UNIVERSITY OF CALIFORNIA SAN DIEGO

Determinants for Effective Neoantigen Based Anti-Tumor Immune Response

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Bioinformatics and Systems Biology
with a Specialization in Biomedical Informatics

by

Andrea Bridget Castro

Committee in charge:

Professor Hannah Carter, Chair
Professor Ludmil Alexandrov, Co-Chair
Professor Olivier Harismendy
Professor Jill Mesirov
Professor Maurizio Zanetti

2022

The dissertation of Andrea Bridget Castro is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2022

DEDICATION

I dedicate this dissertation to my parents.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First and foremost, I would like to acknowledge Hannah Carter for being an outstanding advisor, scientist, and all-around wonderful person. Thank you for patience and guidance as I learned science and cluster best practices, for opening your home to me, and for listening to my thoughts and ideas. Without a doubt, you are the reason why my graduate experience was so positive, and I am very grateful that I had the opportunity to be a part of your lab.

I would like to acknowledge and thank my committee, Ludmil Alexandrov, Olivier Harismendy, Jill Mesirov, and Maurizio Zanetti, for taking the time to meet with me over the years and for providing invaluable feedback on my research. I would like to especially thank Maurizio Zanetti for being a wonderful collaborator and mentor, I have learned so much from you about immunology, and my approach to life is undoubtedly made richer by having known you.

I would like to acknowledge my current and former labmates – Kıvıl, Michelle, Meghana, James, Clarence, Rachel, Adam, Brian, Su, and Cameron – I am privileged to have worked with you all. I appreciate the scientific discussions we had as colleagues and cherish the conversations and outings that we had as friends. You all really are the complete package, and I could not imagine going through this journey with anyone else.

I would like to acknowledge my parents – Mom and Dad – your unconditional support and love enabled me to pursue my dreams and become the best version of myself. Thank you for leading me by example, letting me make my own choices, and helping me learn from my mistakes. Your generosity and selflessness are unmatched, and I am incredibly lucky to have you as my parents. I would also like to acknowledge my brother – Justin – I am so happy that we had the chance to go to the same school at the same time, I have delighted in watching you grow into the young man you are today. I will always see you as the 5-year-old with hair that stands up on end.

You have grown into a talented, intuitive, and well-read person, and I hold your advice in high esteem.

Last, but far from least, I would like to acknowledge my partner Gabe. Thank you for almost a decade of love and companionship, for introducing me to new hobbies and games, for sharing in my happiness, triumphs, highs and lows, and for encouraging me to be humble. Without you, I would not have had the courage to take the leap into the joys of cat ownership. I cannot wait to see what future adventures we will have.

The introduction and chapter forewords include reformatted reprints of the material as it appears in "Neoantigen Controversies" in *Annual Review of Biomedical Data Science*, 2020 by Andrea Castro, Maurizio Zanetti, and Hannah Carter. The dissertation author was a primary author of this review paper.

Chapter 1, in full, includes reformatted reprints of the material as it appears in "Elevated neoantigen levels in tumors with somatic mutations in the HLA-A, HLA-B, HLA-C and B2M genes" in *BMC Medical Genomics,* 2019 by Andrea Castro, Kivilcim Ozturk, Rachel Marty Pyke, Su Xian, Maurizio Zanetti, and Hannah Carter; "MHC-I genotype and tumor mutational burden predict response to immunotherapy" in *Genome Medicine,* 2020 by Aaron M Goodman, Andrea Castro, Rachel Marty Pyke, Ryosuke Okamura, Shumei Kato, Paul Riviere, Garrett Frampton, Ethan Sokol, Xinlian Zhang, Edward D Ball, Hannah Carter, and Razelle Kurzrock; "In silico analysis suggests less effective MHC-II presentation of SARS-CoV-2 RBM peptides: Implication for neutralizing antibody responses" in *Plos one*, 2021 by Andrea Castro, Kivilcim Ozturk, Maurizio Zanetti, and Hannah Carter. The dissertation author was a primary investigator and author of all three papers.

Chapter 2, in full, is a reformatted reprint of the material as it appears as "Strength of immune selection in tumors varies with sex and age" in *Nature Communications*, 2020 by Andrea Castro, Rachel Marty Pyke, Xinlian Zhang, Wesley Kurt Thompson, Chi-Ping Day, Ludmil B. Alexandrov, Maurizio Zanetti and Hannah Carter. The dissertation author was a primary investigator and author of this paper.

Chapter 3, in full, is a reformatted reprint of the material currently being prepared for submission for publication as "Source protein subcellular location is a novel feature to improve prediction of neoantigens for immunotherapy" by Andrea Castro, Saghar Kaabinejadian, Hooman Yari, William Hildebrand, Maurizio Zanetti, and Hannah Carter. The dissertation author was a primary investigator and author of this paper.

VITA

2017 University of California Los Angeles
Bachelor of Science, Microbiology Immunology and Molecular Genetics

2021 University of California San Diego
Master of Science, Computer Science

2022 University of California San Diego
Doctor of Philosophy, Bioinformatics and Systems Biology
with a Specialization in Biomedical Informatics

PUBLICATIONS

Castro, A., Kaabinejadian, S., Hildebrand, W., Zanetti, M. and Carter, H. Immunogenic potential of neopeptides depends on parent protein subcellular location. *bioRxiv* 2021.10.16.464599 (2021) doi:10.1101/2021.10.16.464599.

Zhang, T., Joubert, P., Ansari-Pour, N., Zhao, W., Hoang, P. H., Lokanga, R., Moye, A. L., Rosenbaum, J., Gonzalez-Perez, A., Martínez-Jiménez, F., **Castro, A.**, Muscarella, L. A., Hofman, P., Consonni, D., Pesatori, A. C., Kebede, M., Li, M., Gould Rothberg, B. E., Peneva, I., … Landi, M. T. Genomic and evolutionary classification of lung cancer in never smokers. *Nat. Genet.* **53**, 1348–1359 (2021).

Castro, A., Zanetti, M. and Carter, H. Neoantigen Controversies. (2021) doi:10.1146/annurev-biodatasci-092820-112713.

Castro A., Carter H., Zanetti M. Potential global impact of the N501Y mutation on MHC-II presentation and immune escape. doi:10.1101/2021.02.02.429431

Castro, A., Ozturk, K., Zanetti, M. and Carter, H. In silico analysis suggests less effective MHC-II presentation of SARS-CoV-2 RBM peptides: Implication for neutralizing antibody responses. *PLoS One* **16**, e0246731 (2021).

Talwar, J., Laub, D., Pagadala, M., **Castro, A.**, Lewis, M., Luebeck, GE., et al. Autoimmune Alleles at the Major Histocompatibility Locus Modify Melanoma Susceptibility. bioRxiv. 2021. p. 2021.08.12.456166. doi:10.1101/2021.08.12.456166

Almanza, G., Kouznetsova, V., Clark, AE., Olmedillas, E., **Castro, A.**, Tsigelny, IF., et al. Structure-selected RBM immunogens prime polyclonal memory responses that neutralize SARS-CoV-2 variants of concern. bioRxiv. 2021. p. 2021.10.01.462840. doi:10.1101/2021.10.01.462840

Pagadala, M., Wu, V. H., Pérez-Guijarro, E., Kim, H., **Castro, A.**, Talwar, J., Gonzalez-Colin, C., Cao, S., Schmiedel, B. J., Salem, R. M., Morris, G. P., Harismendy, O., Patel, S. P., Mesirov, J. P., Zanetti, M., Day, C.-P., Fan, C. C., Thompson, W. K., Merlino, G., … Carter, H. Germline

variants that influence the tumor immune microenvironment also drive response to immunotherapy. *bioRxiv* 2021.04.14.436660 (2021) doi:10.1101/2021.04.14.436660.

Castro, A.*, Pyke, R. M.*, Zhang, X., Thompson, W. K., Day, C.-P., Alexandrov, L. B., Zanetti, M., and Carter, H. Strength of immune selection in tumors varies with sex and age. *Nat. Commun.* **11**, 4128 (2020).

Dosset M., **Castro A.**, Carter H., Zanetti M. Telomerase and CD4 T Cell Immunity in Cancer. Cancers . 2020;12. doi:10.3390/cancers12061687

Castro A., Carter H. Mutagenic exposures shape immunotherapy responses. Nature Cancer. 2020;1: 1132–1133.

Goodman, A. M.*, **Castro, A.***, Pyke, R. M., Okamura, R., Kato, S., Riviere, P., Frampton, G., Sokol, E., Zhang, X., Ball, E. D., Carter, H., and Kurzrock, R. MHC-I genotype and tumor mutational burden predict response to immunotherapy. *Genome Med.* **12**, 45 (2020).

Castro, A.*, Ozturk, K.*, Pyke, R. M., Xian, S., Zanetti, M., and Carter, H. Elevated neoantigen levels in tumors with somatic mutations in the HLA-A, HLA-B, HLA-C and B2M genes. *BMC Med. Genomics* **12**, 107 (2019).

*These authors contributed equally to this work.

ABSTRACT OF THE DISSERTATION

Determinants for Effective Neoantigen Based Anti-Tumor Immune Response

by

Andrea Bridget Castro

Doctor of Philosophy in Bioinformatics and Systems Biology
with a Specialization in Biomedical Informatics

University of California San Diego 2022

Professor Hannah Carter, Chair
Professor Ludmil Alexandrov, Co-Chair

Cancer is a complex and heterogeneous disease that can sometimes be effectively targeted with precision and personalized medicine. Advances in next generation sequencing technologies have revolutionized our ability to catalog the landscape of somatic mutations in tumor genomes and have helped identify the role of MHC genotype in tumorigenesis. The MHC is a polymorphic protein complex that acts as a gatekeeper to cellular health by presenting peptides to T cells. This helps the immune system identify and eliminate infected or malignant cells. Tumor-specific neoantigens created from somatic mutations can be presented by the MHC complex, facilitating

immune control of developing tumors. However, tumor interaction with the immune system via this process can result in immunoediting, or pruning of subclones with easily presentable mutations, resulting in an immunologically invisible population of tumor cells, ultimately contributing to escape from immune surveillance. Therapeutic efforts to re-stimulate the immune system to eliminate tumors based on neoantigens have had less success than has been hoped for, and efforts to uncover genomic correlates of immunotherapy response that could serve as predictive biomarkers have had limited success. To identify key aspects of a neoantigen-based contribution to effective anti-tumor immune response, I first analyzed the role and limitations of MHC-based presentation of putative antigen. Next, I investigated the potential for sex and age, which have been tied to differences in immune response, to affect tumor-immune interactions by studying the landscape of presentable driver mutations in a pancancer cohort. Finally, through investigation of the factors that limit the immunogenic potential of tumor mutations, I found a novel factor, parent protein subcellular localization, that improves prediction of immunogenic neoantigens. This body of work provides new insight into key factors limiting the immunogenic potential of the neoantigen landscape in tumors, providing direction for future efforts to improve personalized immunotherapies.

INTRODUCTION

Cancer is a group of heterogeneous, complex diseases that involve uncontrolled cellular growth in any tissue. It has been described and studied for millennia, with the first account dating back to ~3000 BC (*1*), though the actual origin of the word is attributed to Hippocrates in ~400 BC. The 19th century marked the beginning of scientific oncology with Rudolf Virchow's discovery that cancer originates from cells (*2*). Soon thereafter, the discoveries of histocompatibility antigens (1948) (*3*), acquired immunological tolerance (1956) (*4*, *5*), and the development of the immune surveillance theory (1959) (*6*, *7*) contributed to the concept that tumors and the immune system are intrinsically linked (*8*). Based on these foundations, the past few decades have seen the induction of immunotherapy as an important component of treatment, joining the ranks of surgery, radiation, and chemotherapy (*2*, *9*). Significant advances in cancer prevention and treatment (*10–12*), have contributed to improved survival rates, with the risk of death from cancer dropping about 2% per year from 2015-2019 (*13*). Despite this considerable progress, cancer remains a difficult disease to treat, and current efforts are beginning to focus on exploiting patient-specific tumor biological characteristics, their interaction with the surrounding microenvironment (*14*), and how to best leverage existing patient anti-tumor immunity to achieve long-lasting benefit.

Immunotherapy co-opts an individual's own immune system to eliminate tumors. This approach to cancer therapy has sometimes resulted in remarkable responses (*15–17*), motivating significant investment into immunotherapy research and development. However, overall response rates have been disappointing, and the field is racing to understand why immunotherapies succeed or fail (*18*). Recent advances in bioinformatics including whole-exome, -genome, and -transcriptome sequencing, have enabled the rapid and systematic characterization of the tumor

antigenic landscape by cataloging tumor-specific mutant peptides (neopeptides) deriving from nonsynonymous mutations, frameshift mutations, and gene rearrangements (*19–21*). This has allowed researchers to probe the relationship between the MHC and genomic mutations at a level of resolution previously unachievable. Such efforts are revealing the ways in which the complex and dynamic interplay between tumors and the immune system can lead to short-lived or ineffective immune responses (*22*, *23*). Some of the emerging pitfalls that can limit the effectiveness of therapy are not yet widely appreciated but are critical to improving outcomes.

In this dissertation, I investigate the role of the major histocompatibility complex (MHC) in neoantigen presentation in disease. Aim 1 covers three separate analyses that begin by emphasizing the importance and utility of quantifying MHC class I neopeptide presentation in immune checkpoint blockade treated patients. Next, I study the impact of somatic mutations to B2M and HLA-A/B/C on MHC-I integrity, and quantify how this affects effective presentation of antigen in the TCGA. Finally, I explore a potential mechanism of immune evasion by the SARS-CoV-2 virus whereby poor MHC-II presentation of a critical B cell epitope may impact generation of neutralizing antibodies. Aim 2 investigates the extent to which sex and age alone and in conjunction with the MHC, affect the landscape of presentable driver mutations in the TCGA. Finally, in Aim 3, I build upon existing research that ties subcellular protein location to MHC presentation, and identify location as an important, novel feature for prediction of neoantigen immunogenicity. To do this, I developed a novel method for incorporating subcellular parent protein location in a machine learning model by performing dimensionality reduction on pretrained gene ontology cellular component embeddings. Overall, my research helps further define the role and limitations of MHC-I and MHC-II in cancer development, progression, and treatment, and

provides new insights into key factors limiting the immunogenic potential of the neoantigen landscape in tumors.

Acknowledgements

The introduction, in part, includes reformatted reprints of the material as it appears in "Neoantigen Controversies" in *Annual Review of Biomedical Data Science*, 2020 by Andrea Castro, Maurizio Zanetti, and Hannah Carter. The dissertation author was a primary author of this review paper.

References

1. S. I. Hajdu, A note from history: landmarks in history of cancer, part 1. *Cancer*. **117**, 1097–1102 (2011).

2. V. T. DeVita Jr, S. A. Rosenberg, Two hundred years of cancer research. *N. Engl. J. Med.* **366**, 2207–2214 (2012).

3. P. A. Gorer, N. Null, S. Lyman, N. Null, G. D. Snell, N. Null, J. B. S. Haldane, N. Null, Studies on the genetic and antigenic basis of tumour transplantation Linkage between a histocompatibility gene and "fused" in mice. *Proceedings of the Royal Society B: Biological Sciences*. **135**, 499–505 (1948).

4. R. E. Billingham, L. Brent, P. B. Medawar, Quantitative Studies on Tissue Transplantation Immunity. III. Actively Acquired Tolerance. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **239**, 357–414 (1956).

5. A. M. Silverstein, The curious case of the 1960 Nobel Prize to Burnet and Medawar. *Immunology*. **147**, 269–274 (2016).

6. M. Burnet, Cancer: a biological approach. III. Viruses associated with neoplastic conditions. IV. Practical applications. *Br. Med. J.* **1**, 841–847 (1957).

