesting a critical role of immune escape in early evolution.
177 Interestingly, this trend was absent in STAD tumors, suggesting a more homogeneous evolution
178 due to either widespread or rare immune escape.

179 We therefore sought evidence of immune escape in the cancers: alterations in antigen
180 presentation and over-expression of immune checkpoint genes (Methods). Overall, 57% of all
181 cancers showed evidence of at least one escape mechanism, with increased prevalence of
182 escape in MMR (71%) and POLE (98%) cases and significantly different patterns of immune
183 escape (Fig. 2d and Extended Data Fig. 3a), in agreement with previous studies[18,41,42]. STAD
184 cancers in particular had a high proportion of immune escaped cancers – potentially a result of
185 strong early immune predation. Further work is needed to confirm that these differences
186 between mutational subtypes arose from differential selective pressures on immune escape.

187 Consistent with the predictions and previous studies[43], tumors with immune escape had a
188 higher neoantigen burden, and the majority of highly antigenic tumors (neoantigen burden >100)
189 were immune-escaped (Fig. 2e). Increased immune infiltration level was strongly associated
190 with immune escape, even in non-hyper-mutated (MSS) samples (Fig. 2f). We expected
191 neoantigen-associated mutations to be most under-represented amongst high-cancer cell

fraction (CCF) subclonal mutations, as selection had the longest time to act on these mutations. Therefore, we compared the number of neoantigens at high CCF (present in 30%-60% of cells) between MMR cases with and without immune escape, and found greater depletion in non-escaped cancers (Fig. 2g), consistent with immuno-editing shaping the clonal structure of hyper-mutated tumors without immune escape. The above phenomena were also observed in a meta-cohort that combined the three cancer types (Extended Data Fig. 3).

Together, these data suggest that these cancer types usually evolve in the face of stringent immune-selective pressures (analogous to the moderate/high *s* regime in simulated tumors) and consequently immune-escape is frequently selected for at the onset of tumor growth, permitting the development of tumors with high and clonal neoantigen load.

**Subclonal immune escape shapes local neoantigen evolution**

Next, we explored the evidence for subclonal immune escape in a previously published multi-region sequenced colorectal tumor dataset[44]. Overall, loss of heterozygosity (LOH) at HLA loci, called with the LOHHLA tool[37], was found in 5/10 (50%) carcinomas and 1/6 (17%) adenomas, and some of these events were present subclonally, in spatially distinct region(s) of the tumor (Fig. 3a-b).

Simulations of subclonal immune escape in our model predicted that subclones should become proportionally enriched for neoantigens following escape (Fig. 3c), consistent with previous observations[37]. In our primary tumor data, a significantly higher proportion of detected neoantigens were associated with the lost allele in escaped clones than in clones without LOH (Fig. 3d). These results confirm that locally different immune-mediated negative selection pressures shape individual subclones inside a tumor.

To study how subclonal immune escape mechanisms can influence the efficiency of therapy, we

extended our simulations to model immunotherapy. We introduced two different types of escape stochastically during tumor growth, *active* and *passive,* that notionally represented reversible escape mechanisms affecting interactions with the microenvironment (e.g. expression of PDL1) and irreversible cell-intrinsic escape (e.g. genomic loss of an HLA allele) respectively (Methods). After the tumor population grew up to detectable size, we simulated immunotherapy by cancelling the effect of active immune escape, and also increasing the negative selection pressure *s* against neoantigens. The clonal population(s) with active escape rapidly shrank, but clones with passive-type escape continued growing (Fig. 3e). Neoantigens were progressively pruned from the expanding clone, leading eventually to an immune-cold tumor. Thus, the immune landscape of a tumor post-immunotherapy is predicted to be distinct from the original tumor (consistent with observations[45,46]), with potential implications for the choice of the next line of therapy.

**Negative selection leads to neutral VAF distribution**

We sought to explore how negative selection shapes the distribution of subclonal mutation frequencies within an individual cancer. We considered the VAF distribution in simulated tumors with moderate and high negative selection. Evidence for positive selection in the VAF distribution is provided by an over-abundance of passenger mutations at high-frequency that are within the expanding clone[23], whereas under pervasive negative selection, antigenic clones are continually depleted and so rarely grow to a large size (rarely reach high VAF). Thus, the vast majority of higher-VAF mutations are neutral passengers, that evolve according to neutral dynamics and so exhibit a characteristic $1/f^2$ dependence (leading to a 1/f dependence of the cumulative VAF distribution, Fig. 4a)[9]. As negative selection strength increases, the phenomenon is exacerbated: antigenic subclones are more rapidly depleted and so more neutral-like VAF distributions are observed (Fig. 4b). We note that pervasive negative selection

was part of the original neutral theory[47], and our observations are consistent with the classical theory.

The VAF distribution computed of solely neoantigens shows depletion relative to the neutral expectation (red lines in Fig. 4a-b), consistent with population genetics theory of constant-sized models[29,34] (Extended Data Fig. 2c,f). The magnitude of deviation from the neutral curve depends on the strength of negative selection, which means that, in theory, negative selection could be detected from neoantigen-VAF distributions (Extended Data Fig. 4). However, in practice, the few persisting neoantigens are at *very low* VAFs and so are problematic to measure accurately[48], severely hindering the power to quantify negative selection strength directly from neoantigen VAF distributions.

**Negative selection is elusive in VAF distribution**

We performed *in silico* sequencing on simulated tumors, and explored the effect of read depth and false-positive neoantigen identification[49] on the identifiability of negative selection in individual tumors (see Methods). The simulations predicted that very high depth sequencing was required to robustly call negative selection from VAF distributions, and the efficacy strongly depended on the strength of selection against neoantigens (Fig. 4c-d). Erroneously labelling neoantigens also had substantial impact on the power, but could be mitigated by very high-depth sequencing. Detection was mostly limited by the tumors retaining too few neoantigens to reliably evaluate their VAF distribution, a phenomenon further exacerbated when concentrating on strongly immunogenic mutations alone (Extended Data Fig. 5a-d).

In order to overcome the technical issues of limited sequence depth and low antigen numbers, we pooled mutations from groups of identically simulated and comparable TCGA tumors

(Methods) and considered their combined VAF distribution (Fig. 4e), in a similar manner to how cohort-wide positive selection by dN/dS analysis is evaluated[24,25]. In the pooled TCGA cohort, we investigated essential genes[50] that are expected to be constitutively expressed and under selection[25,51]. In cancers with medium T-cell score and no evidence of immune escape, there was a depletion of all neoantigens and neoantigens in essential genes compared to the neutral expectation (Fig. 4f). In contrast, there was no neoantigen depletion in cancers with low T-cell score. Neoantigens in CRC and UCEC cancers individually, as well as frameshift and nonsense mutations in essential genes, showed similar trends (Extended Data Fig. 5e-f), suggesting a more stringent selection in moderately infiltrated tumors and on essential genes.

**Proportional burden can measure negative selection**

Depletion of neoantigens relative to the overall non-synonymous mutation is a well-established signature of immuno-editing[52,53]. We investigated the relationship between the degree of neoantigen depletion and strength of negative selection experienced by neoantigens.

First, we simulated tumors with a known neoantigen production rate ($p_a$*=0.075* per non-synonymous mutation, Supplementary Note*)* to evaluate how the proportion of immunogenic to non-synonymous mutations changed with negative selection strength. As expected, stronger negative selection led to proportionally fewer observed neoantigens in the final tumor (Fig. 5a). We also measured the effective mutation rate (the ratio of the per cell division mutation and survival rate), derived from the linear slope of the neutral VAF curve[9], as a function of increasing negative selection for neoantigens (Supplementary Note). Stronger negative selection caused higher effective mutation rates in antigenic tumors (Fig. 5b), as a consequence of increased death rate. We suggest that the higher cell death rate inferred in hyper-mutated tumors[54] is likely to be, at least in part, a direct consequence of immuno-editing.

Next, we examined the proportional neoantigen burden in TCGA cancers stratified by cancer type and predicted immune escape status. We observed no difference in overall proportional neoantigen burden according to cancer type (Extended Data Fig. 6a), and so combined all data into a single meta-cohort. We detected no significant difference in overall proportional burden between MSS and MMR, and immune escaped or non-escaped cancers (Extended Data Fig. 6b-c). The observed uniformity in overall proportional burden across the cohort is consistent with the lack of neoantigen depletion signal reported in ref[52]. The majority of mutations considered in these analyses were clonal, and so were likely accrued prior to tumor expansion and acquisition of immune escape. To better delineate the decrease in negative selection expected following immune escape, we computed subclonal proportional neoantigen burden for mutations with CCF<0.6. Comparing total and subclonal proportional burden (considering all tumors with >30 subclonal mutations) showed a lower subclonal proportional burden in non-escaped cancers, but no shift was detected in cancers with immune escape (Fig. 5c), consistent with stronger negative selection in non-escaped cancers. When cancer types were considered independently, UCEC and CRC cancers showed a similar pattern, but no subclonal depletion was evident in STAD cancers (Extended Data Fig. 6d).

To examine the potential confounding effect of different mutational processes, we generated synthetic cohorts analogous to real tumors (Methods). Comparing the synthetic cohorts matching the overall mutation composition of CRCs showed no significant difference in proportional burden, suggesting that MMR-specific mutational processes (e.g. Signature 6 from ref[55]) are not strongly biased for neoantigen generation (Extended Data Fig. 6e). A synthetic matched cohort of Fig. 5c confirmed that the observed difference in subclonal proportional neoantigen burden was also independent of mutational processes (Extended Data Fig. 6f). Burden normalized to this synthetic cohort showed a trend for lower than random subclonal neoantigen burden (Extended Data Fig. 6g). These observations imply the presence of active

immune surveillance when escape has not occurred, and highlight the high inter-patient variability in evolutionary dynamics.

## DISCUSSION

Here we have investigated the evolutionary dynamics of neoantigens and immune escape in growing tumors using a mathematical model of tumor evolution. Our analysis shows how negative selection by the immune system (immuno-editing) sculpts the clonal architecture of the tumor: the hallmark of negative selection is the lack of neoantigens at intermediate subclonal frequency within a tumor, and conversely, the presence of numerous neoantigens at intermediate frequency is a hallmark of immune escape. Moreover, strong negative selection for neoantigens inevitably provides a *strong selective pressure* for the evolution of immune escape. Consequently, the observation that many cancers are both (neo)antigenic and have immune escape points to a critical role for immune escape in the genesis of malignancy. Further work directly measuring the immune repertoire at the time invasion first occurs is required.

Our simulations show that under negative selection, the *overall VAF distribution* of a tumor will be effectively-neutral, as it will be dominated by the neutral passenger mutations that are able to spread through the tumor unimpeded by immune predation. In constant size models, neutral mutations linked to disadvantageous alterations show a pattern of background selection[30–33], but in growing populations selection can only be observed on the selected mutations directly. The VAF distribution observable in cancer genome sequencing data becomes more neutral-like as the strength of negative selection increases, as negatively selected clones are pushed to harder-to-detect frequencies leaving only neutrally evolving lineages at high VAF. Furthermore, our analysis suggests that the majority of tumors with high mutational burden – where in theory VAF distributions and so evolutionary dynamics should be easier to resolve – are most likely to be immune escaped and so only exhibit effectively neutral dynamics. Consequently, we suggest that the lack of immune-related selection signal (e.g. as identified by ref[52]) could be due to unclassified immune escape or false-positive neoantigen calls that together mean the mutations

studied are likely to be overall only very weakly negatively selected. Pooling data across cancers increases power to resolve clone size distributions and detect negative selection, and could be combined with dN/dS methods to evaluate selection of gene sets, such as natural HLA-binders[52,56] and MHC-II presented peptides[57].

Our modelling offers insight into the challenges of predicting immunotherapy response using tumour mutation burden (TMB) alone. Strong negative selection (effective immune surveillance) leads to a high rate of cell death, a corresponding increase in the effective mutation rate of tumors, and the net result of high TMB with severe neoantigen depletion. Thus, despite having high TMB, such tumors would be unlikely to respond to immune checkpoint blockade. Assessment of neoantigens should be more predictive: tumors with clonal or numerous subclonal neoantigens are very likely to have evolved immune escape – particularly if the patient's immune system is highly predatory – and to respond to therapies reactivating immune predation. This is consistent with previous studies suggesting that clonal antigens predict sensitivity to immune checkpoint blockade[43]. We illustrate that immune therapies targeted against a specific neoantigen or immune mechanism are vulnerable to intra-tumor heterogeneity, as subclones in which this target is altered or lost (e.g. neoantigen depleted or HLA haplotype mutated) will experience net positive selection when the therapy is applied[58–60]. Relatedly, a subclone that escapes immune blockade therapy and reforms the tumor is predicted to have a different immune landscape due to the action of immune predation during clone emergence, with potential implications for additional lines of therapy.

In summary, our mathematical framework provides insights into the evolutionary dynamics of negatively selected neoantigens in growing tumors and the detectability of these dynamics in genomic data.

## ACKNOWLEDGEMENTS

This work was supported by the Wellcome Trust (202778/B/16/Z to A.S.; 202778/Z/16/Z to T.A.G.; 105104/Z/14/Z to the Centre for Evolution and Cancer, Institute of Cancer Research; 108861/7/15/7 to R.O.S.; 097319/Z/11/Z to C.P.B.) and Cancer Research UK (A22909 to A.S.; A19771 to T.A.G. supporting E.L.). A.R.A.A. and C.G, and A.S and T.A.G. received support from the US National Institutes of Health National Cancer Institute (grant no. U54CA143970) and (U54 CA217376) respectively. R.O.S. was also supported by the Wellcome Centre for Human Genetics (grant no. 203141/7/16/7). B.W. was supported by the Geoffrey W. Lewis Postdoctoral Training fellowship. L.Z. is supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Research Fellowship scheme (846614).

## AUTHOR CONTRIBUTIONS

E.L., A.R.A.A., A.S. and T.A.G. conceptualized the study. A.R.A.A., A.S and T.A.G. acquired funding for the project. E.L., A.S. and T.A.G. led the investigation, analysed data, and wrote the original manuscript. E.L., M.J.W., W.C.H.C., B.W., R.O.S., C.G., J.H., L.Z., M.R.T., and C.P.B. contributed to the mathematical model, computational framework and bioinformatics analysis. All authors reviewed and approved the final manuscript.

## COMPETING INTERESTS

The authors declare no competing interests.