7. F. M. Burnet, The concept of immunological surveillance. *Prog. Exp. Tumor Res.* **13**, 1–27 (1970).

8. S. J. Oiseth, M. S. Aziz, Cancer immunotherapy: a brief review of the history, possibilities, and challenges ahead. *Journal of Cancer Metastasis and Treatment*. **3**, 250–261 (2017).

9. D. G. Maloney, A. J. Grillo-López, C. A. White, D. Bodkin, R. J. Schilder, J. A. Neidhart, N. Janakiraman, K. A. Foon, T. M. Liles, B. K. Dallaire, K. Wey, I. Royston, T. Davis, R. Levy, IDEC-C2B8 (Rituximab) anti-CD20 monoclonal antibody therapy in patients with relapsed low-grade non-Hodgkin's lymphoma. *Blood*. **90**, 2188–2195 (1997).

10. W. K. Hong, S. M. Lippman, L. M. Itri, D. D. Karp, J. S. Lee, R. M. Byers, S. P. Schantz, A. M. Kramer, R. Lotan, L. J. Peters, Prevention of second primary tumors with isotretinoin in squamous-cell carcinoma of the head and neck. *N. Engl. J. Med.* **323**, 795–801 (1990).

11. V. W. Wong, H. Reesink, H. H. Ip, P. N. Lelie, E. Reerink-Brongers, C. Y. Yeung, H. K. Ma, Prevention of the HBsAg carrier state in newborn infants of mothers who are chronic carriers of HBsAg and HBeAg by administration of hepatitis-B vaccine and hepatitis-B immunoglobulin: double-blind randomised placebo-controlled study. *Lancet*. **323**, 921–926 (1984).

12. S. A. Rosenberg, M. T. Lotze, L. M. Muul, S. Leitman, A. E. Chang, S. E. Ettinghausen, Y. L. Matory, J. M. Skibber, E. Shiloni, J. T. Vetto, Observations on the systemic administration

of autologous lymphokine-activated killer cells and recombinant interleukin-2 to patients with metastatic cancer. *N. Engl. J. Med.* **313**, 1485–1492 (1985).

13. ACS Medical Content, News Staff, Risk of dying from cancer continues to drop at an accelerated pace. *American Cancer Society* (2022), (available at https://www.cancer.org/latest-news/facts-and-figures-2022.html).

14. S. Maman, I. P. Witz, A history of exploring cancer in context. *Nat. Rev. Cancer*. **18**, 359–376 (2018).

15. J. J. Melenhorst, G. M. Chen, M. Wang, D. L. Porter, C. Chen, M. A. Collins, P. Gao, S. Bandyopadhyay, H. Sun, Z. Zhao, S. Lundh, I. Pruteanu-Malinici, C. L. Nobles, S. Maji, N. V. Frey, S. I. Gill, L. Tian, I. Kulikovskaya, M. Gupta, D. E. Ambrose, M. M. Davis, J. A. Fraietta, J. L. Brogdon, R. M. Young, A. Chew, B. L. Levine, D. L. Siegel, C. Alanio, E. J. Wherry, F. D. Bushman, S. F. Lacey, K. Tan, C. H. June, Decade-long leukaemia remissions with persistence of CD4+ CAR T cells. *Nature*. **602**, 503–509 (2022).

16. D. S. T. Magalhães, H. M. Magalhães, A. S. A. Mesquita, Long lasting complete response with immunotherapy in a metastatic bladder carcinoma: a case report. *Porto Biomed J.* **6**, e127 (2021).

17. J. Rao, J. Xia, W. Yang, C. Wu, B. Sha, Q. Zheng, F. Cheng, L. Lu, Complete response to immunotherapy combined with an antiangiogenic agent in multiple hepatic metastases after radical surgery for advanced gallbladder cancer: a case report. *Ann Transl Med*. **8**, 1609 (2020).

18. K. Esfahani, L. Roudaia, N. Buhlaiga, S. V. Del Rincon, N. Papneja, W. H. Miller Jr, A review of cancer immunotherapy: from the past, to the present, to the future. *Curr. Oncol.* **27**, S87–S97 (2020).

19. R. Marty, S. Kaabinejadian, D. Rossell, M. J. Slifker, J. van de Haar, H. B. Engin, N. de Prisco, T. Ideker, W. H. Hildebrand, J. Font-Burgada, H. Carter, MHC-I Genotype Restricts the Oncogenic Mutational Landscape. *Cell*. **171**, 1272–1283.e15 (2017).

20. R. Marty Pyke, W. K. Thompson, R. M. Salem, J. Font-Burgada, M. Zanetti, H. Carter, Evolutionary Pressure against MHC Class II Binding Cancer Mutations. *Cell*. **175**, 1991 (2018).

21. S. A. Shukla, M. S. Rooney, M. Rajasagi, G. Tiao, P. M. Dixon, M. S. Lawrence, J. Stevens, W. J. Lane, J. L. Dellagatta, S. Steelman, C. Sougnez, K. Cibulskis, A. Kiezun, N. Hacohen, V. Brusic, C. J. Wu, G. Getz, Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes. *Nat. Biotechnol.* **33**, 1152–1158 (2015).

22. E. Alspach, D. M. Lussier, A. P. Miceli, I. Kizhvatov, M. DuPage, A. M. Luoma, W. Meng, C. F. Lichti, E. Esaulova, A. N. Vomund, D. Runci, J. P. Ward, M. M. Gubin, R. F. V. Medrano, C. D. Arthur, J. M. White, K. C. F. Sheehan, A. Chen, K. W. Wucherpfennig, T. Jacks, E. R. Unanue, M. N. Artyomov, R. D. Schreiber, MHC-II neoantigens shape tumour immunity and response to immunotherapy. *Nature*. **574**, 696–701 (2019).

23. N. McGranahan, R. Rosenthal, C. T. Hiley, A. J. Rowan, T. B. K. Watkins, G. A. Wilson, N. J. Birkbak, S. Veeriah, P. Van Loo, J. Herrero, C. Swanton, TRACERx Consortium, Allele-Specific HLA Loss and Immune Escape in Lung Cancer Evolution. *Cell*. **171**, 1259–1271.e11 (2017).

CHAPTER 1: Limitations and role of peptide-MHC interaction in disease

1.1.1 Foreword

Tumor mutational burden (TMB), or the number of somatic mutations present in a patient's tumor, was recently approved by the FDA as a biomarker for immunotherapy response (*1*). High TMB has been generally associated with immunotherapy benefit in tumor types with higher numbers of somatic mutations such as melanoma (*2*) and tumors with DNA mismatch repair deficiency (*3*, *4*). TMB acts as a proxy for the number of neoantigens, which are relatively rare (*5*) and are important for immunotherapy response (*6*). However, high TMB seems to work well as a predictor of response in only a few tumor types (*7*), and studies have observed unexpected response in low-TMB tumors and non-response in high-TMB tumors. For example, non-small cell lung cancer (NSCLC) tends to have a relatively high TMB, but there is less evidence tying TMB to benefit (*8*–*10*). Interestingly, a recent study suggested that extremely high TMB is associated with a more dysfunctional T cell landscape in NSCLC (*11*), which further complicates relying solely on TMB as a biomarker for response. These conflicting reports suggest that TMB itself is an imperfect proxy for the presence of effective neoantigens.

A recent study that incorporates MHC genotype to more closely examine the relationship between putative neoantigens and response, has identified so-called motif neoepitopes, or mutated peptides with certain amino acid characteristics, that are beneficial only in patients with certain HLA supertypes (*12*). This finding is consistent with the idea that MHC presentation of mutation is required for immune response. However, counterintuitively, other studies have reported that predicted neoantigen load fails to predict outcomes better than TMB (*13*, *14*). This suggests that current methods to identify true neoantigens are still inaccurate (*15*, *16*) and highlights the need for improved methods to quantify presence of targetable tumor antigen. My research in Aim 1.1

assesses the utility of incorporating the ability of a patient's MHC-I to present driver mutations compared to TMB alone and may help explain why some high TMB tumors do not respond well to immunotherapy.

1.1.2 Abstract

Immune checkpoint blockade with antibodies inhibiting cytotoxic T lymphocyte-associated protein-4 (CTLA-4) and programmed cell death protein-1 (PD-1) or its ligand (PD-L1) can stimulate immune responses against cancer and has revolutionized the treatment of tumors. The influence of host germline genetics and its interaction with tumor neoantigens remains poorly defined. We sought to determine the interaction between tumor mutational burden (TMB) and the ability of a patient's major histocompatibility complex class I (MHC-I) to efficiently present mutated driver neoantigens in predicting response ICB. Comprehensive genomic profiling was performed on 83 patients with diverse cancers treated with ICB to determine TMB and HLA-I genotype. The ability of a patient's MHC-I to efficiently present mutated driver neoantigens defined by the PHBR score (with lower PHBR indicating more efficient presentation) was calculated for each patient. The median progression-free survival (PFS) for PHBR score < 0.5 vs. ≥ 0.5 was 5.1 vs. 4.4 months ($P = 0.04$). Using a TMB cutoff of 10 mutations/mb, the stable disease > 6 months/partial response/complete response rate, median PFS, and median overall survival (OS) of TMB high/PHBR high vs. TMB high/PHBR low were 43% vs. 78% ($P = 0.049$), 5.8 vs. 26.8 months ($P = 0.03$), and 17.2 months vs. not reached ($P = 0.23$), respectively. These findings were confirmed in an independent validation cohort of 32 patients. In conclusion, poor presentation of driver mutation neoantigens by MHC-I may explain why some tumors (even with a high TMB) do not respond to ICB.

1.1.3 Introduction

Immune checkpoint blockade (ICB) with antibodies inhibiting cytotoxic T lymphocyte-associated protein-4 (CTLA-4) and programmed cell death protein-1 (PD-1) (or its ligand (PD-L1)) can stimulate immune responses against cancer and has revolutionized the treatment of both solid (*17*) and hematologic malignancies (*18*). Durable remissions after ICB have been reported in patients with diverse advanced cancers including, but not limited to, melanoma (*19*), non-small cell lung cancer (NSCLC) (*20*), renal cell carcinoma (*21*), and Hodgkin lymphoma (*22*). Still, responses to ICB can be variable, toxicity can be serious, resistance is common (*23*), and hyperprogression can occur (*24*). Further, the majority of patients will not benefit from ICB, and there is a need to better select patients for treatment (*25*).

Multiple factors influence the immune response against tumors including tumor T cell infiltration, tumor mutational burden (TMB), PD-L1 expression, interferon signaling, mismatch repair (MMR) deficiency, tumor aneuploidy, and possibly the intestinal microbiota (*26*). Biomarkers that have entered clinical practice include PD-L1 expression measured by immunohistochemistry (IHC) (*27*), PD-L1 amplification (*28*), microsatellite instability (MSI) (*3*, *29*), and TMB (*30*, *31*).

Somatic mutations in tumors can be recognized by the immune system (*32*) resulting in tumor eradication. MMR-deficient/MSI-high tumors have 10 to 100 times as many somatic alterations as MMR-proficient tumors (*29*), resulting in exquisite sensitivity to ICB therapy (*3*). Most cancers harboring MMR alterations are associated with high TMB (*33*). In addition, many cancers harbor high TMB (10–20% depending on the definition of high TMB), even without MMR alterations (*24*, *34*). Higher TMB correlates with better treatment outcomes, including higher

response rates and longer progression-free survival (PFS) and overall survival (OS), in diverse cancers treated with immunotherapies (*30*).

Despite the improved efficacy of ICB in TMB-high tumors, approximately 40–60% of patients with a high TMB will not respond (*30*, *31*). To date, there is no sufficient way to predict which patients with high TMB will or will not respond to ICB. It has been hypothesized that tumors with high TMB and low PD-L1 expression might not respond as well to ICB; however, studies have demonstrated higher response rates and PFS in patients with high TMB versus low TMB, irrespective of PD-L1 expression (*35*).

Major histocompatibility complex class I (MHC-I) molecules, encoded by the human leukocyte antigen-I (HLA-I) locus, present intracellular peptides on the surface of both normal and tumor cells for recognition by CD8+ cytotoxic T cells (*36*). HLA-I genotype has been linked to a variety of different immune responses including infection (*37*), autoimmune diseases (*38*), and the graft versus host/tumor effect seen after allogeneic stem cell transplantation (*39*). There is accumulating experimental evidence suggesting that immunosurveillance shapes the mutational landscapes of cancers through the elimination of early tumor cells (*40–42*). In addition, the predicted number of MHC-I-associated neoantigens has been shown to be low in certain tumors suggesting immune-mediated elimination (*43*), and the anti-tumor activity of ICB is dependent on MHC-I presentation of specific tumor-derived peptides (*44*, *45*).

Marty *et al*. developed a residue-centric patient MHC-I presentation score (termed the Patient Harmonic-mean Best Rank (PHBR) score) that describes a person's ability to present specific cancer mutations to CD8+ T cells, and found that PHBR scores correlated with the likelihood of mutations to emerge in a patient's tumor (*46*). Poor presentation of a mutation across

patients was correlated with higher frequency among tumors. These results support that MHC-I genotype-restricted immunoediting shapes the mutational landscape of malignancies.

It has been suggested that the presence of a high-quality neoantigen is required for response to therapy (*47*) while a high burden of neoantigens has been associated with impaired anti-tumor immune activity (*48*); thus, we focused on neoantigen quality over quantity by using patient minimum PHBR score (i.e., best-presented mutation) to predict whether mutations observed in a patient's tumor are likely to generate effectively presented neoantigens. We assessed the ability of PHBR and TMB to predict response to ICB in diverse solid tumors.

1.1.4 Results

Eighty-three patients with 20 different solid malignancies were identified. The most common malignancies in the cohort included non-small cell lung cancer (NSCLC) ($N = 26$), cutaneous squamous cell carcinoma (SCC) ($N = 10$), and head and neck SCC ($N = 9$). Sixty-six patients were treated with PD-1/L1 monotherapy and 17 with combination therapy. The OBR (stable disease (SD) ≥ 6 months/partial and complete response (PR/CR)) was 43%. Thirty-two patients had at least one PHBR score of < 0.5 and fifty-one patients had a minimum score ≥ 0.5 (lower scores reflecting better neoantigen presentation). A minimum PHBR score ≥ 0.5 was significantly associated in univariate analysis with progressive disease ($P = 0.02$), non-cutaneous SCC malignancies ($P = 0.04$), and a TMB < 50 mutations/mb ($P = 0.05$).

In univariate analysis (**Table 1.1.2**), only higher TMB (≥ 10 mutations/mb) was associated with a better OBR. Caucasian ethnicity, high TMB, and a minimum PHBR score < 0.5 were all significantly associated with longer median PFS while male sex, Caucasian ethnicity, and high TMB were associated with longer median OS. The median PFS for low versus high PHBR scores was 5.1 vs. 4.4 months ($P = 0.04$) (**Figure 1.1.1**). The median PFS for high versus low TMB at

various thresholds (10, 20, 50) was 6.9 vs. 4.0 months (P = 0.001), 14.1 vs. 4.2 months (P = 0.01), and 26.8 vs. 4.4 months (P = 0.03), respectively.

Using a TMB cutoff of 10 mutations/mb, the OBR, median PFS, and median OS of TMB low/PHBR high vs. TMB high/PHBR low were 33% vs. 78% (P = 0.006), 3.5 vs. 26.8 months (P < 0.001), and 10.1 months vs. not reached (P = 0.008), respectively (**Figure 1.1.1 and Table 1.1.3**). Results remain when we exclude patients who had unknown TMB values. Patients with high TMB had greater OBR (43% vs. 78%, P = 0.049), greater PFS (5.8 vs. 26.8 months, P = 0.03), and greater median overall survival (17.2 months vs. not reached, P = 0.23) when accompanied by a well-presented mutation (low PHBR) than their counterparts with less well-presented mutations (high PHBR) (**Table 1.1.3, Figure S1.1.1**).

In a multivariable regression analysis (**Table 1.1.4**) of factors affecting outcome for patients treated with immunotherapy, high TMB (P = 0.01) and treatment with combination therapy (P = 0.006) were significantly associated with a higher OBR. Only high TMB was significantly associated with a prolonged median PFS (P = 0.01) and OS (P = 0.04). However, in stratified Cox regression, which allows for different hazard functions among strata (*49*)of PHBR in the higher TMB (≥ 10 mutations/mb) patients (N = 39), we found that a low PHBR score is significantly predictive of PFS (HR 0.39 (0.16–0.91), P = 0.03). Multivariable regression analysis in this cohort of 39 patients with high TMB showed that PHBR, but not TMB, was selected as an independent factor predicting both OBR and longer PFS (P = 0.049 and 0.03, respectively). In contrast, PHBR had no effect on PFS (P = 0.98) in patients with lower TMB (< 10 mutations/mb) (N = 38). Plotting Kaplan-Meier curves of patients based on lower or higher TMB and low or high PHBR found similar results in the general cohort (i.e., PHBR low versus high is associated with significant separation of the curves in patients with TMB ≥ 10 mutations/mb, but not in patients

with lower TMBs (**Figure 1.1.1**)). Finally, overall, Spearman correlation coefficient between TMB and PHBR was 0.31 with a P value of 0.01, consistent with a higher likelihood of carrying a low PHBR mutation when TMB is high (**Figure S1.1.2**).

Next, we evaluated the added value of PHBR with respect to TMB from another perspective. We first fit a logistic regression model relating OBR to all potential confounders, using a backward selection process where we removed confounders one at a time and compared models using Akaike Information Criterion (AIC) scores (*50*). We kept all confounders for which exclusion did not result in an increased AIC (i.e., the model better explained the data when the confounder was included). The retained confounders included MSI status, ethnicity, and the type of cancer each patient was diagnosed with. Then, we sequentially added TMB and PHBR to the regression model, using AIC once again to compare models (**Table S1.1.1**). We found that with the confounders and TMB in the model, the addition of the PHBR results in a reduction of AIC, indicating added explanatory power of PHBR even when TMB is included. In the final model with all the selected confounders, TMB and PHBR, the PHBR has a negative coefficient with a p-value of 0.08. The AUC values associated with the final models with confounders were 0.64 for both TMB and PHBR models alone, and 0.68 for the model with both TMB and PHBR.

To investigate the generalizability of our analyses across histologies, we revisited Kaplan-Meier analysis for progression-free survival within tumor types with at least 5 patients (NSCLC, SCC, head and neck, breast) and in all tumors excluding NSCLC and SCC, the two most common histologies. In each of these analyses, we observed that low versus high PHBR similarly stratified patients with high TMB. In addition, when we train a logistic regression classifier using the two most frequent histologies (N = 31), NSCLC and SCC, and predict response for the remaining patients (N = 46), we observe that the combination of PHBR and TMB better predicts OBR. These

results suggest that the information provided by TMB and PHBR generalizes beyond high mutation burden tumors such as SCC and NSCLC.

In an external validation cohort of 32 patients with NSCLC treated with pembrolizumab, the results were similar to those in our UCSD cohort: the OBR and median PFS of PHBR < 0.5 vs. ≥ 0.5 was 76% vs. 30% ($P = 0.02$) and 14.5 vs. 2.1 months ($P < 0.001$), respectively (**Figure 1.1.2**). Using a TMB cutoff of 10 mutations/mb, the median PFS of TMB high/PHBR high vs. TMB high/PHBR low was 8.1 months, versus not reached, respectively ($P = 0.02$) (**Figure 1.1.2**). OS data was not available for analysis.

Finally, we compared our findings in an aggregated high-TMB melanoma cohort (*2*, *51–53*) and a low TMB kidney cancer cohort (*54*). While minimum PHBR score did not significantly stratify melanoma patient overall or progression-free survival across all patients (**Figure 1.1.3A-B**), we did find, when also considering sex and age, that lower PHBR scores (i.e better presented mutations) were significantly associated with better overall and progression-free survival outcomes in high-TMB patients, consistent with our reported findings. As expected in the low TMB kidney tumors, there was no correlation between mutation burden and increased progression-free or overall survival (**Figure S1.1.1A-B**). Interestingly, while we did not see significant survival stratification with min-PHBR (**Figure 1.1.4C-D**), we did find that responders tended to have lower PHBR scores (i.e., better presented mutations) than non-responders, although the trends did not reach statistical significance.

1.1.5 Discussion

In a cohort of 83 patients with diverse solid tumors, we demonstrate that both TMB and efficient neoantigen presentation (defined by at least one PHBR score < 0.5) predict better response (as defined by SD ≥ 6 months/PR/CR rate) and longer PFS and OS after treatment with

ICB. This finding was confirmed in an independent cohort of 32 patients with NSCLC treated with PD-1 blockade. Further, by incorporating the PHBR score, we were able to identify a group of higher TMB tumors (≥ 10 mutations/mb) that are less likely to benefit from ICB. Specifically, patients with tumors that poorly present driver neoantigens are less likely to respond to ICB, even in tumors with a higher mutational load. Numerous studies show that a significant proportion of patients with a higher TMB do not respond to ICB and there is a need to better identify this group of patients (*30*, *31*, *34*).

Chowell *et al*. demonstrated that HLA-I homozygosity and somatic loss of heterozygosity (LOH) are predictive of poor outcomes in two independent cohorts treated with ICB (*55*). In addition, McGranahan *et al.* observed that 40% of early-stage NSCLC tumors had HLA loss of heterozygosity (*47*). It was hypothesized that patients homozygous in at least one HLA-I locus would be predicted to present a smaller and less diverse tumor-derived neoantigen repertoire to CD8+ cytotoxic T cells and that the diversity of HLA molecules in a given patient influences the selection and clonal expansion of T cells following ICB (*56*).

Our report differs from the Chowell *et al.* in several ways. We assessed patient-specific MHC-I ability to bind to tumor neoantigens (PHBR score), not HLA-I diversity. Furthermore, by evaluating the interaction between TMB and the PHBR score, we demonstrated that tumors that present neoantigens efficiently respond to ICB, at least in the case of higher TMB (≥ 10 mutations/mb). However, in patients with lower TMB, the presentation of neoantigens as reflected by PHBR had no association with outcome. We hypothesize that, when there are multiple neoantigens produced by the mutanome (i.e., in patients with higher TMB), there is the opportunity for MHC-I to present them (or at least one of them) in such a way that is critical to the response. However, when there are few neoantigens, the opportunity to present them may be diminished to

such an extent that the PHBR is not impactful. Additional studies will be required to better understand the neoantigen landscape as it relates to host anti-tumor immunity, in addition to the optimal method to combine information across multiple neoantigen for predicting response to therapy.

In our study, all data gathered to identify possible biomarkers to ICB was obtained via one NGS test at one time point. Prediction scores and gene signatures that take into count numerous variables including T cell infiltration into tumors, mutational load, and PD-L1 level have also been developed (*57*, *58*). Here we show that, with further validation, the PHBR score and TMB obtained via NGS, both of which are easy to assay, provide the ability to deliver data in real time for clinicians to make treatment decisions.

Our study has several limitations. It was a retrospective study that included a non-uniform group of patients with different malignancies treated with different checkpoint inhibitors. However, similar results were obtained in our validation cohort of NSCLC all treated with the same therapy. Our study excluded melanoma and included only small subsets of patients with individual tumor types; while our specific analyses for tumor types with ≥ 5 patients and leave-one-out analyses suggest generalizability, much larger sample sizes will be required to determine whether these findings generalize to specific histologies. Our study did not assess T cell receptor (TCR) specificity and diversity. TCR specificity for MHC-I/peptide complex is essential for CD8+ T cell cellular-mediated cytotoxicity. A strong correlation between TCR CDR3 diversity and TMB has been reported (*56*). Finally, we only assessed the PHBR score for MHC-I and not MHC-II. MHC-II presentation of neoantigens is possibly an important determinant of an immune response against a tumor. Frequent cancer driver mutations are poorly presented by MHC-II, and MHC-II shows less inter-patient variability but stronger selective effects than MHC-I (*59*).

In summary, the ability of patient-specific MHC-I complexes to bind and present neoantigens represented by the PHBR score can predict who is most likely to respond to ICB within the subgroup of patients with higher TMB. These results need to be extensively validated prior to incorporation into routine clinical use. Future studies are needed to clarify the role of PHBR score in predicting response to ICB in specific malignancies. Patients with high PHBR scores may benefit from immunotherapies that circumvent antigen presentation by MHC-I (e.g., chimeric antigen receptor T cells). Finally, much effort will be needed to decipher how to best incorporate MHC-I-related PHBR, reflecting neoantigen presentation by HLA-I, in the context of PD-L1 expression, TCR repertoire, and HLA-II genotype.

1.1.6 Materials and Methods

Patient selection. Three hundred and twenty-eight patients with diverse solid tumors treated with ICB (4/2010–5/2018) at a single institution were reviewed. Patients with melanoma, tumors that were not sequenced by Foundation Medicine (FM), and patients without an identified missense alteration by NGS were excluded. We excluded patients without next-generation sequencing or those with sequencing, but no identified missense alterations, because PHBR cannot be calculated in those cases; we omitted melanoma because melanoma patients have disproportionately high TMBs and high response rates to immunotherapy as compared to the majority of other cancers. All patients were treated with anti-PD-1/L1 monotherapy (or in combination with a second agent). The validation cohort was composed of thirty-two NSCLC patients treated with pembrolizumab (starting from 2012 to 2013) at Memorial Sloan Kettering and the University of California Los Angeles. All validation patients had consented to Institutional Review Board-approved protocols regarding tissue collection and sequencing.

TMB and HLA-I sequencing. Patients had NGS performed on tumor samples to determine genetic alterations, TMB, and HLA-I genotype (*60*). Formalin-fixed paraffin-embedded tumor samples were submitted for NGS to FM [clinical laboratory improvement amendments (CLIA)-certified lab]. The FoundationOne assay was used (hybrid-capture-based NGS; 236 or 315 genes; http://www.foundationone.com/). The methods have been previously described (*60*). Average sequencing depth of coverage was greater than 250X, with > 100X at > 99% of exons. For TMB, the number of somatic mutations detected on NGS (interrogating 1.2 mb of the genome) is quantified and that value extrapolated to the whole exome using a validated algorithm (*61*). Alterations likely or known to be bona fide oncogenic drivers and germline polymorphisms are excluded. TMB was measured in mutations per megabase (mb). Sequence-derived HLA-A/B/C typing was conducted by back-converting BAM files to fastq, then performing HLA realignment and typing using OptiType (*62*).

PHBR. The Patient Harmonic-mean Best Rank (PHBR) score as previously described (*46*), is a metric that represents how well the specific HLA-I genotype of an individual can bind and present a specific missense mutation. Each patient was assigned the PHBR score of his or her best-presented missense driver mutation. For patients with two or more missense mutations, only the mutation with the lowest PHBR score was selected. PHBR low (strong presentation) and high (poor presentation) were defined as < 0.5 and ≥ 0.5, respectively.

Mapping Foundation Medicine mutations to peptides. RefSeq transcript IDs from the FM variant spreadsheet were mapped to corresponding Ensembl transcript IDs with coding (CDS) sequences. For evaluation of missense mutations, we replaced the native amino acid residue with the mutated residue and selected all 38 possible peptides of length 8–11 that covered the mutated amino acid residue. For evaluation of in-frame insertion and deletion mutations, bases were

inserted or deleted from the CDS sequence according to the "cds effect" column from the FM data. The new CDS sequence was then translated into an amino acid sequence using the Seq.translate function from Biopython (Bio) package (*63*). We then selected any resulting novel peptides of length 8–11 for affinity analysis.

Affinity analysis. We calculated the allele-specific binding affinities of the previously described mutated peptides using NetMHCpan4.0 (*64*). Conventionally, a NetMHCpan4.0 binding affinity percentile rank less than 2 indicates weak peptide-MHC binding, while a binding affinity percentile rank less than 0.5 indicates strong peptide-MHC binding (*65*). Patient Harmonic-mean Best Rank PHBR scores (*46*)were used to represent a patient's ability to present the mutations in their tumor. HLA-A, HLA-B, and HLA-C alleles were obtained from FM. We evaluated the binding affinity of each HLA allele for 38 possible peptides of length 8–11 overlapping each mutation using NetMHCpan4.0. For individual alleles, the best rank percentile from NetMHCpan4.0 out of the 38 possible peptides was assigned. Best rank percentiles for all 6 alleles were aggregated into the PHBR score using a harmonic mean. High PHBR scores are indicative of poor affinity of peptides overlapping a mutation with the patient's MHC-I molecules and vice versa.

Validation. Matched tumor-normal exome sequencing fastq files obtained from (*66*) (dbGaP study accession phs000980.v1.p1.c1) were preprocessed and mutations called according to the GATK best practice workflow. Only mutations occurring in the 309 genes from the Foundation Medicine gene panel were retained. HLA typing was done in silico using the OptiType software package (*62*). Mutated peptides were created using the same method as described above. Similarly, PHBR scores were generated as described previously.

Statistical analysis. We used the Fisher exact test to assess categorical variables. P values < 0.05 were considered significant (values < 0.10 were included in the multivariable regression analyses). Overall benefit rate (OBR) (stable disease for ≥ 6 months and partial or complete response) was determined (RECIST criteria). Median PFS and OS were calculated from the start of checkpoint blockade and data was censored at the last visit for patients still progression free or alive, respectively, for PFS and OS. For the outcome analysis, comparisons were made between TMB low vs. high and PHBR low vs. high. Patients with no TMB values were assigned to the low TMB category for discrete analyses, and a pseudocount of 0.001 was added to TMB for all patients. We performed a Cox proportional hazards regression stratified by high (≥ 10 mutations/mb) or low (< 10 mutations/mb) TMB to quantify the specific effect of PHBR on PFS. These findings were visualized using Kaplan-Meier curves. Statistical analysis was performed on R version 3.5.2 and IBM SPSS Statistics version 24.