## REFERENCES

1. Schumacher, T. N. & Schreiber, R. D. Neoantigens in cancer immunotherapy. *Science (80-. ).* **348**, 69 (2015).
2. Lu, Y.-C. & Robbins, P. F. Cancer immunotherapy targeting neoantigens. *Semin Immunol* **28**, 22–27 (2016).
3. Galon, J. *et al.* Towards the introduction of the 'Immunoscore' in the classification of malignant tumours. *J. Pathol.* **232**, 199–209 (2014).
4. Sharma, P. & Allison, J. P. The future of immune checkpoint therapy. *Science (80-. ).* **348**, 56–61 (2015).
5. Larkin, J. *et al.* Combined Nivolumab and Ipilimumab or Monotherapy in Untreated Melanoma. *N. Engl. J. Med.* **373**, 23–34 (2015).
6. Milo, I. *et al.* The immune system profoundly restricts intratumor genetic heterogeneity. *Sci. Immunol.* **3**, eaat1435 (2018).
7. Dunn, G. P., Bruce, A. T., Ikeda, H., Old, L. J. & Schreiber, R. D. Cancer immunoediting: from

immunosurveillance to tumor escape. *Nat. Immunol.* **3**, 991–998 (2002).
8. DuPage, M., Mazumdar, C., Schmidt, L. M., Cheung, A. F. & Jacks, T. Expression of tumour-specific antigens underlies cancer immunoediting. *Nature* **482**, 405–409 (2012).
9. Williams, M. J., Werner, B., Barnes, C. P., Graham, T. A. & Sottoriva, A. Identification of neutral tumor evolution across cancer types. *Nat. Genet.* **48**, 238–244 (2016).
10. Koebel, C. M. *et al.* Adaptive immunity maintains occult cancer in an equilibrium state. *Nature* **450**, 903–907 (2007).
11. Hanahan, D. & Weinberg, R. A. Hallmarks of Cancer: The Next Generation. *Cell* **144**, 646–674 (2011).
12. Koebel, C. M. *et al.* Adaptive immunity maintains occult cancer in an equilibrium state. *Nature* **450**, 903 EP- (2007).
13. Marty, R. *et al.* MHC-I Genotype Restricts the Oncogenic Mutational Landscape. *Cell* **171**, 1272-1283.e15 (2017).
14. Rosenthal, R. *et al.* Neoantigen-directed immune escape in lung cancer evolution. *Nature* **567**, 479–485 (2019).
15. Yarchoan, M., Johnson III, B. A., Lutz, E. R., Laheru, D. A. & Jaffee, E. M. Targeting neoantigens to augment antitumour immunity. *Nat. Rev. Cancer* **17**, 209–222 (2017).
16. Rizvi, N. A. *et al.* Mutational landscape determines sensitivity to PD-1 blockade in non--small cell lung cancer. *Science (80-. ).* **348**, 124–128 (2015).
17. Lennerz, V. *et al.* The response of autologous T cells to a human melanoma is dominated by mutated neoantigens. *Proc Natl Acad Sci U S A* **102**, 16013–16018 (2005).
18. Le, D. T. *et al.* Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade. *Science (80-. ).* **357**, 409–413 (2017).
19. Łuksza, M. *et al.* A neoantigen fitness model predicts tumour response to checkpoint blockade immunotherapy. *Nature* **551**, 517–520 (2017).
20. Balachandran, V. P. *et al.* Identification of unique neoantigen qualities in long-term survivors of pancreatic cancer. *Nature* **551**, 512 (2017).
21. Gibney, G. T., Weiner, L. M. & Atkins, M. B. Predictive biomarkers for checkpoint inhibitor-based immunotherapy. *Lancet. Oncol.* **17**, e542–e551 (2016).
22. Turajlic, S., Sottoriva, A., Graham, T. & Swanton, C. Resolving genetic heterogeneity in cancer. *Nat. Rev. Genet.* **20**, 404–416 (2019).
23. Williams, M. J. *et al.* Quantification of subclonal selection in cancer from bulk sequencing data. *Nat. Genet.* **50**, 895–903 (2018).
24. Ostrow, S. L., Barshir, R., DeGregori, J., Yeger-Lotem, E. & Hershberg, R. Cancer Evolution Is Associated with Pervasive Positive Selection on Globally Expressed Genes. *PLOS Genet.* **10**, e1004239 (2014).
25. Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171**, 1029-1041.e21 (2017).
26. Temko, D., Tomlinson, I. P. M., Severini, S., Schuster-Böckler, B. & Graham, T. A. The effects of mutational processes and selection on driver mutations across cancer types. *Nat. Commun.* **9**, 1857 (2018).
27. Cannataro, V. L., Gaffney, S. G. & Townsend, J. P. Effect Sizes of Somatic Mutations in Cancer. *JNCI J. Natl. Cancer Inst.* **110**, 1171–1177 (2018).
28. Williams, M. J. *et al.* Measuring the distribution of fitness effects in somatic evolution by combining clonal dynamics with dN/dS ratios. *Elife* **9**, e48714 (2020).
29. Cvijović, I., Good, B. H. & Desai, M. M. The Effect of Strong Purifying Selection on Genetic Diversity. *Genetics* **209**, 1235 (2018).
30. Good, B. H., Walczak, A. M., Neher, R. A. & Desai, M. M. Genetic Diversity in the Interference Selection Limit. *PLOS Genet.* **10**, e1004222 (2014).
31. Neher, R. A. & Hallatschek, O. Genealogies of rapidly adapting populations. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 437–442 (2013).
32. Charlesworth, B., Morgan, M. T. & Charlesworth, D. The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**, 1289–1303 (1993).
33. Haigh, J. The accumulation of deleterious genes in a population—Muller's Ratchet. *Theor. Popul. Biol.* **14**, 251–267 (1978).
34. Kessler, D. A. & Levine, H. Scaling solution in the large population limit of the general asymmetric

stochastic Luria-Delbrück evolution process. *J. Stat. Phys.* **158**, 783–805 (2015).
35. Antal, T. & Krapivsky, P. L. Exact solution of a two-type branching process: models of tumor progression. *J. Stat. Mech. Theory Exp.* **2011**, P08018 (2011).
36. Vinay, D. S. *et al.* Immune evasion in cancer: Mechanistic basis and therapeutic strategies. *Semin. Cancer Biol.* **35**, S185--S198 (2015).
37. McGranahan, N. *et al.* Allele-Specific HLA Loss and Immune Escape in Lung Cancer Evolution. *Cell* **171**, 1259--1271.e11 (2017).
38. Kather, J. N., Halama, N. & Jaeger, D. Genomics and emerging biomarkers for immunotherapy of colorectal cancer. *Semin. Cancer Biol.* **52**, 189–197 (2018).
39. Shukla, S. A. *et al.* Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes. *Nat. Biotechnol.* **33**, 1152–1158 (2015).
40. Schenck, R. O., Lakatos, E., Gatenbee, C., Graham, T. A. & Anderson, A. R. A. NeoPredPipe: high-throughput neoantigen prediction and recognition potential pipeline. *BMC Bioinformatics* **20**, 264 (2019).
41. Grasso, C. S. *et al.* Genetic mechanisms of immune evasion in colorectal cancer. *Cancer Discov.* **8**, 730–749 (2018).
42. Xie, T. *et al.* A Comprehensive Characterization of Genome-Wide Copy Number Aberrations in Colorectal Cancer Reveals Novel Oncogenes and Patterns of Alterations. *PLoS One* **7**, e42001 (2012).
43. McGranahan, N. *et al.* Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. *Science (80-. ).* **351**, 1463–1469 (2016).
44. Cross, W. *et al.* The evolutionary landscape of colorectal tumorigenesis. *Nat. Ecol. Evol.* **2**, 1661–1672 (2018).
45. Riaz, N. *et al.* Tumor and Microenvironment Evolution during Immunotherapy with Nivolumab. *Cell* **171**, 934-949.e16 (2017).
46. Anagnostou, V. *et al.* Evolution of Neoantigen Landscape during Immune Checkpoint Blockade in Non-Small Cell Lung Cancer. *Cancer Discov.* **7**, 264–276 (2017).
47. Kimura, M. *The Neutral Theory of Molecular Evolution*. (Cambridge University Press, 1983). doi:DOI: 10.1017/CBO9780511623486
48. Stead, L. F., Sutton, K. M., Taylor, G. R., Quirke, P. & Rabbitts, P. Accurately Identifying Low-Allelic Fraction Variants in Single Samples with Next-Generation Sequencing: Applications in Tumor Subclone Resolution. *Hum. Mutat.* **34**, 1432–1438 (2013).
49. Yadav, M. *et al.* Predicting immunogenic tumour mutations by combining mass spectrometry and exome sequencing. *Nature* **515**, 572–576 (2014).
50. Blomen, V. A. *et al.* Gene essentiality and synthetic lethality in haploid human cells. *Science (80-. ).* **350**, 1092–1096 (2015).
51. Van den Eynden, J., Basu, S. & Larsson, E. Somatic Mutation Patterns in Hemizygous Genomic Regions Unveil Purifying Selection during Tumor Evolution. *PLOS Genet.* **12**, e1006506 (2016).
52. Van den Eynden, J., Jiménez-Sánchez, A., Miller, M. L. & Larsson, E. Lack of detectable neoantigen depletion signals in the untreated cancer genome. *Nat. Genet.* **51**, 1741–1748 (2019).
53. 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).
54. Werner, B. *et al.* Measuring single cell divisions in human tissues from multi-region sequencing data. *Nat. Commun.* **11**, 1035 (2020).
55. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
56. Zapata, L. *et al.* Negative selection in tumor genome evolution acts on essential cellular functions and the immunopeptidome. *Genome Biol.* **19**, 67 (2018).
57. Marty Pyke, R. *et al.* Evolutionary Pressure against MHC Class II Binding Cancer Mutations. *Cell* **175**, 416-428.e13 (2018).
58. Kim, J. M. & Chen, D. S. Immune escape to PD-L1/PD-1 blockade: seven steps to success (or failure). *Ann. Oncol.* **27**, 1492–1504 (2016).
59. Sharma, P., Hu-Lieskovan, S., Wargo, J. A. & Ribas, A. Primary, Adaptive, and Acquired Resistance to Cancer Immunotherapy. *Cell* **168**, 707–723 (2017).
60. Iorgulescu, J. B., Braun, D., Oliveira, G., Keskin, D. B. & Wu, C. J. Acquired mechanisms of immune escape in cancer following immunotherapy. *Genome Med.* **10**, 87 (2018).

## FIGURE LEGENDS

**Figure 1: Tumor growth model predicts two distinct types of immune phenotypes and the necessity of immune escape. (a)** Schematic representation of the model. Left panel: tumor growth for four generations. Filled circles represent cells, colored by immunogenicity. Related cells are connected with lines. Middle panel: cell division/mutation process. Right panel: prior distribution of newly generated neoantigenicities. For details, see Methods. **(b)** Growth curve of six simulated tumors at s=-0.8. Line color shows the antigen score of the tumor population over time. **(c)** Cancer cell fraction (CCF) of the most common antigenic mutation of n=100 tumors at the final time-point. **(d)** Distribution of antigenicity values of all neoantigens generated (grey) and only neoantigens present in >10 surviving cells (blue). Thin lines: individual simulations; thick dashed line: ensemble mean. Inset: Mann-Whitney two-sided test. **(e)** Distribution of maximum tumor size reached by hyper-mutated tumors at s=-0.8. Inset: growth curve of a single tumor colored by antigenic score as in (b), blue line: number of non-immunogenic cells. **(f)** Neoantigen scores in n=100 tumors at s=-0.8, without (left) and with (right) clonal immune escape. **(g-h)** Number of detectable neoantigens (read depth ~50x) in n=50 simulated tumors as a function of negative selection strength. Middle panel: mean clonal neoantigen burden. Bottom panel: clonality of immune escape. Only non-hyper-mutated (g) and hyper-mutated (h) tumors that reached a detectable size are shown. Violin widths represent raw data density.

**Figure 2: Colorectal, stomach and endometrial tumors from TCGA are antigen-hot and enriched for immune escape. (a)** Cancer type and mutator subtype of the TCGA cancers analyzed. The size and shade of each circle represent the number of tumors (also shown) in that sub-category. **(b)** Distribution of normalized binding strength of neoantigens in TCGA cancers with low, medium and high immune infiltration. The thick line shows the mean density of all distributions from tumors in each category, the shaded regions represent $\pm 1$ standard deviation around this mean. **(c)** Distribution of the number of subclonally detected (in <60% of the tumor) neoantigen-associated mutations in cancers according to immune infiltration (T-cell average) score. Two-sided Mann-Whitney tests are reported on each plot. **(d)** Prevalence of immune escape in MSS, MMR and POLE samples. Two-sided chi-squared test is indicated on top of each panel. **(e)** Distribution of the number of subclonal antigenic mutations in cancers with and without immune escape (magenta and grey, respectively) Two-sided Mann-Whitney test is reported on each panel. **(f)** Prevalence of immune escape in MSS cancers according to their immune infiltration level. Two-sided chi-squared test is indicated on top of each panel. **(g)** Number of antigenic mutations present in large subclones (>30% and <60% of cells) in MMR samples with and without immune escape. One-sided Mann-Whitney test is reported above each plot. Violin widths in (c), (e) & (g) represent raw data density with binned individual data points overlaid on top.

**Figure 3: Subclonal immune escape shapes neoantigen landscape and tumor growth after therapy. (a)** Immune escape through loss of heterozygosity (LOH) at an HLA locus in the multi-region sequenced colorectal cohort. LOH events are divided up according to whether the alteration is detected in all (clonal) or not all (subclonal) biopsies. **(b)** HLA LOH in individual biopsies in tumors with at least one subclonal or clonal loss event. Unfilled boxes represent homozygous HLA alleles. **(c)** The number of antigenic mutations detected in two distinct (with and without immune escape) subclones of n=25 simulated tumor. Antigenic mutations are detected at simulated read depth of 100x. Visual elements of the boxplot correspond to the following summary statistics: centre line, median; box limits, upper and lower quartiles; whiskers, 1.5x inter-quartile range. **(d)** The proportion of all neoantigens binding to the HLA allele lost in the LOH event in the colorectal tumors that show subclonal HLA LOH (n=6). One-sided Wilcoxon signed-rank tests are reported on (c) and (d). **(e)** Growth curve of simulated tumors following anti-PD-L1-type immunotherapy. The tumors have previously developed active immune escape, but also harbor a small subclone with different escape mechanism. Black dashed lines show the number of cells in this subclone over time. The inset shows growth around the point when the subclone takes over, on a logarithmic scale.

**Figure 4: Negative selection leads to characteristic depletion of neoantigens and effectively-neutral overall VAF distributions. (a-b)** Cumulative number of mutations as a function of the inverse of the frequency for all mutations (grey, left axis) and neoantigen-associated mutations (red, right axis) harbored in at least 30 cells in (a) a tumor with s=-0.8; (b) a tumor with s=-1.2. **(c)** Power to detect negative selection from the VAF distribution as a function of sequencing read depth (x axis) and false neoantigen rate (y axis). Power is the proportion of 100 simulated tumors with significant difference (two-sided Kolmogorov-Smirnov test, $\alpha$=0.1) between the distribution of all mutations and neoantigen-associated mutations. **(d)** Power (in n=100 tumors) to identify negative selection as a function of selection strength (x axis) and the stringency of the two-sided Kolmogorov-Smirnov test used for detection ($\alpha$=0.1, $\alpha$=0.05, and $\alpha$=0.01, shown in black, maroon and red, respectively). **(e)** Cumulative VAF distribution as a function of the inverse of the frequency for all (in grey) and neoantigen-associated mutations (in red) detected with a sequencing depth of 800x in antigen cold tumors from a simulated set of n=100. The y axis shows proportion of mutations. The mutation-antigenicity threshold 0.2 is used in all cases in (a)-(e). **(f)** Cumulative VAF distribution of mutations detected in any low- and medium-immune infiltrated TCGA MSS cancers without immune escape. The distribution is shown for all mutations (grey), exonic mutations (blue), exonic mutations in essential genes (purple), antigenic mutations (pink) and neoantigen-associated mutations in essential genes (red).

**Figure 5: Proportional neoantigen burden as a measure of selection. (a)** The proportion of neoantigen-associated mutations (the percentage of all mutations) as a function of negative selection pressure, computed from n=100 tumors each, with a simulated read depth of 200x. The expected value of antigens per mutation is indicated with a horizontal dashed line. The mutation-antigenicity threshold of 0.2 is used. **(b)** Effective mutation rate (per cell division mutation rate divided by per cell division death rate) computed from the VAF distribution of mutations in antigen-hot tumors as a function of negative selection pressure. Read depth = 200x. Colors in (a) & (b) indicate selection strength also shown on x axis **(c)** Proportional neoantigen burden of escaped and non-escaped TCGA samples, computed from all mutations (red) and only subclonal mutations (CCF<0.6, colored salmon). Lines connect total and subclonal proportional burdens measured in the same sample. Paired two-sided Wilcoxon test is reported above the violin plots. Violin widths represent raw data density with individual data points in (c) also indicated by end-points of connecting lines.

## METHODS

### *Mathematical model of tumor growth and mutation accumulation*

We created a minimal stochastic branching process model to represent tumor growth and accumulation of mutations under selection pressure from the environment[61]. The model described the proliferation, death and mutation accumulation of tumor cells, and environmental factors (e.g. the level of T-cell infiltration) were described implicitly through parameters that quantified the strength of selection against tumor cells.

We made use of a rejection-kinetic Monte Carlo algorithm[62] to permit efficient simulation of large populations of cells. Tumor evolution was initiated by a single transformed cell that produced two surviving offspring at birth rate $b$ per unit time. Cells in clone $i$ died at rate $d_i$ per unit time, where the death rate increased with the neoantigen burden of the clone. Each time a cell divided, it acquired new unique mutations at overall rate $\mu$ (Poisson distribution), which were assigned as neoantigens at rate $p_a$, or as passengers (evolutionary neutral) at rate 1-$p_a$. Each antigenic mutation was assigned an antigenicity value (denoted $A_j$ for the j$^{th}$ antigen in a given cell) sampled from an exponential distribution with the rate parameter set to 5 to produce a skewed distribution wherein >99% of antigenicity values fall between 0 and 1, and most neoantigens are only negligibly immunogenic (Fig. 1a). Neoantigens caused the death rate $d_i$ of the lineage to increase from a basal rate of $d_b$=0.1 to a higher value determined by the strength of negative selection against each new neoantigen, controlled by the parameter $s$. The overall effect on the birth/death rate of cells was determined by the cumulative antigenicity of neoantigens harbored in the lineage, $\sum A_j$. The death rate of a subclone was computed as:

$$d_i = (1 + s * \sum_{j=1}^{n_i} A_j^i)(d_b - 1) + 1. \quad (1)$$

And we defined the selective (dis)advantage of a subclone by its effective proliferation rate (the difference of its birth and death rate), as compared to a non-immunogenic clone:

$$1 + s * \sum_{j=1}^{n_i} A_j^i = fitness = \frac{b - d_i}{b - d_b}, \quad (2)$$

where $A_j^i$ denotes the jth neoantigen in lineage i; *s*=0 stands for neutral evolution with no neoantigen-associated selection and negative selection is represented by *s*<0.