1.1.7 Figures

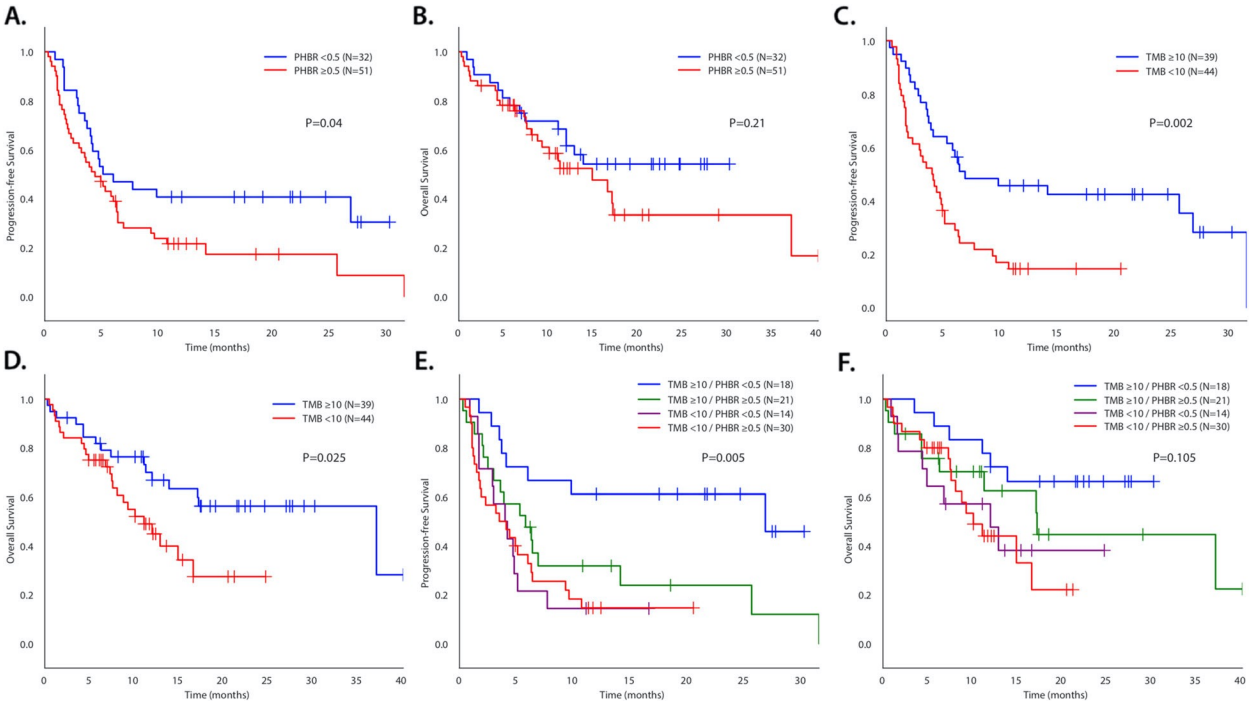


Figure 1.1.1 Kaplan-Meier PFS and OS for patients treated with immunotherapy. P-values in Figure 1.1.1 compare all four categories. They differ slightly from P values in Table 1.1.3, which compares value to the reference. PFS (**A**) and OS (**B**) dichotomized by PHBR < 0.5 and ≥ 0.5 (N = 83). PFS (**C**) and OS (**D**) dichotomized by TMB < 10 and ≥ 10 mutations/mb (N = 83). PFS (e) and OS (f) separated by TMB < 10 and ≥ 10 and PHBR < 0.5 and ≥ 0.5 (N = 83). For PFS (**E**), P = 0.005 for difference between all four curves. Curve for TMB ≥ 10/PHBR < 0.5 versus TMB ≥ 10/PHBR ≥ 0.5 was significantly different (P = 0.025); TMB ≥ 10/PHBR ≥ 0.5 did not differ significantly from TMB < 10/PHBR ≥ 0.5 (P = 0.19) or from TMB < 10/PHBR < 0.5 (P = 0.26); TMB < 10/PHBR ≥ 0.5 did not differ significantly from TMB < 10/PHBR < 0.5 (P = 0.91). For OS (**F**), P = 0.1 for difference between all four curves. Differences between individual curves were not statistically different.

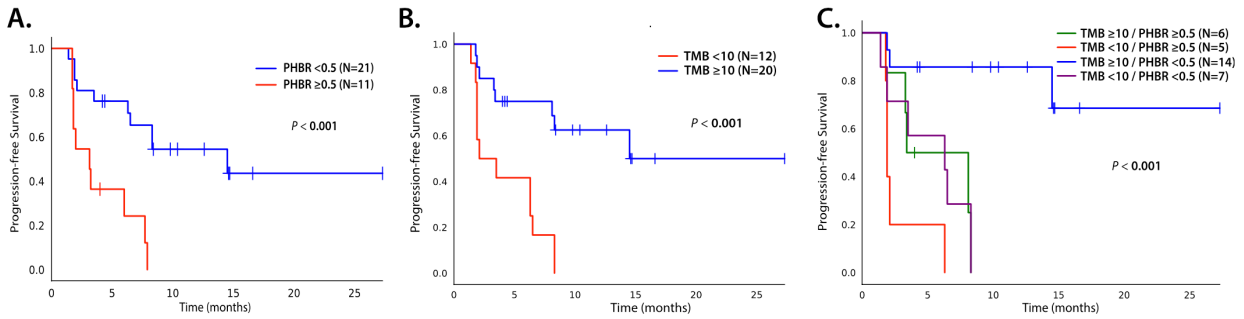


Figure 1.1.2. PFS for patients treated with immunotherapy in the validation dataset (N = 32). P values in the figure compare all four categories. (**A**) PFS dichotomized by PHBR < 0.5 and ≥ 0.5. (**B**) PFS dichotomized by TMB < 10 and ≥ 10 mutations/mb. (**C**)PFS separated by TMB < 10 and ≥ 10 and PHBR < 0.5 and ≥ 0.5

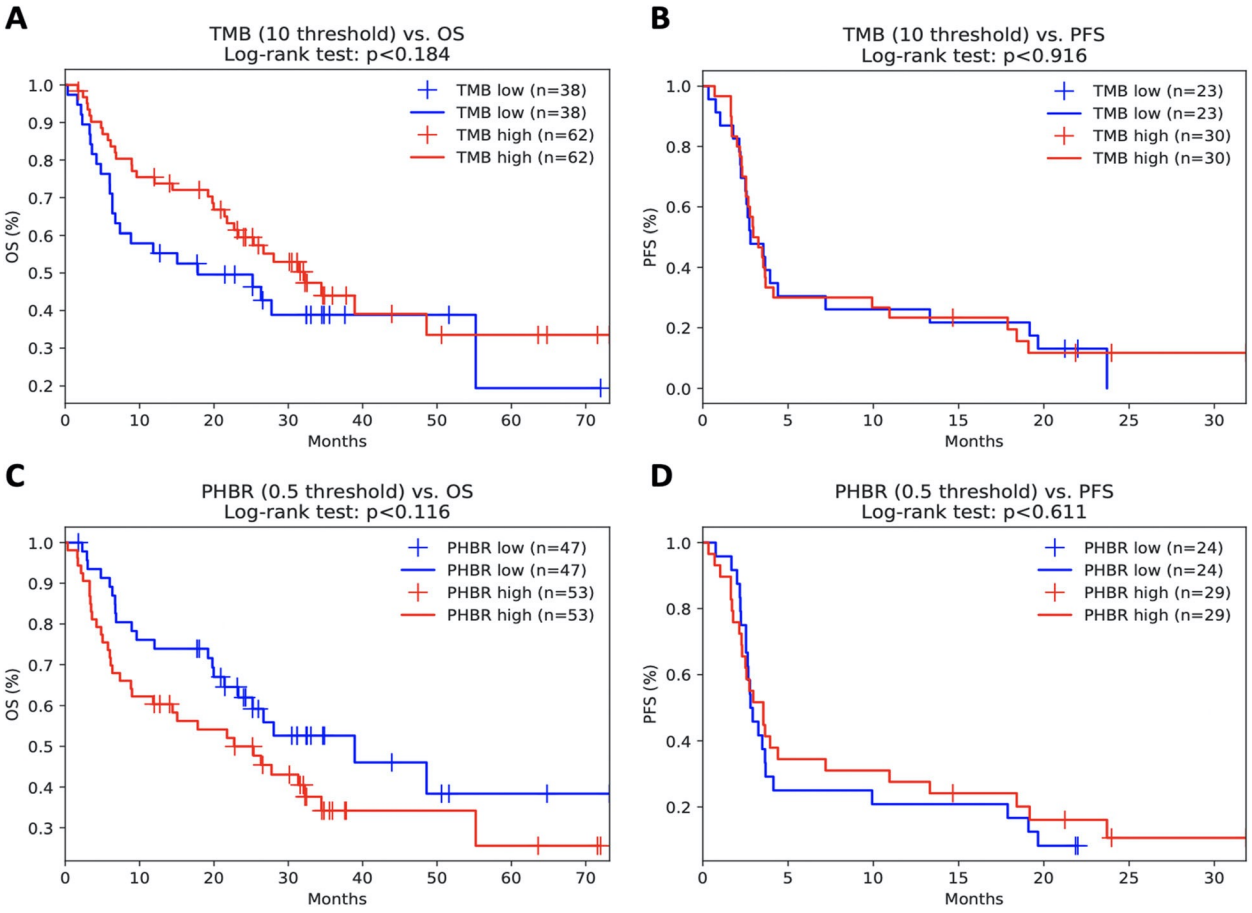


Figure 1.1.3. Kaplan-Meier curves showing the effects of TMB and presentable mutations on survival. (**A**) TMB versus overall survival, (B) TMB versus progression-free survival, (**C**) minimum PHBR score versus overall survival and (**D**) minimum PHBR score versus progression-free survival in the combined melanoma cohort.

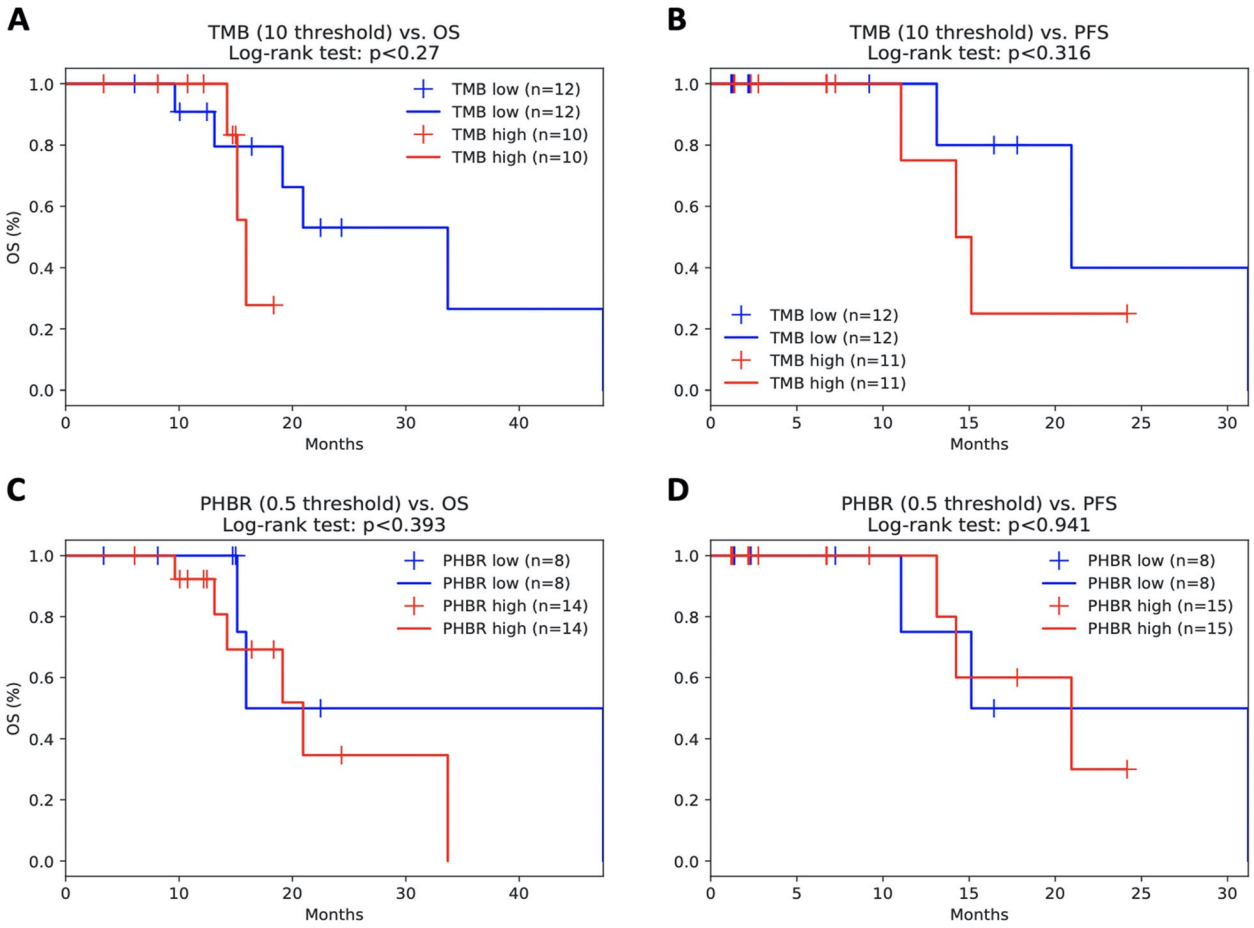


Figure 1.1.4. Kaplan-Meier curves showing the effects of TMB and mutation presentability in the Miao kidney cohort. (**A**) TMB versus overall survival, (**B**) TMB versus progression-free survival, (**C**) minimum PHBR score versus overall survival, and (**D**) progression-free survival

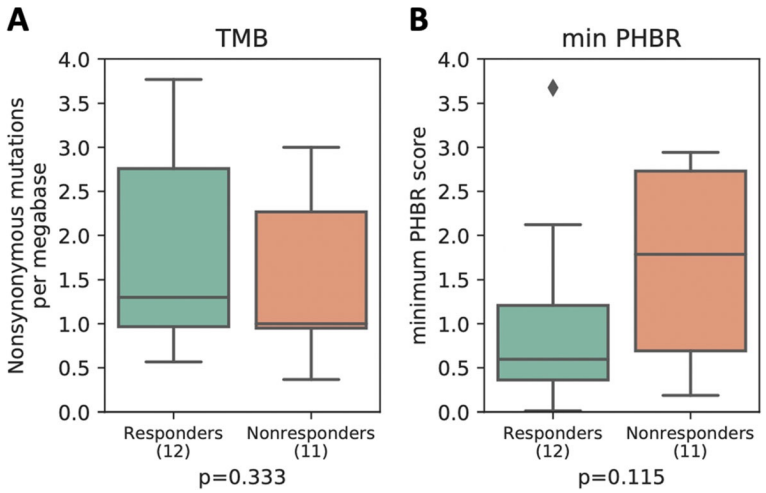


Figure 1.1.5. Analysis of responders and non-responders in the Miao kidney cohort. Boxplots showing the distribution of (**A**) TMB and (**B**)minimum PHBR score for responders and non-responders in the Miao cohort. P values were calculated by the Mann-Whitney U test.

1.1.8 Tables

Table 1.1.1. Patient demographics by PHBR score (< 0.5 vs. ≥ 0.5) (N=83). [1]Relative risk for PHBR < 0.5. [2]Calculated using Fisher's exact test. [3]Others: African American (N=2), Asian (N=4), Hispanic (N=5), and unknown (N=1). [4]At time of initiation of treatment with immunotherapy. [5]Others: adrenal (N=1), appendix (N=4), basal cell carcinoma (N=3), breast cancer (N=6), cervical (N=1), cholangiocarcinoma (N=1), colorectal (N=2), duodenal (N=1), gastroesophageal (N=5), glioblastoma (N=2), thyroid (N=1), prostate (N=1), rectal squamous cell carcinoma (N=1), renal cell carcinoma (N=1), sarcoma (N=3), urothelial (N=4), and urethral squamous cell carcinoma (N=1). [6]TMB was performed on 77 patients. [7]One patient had SD, but had not reached to 6 months. Only 82 patients were evaluable for this comparison. **Abbreviations**: CR: complete response, HR: hazard ratio, NR: not reached to 50%, NSCLC: non-small cell lung cancer, OS: overall survival, PFS: progression-free survival, PD: progressive disease, PHBR: Patient Harmonic-mean Best Rank, PR: partial response, RR: relative risk, SCC: squamous cell carcinoma, SD: stable disease, TMB: tumor mutational burden.