This antigenicity-dependent increase in the clone death rate represented an aggregate of the many stochastic factors that lead to the negative selection of neoantigens, including; (i) sufficient presentation of neoantigens on the cell surface; (ii) recognition of neoantigens by T-cell; (iii) antigen-mediated recruitment of further T-cells; and (iv) T-cell killing efficiency. We chose to integrate all variability into a single probabilistic rate to be able to observe general qualities of the tumor-immune interaction without the need for precise parametrisation. For details on the steps of *in silico* simulations, see Supplementary Note and code at https://zenodo.org/record/3601322#.XvKCGJJKii4.

We also modelled the acquisition of immune escape during tumor growth. Known immune escape mechanisms include mutations affecting the antigen presenting machinery and expression of immune checkpoint molecules[36,37]. Immune escape was modelled as a heritable property of a cell (representing e.g. copy number alteration of the PD-L1 or HLA gene). Immune escape occurred as a result of a mutation with probability $p_e$ per nonsynonymous mutation; or through manual introduction of the escape alteration at a pre-determined clone size to achieve clonal or subclonal immune escape. We considered two different types of escape mechanism: (i) *active escape*, which shields the clone from negative selection (decreasing the clone death probability to $d_b$) but does not decrease the neoantigen burden of the cell (corresponding to escape mechanisms such as PD-L1 overexpression); and (ii) *passive escape*, which renders a portion of neoantigenic mutations neutral (by rendering their antigenicity, $A_j$ to *0*; representing, for example, loss of a HLA allele that predicts a subset of neoantigenic peptides being presented).

We also incorporated therapeutic intervention in our model by time-dependently changing model parameters. The most commonly used agents in immunotherapy target and inhibit immune checkpoint pathways, helping the immune system to overcome immune escape achieved by checkpoint over-expression and re-activate immune predation of neoantigenic cancer cells. We simulated this effect by rendering active type immune escape ineffective (death rate of escaped cells is increased by antigenic load) and simultaneously increasing the negative selection strength *s* experienced by each neoantigen.

We chose model parameters to represent a wide range of possible tumor-immune environments, and correspond to phenotypic properties of real cancers (Extended Data Fig. 6). The following parameters were used in all simulations: $b = 1$; $d_b = 0.1$; $\mu = 1$ (not hyper-mutated) and $\mu = 10$ (hyper-mutated); $-2 \leq s \leq 0$ (as indicated in figures or in caption); $p_a = 0.075$ and $p_e = 10^{-6}$ (where applicable). For analyses where cells and mutations were classified as antigenic or not, the cell- and mutation-antigenicity thresholds $T_c = 0.5$ and $T_m = 0.2$ were used, unless stated otherwise. For further discussion on the simplifications applied in the model, and the choices of simulation parameters and how they influence results, see the Supplementary Note and Extended Data Figs. 7-9.

***Simulation of VAF/CCF distributions and power calculation***

To evaluate the mutation spectrum of simulated tumors, mutations harbored in at least 10 cells out of $10^5$ (0.01%) were collected at the end of each simulation and the number of carrier cells reported. Real sequencing data naturally introduces uncertainty about mutated allele frequency due to limited sequencing depth and several sources of sampling bias[22]. To account for imperfect measurements, CCF values were either computed by taking the raw frequency values or via a simulated sequencing step introducing noise to these frequencies with indicated read depth. For a given read depth, *D*, each frequency value, *f*, was substituted by a new frequency

sampled from a binomial distribution with parameters *D* and *f*: $\bar{f} \sim Binom(D, f)/D$. We filtered for mutations with $\bar{f}$ above 0, to discard mutations that are not picked up due to limited detection power.

In addition to sequencing limitations, neoantigen identification from DNA sequencing alone has a high rate of false-positive calls[49], and therefore the VAF distribution of neoantigens is expected to be 'contaminated' with a large proportion of neutrally-evolving passenger mutations. To simulate this effect when evaluating the power of detecting selection, we randomly sampled non-antigenic mutations of simulated tumors (varied between 5% to 500% of the number of true neoantigens, Fig. 4c) that were falsely labelled as neoantigens and included in the neoantigen-based VAF distribution.

We computed the power to detect selection by comparing the distribution of all detected mutations to that of the neoantigen-labelled subset using a two-sample Kolmogorov-Smirnov test, and identified any samples as under selection in which the p-value of the test was below 0.1 (Fig. 4c) or a pre-defined value (Fig. 4d).

***TCGA sample acquisition and processing***

All samples from the TCGA COAD and READ (merged together as CRC), STAD and UCEC domains were retrieved through the NCI Genomics Data Commons (GDC) portal[63] between 15/06/2018 and 13/11/2019. Only patients with matched germline (from blood samples) and primary tumor information available were considered. For each sample, purity (fraction of tumor cells in the sample) and overall ploidy were evaluated using ASCAT[64] on Affymetrix SNP array data. Samples with purity below 0.4 and ploidy above 3.6 were excluded from the analysis, leaving 363 CRC, 146 STAD and 370 UCEC samples for which HLA typing and neoantigen calls were performed (Supplementary Table 1 and Fig. 2a).

For analyzing immune escape, the cohort was narrowed down to patients for whom gene expression data was available in GDC; and at least one pair of their HLA A/B/C alleles were heterozygous and distinct enough to allow for loss of heterozygosity calls (n(CRC) = 341, n(STAD)=118, n(UCEC)=362).

For each patient considered, the following information was downloaded: blood derived normal bam files; primary tumor bam files; unfiltered variant call (vcf) files processed with Mutect2; SNP array files; gene expression HTSeq counts (where available); and clinical information. We used the unfiltered controlled-access variant call format (vcf) files to avoid over-filtering and missing antigenic variants. The variants were filtered to only include variants that passed all filters of the vcf files and not present (allelic depth of 0 or 1 for bases covered with over 30 reads) in normal samples.

Samples were divided into MSS, MMR and POLE subtypes using data integrated from (i) clinical TCGA annotation[65]; (ii) calls retrieved from ref[66] that used the computational tool MANTIS to analyze repetitions in tumor-normal sample pairs over microsatellite loci; (iii) and mutational signature activities computed using non-negative least squares regression[26,55]. Samples with a MANTIS score ≥ 0.5 and TCGA annotation of 'MSI-H' ('microsatellite instability', where available) were considered MMR, and those with MANTIS < 0.5 and 'MSI-L'/'MSS' were labelled MSS. In case the two sources of information contradicted each other, neither of the categories was assigned. Samples with at least 1,000 mutations inferred to originate from the characteristic POLE signature (signature 10 in ref[55]) were labelled as POLE tumors regardless of their MMR status.

***Multi-region sequenced dataset processing***

The multi-region sequenced colorectal dataset was accessed from Cross *et al.*[44] (raw data

712 available from the European Genome-Phenome Archive (https://ega-archive.org/) at accession
713 code: EGAS00001003066). Bam files with marked duplicates were used for HLA calling and
714 HLA variant detection. As in the original work, variants were called using Platypus[67], annotated
715 by ANNOVAR[68], and filtered to only contain somatic single nucleotide variations that were
716 present in at least 1 tumor sample and in either 0 reads in the normal sample (for normal
717 coverage <=30 reads) or in at most 1 read (for normal coverage above 30 reads).

718 ***HLA haplotyping and calling immune escape***

719 HLA-A, -B and -C haplotyping was performed on blood derived normal bam files using
720 POLYSOLVER[39]. As POLYSOLVER takes into account the individual's race to compute the
721 likelihood of each allele haplotype, we supplied ethnicity data, where available from clinical
722 TCGA information, and ran haplotyping with race 'Unknown' otherwise.

723 Using exome and RNAseq data, we tested for the presence of three types of immune escape
724 mechanisms: (i) somatic mutations in either one of the HLA alleles or in the B2M gene[39,41]; (ii)
725 loss of an HLA haplotype through loss of heterozygosity (LOH) in the corresponding genomic
726 locus[37]; and (iii) PD-L1 or CTLA-4 over-expression[69].

727 Mutations in HLA alleles were called using the previously called HLA haplotypes and the
728 corresponding functionality of POLYSOLVER[39]. Variant calling was run using default settings
729 and HLA was considered mutated if at least one allele had a nonsynonymous somatic mutation
730 located in an exon or at a splice-site. Mutations in B2M were called if the sample contained a
731 nonsynonymous somatic mutation located inside one of the exons of the B2M gene, as
732 annotated by ANNOVAR[68] and confirmed using Variant Effect Predictor[70]. Loss of
733 heterozygosity at the HLA locus was assessed using the software LOHHLA[37], using blood
734 derived normal, and tumor bam files were used. Tumor purity and ploidy estimates were derived
735 from ASCAT (for TCGA data) and from Sequenza[71] (for the multi-region sequenced colorectal

tumors). A sample was considered to have Allelic Imbalance at an HLA locus if the corresponding p-value was below 0.01 and LOH if, in addition, the copy number prediction of that allele was below 0.5, with the confidence interval strictly below 0.7. Immune checkpoint over-expression was assessed using RNA-seq data. Normal expression values (in transcripts per million (TPM)) of PD-L1 and CTLA-4 were established for each cohort from TCGA based on RNA-seq counts of the two proteins in 'solid tissue normal' samples. Checkpoint over-expression was called if either PD-L1 or CTLA-4 expression in the tumor was higher than the mean plus two standard deviations of normal expression. Immune checkpoint over-expression could not be inferred for the multi-region sequenced dataset as only genomic data were available.

We note that the extent of the impact of these escape alterations is not always known – especially for mutations altering antigen presenting proteins – but we argued that nonetheless they represent a level of impairment in the tumor-immune interaction.

Immune infiltration levels were computed from RNA-seq data based on the method of Grasso et al.[41]: a signature of 12 genes (*CCL2, CCL3, CCL4, CXCL9, CXCL10, CD8A, HLA-DOB, HLA-DMB, HLA-DOA, GZMK, ICOS*, and *IRF1*) was extracted, and a continuous T-cell score derived as their log(TPM) average. The continuous score was then divided into three equal sized intervals (based on all cancers) to provide low, medium and high T-cell score levels.

***Neoantigen prediction***

Neoantigens were predicted from variant call tables and HLA types using NeoPredPipe[40], a neoantigen prediction and evaluation pipeline designed for parallel analysis of single- and multi-region samples. We only evaluated single nucleotide variants leading to a single amino acid change, and novel peptides of 9 and 10 amino acids were considered. The pipeline was run

with default analysis settings and preserving intermediate files (–p flag), using hg38 and hg19 ANNOVAR[68] reference files for annotation of the TCGA and multi-region CRC samples, respectively. The analysis outputted a table of novel peptides binding the patient's MHC-I molecules (considering all six alleles independently) and their respective recognition potential calculated from their MHC-binding affinity and similarity to pathogenic peptides, as described in ref[19]. For evaluating the recognizability (R) part of the recognition potential, we used the parameter values derived in ref[19]. Unless stated otherwise, we labelled a peptide as neoantigen if its recognition potential was >= $10^{-1}$ (with respect to any of the patient's HLA types) to focus on antigens with the highest predicted probability of eliciting an immune response: both similar to known pathogens and similar or stronger MHC-binders than their wild-type counterpart. A mutation was considered (neo)antigenic if there was at least a single peptide produced from the mutated base that got labelled as neoantigen.

To evaluate the antigenicity distribution of tumors, we used the predicted percentile rank of neoantigens that ranks a putative antigen against a large set of random substrates to the same HLA molecule, and thus eliminates bias introduced by structural properties of HLA alleles[72], that might be present in plain binding affinity values (considered in the recognition potential pipeline). We inverted this value to obtain a normalized binding score that correlates with the importance ranking of peptides, where values above ~1.3 represented strong putative antigens.

***Computation of VAF and CCF values***

For each mutation, we calculated the VAF as the number of mutant reads spanning the position, divided by the number of total reads of the position. The proportion of cancer cells carrying a particular mutation (CCF) was calculated from the VAF of the mutation, sample purity (tumor content), and copy number (CN) of the mutation's genomic locus as: $(VAF * CN)/purity$. CCF values above 1 (arising from sequencing noise and copy-neutral loss-of-heterozygosity events)

were assumed to be 1. We only considered a mutation as subclonal if it had CCF<0.6, to account for the possibility of 'bleeding' of clonal mutations into the subclonal frequency range because of the limited sequence depth of TCGA samples.

For pooling together VAF distributions of a cohort of samples (Fig. 4f), we first filtered the set of TCGA cancers: cancers with any evidence of immune escape (including allelic imbalance of HLA locus), MMR or POLE cancers and cancers with purity <50% were discarded. The remaining cancers were divided into low and medium immune infiltration groups (all highly T-cell score cancers were immune escaped and previously discarded). Total and neoantigen-associated cumulative VAF distributions were computed from all mutations detected at subclonal frequencies in the two groups. In a similar manner, TCGA MSS cancers with purity >70% (to ensure more accurate VAF and ploidy calls) were combined into a cohort to study mutations in essential genes (Extended Data Fig. 5f). Essential genes, and antigenic mutations located in essential genes were identified using the list of shared genes in ref[50].

***Synthetic cohorts***

In order to evaluate the antigen-producing capacity of different mutational processes, we generated synthetic tumor cohorts matching the mutation number and tri-nucleotide composition of real cancers. We measured the average composition (as measured by 96-channel-composition[55]) of the real cohort (e.g. TCGA CRCs, Extended Data Fig. 6d), and randomly sampled a matching number of exonic mutations at probability specified by the respective channel intensities. Six HLA haplotypes were also randomly sampled from the complete list of alleles in the real cohort. Sampling was repeated independently 100 times to generate a synthetic cohort.

***Statistical analysis***

Details of statistical analysis performed are summarized in the Life Science Reporting Summary. All data processing and statistical tests were performed in R (version 3.5.0) using built-in functions. The tests and functions used were as follows: Figs. 1d, 2c,e, Extended Data Figs. 3c, 6a,b,c,e: Mann-Whitney U-test/ Wilcoxon sum-rank test (*wilcox.test,* default settings). Figs. 2d,f and Extended Data Fig. 3a,b,d: Chi-squared test (*chisq.test*). Fig. 2g and Extended Data Fig. 3e: One-sided Mann-Whitney U-test (*wilcox.test* with option *alternative='greater'*). Fig. 3c-d: One-sided paired Wilcoxon signed-rank test (*wilcox.test* with options *paired=TRUE* and *alternative='greater'*). Fig. 4c-d and S5b-c: Kolmogorov-Smirnov test (*ks.test*) between the raw VAF distribution of neoantigens and all mutations. The two distributions were deemed with significance level $p<0.1$ or as indicated in Fig. 4c-d and Extended Data Fig. 5b-c. Fig. 5c and Extended Data Fig. 6d,f: Paired Wilcoxon signed-rank test (*wilcox.test*, option *paired=TRUE*). Extended Data Fig. 6g: Students t-test against mean of 1 (*t.test, mu=1*).

All violin plots were generated with automatic smoothing bandwidth value of *geom_violin*. Individual observations for TCGA samples are shown on top of violins, generated with *geom_dotplot*.

## DATA AVAILABILITY

The datasets analyzed during the current study are available from the NCI Genomics Data Commons Portal (https://portal.gdc.cancer.gov) COAD, READ, STAD and UCEC domains, and from the European Genome-Phenome Archive (https://ega-archive.org/) at accession code: EGAS00001003066.

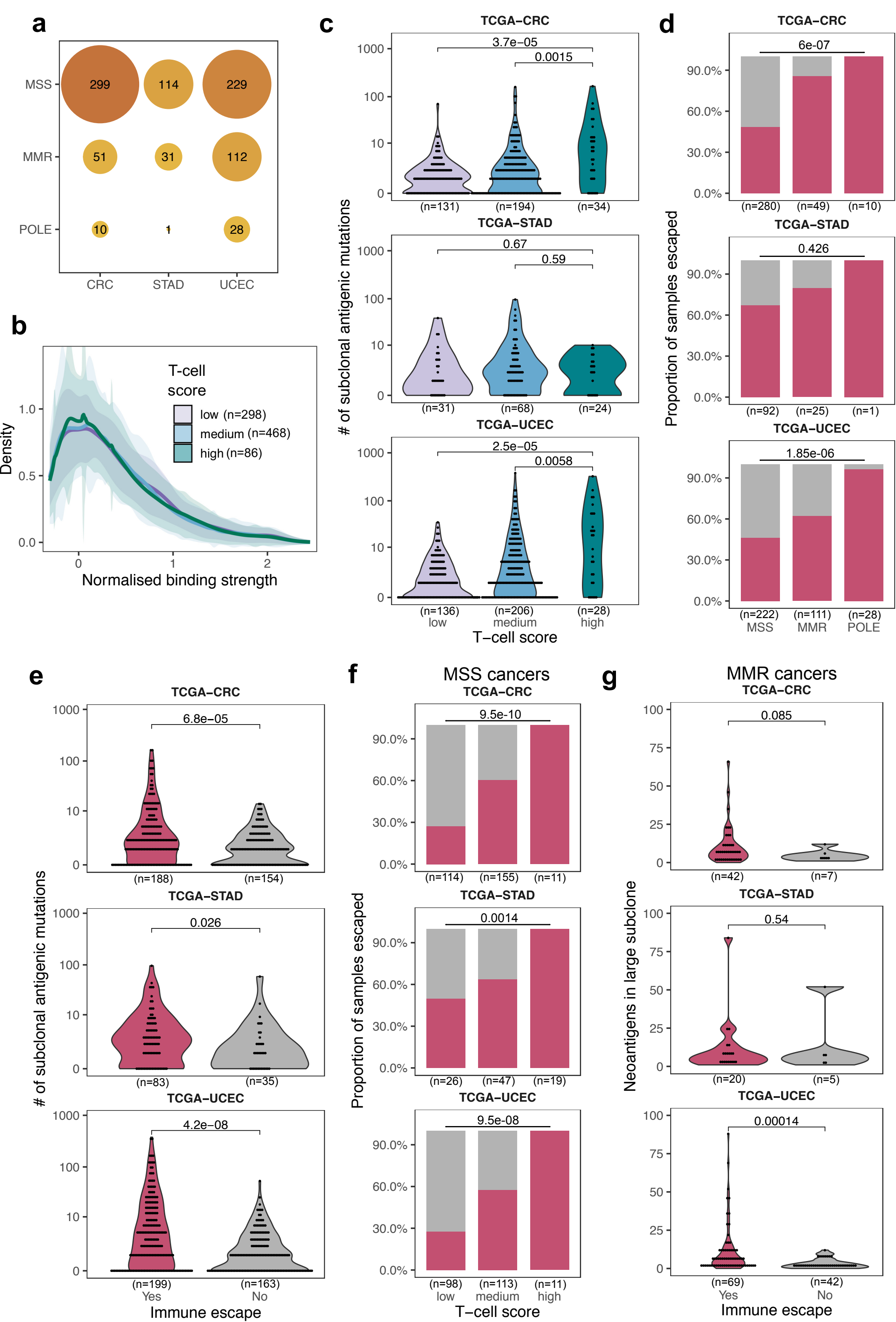
## CODE AVAILABILITY

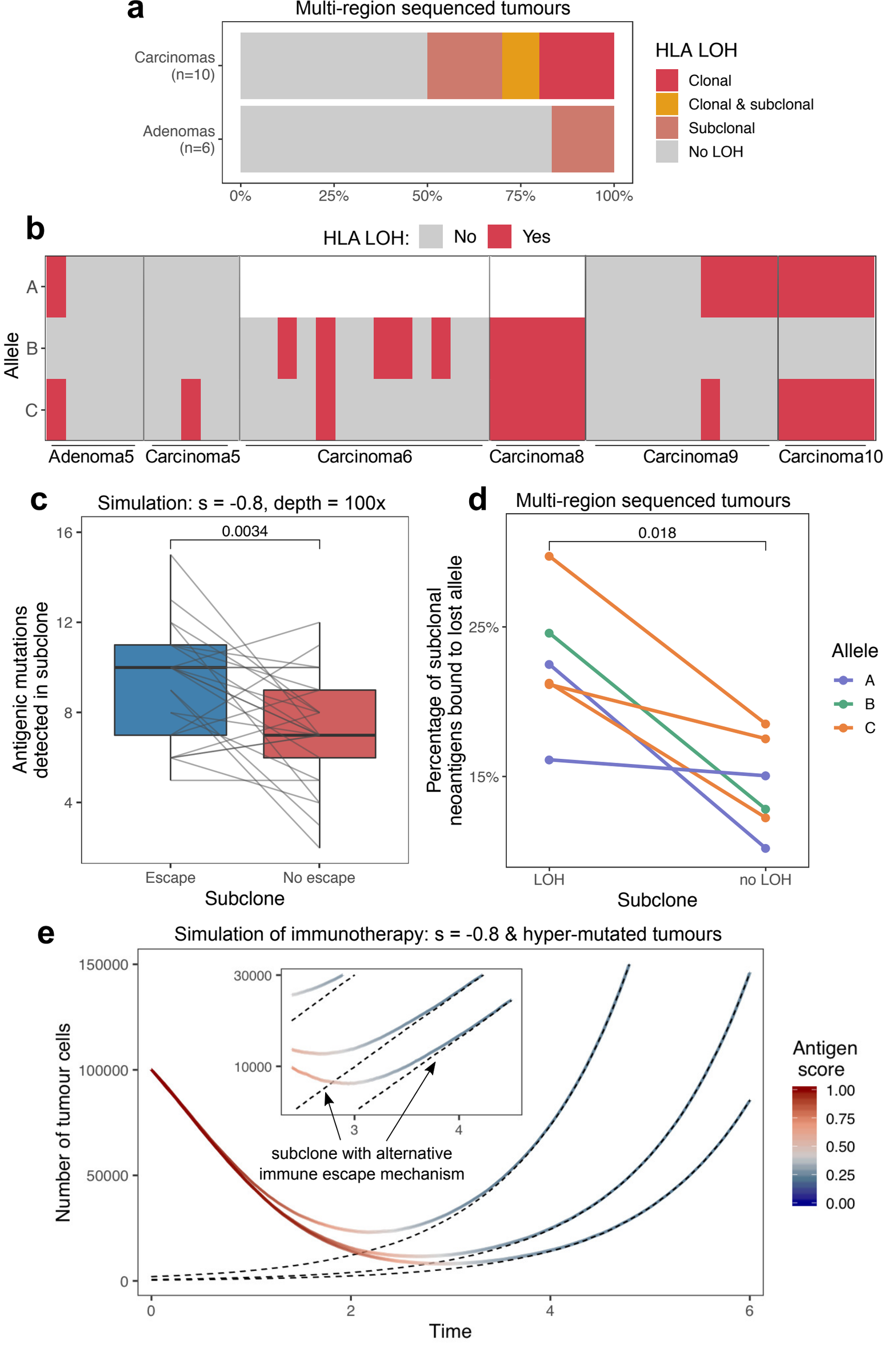
Julia (https://julialang.org/, version 0.5+) code implementing simulations of the tumor growth model is available from Zenodo (doi: 10.5281/zenodo.3601322)[61].

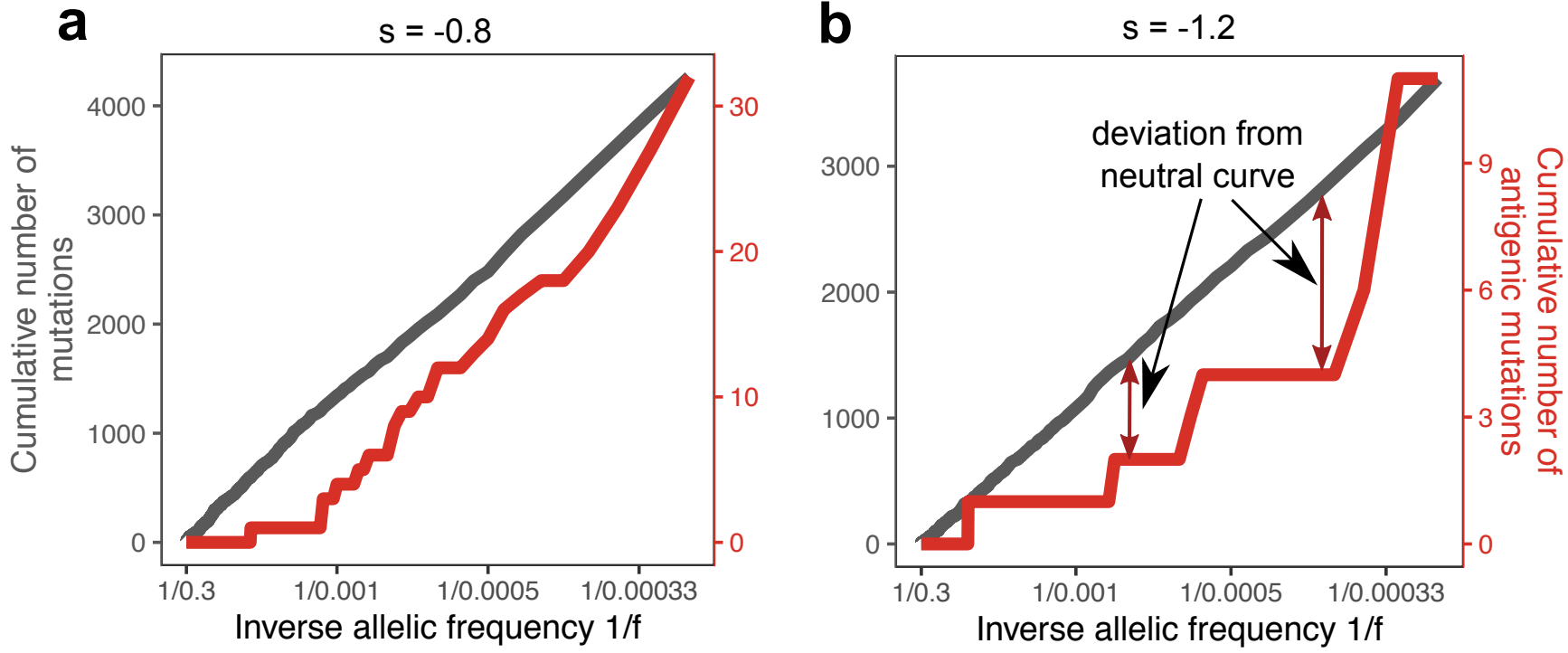
## REFERENCES (CONTINUED)

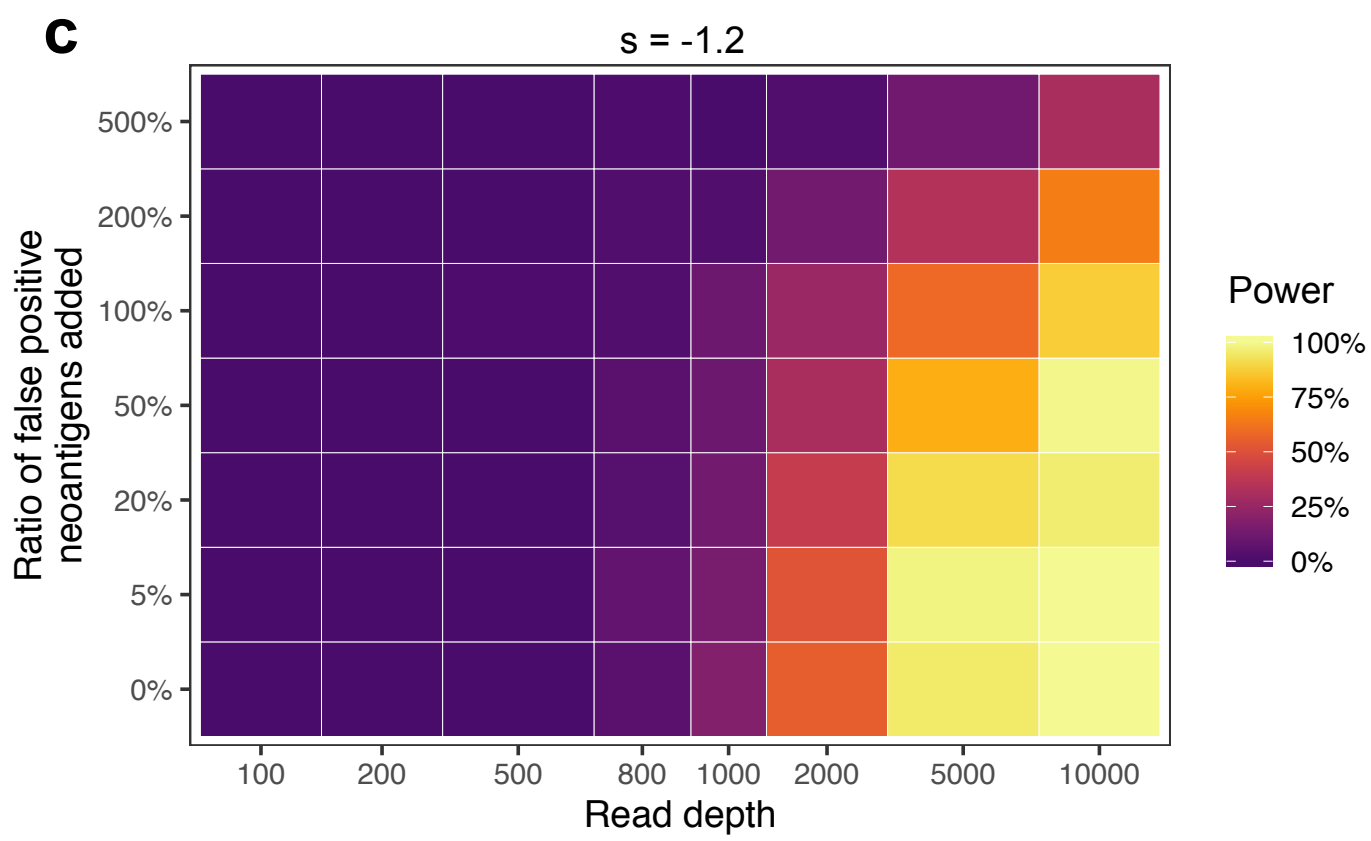

61. Lakatos, E. CloneGrowthSimulation. (2020). doi:10.5281/zenodo.3601322
62. Gillespie, D. T. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comput. Phys.* **22**, 403–434 (1976).
63. Grossman, R. L. *et al.* Toward a Shared Vision for Cancer Genomic Data. *N. Engl. J. Med.* **375**, 1109–1112 (2016).
64. Van Loo, P. *et al.* Allele-specific copy number analysis of tumors. *Proc. Natl. Acad. Sci.* **107**, 16910–16915 (2010).
65. The Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330–337 (2012).
66. Kautto, E. A. *et al.* Performance evaluation for rapid detection of pan-cancer microsatellite instability with MANTIS. *Oncotarget* **8**, 7452–7463 (2016).
67. Rimmer, A. *et al.* Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet.* **46**, 912–918 (2014).
68. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
69. Alsaab, H. O. *et al.* PD-1 and PD-L1 Checkpoint Signaling Inhibition for Cancer Immunotherapy: Mechanism, Combinations, and Clinical Outcome. *Front. Pharmacol.* **8**, 561 (2017).
70. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).
71. Favero, F. *et al.* Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. *Ann. Oncol. Off. J. Eur. Soc. Med. Oncol.* **26**, 64–70 (2015).
72. Jurtz, V. *et al.* NetMHCpan-4.0: Improved Peptide--MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. *J. Immunol.* **199**, 3360–3368 (2017).