Variable	Group	*N* (82)	PHBR < 0.5 (*N* = 32)	PHBR ≥ 0.5 (*N* = 51)	Relative risk (95% CI)[1]	*P* value[2]
Sex	Male	46	22 (48%)	24 (52%)	1.77 (0.96–3.26)	0.07
	Female	37	10 (27%)	27 (73%)		
Ethnicity	Caucasian	71	27 (38%)	44 (62%)	0.91 (0.44–1.90)	> 0.99
	Others[3]	12	5 (42%)	7 (58%)		
Age[4] (years)	< 60	17	6 (35%)	11 (65%)	0.90 (0.44–1.82)	> 0.99
	≥ 60	66	26 (39%)	40 (61%)		
Tumor type	Head and neck SCC	9	4 (44%)	5 (56%)	1.18 (0.54–2.58)	0.73
	Others	74	28 (38%)	46 (62%)		
	NSCLC	26	7 (27%)	19 (73%)	0.61 (0.31–1.23)	0.16
	Others	57	25 (44%)	32 (56%)		
	Cutaneous SCC	10	7 (70%)	3 (30%)	2.04 (1.22–3.42)	**0.04**
	Others	73	26 (34%)	48 (66%)		
	Others[5]	38	14 (37%)	24 (63%)	0.92 (0.53–1.60)	0.82
	Head and neck SCC, NSCLC, and cutaneous SCC	45	18 (40%)	27 (60%)		
TMB[6] (mutations/mb)	< 50	65	21 (32%)	44 (68%)	0.49 (0.28–0.83)	**0.048**
	≥ 50	12	8 (67%)	4 (33%)		
	< 20	56	18 (32%)	38 (68%)	0.61 (0.35–1.07)	0.12
	≥ 20	21	11 (52%)	10 (48%)		
	< 10	38	11 (29%)	27 (71%)	0.63 (0.34–1.15)	0.16
	≥ 10	39	18 (46%)	21 (54%)		
PD-1/L1 Therapy	Monotherapy	66	26 (39%)	40 (61%)	1.12 (0.55–2.27)	> 0.99
	Combination	17	6 (35%)	11 (65%)		
Overall benefit rate	SD ≥ 6 months/PR/CR[7]	36	17 (47%)	19 (53%)	1.45 (0.84–2.49)	0.25
	Others	46	15 (33%)	31 (67%)		
	PD	32	7 (22%)	25 (78%)	0.45 (0.22–0.91)	**0.02**
	Others	51	25 (49%)	26 (51%)		

Table 1.1.2. Univariate analysis of factors affecting outcome for patients treated with immune checkpoint blockade (N = 83). [1]Thirty-six patients achieved SD with ≥ 6 months/PR/CR. One patient attained ongoing SD, but has not yet reached 6-month follow-up and is therefore not considered evaluable for this parameter; only 82 patients were evaluable for this comparison. [2]Calculated using Fisher's exact test. [3]Calculated using the log-rank test. [4]Others: African American (N=2), Asian (N=4), Hispanic (N=5), and unknown (N=1). [5]At time of initiation of treatment with immunotherapy. [6]Others: adrenal (N=1), appendix (N=4), basal cell carcinoma (N=3), breast cancer (N=6), cervical (N=1), cholangiocarcinoma (N=1), colorectal (N=2), duodenal (N=1), gastroesophageal (N=5), glioblastoma (N=2), thyroid (N=1), prostate (N=1), rectal squamous cell carcinoma (N=1), renal cell carcinoma (N=1), sarcoma (N=3), urothelial (N = 4), and urethral squamous cell carcinoma (N=1). [7]Seventy-seven patients with TMB were evaluable for the response rate, PFS, and OS. **Abbreviations**: HR: hazard ratio, NR: not reached to 50%, NSCLC: non-small cell lung cancer, OS: overall survival, PFS: progression-free survival, PHBR: Patient Harmonic-mean Best Rank, SCC: squamous cell carcinoma, TMB: tumor mutational burden.

	Rate of SD ≥ 6 month/PR/CR[1]		PFS			OS		
Variable	***N*** (%)	***P*** value[2]	Median, months	HR (95% CI)	***P*** value[3]	Median, months	HR (95% CI)	***P*** value[3]
Sex								
Male (*N* = 46) vs. female (*N* = 37)	23 (51%) vs. 13 (35%)	0.18	6.3 vs. 4.1	0.63 (0.38–1.04)	0.07	NR (MFU, 19.1) vs. 12.0	0.51 (0.27–0.95)	**0.03**
Ethnicity								
Caucasian (*N* = 71) vs. others[4] (*N* = 12)	32 (45%) vs. 4 (36%)	0.75	4.9 vs. 2.9	0.52 (0.26–1.00)	**0.045**	18.5 vs. 8.2	0.45 (0.19–1.06)	**0.004**
Age [5], years								
< 60 (*N* = 17) vs. ≥ 60 (*N* = 66)	6 (35%) vs. 30 (46%)	0.58	3.5 vs. 5.1	1.29 (0.70–2.39)	0.41	12.0 vs. 14.9	0.86 (0.36–2.06)	0.73
Tumor type								
Head and neck SCC (*N* = 9) vs. not (*N* = 74)	4 (44%) vs. 32 (44%)	> 0.99	4.8 vs. 4.9	1.01 (0.46–2.22)	0.99	12.9 vs. 16.6	1.11 (0.43–2.84)	0.83
NSCLC (*N* = 26) vs. not (*N* = 57)	8 (31%) vs. 28 (50%)	0.15	3.0 vs. 6.0	1.67 (0.99–2.81)	**0.05**	9.3 vs. 16.6	1.37 (0.71–2.64)	0.34
Cutaneous SCC (*N* = 10) vs. not (*N* = 73)	7 (70%) vs. 29 (40%)	0.10	26.8 vs. 4.7	0.43 (0.17–1.08)	0.06	NR (median follow-up, 21.7) vs. 13.9 14.9 vs. 17.1	0.43 (0.13–1.40)	0.15
Others[6] (*N* = 38) vs. head and neck SCC, NSCLC, and cutaneous SCC (*N* = 45)	17 (46%) vs. 19 (42%)	0.82	5.1 vs. 4.8	0.91 (0.55–1.52)	0.72		1.02 (0.54–1.93)	0.95
TMB[7], mutations/mb								
≥ 50 (*N* = 12) vs. < 50 (*N* = 65)	9 (75%) vs. 25 (39%)	**0.03**	26.8 vs. 4.4	0.40 (0.17–0.94)	**0.03**	NR (median follow-up, 17.5) vs. 12.9	0.39 (0.12–1.27)	0.10
≥ 20 (*N* = 21) vs. < 20 (*N* = 56)	14 (67%) vs. 20 (36%)	**0.02**	14.1 vs. 4.2	0.45 (0.23–0.85)	**0.01**	NR (median follow-up, 22.4) vs. 12.0	0.42 (0.19–0.96)	**0.03**
≥ 10 (*N* = 39) vs. < 10 (N = 38)	23 (59%) vs. 11 (29%)	**0.01**	6.9 vs. 4.0	0.40 (0.23–0.68)	**0.001**	37.1 vs. 10.1	0.42 (0.21–0.82)	**0.009**
PHBR								
< 0.5 (*N* = 32) vs. ≥ 0.5 (N = 51)	17 (53%) vs. 19 (38%)	0.25	5.1 vs. 4.4	0.58 (0.34–0.99)	**0.04**	NR (median follow-up, 21.7) vs. 14.9	0.66 (0.34–1.27)	0.21
PD-1/L1 therapy								
Monotherapy (*N* = 66) vs. combination (*N* = 17)	25 (39%) vs. 11 (65%)	0.06	4.1 vs. 6.3	1.17 (0.63–2.16)	0.63	17.1 vs. 11.3	0.78 (0.37–1.66)	0.51

Table 1.1.3. Overall response rate, PFS, and OS segregated by TMB low/high and PHBR low/high among patients treated with immunotherapy patients (N = 77 with TMB available). [1]Thirty-six patients achieved SD with ≥ 6 month/PR/CR. [2]P-values in Figure 1.1.1 are different as they compare all four categories at the same time. [3]Not reached to the median (median follow-up duration, 23.0 months). [4]Not reached to the median (median follow-up duration, 24.6 months). [5]Not reached to the median (median follow-up duration, 27.0 months). Abbreviations: HR: hazard ratio, NR: not reached to 50%, OS: overall survival, PFS: progression-free survival, PHBR: Patient Harmonic-mean Best Rank, TMB: tumor mutational burden.

	Rate of SD with ≥ 6 month/PR/CR[1]		PFS			OS		
Group	***N*** (%)	***P*** value	Median (months)	HR (95% CI)	***P*** value[2]	Median (months)	HR (95% CI)	***P*** value[2]
TMB/PHBR (TMB cutoff = 10 mutations/mb)								
Low/high (*N* = 27) vs. low/low (*N* = 11)	9 (33%) vs. 2 (18%)	0.45	3.5 vs. 4.2	1.01 (0.48–2.12)	0.99	10.1 vs. 12.0	0.90 (0.37–2.22)	0.82
Low/high (*N* = 27) vs. high/high (*N* = 21)	9 (33%) vs. 9 (43%)	0.56	3.5 vs. 5.8	0.76 (0.54–1.05)	0.09	10.1 vs. 17.2	0.72 (0.47–1.10)	0.12
Low/high (*N* = 27) vs. high/low (*N* = 18)	9 (33%) vs. 14 (78%)	**0.006**	3.5 vs. 26.8	0.62 (0.47–0.83)	**< 0.001**	10.1 vs. NR[3]	0.66 (0.47–0.91)	**0.008**
Low/low (*N* = 11) vs. high/high (*N* = 21)	2 (18%) vs. 9 (43%)	0.25	4.2 vs. 5.8	0.58 (0.25–1.31)	0.18	12.0 vs. 17.2	0.62 (0.23–1.69)	0.34
Low/low (*N* = 11) vs. high/low (*N* = 18)	2 (18%) vs. 14 (78%)	**0.003**	4.2 vs. 26.8	0.50 (0.30–0.83)	**0.003**	12.0 vs. NR[3]	0.59 (0.34–1.02)	**0.049**
High/high (*N* = 21) vs. high/low (*N* = 18)	9 (43%) vs. 14 (78%)	**0.049**	5.8 vs. 26.8	0.39 (0.16–0.91)	**0.03**	17.2 vs. NR[3]	0.53 (0.19–1.50)	0.23
TMB/PHBR (TMB cutoff = 20 mutations/mb)								
Low/high (*N* = 38) vs. low/low (*N* = 18)	13 (34%) vs. 7 (39%)	0.77	4.1 vs. 4.2	0.89 (0.47–1.67)	0.71	11.1 vs. 12.0	0.81 (0.38–1.73)	0.58
Low/high (*N* = 38) vs. high/high (*N* = 10)	13 (34%) vs. 5 (50%)	0.47	4.1 vs. 3.6	0.96 (0.65–1.41)	0.82	11.1 vs. 17.2	0.76 (0.45–1.31)	0.32
Low/high (*N* = 38) vs. high/low (*N* = 11)	13 (34%) vs. 9 (82%)	**0.007**	4.1 vs. NR[4]	0.59 (0.41–0.84)	**0.001**	11.1 vs. NR[5]	0.66 (0.44–0.99)	**0.03**
Low/low (*N* = 18) vs. high/high (*N* = 10)	7 (39%) vs. 5 (50%)	0.70	4.2 vs. 3.6	1.06 (0.44–2.53)	0.90	12.0 vs. 17.2	0.82 (0.26–2.62)	0.74
Low/low (*N* = 18) vs. high/low (*N* = 11)	7 (39%) vs. 9 (82%)	0.052	4.2 vs. NR[4]	0.46 (0.24–0.86)	**0.007**	12.0 vs. NR[5]	0.60 (0.31–1.15)	0.11
High/high (*N* = 10) vs. high/low (*N* = 11)	5 (50%) vs. 9 (82%)	0.18	3.6 vs. NR[4]	0.16 (0.04–0.64)	**0.004**	17.2 vs. NR[5]	0.37 (0.08–1.70)	0.19

Table 1.1.4. Multivariable regression analysis of factors affecting outcome for patients treated with immunotherapy (N = 77 with TMB available). Variables with p-value of ≤ 0.1 in univariate (**Table 1.1.2**) were included in the multivariable regression analysis. **Abbreviations**: CR: complete response, HR: hazard ratio, NSCLC: non-small cell lung cancer, OR: odds ratio, PHBR: Patient Harmonic-mean Best Rank, PR: partial response, SCC: squamous cell carcinoma, SD: stable disease, TMB: tumor mutational burden.

Group	OR (95% CI)	*P* value
Rate of SD ≥ 6 month/PR/CR		
Cutaneous SCC versus others	3.96 (0.69–22.64)	0.12
TMB ≥ 10 mutations/mb versus < 10	4.51 (1.40–14.61)	**0.01**
PD-1/L1 monotherapy versus combination	0.15 (0.04–0.58)	**0.006**
Progression-free survival		
Male versus female	0.94 (0.53–1.68)	0.83
Caucasian versus others	0.69 (0.33–1.43)	0.32
NSCLC versus others	1.52 (0.86–2.67)	0.15
Cutaneous SCC versus others	0.71 (0.22–2.26)	0.56
TMB ≥ 10 mutations/mb versus others	0.47 (0.26–0.86)	**0.01**
PHBR < 0.5 versus ≥ 0.5	0.75 (0.41–1.38)	0.36
Overall survival		
Male versus female	0.64 (0.33–1.26)	0.20
Caucasian versus others	0.68 (0.27–1.72)	0.42
TMB ≥ 10 mutations/mb versus < 10	0.48 (0.24–0.970)	**0.04**

Table 1.1.5. Cox proportional hazards regression for high-TMB patients in combined melanoma cohorts.

Variables	Coefficients	*P* value	Confidence interval (95%)
Age	OS – 0.01	OS 0.59	OS (– 0.04, 0.02)
	PFS 0.06	PFS 0.13	PFS (– 0.02, 0.15)
Sex	OS – 0.33	OS 0.40	OS (– 1.09, 0.44)
	PFS – 0.10	PFS 0.90	PFS (–1.67, 1.47)
TMB	OS – 0.03	OS 0.05	OS (– 0.05, 0.00)
	PFS 0.03	PFS 0.24	PFS (– 0.02, 0.07)
min-PHBR	**OS 0.28**	**OS 0.03***	**OS (0.02, 0.54)**
	PFS 0.82	**PFS 0.02***	**PFS (0.15, 1.49)**

1.1.9 Supplemental Data, Tables and Figures

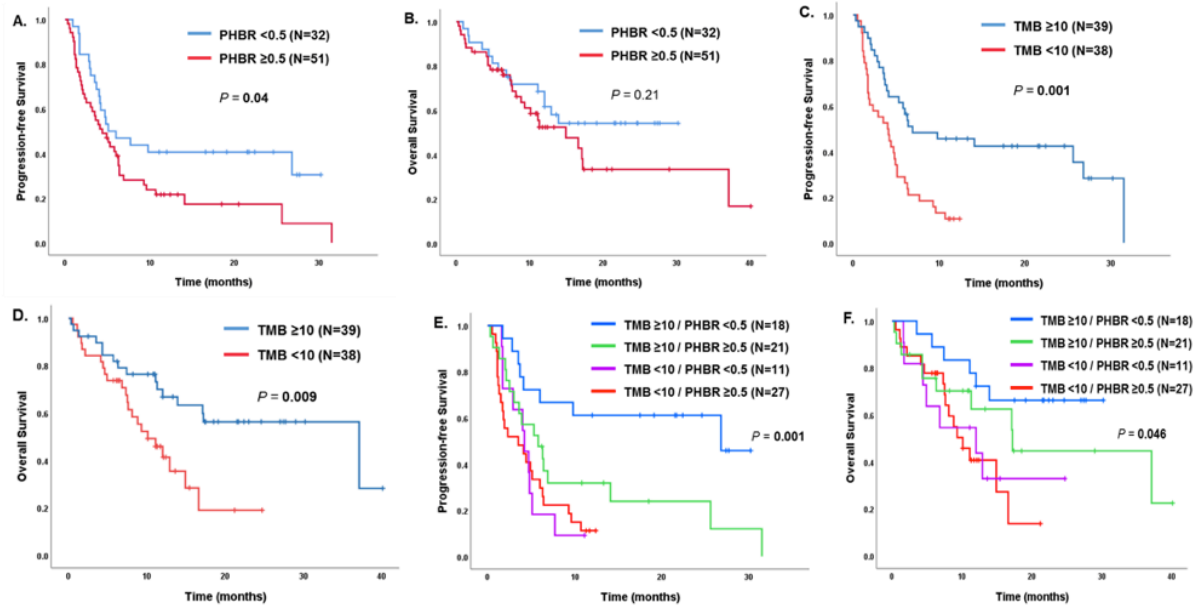


Figure S1.1.1. Kaplan and Meier PFS and OS for patients treated with immunotherapy, excluding patients with TMB=0. PFS (A) and OS (B) dichotomized by PHBR <0.5 and ≥0.5 (N=83). PFS (C) and OS (D) dichotomized by TMB <10 and ≥10 mutations/mb (N=77). PFS (E) and OS (F) separated by TMB <10 and ≥10 and PHBR <0.5 and ≥0.5 (N=77 with TMB available).

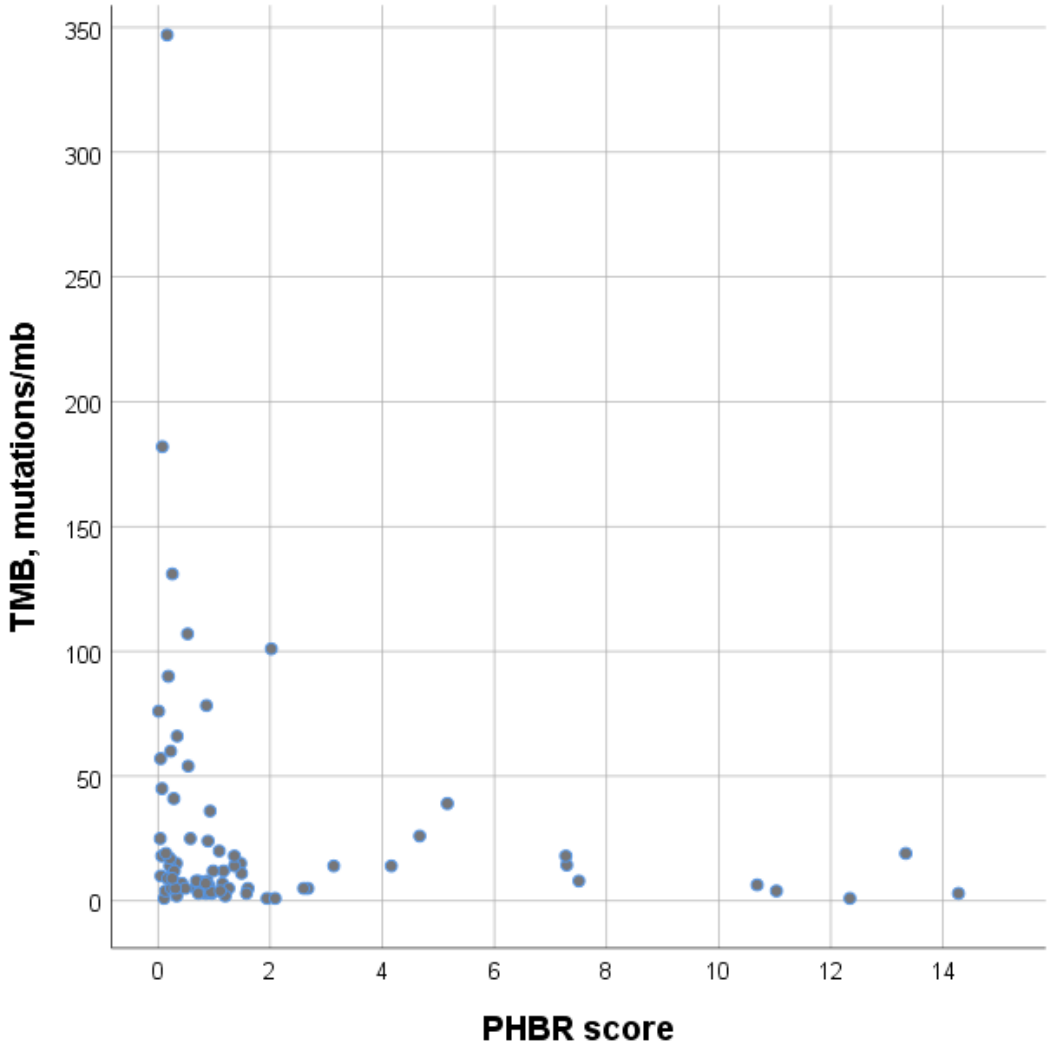


Figure S1.1.2. Correlation between PHBR score and TMB (N=77 with TMB available).

Table S1.1.1: Covariates retained after the backwards selection process. The coefficients and respective p-values for the covariates including TMB and PHBR in the final model are shown.

Covariates selected by backwards selection	*Coefficient*	*P-value*	*Confidence Interval (95%)*
Intercept	-1.29	0.09	(-2.79, 0.20)
MSI Status (MSS)	1.46	0.04	(0.04, 2.87)
Ethnicity (African American)	10.96	0.83	(-89.16, 111.08)
Ethnicity (Hispanic)	-9.59	0.85	(-106.77, 87.60)
Diagnosis (Adrenal cortical carcinoma)	10.25	0.75	(-53.39, 73.89)
Diagnosis (Colorectal adenocarcinoma)	2.84	0.59	(-7.44, 13.13)
Diagnosis (NSCLC) (squamous cell carcinoma)	-9.31	0.89	(-141.53, 122.92)
Diagnosis (Skin squamous cell carcinoma)	0.85	0.39	(-1.10, 2.80)
Diagnosis (Spleen angiosarcoma)	8.52	0.78	(-52.56, 69.60)
TMB (mutations per megabase)	0.013	0.25	(-0.0091, 0.035)
min PHBR score	-0.31	0.08	(-0.66, 0.042)

1.1.10 Author Contributions

Original concept: Aaron M Goodman, Hannah Carter, and Razelle Kurzrock

Curation and data analysis of clinical data: Aaron Goodman, Ryosuke Okamura, Shumei Kato, and Paul Riviere

Statistical analysis: Xinlian Zhang, Ryosuke Okamura, Andrea Castro, and Hannah Carter

PHBR score analysis: Andrea Castro, Rachel Marty Pyke, and Hannah Carter

Curation and analysis of Foundation Medicine data: Garrett Frampton, Ethan Sokol

Manuscript writing: Aaron M Goodman, Andrea Castro, Hannah Carter, Edward D Ball, and Razelle Kurzrock

1.1.11 Acknowledgements

We thank the members of the Thoracic Oncology Service and the Chan and Wolchok labs at MSKCC for the helpful discussions. We thank the Immune Monitoring Core at MSKCC, including L. Caro, R. Ramsawak, and Z. Mu, for the exceptional support with processing and banking peripheral blood lymphocytes. We thank P. Worrell and E. Brzostowski for the help in identifying tumor specimens for analysis. We thank A. Viale for the superb technical assistance. We thank D. Philips, M. van Buuren, and M. Toebes for the help in performing the combinatorial coding screens. The data presented in this paper are tabulated in the main paper and in the supplementary materials. This work was supported by the Geoffrey Beene Cancer Research Center (MDH, NAR, TAC, JDW, AS), the Society for Memorial Sloan Kettering Cancer Center (MDH), Lung Cancer Research Foundation (WL), Frederick Adler Chair Fund (TAC), The One Ball Matt Memorial Golf Tournament (EBG), Queen Wilhelmina Cancer Research Award (TNS), The STARR Foundation (TAC, JDW), the Ludwig Trust (JDW), and a Stand Up To Cancer- Cancer Research Institute Cancer Immunology Translational Cancer Research Grant (JDW, TNS, TAC). Stand Up To Cancer is a program of the Entertainment Industry Foundation administered by the American Association for Cancer Research.

We thank Martin Miller at the Memorial Sloan Kettering Cancer Center (MSKCC) for his assistance with the NetMHC server, Agnes Viale and Kety Huberman at the MSKCC Genomics Core, Annamalai Selvakumar and Alice Yeh at the MSKCC HLA typing laboratory for their technical assistance, and John Khoury for the assistance in chart review.

Chapter 1.1, in full, includes reformatted reprints of the material as it appears in "MHC-I genotype and tumor mutational burden predict response to immunotherapy" in *Genome Medicine,* 2020 by Aaron M Goodman, Andrea Castro, Rachel Marty Pyke, Ryosuke Okamura, Shumei

Kato, Paul Riviere, Garrett Frampton, Ethan Sokol, Xinlian Zhang, Edward D Ball, Hannah Carter, and Razelle Kurzrock. The dissertation author was a primary investigator and author of this paper.

1.1.12 References

1. L. Marcus, L. A. Fashoyin-Aje, M. Donoghue, M. Yuan, L. Rodriguez, P. S. Gallagher, R. Philip, S. Ghosh, M. R. Theoret, J. A. Beaver, R. Pazdur, S. J. Lemery, FDA Approval Summary: Pembrolizumab for the Treatment of Tumor Mutational Burden-High Solid Tumors. *Clin. Cancer Res.* **27**, 4685–4689 (2021).

2. A. Snyder, V. Makarov, T. Merghoub, J. Yuan, J. M. Zaretsky, A. Desrichard, L. A. Walsh, M. A. Postow, P. Wong, T. S. Ho, T. J. Hollmann, C. Bruggeman, K. Kannan, Y. Li, C. Elipenahli, C. Liu, C. T. Harbison, L. Wang, A. Ribas, J. D. Wolchok, T. A. Chan, Genetic basis for clinical response to CTLA-4 blockade in melanoma. *N. Engl. J. Med.* **371**, 2189–2199 (2014).

3. D. T. Le, J. N. Durham, K. N. Smith, H. Wang, B. R. Bartlett, L. K. Aulakh, S. Lu, H. Kemberling, C. Wilt, B. S. Luber, F. Wong, N. S. Azad, A. A. Rucki, D. Laheru, R. Donehower, A. Zaheer, G. A. Fisher, T. S. Crocenzi, J. J. Lee, T. F. Greten, A. G. Duffy, K. K. Ciombor, A. D. Eyring, B. H. Lam, A. Joe, S. P. Kang, M. Holdhoff, L. Danilova, L. Cope, C. Meyer, S. Zhou, R. M. Goldberg, D. K. Armstrong, K. M. Bever, A. N. Fader, J. Taube, F. Housseau, D. Spetzler, N. Xiao, D. M. Pardoll, N. Papadopoulos, K. W. Kinzler, J. R. Eshleman, B. Vogelstein, R. A. Anders, L. A. Diaz Jr, Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade. *Science*. **357**, 409–413 (2017).

4. H. Yamashita, K. Nakayama, M. Ishikawa, K. Nakamura, T. Ishibashi, K. Sanuki, R. Ono, H. Sasamori, T. Minamoto, K. Iida, R. Sultana, N. Ishikawa, S. Kyo, Microsatellite instability is a biomarker for immune checkpoint inhibitors in endometrial cancer. *Oncotarget*. **9**, 5652–5664 (2018).

5. J. S. Nielsen, A. R. Chang, D. A. Wick, C. G. Sedgwick, Z. Zong, A. J. Mungall, S. D. Martin, N. N. Kinloch, S. Ott-Langer, Z. L. Brumme, S. P. Treon, J. M. Connors, R. D. Gascoyne, J. R. Webb, B. R. Berry, R. D. Morin, N. Macpherson, B. H. Nelson, Mapping the human T cell repertoire to recurrent driver mutations in MYD88 and EZH2 in lymphoma. *Oncoimmunology*. **6**, e1321184 (2017).

6. T. N. Schumacher, R. D. Schreiber, Neoantigens in cancer immunotherapy. *Science*. **348**, 69–74 (2015).

7. K. Litchfield, J. L. Reading, C. Puttick, K. Thakkar, C. Abbosh, R. Bentham, T. B. K. Watkins, R. Rosenthal, D. Biswas, A. Rowan, E. Lim, M. Al Bakir, V. Turati, J. A. Guerra-Assunção, L. Conde, A. J. S. Furness, S. K. Saini, S. R. Hadrup, J. Herrero, S.-H. Lee, P. Van Loo, T. Enver, J. Larkin, M. D. Hellmann, S. Turajlic, S. A. Quezada, N. McGranahan, C. Swanton, Meta-analysis of tumor- and T cell-intrinsic mechanisms of sensitization to checkpoint inhibition. *Cell*. **184**, 596–614.e14 (2021).

8. M. D. Hellmann, L. Paz-Ares, R. Bernabe Caro, B. Zurawski, S.-W. Kim, E. Carcereny Costa, K. Park, A. Alexandru, L. Lupinacci, E. de la Mora Jimenez, H. Sakai, I. Albert, A. Vergnenegre, S. Peters, K. Syrigos, F. Barlesi, M. Reck, H. Borghaei, J. R. Brahmer, K. J. O'Byrne, W. J. Geese, P. Bhagavatheeswaran, S. K. Rabindran, R. S. Kasinathan, F. E.

Nathan, S. S. Ramalingam, Nivolumab plus Ipilimumab in Advanced Non-Small-Cell Lung Cancer. *N. Engl. J. Med.* **381**, 2020–2031 (2019).

9. M. Garassino, D. Rodriguez-Abreu, S. Gadgeel, E. Esteban, E. Felip, G. Speranza, M. Reck, R. Hui, M. Boyer, R. Cristescu, D. Aurora-Garg, A. Albright, A. Loboda, J. Kobie, J. Lunceford, M. Ayers, G. Lubiniecki, B. Piperdi, M. C. Pietanza, E. Garon, OA04.06 Evaluation of TMB in KEYNOTE-189: Pembrolizumab Plus Chemotherapy vs Placebo Plus Chemotherapy for Nonsquamous NSCLC. *J. Thorac. Oncol.* **14**, S216–S217 (2019).

10. C. Langer, S. Gadgeel, H. Borghaei, A. Patnaik, S. Powell, R. Gentzler, J. C. Yang, M. Gubens, L. Sequist, M. Awad, R. Cristescu, D. Aurora-Garg, A. Albright, A. Loboda, J. Kobie, J. Lunceford, M. Ayers, G. Lubiniecki, B. Piperdi, M. C. Pietanza, V. Papadimitrakopoulou, OA04.05 KEYNOTE-021: TMB and Outcomes for Carboplatin and Pemetrexed With or Without Pembrolizumab for Nonsquamous NSCLC. *J. Thorac. Oncol.* **14**, S216 (2019).

11. E. Ghorani, J. L. Reading, J. Y. Henry, M. R. de Massy, R. Rosenthal, V. Turati, K. Joshi, A. J. S. Furness, A. B. Aissa, S. K. Saini, S. Ramskov, A. Georgiou, M. W. Sunderland, Y. N. S. Wong, M. V. De Mucha, W. Day, F. Galvez-Cancino, P. D. Becker, I. Uddin, M. Ismail, T. Ronel, A. Woolston, M. Jamal-Hanjani, S. Veeriah, N. J. Birkbak, G. A. Wilson, K. Litchfield, L. Conde, J. A. Guerra-Assunção, K. Blighe, D. Biswas, R. Salgado, T. Lund, M. Al Bakir, D. A. Moore, C. T. Hiley, S. Loi, Y. Sun, Y. Yuan, K. AbdulJabbar, S. Turajilic, J. Herrero, T. Enver, S. R. Hadrup, A. Hackshaw, K. S. Peggs, N. McGranahan, B. Chain, C. Swanton, S. A. Quezada, The T cell differentiation landscape is shaped by tumour mutations in lung cancer. *Nat Cancer*. **1**, 546–561 (2020).

12. A. L. Cummings, J. Gukasyan, H. Y. Lu, T. Grogan, G. Sunga, C. M. Fares, N. Hornstein, J. Zaretsky, J. Carroll, B. Bachrach, W. O. Akingbemi, D. Li, Z. Noor, A. Lisberg, J. W. Goldman, D. Elashoff, A. A. T. Bui, A. Ribas, S. M. Dubinett, M. Rossetti, E. B. Garon, Mutational landscape influences immunotherapy outcomes among patients with non-small-cell lung cancer with human leukocyte antigen supertype B44. *Nature Cancer* (2020), doi:10.1038/s43018-020-00140-1.