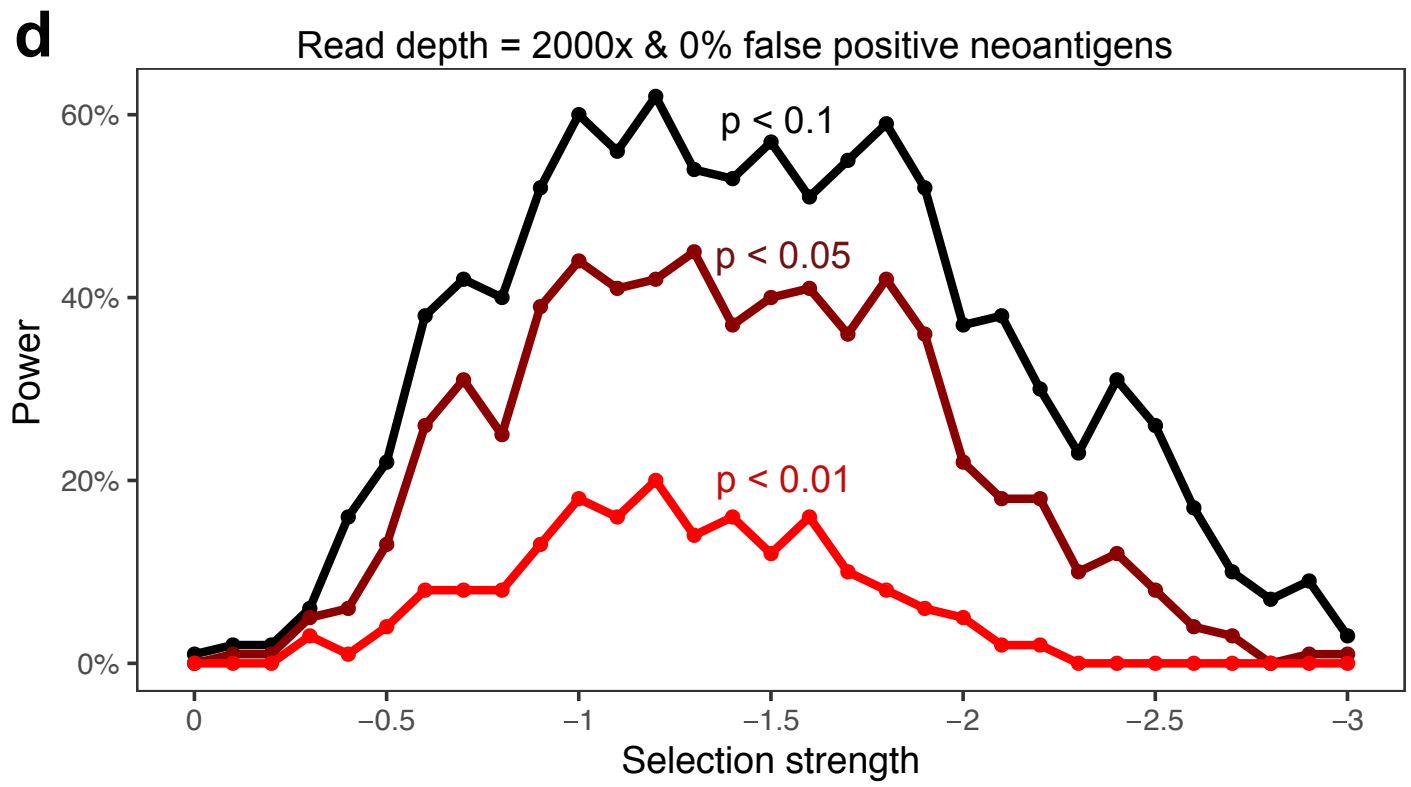
a
: neutral mutations
: antigenic mutations
Mutation antigenicity: $A_j$
Cell immunogenicity : $\sum A_j$ ~ Cell death
selection strength: s
Density
$T_m$
Antigenicity
Analysis:
Mutation is neoantigen if: $A_j > T_m$
Cell is immunogenic if: $\sum A_j > T_c$
b
s = -0.8
Number of tumour cells
Time
Antigen score
c
s = -0.8
CCF of most common neoantigen
Simulation index
d
s = -0.8
Density
Antigenicity
0.003
prior
posterior
e
s = -0.8 & hyper-mutated
Density
Number of tumour cells
Number of non−imm cells
Time
Maximum tumour size [no. of tumour cells]
f
s = -0.8
s = -0.8 & clonal escape
Proportion of tumours
antigen-cold
antigen-hot
antigen-warm
Final antigen score
g
probabilistic escape & detectable size
# of subclonal antigenic mutations
mean # of clonal ant. mutations
0.34
0.16
0.34
0.22
% of tumours
Selection strength
h
hyper-mutated & probabilistic escape & detectable size
# of subclonal antigenic mutations
mean # of clonal ant. mutations
0.70
4.50
2.96
2.22
% of tumours
Selection strength
no escape
clonal escape
Clonality of immune escape

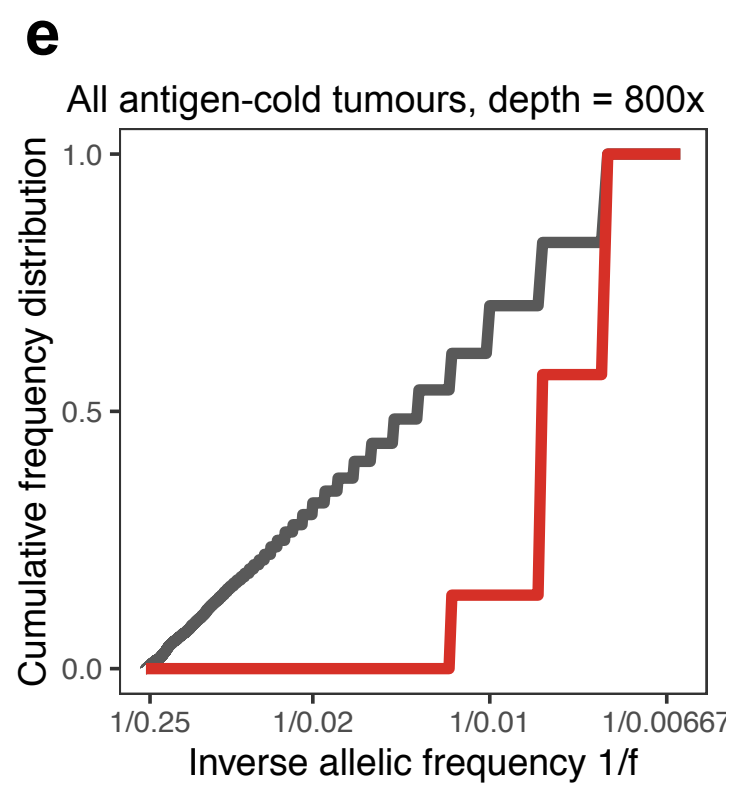

a
MSS
MMR
POLE
299
114
229
51
31
112
10
1
28
CRC
STAD
UCEC
b
T-cell score
low (n=298)
medium (n=468)
high (n=86)
Density
1.0
0.5
0.0
0
1
2
Normalised binding strength
c
TCGA–CRC
3.7e−05
0.0015
TCGA–STAD
0.67
0.59
TCGA–UCEC
2.5e−05
0.0058
# of subclonal antigenic mutations
1000
100
10
0
(n=131)
(n=194)
(n=34)
(n=31)
(n=68)
(n=24)
(n=136)
low
(n=206)
medium
(n=28)
high
T−cell score
d
TCGA–CRC
6e-07
TCGA–STAD
0.426
TCGA–UCEC
1.85e-06
Proportion of samples escaped
90.0%
60.0%
30.0%
0.0%
(n=280)
(n=49)
(n=10)
(n=92)
(n=25)
(n=1)
(n=222)
MSS
(n=111)
MMR
(n=28)
POLE
e
TCGA–CRC
6.8e−05
TCGA–STAD
0.026
TCGA–UCEC
4.2e−08
# of subclonal antigenic mutations
1000
100
10
0
(n=188)
(n=154)
(n=83)
(n=35)
(n=199)
Yes
(n=163)
No
Immune escape
f
MSS cancers
TCGA–CRC
9.5e-10
TCGA–STAD
0.0014
TCGA–UCEC
9.5e-08
Proportion of samples escaped
90.0%
60.0%
30.0%
0.0%
(n=114)
(n=155)
(n=11)
(n=26)
(n=47)
(n=19)
(n=98)
low
(n=113)
medium
(n=11)
high
T−cell score
g
MMR cancers
TCGA–CRC
0.085
TCGA–STAD
0.54
TCGA–UCEC
0.00014
Neoantigens in large subclone
100
75
50
25
0
(n=42)
(n=7)
(n=20)
(n=5)
(n=69)
Yes
(n=42)
No
Immune escape

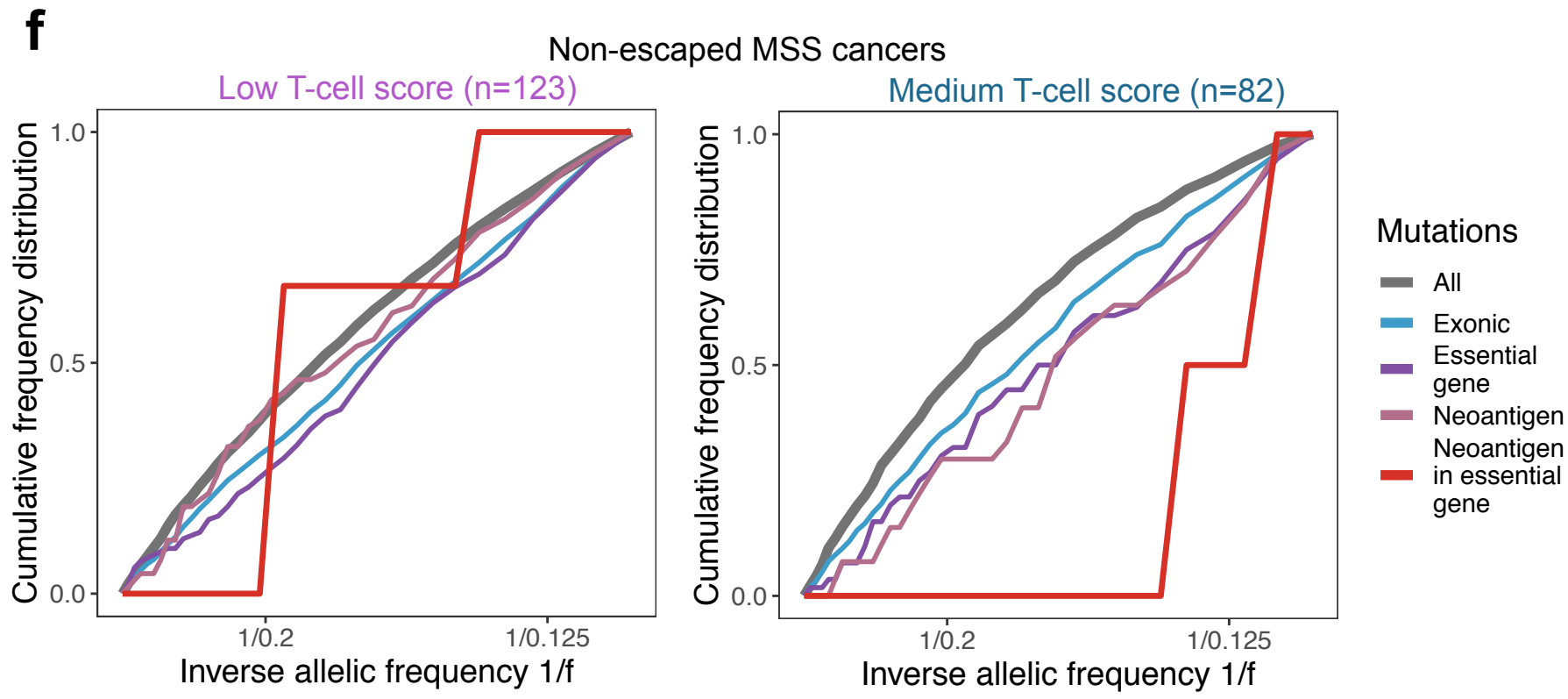

a
Multi-region sequenced tumours
Carcinomas (n=10)
Adenomas (n=6)
0%
25%
50%
75%
100%
HLA LOH
Clonal
Clonal & subclonal
Subclonal
No LOH
b
HLA LOH: No Yes
Allele
A
B
C
Adenoma5
Carcinoma5
Carcinoma6
Carcinoma8
Carcinoma9
Carcinoma10
c
Simulation: s = -0.8, depth = 100x
0.0034
Antigenic mutations detected in subclone
16
12
8
4
Escape
No escape
Subclone
d
Multi-region sequenced tumours
0.018
Percentage of subclonal neoantigens bound to lost allele
25%
15%
LOH
no LOH
Subclone
Allele
A
B
C
e
Simulation of immunotherapy: s = -0.8 & hyper-mutated tumours
Number of tumour cells
150000
100000
50000
0
30000
10000
3
4
subclone with alternative immune escape mechanism
0
2
4
6
Time
Antigen score
1.00
0.75
0.50
0.25
0.00

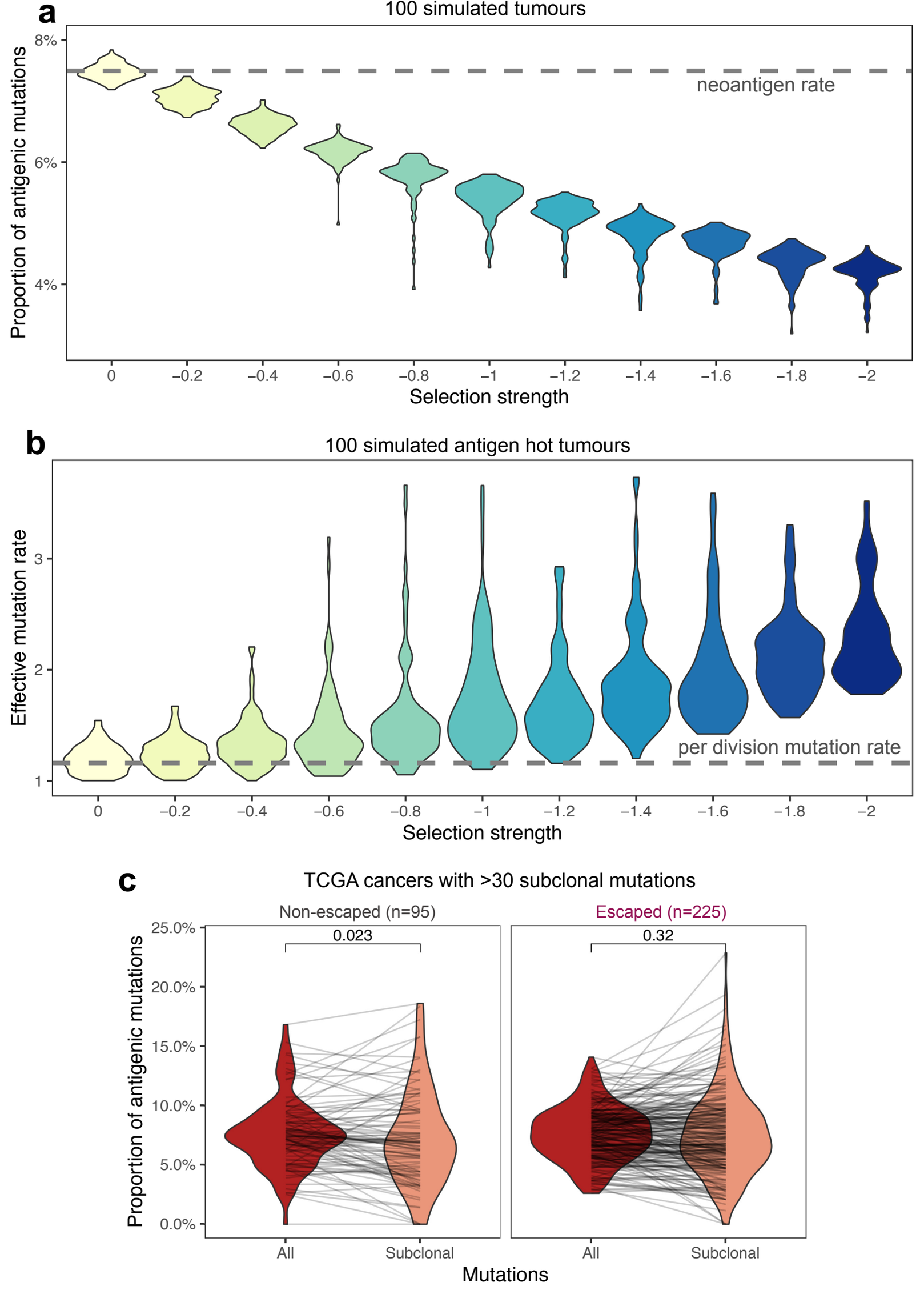

a
s = -0.8
Cumulative number of mutations
4000
3000
2000
1000
0
30
20
10
0
1/0.3
1/0.001
1/0.0005
1/0.00033
Inverse allelic frequency 1/f
b
s = -1.2
deviation from neutral curve
3000
2000
1000
0
9
6
3
0
Cumulative number of antigenic mutations
1/0.3
1/0.001
1/0.0005
1/0.00033
Inverse allelic frequency 1/f

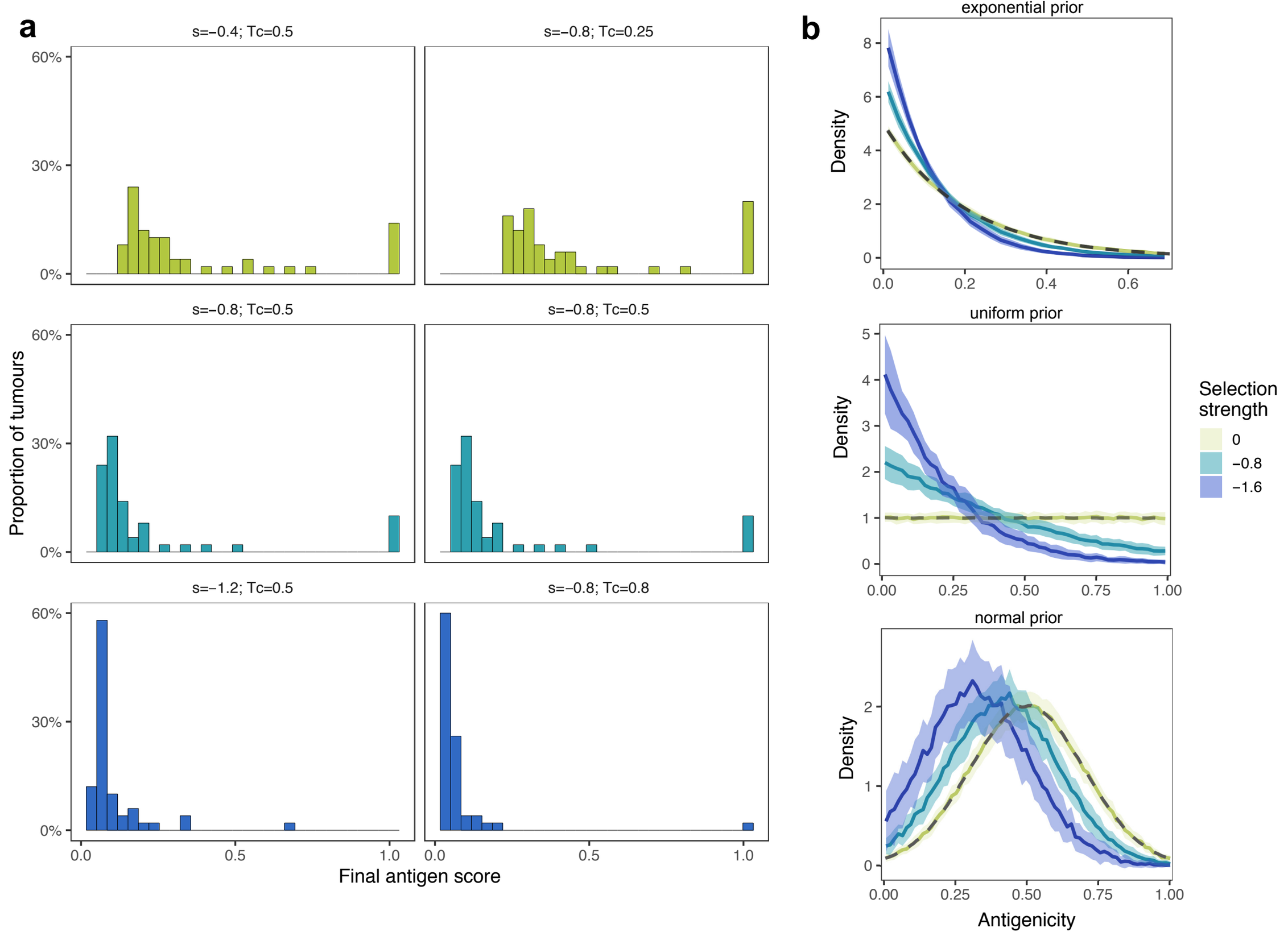

c
s = -1.2
Ratio of false positive neoantigens added
500%
200%
100%
50%
20%
5%
0%
100
200
500
800
1000
2000
5000
10000
Read depth
Power
100%
75%
50%
25%
0%


d
Read depth = 2000x & 0% false positive neoantigens
60%
40%
20%
0%
Power
p < 0.1
p < 0.05
p < 0.01
0
−0.5
−1
−1.5
−2
−2.5
−3
Selection strength


e
All antigen-cold tumours, depth = 800x
Cumulative frequency distribution
1.0
0.5
0.0
1/0.25
1/0.02
1/0.01
1/0.00667
Inverse allelic frequency 1/f


f
Non-escaped MSS cancers
Low T-cell score (n=123)
Medium T-cell score (n=82)
Cumulative frequency distribution
1.0
0.5
0.0
1/0.2
1/0.125
Inverse allelic frequency 1/f
Mutations
All
Exonic
Essential gene
Neoantigen
Neoantigen in essential gene


a
100 simulated tumours
Proportion of antigenic mutations
8%
6%
4%
neoantigen rate
0
−0.2
−0.4
−0.6
−0.8
−1
−1.2
−1.4
−1.6
−1.8
−2
Selection strength
b
100 simulated antigen hot tumours
Effective mutation rate
3
2
1
per division mutation rate
0
−0.2
−0.4
−0.6
−0.8
−1
−1.2
−1.4
−1.6
−1.8
−2
Selection strength
c
TCGA cancers with >30 subclonal mutations
Non-escaped (n=95)
Escaped (n=225)
0.023
0.32
Proportion of antigenic mutations
25.0%
20.0%
15.0%
10.0%
5.0%
0.0%
All
Subclonal
All
Subclonal
Mutations


a
s=−0.4; Tc=0.5
s=−0.8; Tc=0.25
s=−0.8; Tc=0.5
s=−0.8; Tc=0.5
s=−1.2; Tc=0.5
s=−0.8; Tc=0.8
60%
30%
0%
Proportion of tumours
0.0
0.5
1.0
Final antigen score
b
exponential prior
uniform prior
normal prior
Density
Antigenicity
Selection strength
0
−0.8
−1.6

growing population

constant size population

a

b

c

d

e

f

case (i)

case (ii)

case (iii)

s = -1 ; mu = 1

s = -0.1 ; mu = 10

s = -1 ; mu = 6

s = -1 ; mu = 10

s = -0.1 ; mu = 100

Mean antigenicity

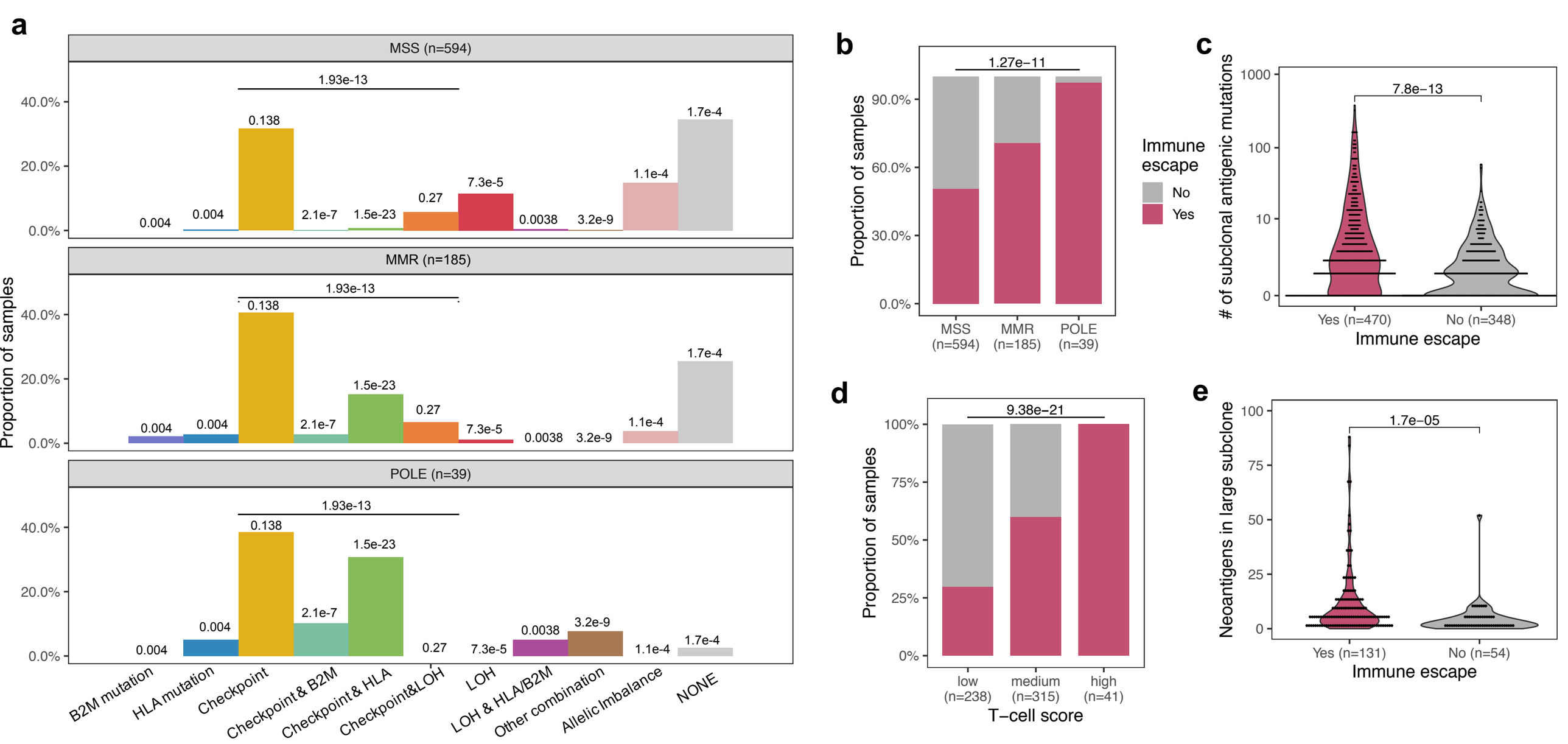
Time

Count

CCF of most common antigen

Cumulative distribution

Inverse allelic frequency (1/f)


a
MSS (n=594)
MMR (n=185)
POLE (n=39)
Proportion of samples
40.0%
20.0%
0.0%
1.93e-13
0.138
0.004
0.004
2.1e-7
1.5e-23
0.27
7.3e-5
0.0038
3.2e-9
1.1e-4
1.7e-4
B2M mutation
HLA mutation
Checkpoint
Checkpoint & B2M
Checkpoint & HLA
Checkpoint&LOH
LOH
LOH & HLA/B2M
Other combination
Allelic Imbalance
NONE
b
1.27e−11
Proportion of samples
90.0%
60.0%
30.0%
0.0%
MSS (n=594)
MMR (n=185)
POLE (n=39)
Immune escape
No
Yes
c
7.8e−13
# of subclonal antigenic mutations
1000
100
10
0
Yes (n=470)
No (n=348)
Immune escape
d
9.38e−21
Proportion of samples
100%
75%
50%
25%
0%
low (n=238)
medium (n=315)
high (n=41)
T−cell score
e
1.7e−05
Neoantigens in large subclone
100
75
50
25
0
Yes (n=131)
No (n=54)
Immune escape

s=−0.4; Tm=0.1

s=−0.4; Tm=0.2

s=−0.4; Tm=0.35

s=−0.8; Tm=0.1

s=−0.8; Tm=0.2

s=−0.8; Tm=0.35

s=−1.2; Tm=0.1

s=−1.2; Tm=0.2

s=−1.2; Tm=0.35

Cumulative distribution

1.0

0.5

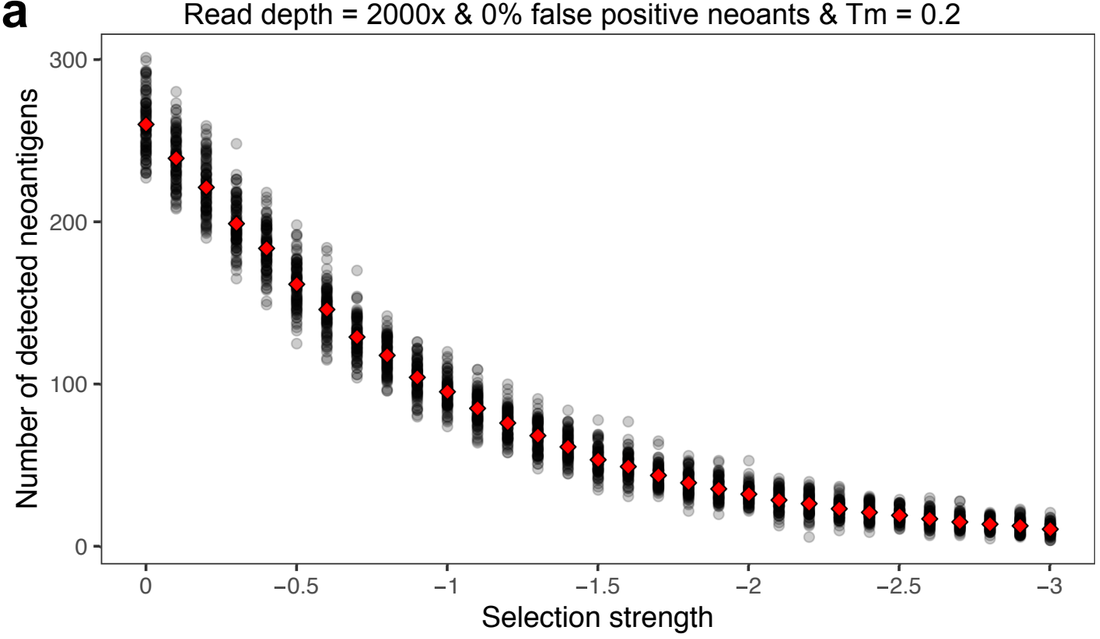
0.0

1/0.3

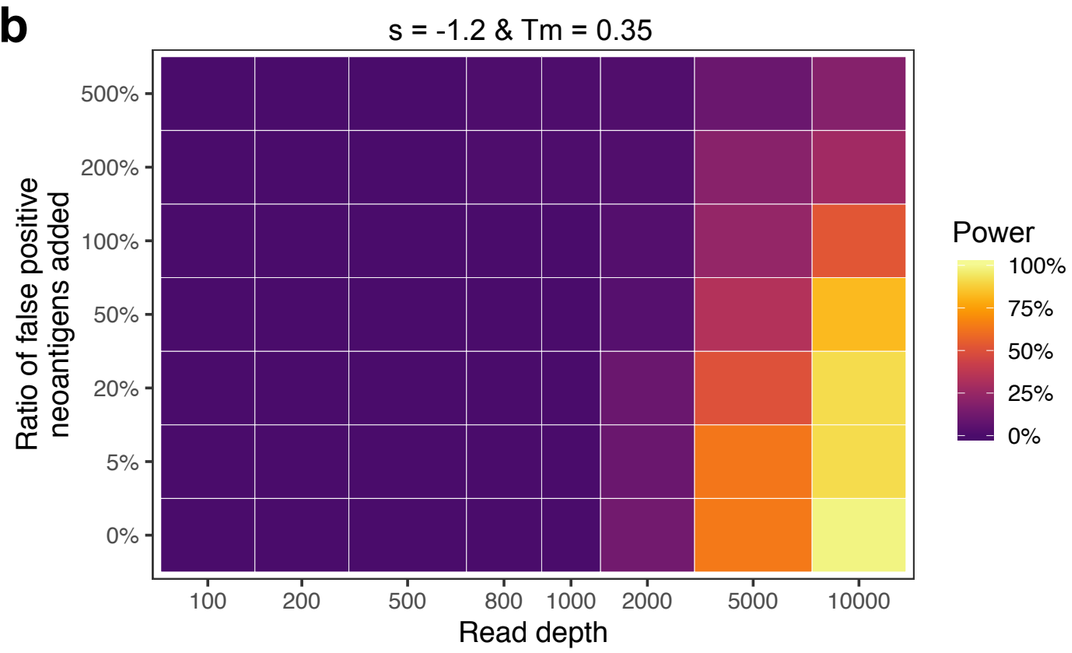
1/0.0005

Inverse allelic frequency (1/f)

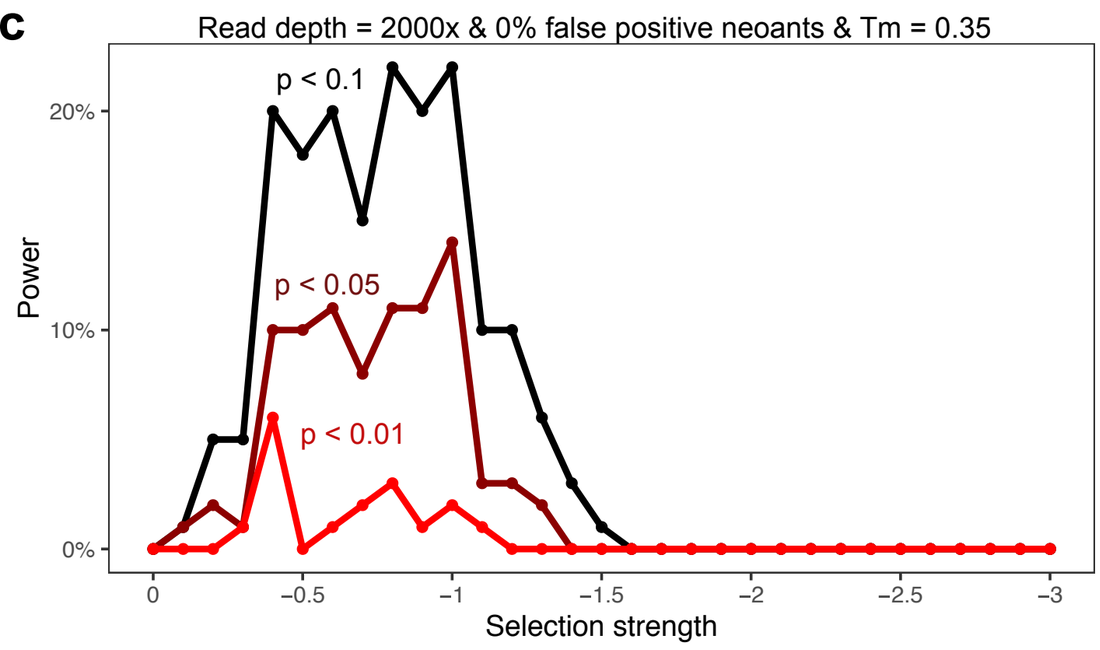
a


Read depth = 2000x & 0% false positive neoants & Tm = 0.2
Number of detected neoantigens
300
200
100
0
0
−0.5
−1
−1.5
−2
−2.5
−3
Selection strength