13. M. A. Wood, B. R. Weeder, J. K. David, A. Nellore, R. F. Thompson, Burden of tumor mutations, neoepitopes, and other variants are weak predictors of cancer immunotherapy response and overall survival. *Genome Med.* **12**, 33 (2020).

14. M. D. Hellmann, T. Nathanson, H. Rizvi, B. C. Creelan, F. Sanchez-Vega, A. Ahuja, A. Ni, J. B. Novik, L. M. B. Mangarin, M. Abu-Akeel, C. Liu, J. L. Sauter, N. Rekhtman, E. Chang, M. K. Callahan, J. E. Chaft, M. H. Voss, M. Tenet, X.-M. Li, K. Covello, A. Renninger, P. Vitazka, W. J. Geese, H. Borghaei, C. M. Rudin, S. J. Antonia, C. Swanton, J. Hammerbacher, T. Merghoub, N. McGranahan, A. Snyder, J. D. Wolchok, Genomic Features of Response to Combination Immunotherapy in Patients with Advanced Non-Small-Cell Lung Cancer. *Cancer Cell*. **33**, 843–852.e4 (2018).

15. M. Yadav, S. Jhunjhunwala, Q. T. Phung, P. Lupardus, J. Tanguay, S. Bumbaca, C. Franci, T. K. Cheung, J. Fritsche, T. Weinschenk, Z. Modrusan, I. Mellman, J. R. Lill, L. Delamarre,

Predicting immunogenic tumour mutations by combining mass spectrometry and exome sequencing. *Nature*. **515**, 572–576 (2014).

16. D. K. Wells, M. M. van Buuren, K. K. Dang, V. M. Hubbard-Lucey, K. C. F. Sheehan, K. M. Campbell, A. Lamb, J. P. Ward, J. Sidney, A. B. Blazquez, A. J. Rech, J. M. Zaretsky, B. Comin-Anduix, A. H. C. Ng, W. Chour, T. V. Yu, H. Rizvi, J. M. Chen, P. Manning, G. M. Steiner, X. C. Doan, Tumor Neoantigen Selection Alliance, T. Merghoub, J. Guinney, A. Kolom, C. Selinsky, A. Ribas, M. D. Hellmann, N. Hacohen, A. Sette, J. R. Heath, N. Bhardwaj, F. Ramsdell, R. D. Schreiber, T. N. Schumacher, P. Kvistborg, N. A. Defranoux, Key Parameters of Tumor Epitope Immunogenicity Revealed Through a Consortium Approach Improve Neoantigen Prediction. *Cell*. **183**, 818–834.e13 (2020).

17. S. C. Wei, C. R. Duffy, J. P. Allison, Fundamental Mechanisms of Immune Checkpoint Blockade Therapy. *Cancer Discov.* **8**, 1069–1086 (2018).

18. A. Goodman, S. P. Patel, R. Kurzrock, PD-1–PD-L1 immune-checkpoint blockade in B-cell lymphomas. *Nat. Rev. Clin. Oncol.* **14**, 203–220 (2016).

19. F. S. Hodi, V. Chiarion-Sileni, R. Gonzalez, J.-J. Grob, P. Rutkowski, C. L. Cowey, C. D. Lao, D. Schadendorf, J. Wagstaff, R. Dummer, P. F. Ferrucci, M. Smylie, A. Hill, D. Hogg, I. Marquez-Rodas, J. Jiang, J. Rizzo, J. Larkin, J. D. Wolchok, Nivolumab plus ipilimumab or nivolumab alone versus ipilimumab alone in advanced melanoma (CheckMate 067): 4-year outcomes of a multicentre, randomised, phase 3 trial. *Lancet Oncol.* **19**, 1480–1492 (2018).

20. S. Gettinger, L. Horn, D. Jackman, D. Spigel, S. Antonia, M. Hellmann, J. Powderly, R. Heist, L. V. Sequist, D. C. Smith, P. Leming, W. J. Geese, D. Yoon, A. Li, J. Brahmer, Five-year follow-up of nivolumab in previously treated advanced non–small-cell lung cancer: Results from the CA209-003 study. *J. Clin. Oncol.* **36**, 1675–1684 (2018).

21. R. J. Motzer, N. M. Tannir, D. F. McDermott, O. Arén Frontera, B. Melichar, T. K. Choueiri, E. R. Plimack, P. Barthélémy, C. Porta, S. George, T. Powles, F. Donskov, V. Neiman, C. K. Kollmannsberger, P. Salman, H. Gurney, R. Hawkins, A. Ravaud, M.-O. Grimm, S. Bracarda, C. H. Barrios, Y. Tomita, D. Castellano, B. I. Rini, A. C. Chen, S. Mekan, M. B. McHenry, M. Wind-Rotolo, J. Doan, P. Sharma, H. J. Hammers, B. Escudier, CheckMate 214 Investigators, Nivolumab plus Ipilimumab versus Sunitinib in Advanced Renal-Cell Carcinoma. *N. Engl. J. Med.* **378**, 1277–1290 (2018).

22. P. Armand, A. Engert, A. Younes, M. Fanale, A. Santoro, P. L. Zinzani, J. M. Timmerman, G. P. Collins, R. Ramchandren, J. B. Cohen, J. P. De Boer, J. Kuruvilla, K. J. Savage, M. Trneny, M. A. Shipp, K. Kato, A. Sumbul, B. Farsaci, S. M. Ansell, Nivolumab for relapsed/refractory classic Hodgkin lymphoma after failure of autologous hematopoietic cell transplantation: Extended follow-up of the multicohort single-arm phase II CheckMate 205 trial. *J. Clin. Oncol.* **36**, 1428–1439 (2018).

23. J. M. Zaretsky, A. Garcia-Diaz, D. S. Shin, H. Escuin-Ordinas, W. Hugo, S. Hu-Lieskovan, D. Y. Torrejon, G. Abril-Rodriguez, S. Sandoval, L. Barthly, J. Saco, B. Homet Moreno, R. Mezzadra, B. Chmielowski, K. Ruchalski, I. P. Shintaku, P. J. Sanchez, C. Puig-Saus, G.

Cherry, E. Seja, X. Kong, J. Pang, B. Berent-Maoz, B. Comin-Anduix, T. G. Graeber, P. C. Tumeh, T. N. M. Schumacher, R. S. Lo, A. Ribas, Mutations Associated with Acquired Resistance to PD-1 Blockade in Melanoma. *N. Engl. J. Med.* **375**, 819–829 (2016).

24. S. Kato, A. Goodman, V. Walavalkar, D. A. Barkauskas, A. Sharabi, R. Kurzrock, Hyperprogressors after Immunotherapy: Analysis of Genomic Alterations Associated with Accelerated Growth Rate. *Clin. Cancer Res.* **23**, 4242–4250 (2017).

25. V. Subbiah, R. Kurzrock, The Marriage Between Genomics and Immunotherapy: Mismatch Meets Its Match. *Oncologist*. **24**, 1–3 (2019).

26. P. Sharma, S. Hu-Lieskovan, J. A. Wargo, A. Ribas, Primary, Adaptive, and Acquired Resistance to Cancer Immunotherapy. *Cell*. **168**, 707–723 (2017).

27. S. P. Patel, R. Kurzrock, PD-L1 Expression as a Predictive Biomarker in Cancer Immunotherapy. *Mol. Cancer Ther.* **14**, 847–856 (2015).

28. A. M. Goodman, D. Piccioni, S. Kato, A. Boichard, H.-Y. Wang, G. Frampton, S. M. Lippman, C. Connelly, D. Fabrizio, V. Miller, J. K. Sicklick, R. Kurzrock, Prevalence of PDL1 Amplification and Preliminary Response to Immune Checkpoint Blockade in Solid Tumors. *JAMA Oncol*. **4**, 1237–1244 (2018).

29. D. T. Le, J. N. Uram, H. Wang, B. R. Bartlett, H. Kemberling, A. D. Eyring, A. D. Skora, B. S. Luber, N. S. Azad, D. Laheru, B. Biedrzycki, R. C. Donehower, A. Zaheer, G. A. Fisher, T. S. Crocenzi, J. J. Lee, S. M. Duffy, R. M. Goldberg, A. de la Chapelle, M. Koshiji, F. Bhaijee, T. Huebner, R. H. Hruban, L. D. Wood, N. Cuka, D. M. Pardoll, N. Papadopoulos, K. W. Kinzler, S. Zhou, T. C. Cornish, J. M. Taube, R. A. Anders, J. R. Eshleman, B. Vogelstein, L. A. Diaz Jr, PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. *N. Engl. J. Med.* **372**, 2509–2520 (2015).

30. A. M. Goodman, S. Kato, L. Bazhenova, S. P. Patel, G. M. Frampton, V. Miller, P. J. Stephens, G. A. Daniels, R. Kurzrock, Tumor Mutational Burden as an Independent Predictor of Response to Immunotherapy in Diverse Cancers. *Mol. Cancer Ther.* **16**, 2598–2608 (2017).

31. R. M. Samstein, C.-H. Lee, A. N. Shoushtari, M. D. Hellmann, R. Shen, Y. Y. Janjigian, D. A. Barron, A. Zehir, E. J. Jordan, A. Omuro, T. J. Kaley, S. M. Kendall, R. J. Motzer, A. A. Hakimi, M. H. Voss, P. Russo, J. Rosenberg, G. Iyer, B. H. Bochner, D. F. Bajorin, H. A. Al-Ahmadie, J. E. Chaft, C. M. Rudin, G. J. Riely, S. Baxi, A. L. Ho, R. J. Wong, D. G. Pfister, J. D. Wolchok, C. A. Barker, P. H. Gutin, C. W. Brennan, V. Tabar, I. K. Mellinghoff, L. M. DeAngelis, C. E. Ariyan, N. Lee, W. D. Tap, M. M. Gounder, S. P. D'Angelo, L. Saltz, Z. K. Stadler, H. I. Scher, J. Baselga, P. Razavi, C. A. Klebanoff, R. Yaeger, N. H. Segal, G. Y. Ku, R. P. DeMatteo, M. Ladanyi, N. A. Rizvi, M. F. Berger, N. Riaz, D. B. Solit, T. A. Chan, L. G. T. Morris, Tumor mutational load predicts survival after immunotherapy across multiple cancer types. *Nat. Genet.* **51**, 202–206 (2019).

32. N. H. Segal, D. W. Parsons, K. S. Peggs, V. Velculescu, K. W. Kinzler, B. Vogelstein, J. P. Allison, Epitope landscape in breast and colorectal cancer. *Cancer Res.* **68**, 889–892 (2008).

33. A. Vanderwalde, D. Spetzler, N. Xiao, Z. Gatalica, J. Marshall, Microsatellite instability status determined by next-generation sequencing and compared with PD-L1 and tumor mutational burden in 11,348 patients. *Cancer Medicine*. **7** (2018), pp. 746–756.

34. T. A. Chan, M. Yarchoan, E. Jaffee, C. Swanton, S. A. Quezada, A. Stenzinger, S. Peters, Development of tumor mutation burden as an immunotherapy biomarker: utility for the oncology clinic. *Ann. Oncol.* **30**, 44–56 (2019).

35. N. Ready, M. D. Hellmann, M. M. Awad, G. A. Otterson, M. Gutierrez, J. F. Gainor, H. Borghaei, J. Jolivet, L. Horn, M. Mates, J. Brahmer, I. Rabinowitz, P. S. Reddy, J. Chesney, J. Orcutt, D. R. Spigel, M. Reck, K. J. O'Byrne, L. Paz-Ares, W. Hu, K. Zerba, X. Li, B. Lestini, W. J. Geese, J. D. Szustakowski, G. Green, H. Chang, S. S. Ramalingam, First-Line Nivolumab Plus Ipilimumab in Advanced Non-Small-Cell Lung Cancer (CheckMate 568): Outcomes by Programmed Death Ligand 1 and Tumor Mutational Burden as Biomarkers. *J. Clin. Oncol.* **37**, 992–1000 (2019).

36. D. Chowell, S. Krishna, P. D. Becker, C. Cocita, J. Shu, X. Tan, P. D. Greenberg, L. S. Klavinskis, J. N. Blattman, K. S. Anderson, TCR contact residue hydrophobicity is a hallmark of immunogenic CD8+ T cell epitopes. *Proc. Natl. Acad. Sci. U. S. A.* **112**, E1754–62 (2015).

37. International HIV Controllers Study, F. Pereyra, X. Jia, P. J. McLaren, A. Telenti, P. I. W. de Bakker, B. D. Walker, S. Ripke, C. J. Brumme, S. L. Pulit, M. Carrington, C. M. Kadie, J. M. Carlson, D. Heckerman, R. R. Graham, R. M. Plenge, S. G. Deeks, L. Gianniny, G. Crawford, J. Sullivan, E. Gonzalez, L. Davies, A. Camargo, J. M. Moore, N. Beattie, S. Gupta, A. Crenshaw, N. P. Burtt, C. Guiducci, N. Gupta, X. Gao, Y. Qi, Y. Yuki, A. Piechocka-Trocha, E. Cutrell, R. Rosenberg, K. L. Moss, P. Lemay, J. O'Leary, T. Schaefer, P. Verma, I. Toth, B. Block, B. Baker, A. Rothchild, J. Lian, J. Proudfoot, D. M. L. Alvino, S. Vine, M. M. Addo, T. M. Allen, M. Altfeld, M. R. Henn, S. Le Gall, H. Streeck, D. W. Haas, D. R. Kuritzkes, G. K. Robbins, R. W. Shafer, R. M. Gulick, C. M. Shikuma, R. Haubrich, S. Riddler, P. E. Sax, E. S. Daar, H. J. Ribaudo, B. Agan, S. Agarwal, R. L. Ahern, B. L. Allen, S. Altidor, E. L. Altschuler, S. Ambardar, K. Anastos, B. Anderson, V. Anderson, U. Andrady, D. Antoniskis, D. Bangsberg, D. Barbaro, W. Barrie, J. Bartczak, S. Barton, P. Basden, N. Basgoz, S. Bazner, N. C. Bellos, A. M. Benson, J. Berger, N. F. Bernard, A. M. Bernard, C. Birch, S. J. Bodner, R. K. Bolan, E. T. Boudreaux, M. Bradley, J. F. Braun, J. E. Brndjar, S. J. Brown, K. Brown, S. T. Brown, J. Burack, L. M. Bush, V. Cafaro, O. Campbell, J. Campbell, R. H. Carlson, J. K. Carmichael, K. K. Casey, C. Cavacuiti, G. Celestin, S. T. Chambers, N. Chez, L. M. Chirch, P. J. Cimoch, D. Cohen, L. E. Cohn, B. Conway, D. A. Cooper, B. Cornelson, D. T. Cox, M. V. Cristofano, G. Cuchural Jr, J. L. Czartoski, J. M. Dahman, J. S. Daly, B. T. Davis, K. Davis, S. M. Davod, E. DeJesus, C. A. Dietz, E. Dunham, M. E. Dunn, T. B. Ellerin, J. J. Eron, J. J. W. Fangman, C. E. Farel, H. Ferlazzo, S. Fidler, A. Fleenor-Ford, R. Frankel, K. A. Freedberg, N. K. French, J. D. Fuchs, J. D. Fuller, J. Gaberman, J. E. Gallant, R. T. Gandhi, E. Garcia, D. Garmon, J. C. Gathe Jr, C. R. Gaultier, W. Gebre, F. D. Gilman, I. Gilson, P. A. Goepfert, M. S. Gottlieb, C. Goulston, R. K. Groger, T. D. Gurley, S. Haber, R. Hardwicke, W. D. Hardy, P. R. Harrigan, T. N. Hawkins, S. Heath, F. M. Hecht, W. K. Henry, M. Hladek, R. P. Hoffman, J. M. Horton, R. K. Hsu, G. D. Huhn, P. Hunt, M. J. Hupert, M. L. Illeman, H. Jaeger, R. M. Jellinger, M. John, J. A. Johnson, K. L. Johnson, H. Johnson, K. Johnson, J. Joly, W. C. Jordan, C. A. Kauffman, H. Khanlou, R.