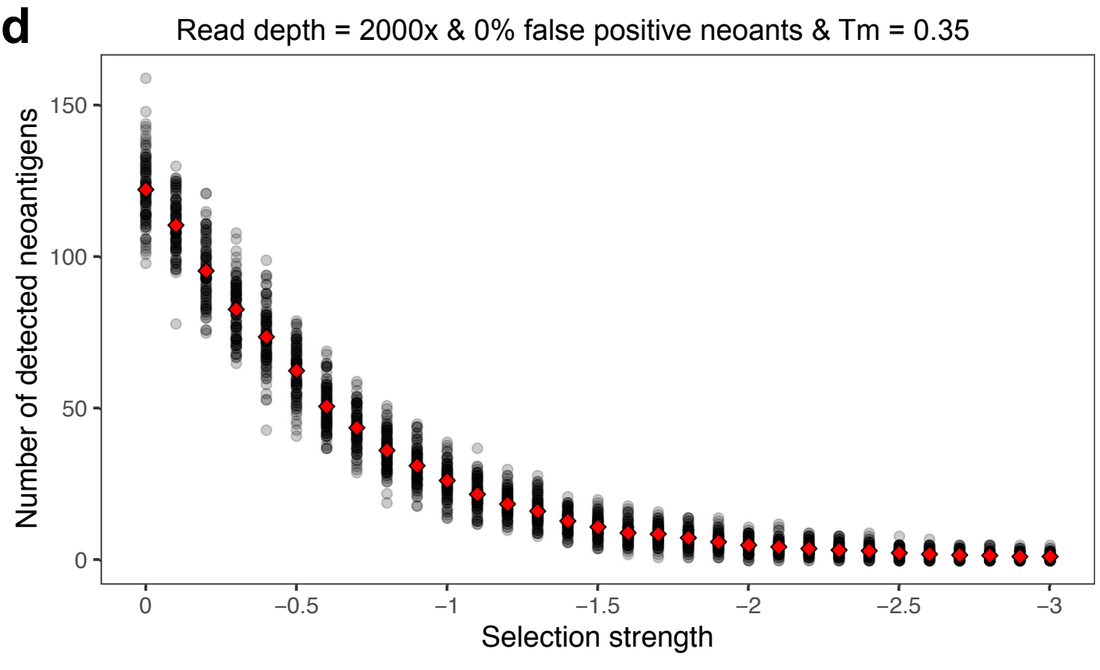
b


s = -1.2 & Tm = 0.35
Ratio of false positive neoantigens added
500%
200%
100%
50%
20%
5%
0%
100
200
500
800
1000
2000
5000
10000
Read depth
Power
100%
75%
50%
25%
0%

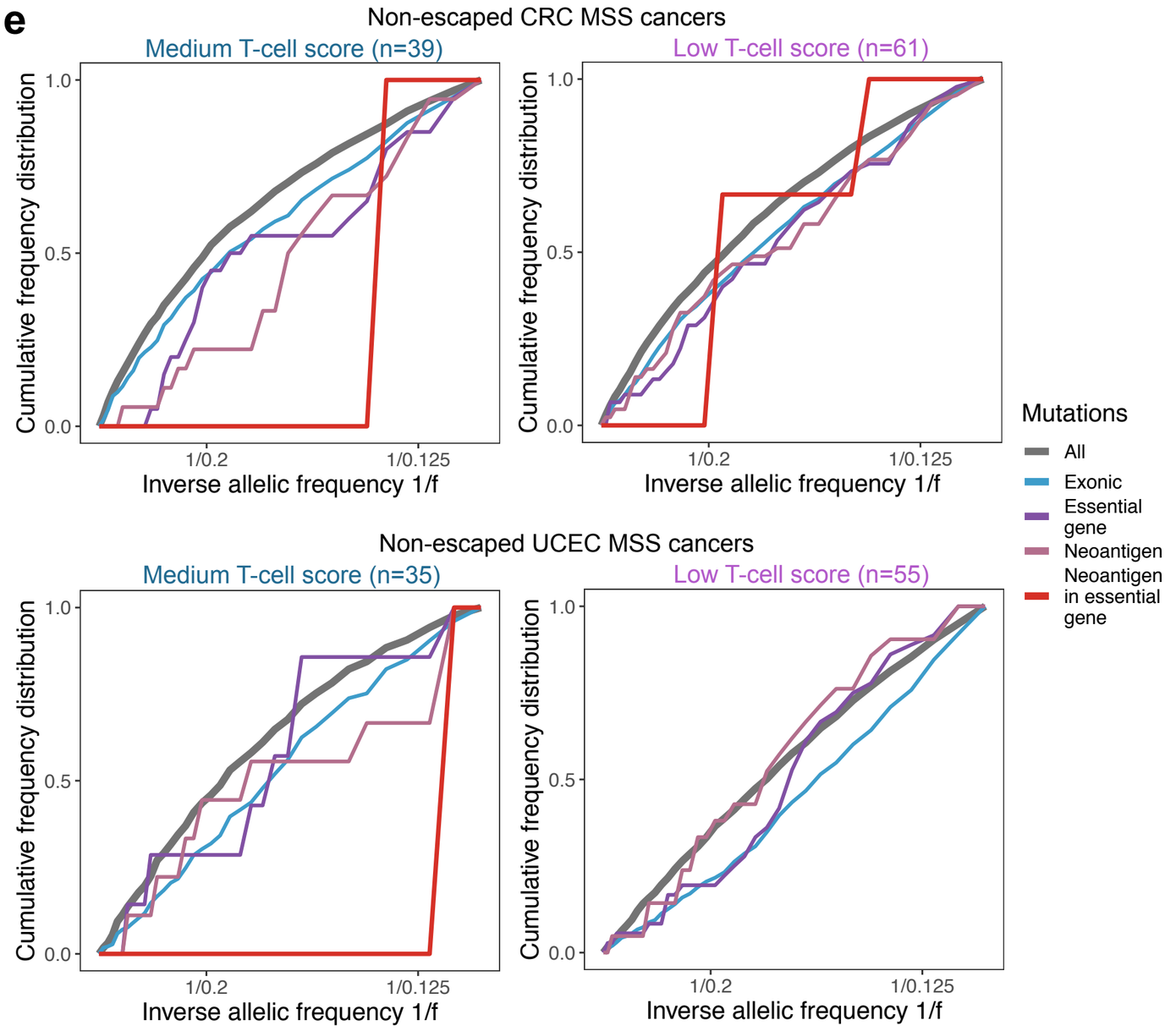
c


Read depth = 2000x & 0% false positive neoants & Tm = 0.35
Power
20%
10%
0%
p < 0.1
p < 0.05
p < 0.01
0
−0.5
−1
−1.5
−2
−2.5
−3
Selection strength

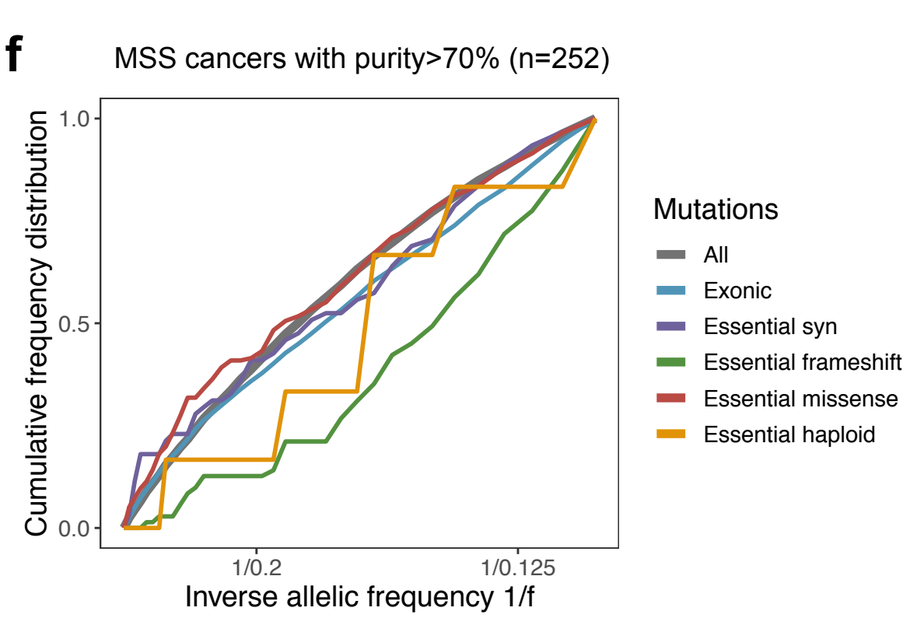
d

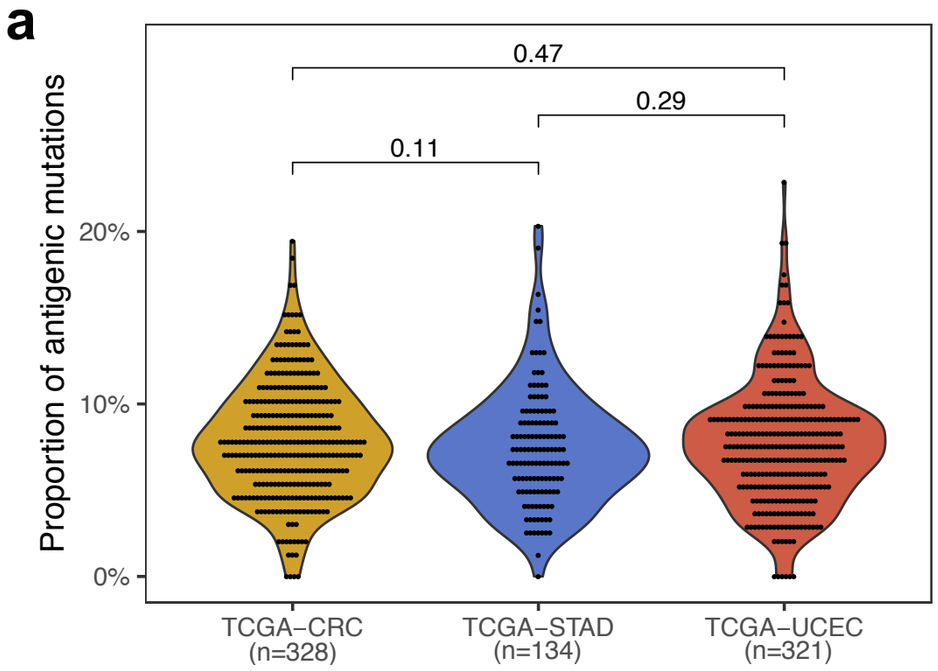

Read depth = 2000x & 0% false positive neoants & Tm = 0.35
Number of detected neoantigens
150
100
50
0
0
−0.5
−1
−1.5
−2
−2.5
−3
Selection strength

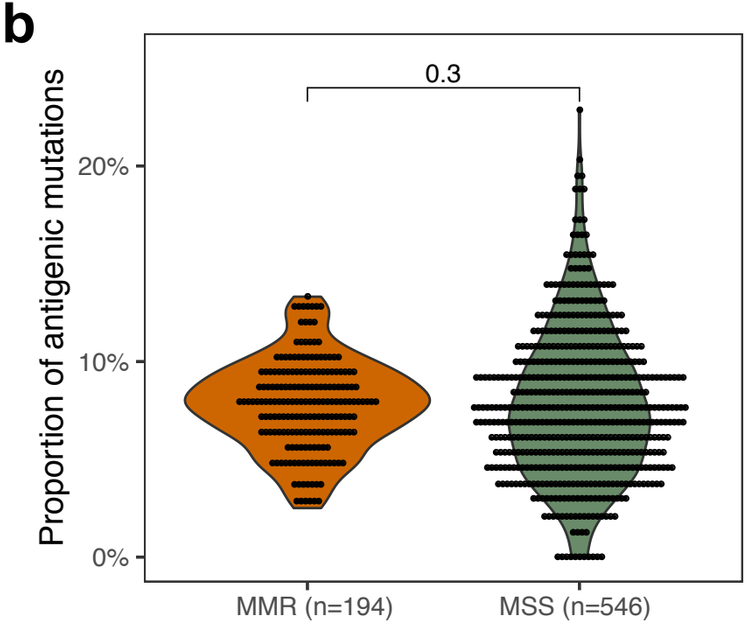
e

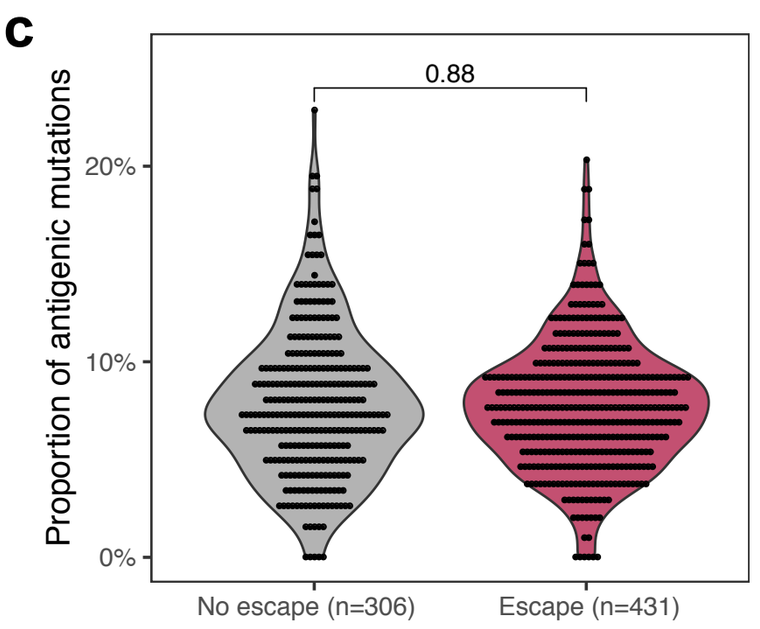

Non-escaped CRC MSS cancers
Medium T-cell score (n=39)
Low T-cell score (n=61)
Cumulative frequency distribution
1.0
0.5
0.0
1/0.2
1/0.125
Inverse allelic frequency 1/f
Non-escaped UCEC MSS cancers
Medium T-cell score (n=35)
Low T-cell score (n=55)
Mutations
All
Exonic
Essential gene
Neoantigen
Neoantigen in essential gene

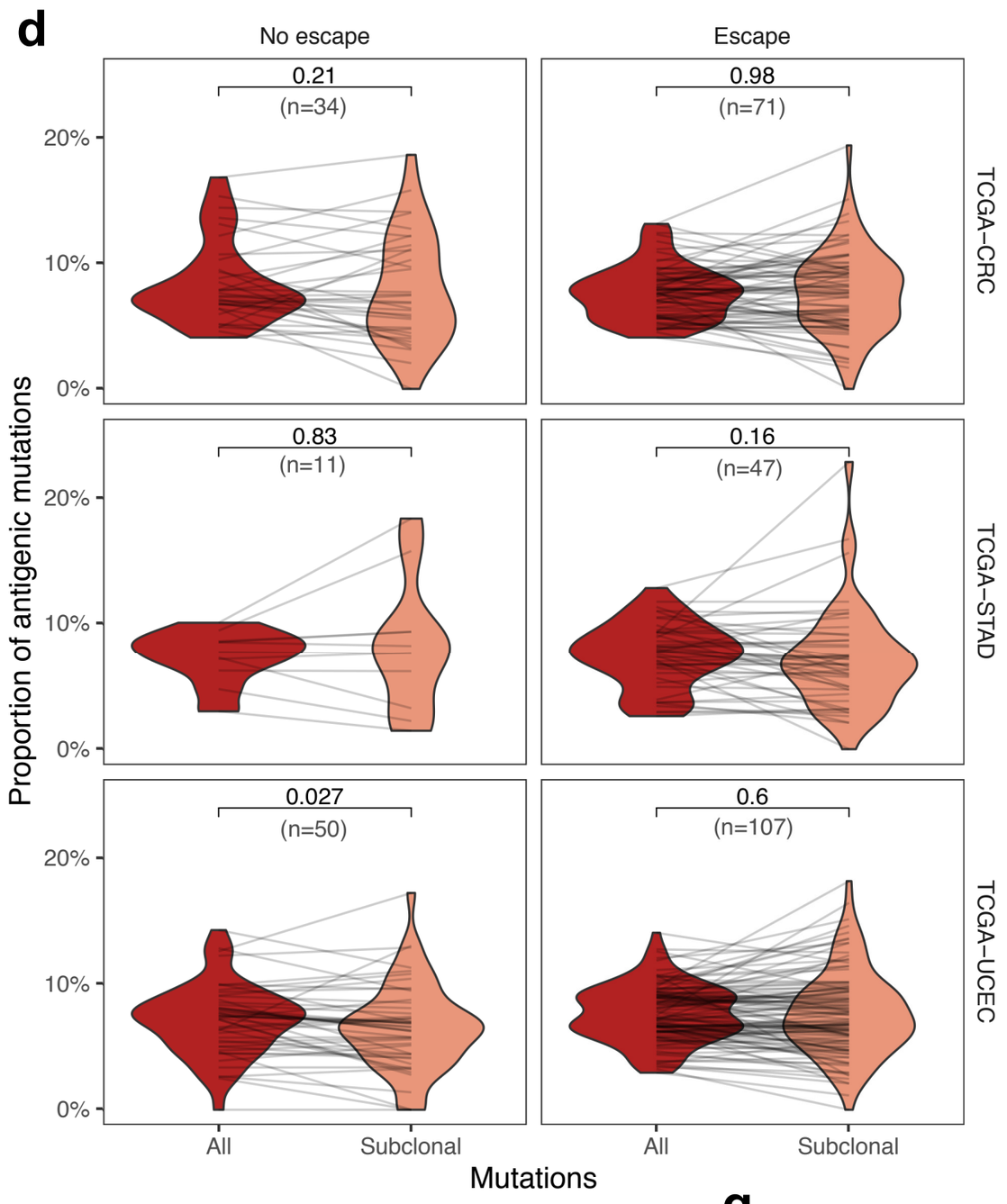
f

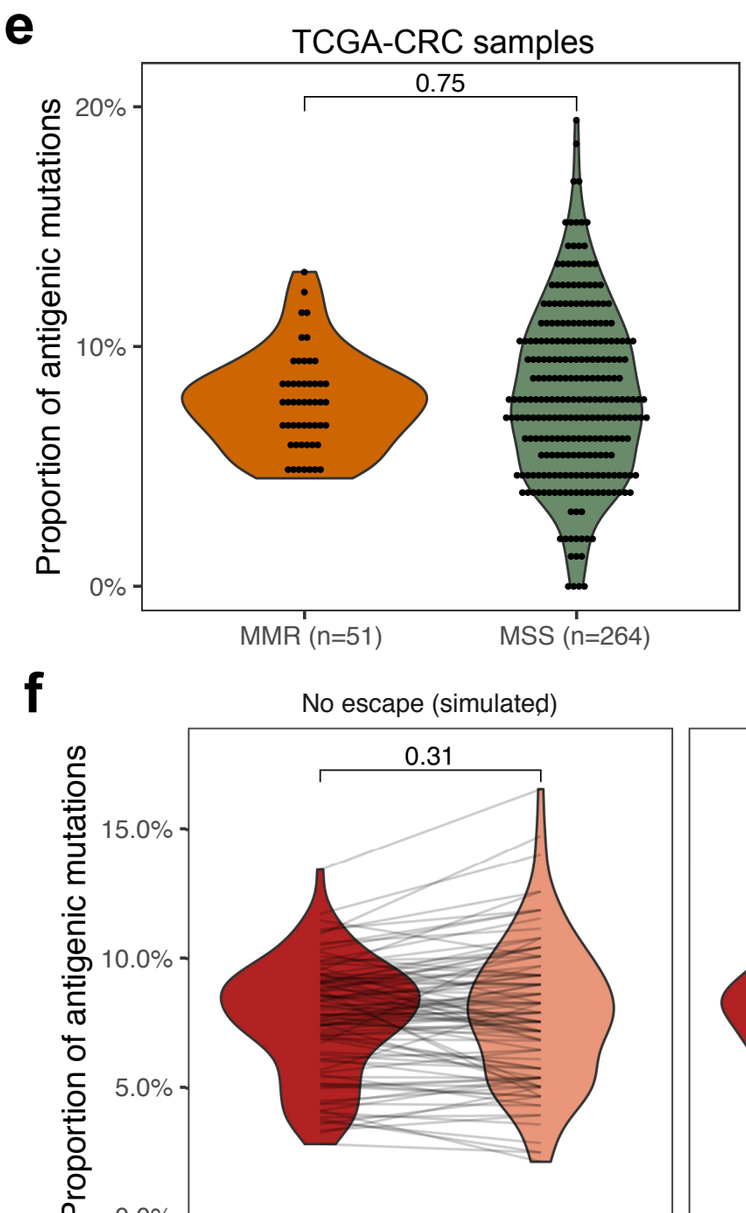

MSS cancers with purity>70% (n=252)
Cumulative frequency distribution
1.0
0.5
0.0
1/0.2
1/0.125
Inverse allelic frequency 1/f
Mutations
All
Exonic
Essential syn
Essential frameshift
Essential missense
Essential haploid


a
Proportion of antigenic mutations
0.47
0.29
0.11
20%
10%
0%
TCGA−CRC
(n=328)
TCGA−STAD
(n=134)
TCGA−UCEC
(n=321)

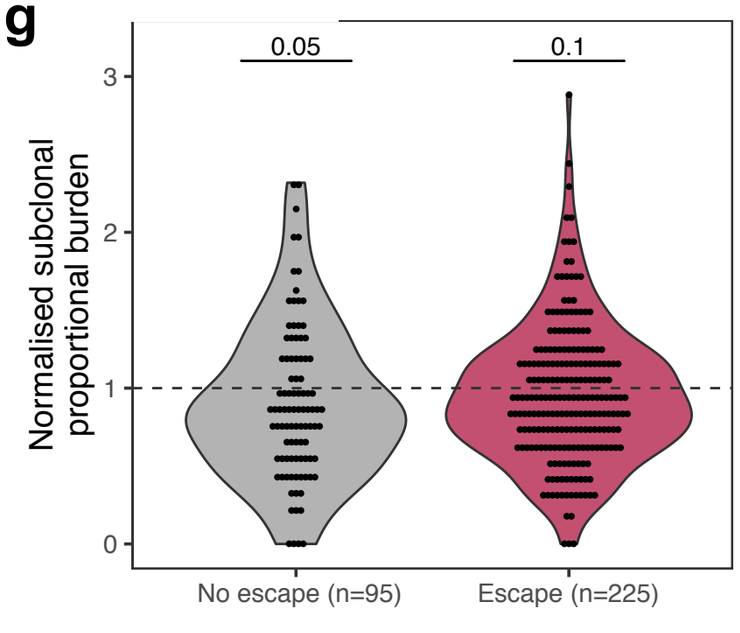

b
Proportion of antigenic mutations
0.3
20%
10%
0%
MMR (n=194)
MSS (n=546)

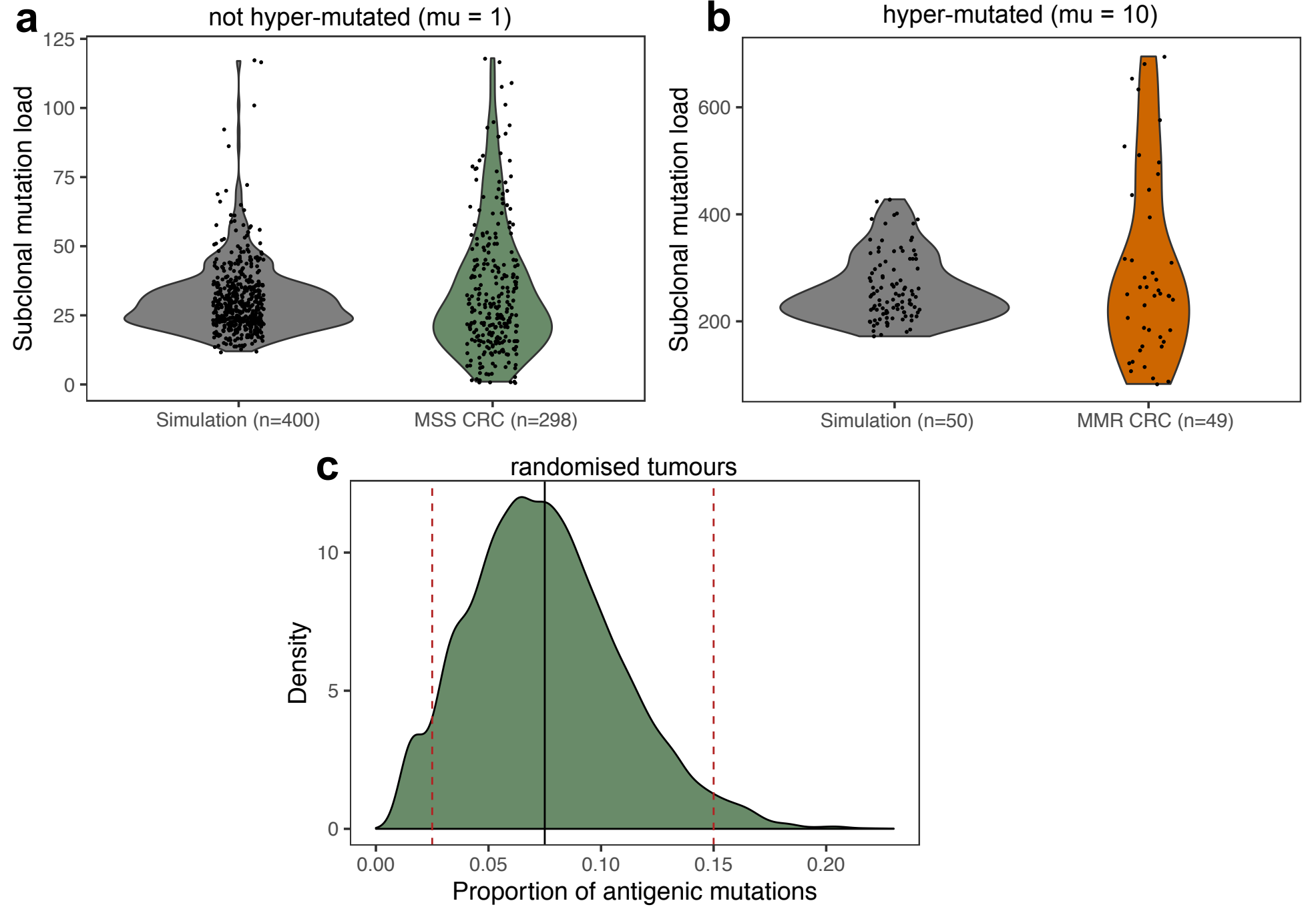

c
Proportion of antigenic mutations
0.88
20%
10%
0%
No escape (n=306)
Escape (n=431)

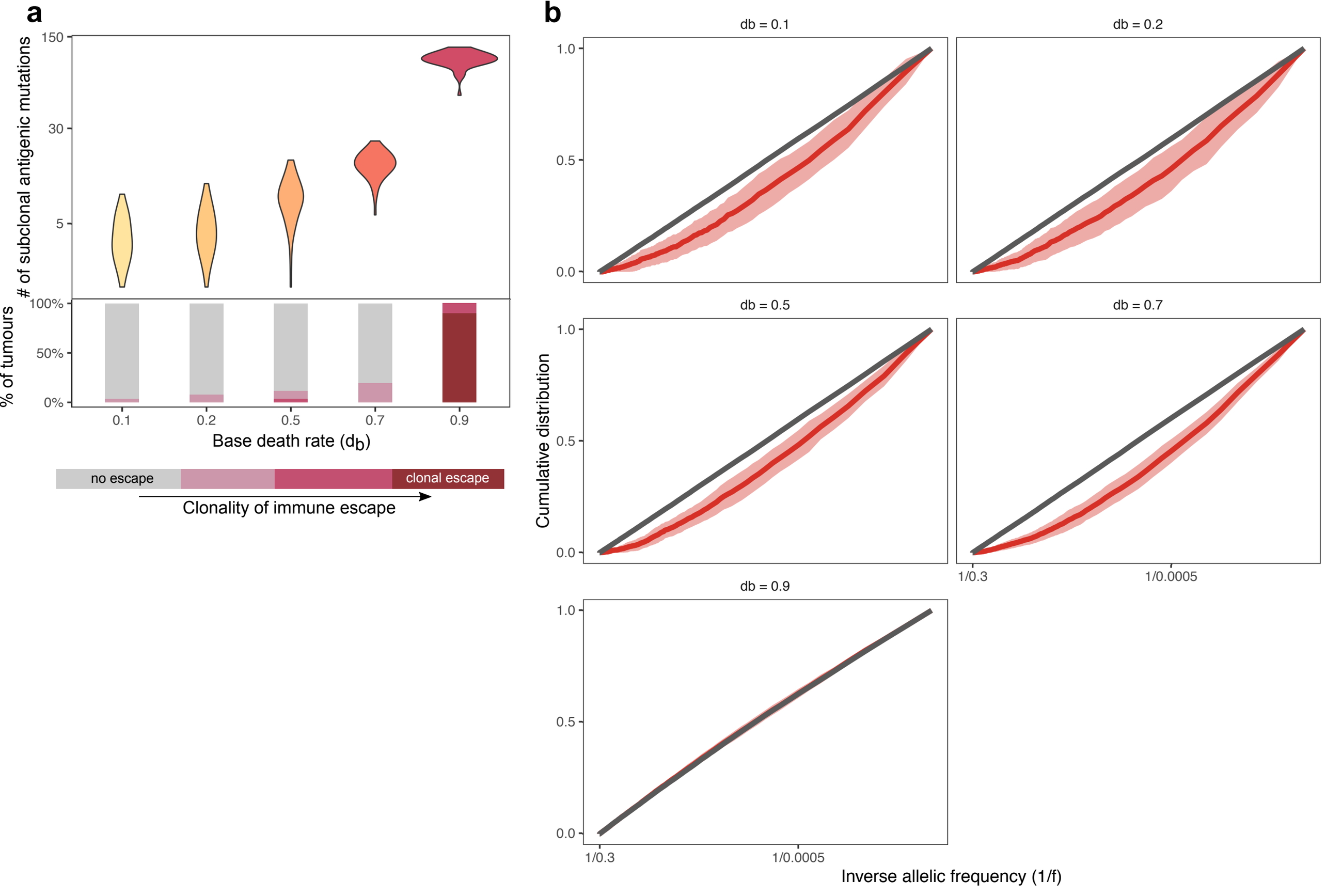

d
No escape
Escape
0.21
(n=34)
0.98
(n=71)
0.83
(n=11)
0.16
(n=47)
0.027
(n=50)
0.6
(n=107)
20%
10%
0%
TCGA−CRC
TCGA−STAD
TCGA−UCEC
Proportion of antigenic mutations
All
Subclonal
Mutations


e
TCGA-CRC samples
Synthetic CRC dataset
0.75
0.48
20%
10%
0%
Proportion of antigenic mutations
MMR (n=51)
MSS (n=264)
MMR (n=100)
MSS (n=100)


f
No escape (simulated)
Escape (simulated)
0.31
0.3
15.0%
10.0%
5.0%
0.0%
Proportion of antigenic mutations
All
Subclonal
Mutations


g
0.05
0.1
3
2
1
0
Normalised subclonal
proportional burden
No escape (n=95)
Escape (n=225)


a
not hyper-mutated (mu = 1)
Subclonal mutation load
125
100
75
50
25
0
Simulation (n=400)
MSS CRC (n=298)
b
hyper-mutated (mu = 10)
Subclonal mutation load
600
400
200
Simulation (n=50)
MMR CRC (n=49)
c
randomised tumours
Density
10
5
0
0.00
0.05
0.10
0.15
0.20
Proportion of antigenic mutations


a
# of subclonal antigenic mutations
150
30
5
% of tumours
100%
50%
0%
0.1
0.2
0.5
0.7
0.9
Base death rate ($d_b$)
no escape
clonal escape
Clonality of immune escape
b
db = 0.1
db = 0.2
db = 0.5
db = 0.7
db = 0.9
Cumulative distribution
1.0
0.5
0.0
1/0.3
1/0.0005
Inverse allelic frequency (1/f)

p_a = 0.025; s = −0.4

p_a = 0.075; s = −0.4

p_a = 0.15; s = −0.4

p_a = 0.025; s = −0.8

p_a = 0.075; s = −0.8

p_a = 0.15; s = −0.8

p_a = 0.025; s = −1.2

p_a = 0.075; s = −1.2

p_a = 0.15; s = −1.2

Cumulative distribution

1.0

0.5

0.0

1/0.3

1/0.0005

Inverse allelic frequency (1/f)