K. Killian, A. Y. Kim, D. D. Kim, C. A. Kinder, J. T. Kirchner, L. Kogelman, E. M. Kojic, P. T. Korthuis, W. Kurisu, D. S. Kwon, M. LaMar, H. Lampiris, M. Lanzafame, M. M. Lederman, D. M. Lee, J. M. L. Lee, M. J. Lee, E. T. Y. Lee, J. Lemoine, J. A. Levy, J. M. Llibre, M. A. Liguori, S. J. Little, A. Y. Liu, A. J. Lopez, M. R. Loutfy, D. Loy, D. Y. Mohammed, A. Man, M. K. Mansour, V. C. Marconi, M. Markowitz, R. Marques, J. N. Martin, H. L. Martin Jr, K. H. Mayer, M. J. McElrath, T. A. McGhee, B. H. McGovern, K. McGowan, D. McIntyre, G. X. Mcleod, P. Menezes, G. Mesa, C. E. Metroka, D. Meyer-Olson, A. O. Miller, K. Montgomery, K. C. Mounzer, E. H. Nagami, I. Nagin, R. G. Nahass, M. O. Nelson, C. Nielsen, D. L. Norene, D. H. O'Connor, B. O. Ojikutu, J. Okulicz, O. O. Oladehin, E. C. Oldfield 3rd, S. A. Olender, M. Ostrowski, W. F. Owen Jr, E. Pae, J. Parsonnet, A. M. Pavlatos, A. M. Perlmutter, M. N. Pierce, J. M. Pincus, L. Pisani, L. J. Price, L. Proia, R. C. Prokesch, H. C. Pujet, M. Ramgopal, A. Rathod, M. Rausch, J. Ravishankar, F. S. Rhame, C. S. Richards, D. D. Richman, B. Rodes, M. Rodriguez, R. C. Rose 3rd, E. S. Rosenberg, D. Rosenthal, P. E. Ross, D. S. Rubin, E. Rumbaugh, L. Saenz, M. R. Salvaggio, W. C. Sanchez, V. M. Sanjana, S. Santiago, W. Schmidt, H. Schuitemaker, P. M. Sestak, P. Shalit, W. Shay, V. N. Shirvani, V. I. Silebi, J. M. Sizemore Jr, P. R. Skolnik, M. Sokol-Anderson, J. M. Sosman, P. Stabile, J. T. Stapleton, S. Starrett, F. Stein, H.-J. Stellbrink, F. L. Sterman, V. E. Stone, D. R. Stone, G. Tambussi, R. A. Taplitz, E. M. Tedaldi, A. Telenti, W. Theisen, R. Torres, L. Tosiello, C. Tremblay, M. A. Tribble, P. D. Trinh, A. Tsao, P. Ueda, A. Vaccaro, E. Valadas, T. J. Vanig, I. Vecino, V. M. Vega, W. Veikley, B. H. Wade, C. Walworth, C. Wanidworanun, D. J. Ward, D. A. Warner, R. D. Weber, D. Webster, S. Weis, D. A. Wheeler, D. J. White, E. Wilkins, A. Winston, C. G. Wlodaver, A. van't Wout, D. P. Wright, O. O. Yang, D. L. Yurdin, B. W. Zabukovic, K. C. Zachary, B. Zeeman, M. Zhao, The major genetic determinants of HIV-1 control affect HLA class I peptide presentation. *Science*. **330**, 1551–1557 (2010).

38. M. J. Simmonds, S. C. L. Gough, The HLA Region and Autoimmune Disease: Associations and Mechanisms of Action. *Curr. Genomics*. **8**, 453–465 (2007).

39. H.-J. Kolb, Graft-versus-leukemia effects of transplantation and donor lymphocytes. *Blood*. **112**, 4371–4383 (2008).

40. M. DuPage, C. Mazumdar, L. M. Schmidt, A. F. Cheung, T. Jacks, Expression of tumour-specific antigens underlies cancer immunoediting. *Nature*. **482**, 405–409 (2012).

41. A. Garcia-Lora, I. Algarra, F. Garrido, MHC class I antigens, immune surveillance, and tumor immune escape. *J. Cell. Physiol.* **195**, 346–355 (2003).

42. H. Matsushita, M. D. Vesely, D. C. Koboldt, C. G. Rickert, R. Uppaluri, V. J. Magrini, C. D. Arthur, J. M. White, Y.-S. Chen, L. K. Shea, J. Hundal, M. C. Wendl, R. Demeter, T. Wylie, J. P. Allison, M. J. Smyth, L. J. Old, E. R. Mardis, R. D. Schreiber, Cancer exome analysis reveals a T-cell-dependent mechanism of cancer immunoediting. *Nature*. **482**, 400–404 (2012).

43. M. S. Rooney, S. A. Shukla, C. J. Wu, G. Getz, N. Hacohen, Molecular and genetic properties of tumors associated with local immune cytolytic activity. *Cell*. **160**, 48–61 (2015).

44. M. M. Gubin, X. Zhang, H. Schuster, E. Caron, J. P. Ward, T. Noguchi, Y. Ivanova, J. Hundal, C. D. Arthur, W.-J. Krebber, G. E. Mulder, M. Toebes, M. D. Vesely, S. S. K. Lam, A. J. Korman, J. P. Allison, G. J. Freeman, A. H. Sharpe, E. L. Pearce, T. N. Schumacher, R. Aebersold, H.-G. Rammensee, C. J. M. Melief, E. R. Mardis, W. E. Gillanders, M. N. Artyomov, R. D. Schreiber, Checkpoint blockade cancer immunotherapy targets tumour-specific mutant antigens. *Nature*. **515**, 577–581 (2014).

45. E. Tran, M. Ahmadzadeh, Y.-C. Lu, A. Gros, S. Turcotte, P. F. Robbins, J. J. Gartner, Z. Zheng, Y. F. Li, S. Ray, J. R. Wunderlich, R. P. Somerville, S. A. Rosenberg, Immunogenicity of somatic mutations in human gastrointestinal cancers. *Science*. **350**, 1387–1390 (2015).

46. R. Marty, S. Kaabinejadian, D. Rossell, M. J. Slifker, J. van de Haar, H. B. Engin, N. de Prisco, T. Ideker, W. H. Hildebrand, J. Font-Burgada, H. Carter, MHC-I Genotype Restricts the Oncogenic Mutational Landscape. *Cell*. **171**, 1272–1283.e15 (2017).

47. N. McGranahan, R. Rosenthal, C. T. Hiley, A. J. Rowan, T. B. K. Watkins, G. A. Wilson, N. J. Birkbak, S. Veeriah, P. Van Loo, J. Herrero, C. Swanton, TRACERx Consortium, Allele-Specific HLA Loss and Immune Escape in Lung Cancer Evolution. *Cell*. **171**, 1259–1271.e11 (2017).

48. A. Castro, K. Ozturk, R. M. Pyke, S. Xian, M. Zanetti, H. Carter, Elevated neoantigen levels in tumors with somatic mutations in the HLA-A, HLA-B, HLA-C and B2M genes. *BMC Med. Genomics*. **12**, 107 (2019).

49. T. Therneau, P. Grambsch, Modeling survival data: Extending the Cox model (pp. 69--75) (2013).

50. Y. Sakamoto, M. Ishiguro, G. Kitagawa, Akaike information criterion statistics. *Dordrecht, The Netherlands: D. Reidel*. **81**, 26853 (1986).

51. N. Riaz, J. J. Havel, V. Makarov, A. Desrichard, W. J. Urba, J. S. Sims, F. S. Hodi, S. Martín-Algarra, R. Mandal, W. H. Sharfman, S. Bhatia, W.-J. Hwu, T. F. Gajewski, C. L. Slingluff Jr, D. Chowell, S. M. Kendall, H. Chang, R. Shah, F. Kuo, L. G. T. Morris, J.-W. Sidhom, J. P. Schneck, C. E. Horak, N. Weinhold, T. A. Chan, Tumor and Microenvironment Evolution during Immunotherapy with Nivolumab. *Cell*. **171**, 934–949.e16 (2017).

52. E. M. Van Allen, D. Miao, B. Schilling, S. A. Shukla, C. Blank, L. Zimmer, A. Sucker, U. Hillen, M. H. G. Foppen, S. M. Goldinger, J. Utikal, J. C. Hassel, B. Weide, K. C. Kaehler, C. Loquai, P. Mohr, R. Gutzmer, R. Dummer, S. Gabriel, C. J. Wu, D. Schadendorf, L. A. Garraway, Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. *Science*. **350**, 207–211 (2015).

53. W. Hugo, J. M. Zaretsky, L. Sun, C. Song, B. H. Moreno, S. Hu-Lieskovan, B. Berent-Maoz, J. Pang, B. Chmielowski, G. Cherry, E. Seja, S. Lomeli, X. Kong, M. C. Kelley, J. A. Sosman, D. B. Johnson, A. Ribas, R. S. Lo, Genomic and Transcriptomic Features of Response to Anti-PD-1 Therapy in Metastatic Melanoma. *Cell*. **168**, 542 (2017).

54. D. Miao, C. A. Margolis, W. Gao, M. H. Voss, W. Li, D. J. Martini, C. Norton, D. Bossé, S. M. Wankowicz, D. Cullen, C. Horak, M. Wind-Rotolo, A. Tracy, M. Giannakis, F. S. Hodi, C. G. Drake, M. W. Ball, M. E. Allaf, A. Snyder, M. D. Hellmann, T. Ho, R. J. Motzer, S. Signoretti, W. G. Kaelin Jr, T. K. Choueiri, E. M. Van Allen, Genomic correlates of response to immune checkpoint therapies in clear cell renal cell carcinoma. *Science*. **359**, 801–806 (2018).

55. D. Chowell, L. G. T. Morris, C. M. Grigg, J. K. Weber, R. M. Samstein, V. Makarov, F. Kuo, S. M. Kendall, D. Requena, N. Riaz, B. Greenbaum, J. Carroll, E. Garon, D. M. Hyman, A. Zehir, D. Solit, M. Berger, R. Zhou, N. A. Rizvi, T. A. Chan, Patient HLA class I genotype influences cancer response to checkpoint blockade immunotherapy. *Science*. **359**, 582–587 (2018).

56. B. Li, T. Li, J.-C. Pignon, B. Wang, J. Wang, S. A. Shukla, R. Dou, Q. Chen, F. S. Hodi, T. K. Choueiri, C. Wu, N. Hacohen, S. Signoretti, J. S. Liu, X. S. Liu, Landscape of tumor-infiltrating T cell repertoire of human cancers. *Nat. Genet.* **48**, 725–732 (2016).

57. N. Auslander, G. Zhang, J. S. Lee, D. T. Frederick, B. Miao, T. Moll, T. Tian, Z. Wei, S. Madan, R. J. Sullivan, G. Boland, K. Flaherty, M. Herlyn, E. Ruppin, Publisher Correction: Robust prediction of response to immune checkpoint blockade therapy in metastatic melanoma. *Nat. Med.* **24**, 1942 (2018).

58. P. Jiang, S. Gu, D. Pan, J. Fu, A. Sahu, X. Hu, Z. Li, N. Traugh, X. Bu, B. Li, J. Liu, G. J. Freeman, M. A. Brown, K. W. Wucherpfennig, X. S. Liu, Signatures of T cell dysfunction and exclusion predict cancer immunotherapy response. *Nat. Med.* **24**, 1550–1558 (2018).

59. R. Marty Pyke, W. K. Thompson, R. M. Salem, J. Font-Burgada, M. Zanetti, H. Carter, Evolutionary Pressure against MHC Class II Binding Cancer Mutations. *Cell*. **175**, 1991 (2018).

60. G. M. Frampton, A. Fichtenholtz, G. A. Otto, K. Wang, S. R. Downing, J. He, M. Schnall-Levin, J. White, E. M. Sanford, P. An, J. Sun, F. Juhn, K. Brennan, K. Iwanik, A. Maillet, J. Buell, E. White, M. Zhao, S. Balasubramanian, S. Terzic, T. Richards, V. Banning, L. Garcia, K. Mahoney, Z. Zwirko, A. Donahue, H. Beltran, J. M. Mosquera, M. A. Rubin, S. Dogan, C. V. Hedvat, M. F. Berger, L. Pusztai, M. Lechner, C. Boshoff, M. Jarosz, C. Vietz, A. Parker, V. A. Miller, J. S. Ross, J. Curran, M. T. Cronin, P. J. Stephens, D. Lipson, R. Yelensky, Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. *Nat. Biotechnol.* **31**, 1023–1031 (2013).

61. Z. R. Chalmers, C. F. Connelly, D. Fabrizio, L. Gay, S. M. Ali, R. Ennis, A. Schrock, B. Campbell, A. Shlien, J. Chmielecki, F. Huang, Y. He, J. Sun, U. Tabori, M. Kennedy, D. S. Lieber, S. Roels, J. White, G. A. Otto, J. S. Ross, L. Garraway, V. A. Miller, P. J. Stephens, G. M. Frampton, Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. *Genome Med.* **9**, 34 (2017).

62. A. Szolek, B. Schubert, C. Mohr, M. Sturm, M. Feldhahn, O. Kohlbacher, OptiType: precision HLA typing from next-generation sequencing data. *Bioinformatics*. **30**, 3310–3316 (2014).

63. P. J. A. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, M. J. L. de Hoon, Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*. **25**, 1422–1423 (2009).

64. V. Jurtz, S. Paul, M. Andreatta, P. Marcatili, B. Peters, M. Nielsen, NetMHCpan-4.0: Improved Peptide-MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. *J. Immunol.* **199**, 3360–3368 (2017).

65. M. Nielsen, M. Andreatta, NetMHCpan-3.0; improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets. *Genome Med.* **8**, 33 (2016).

66. N. A. Rizvi, M. D. Hellmann, A. Snyder, P. Kvistborg, V. Makarov, J. J. Havel, W. Lee, J. Yuan, P. Wong, T. S. Ho, M. L. Miller, N. Rekhtman, A. L. Moreira, F. Ibrahim, C. Bruggeman, B. Gasmi, R. Zappasodi, Y. Maeda, C. Sander, E. B. Garon, T. Merghoub, J. D. Wolchok, T. N. Schumacher, T. A. Chan, Mutational landscape determines sensitivity to PD-1 blockade in non–small cell lung cancer. *Science*. **348**, 124–128 (2015).

1.2.1 Foreword

Much emphasis has been placed on high affinity peptide-MHC interactions as a requirement for effective immune targeting of tumors. Indeed, my work described in Aim 1.1 has demonstrated the requisite of highly presentable mutations for improved outcomes in immune checkpoint blockade treated patients. However, over time, tumor interaction with the immune system results in immunoediting, or pruning of susceptible subclones, and eventual escape from immune detection and elimination.

Impact to the integrity of the neopeptide-MHC complex, whether to the antigen processing pathway or to the MHC itself, could aid in tumor escape. B2M mutations and loss of heterozygosity (LOH) are linked to decreased MHC class I expression and decreased patient survival (*1*, *2*). Somatic mutations to HLA-A and HLA-B have also been shown to be under positive selection during tumorigenesis and are more frequent when tumor immune cell infiltration and cytotoxicity are high (*3*). Somatic LOH in the human leukocyte antigen (HLA) locus was reported to occur in 40% of adult non-small-cell lung cancers and contributes to impaired immune surveillance (*4*). In addition, our recent characterizations of multi-cancer pediatric cohorts have observed similar rates of somatic mutations and LOH affecting MHC-I.

At a time when immunotherapy treatments are becoming more widely used, it remains critical to assess the integrity and subsequent utility of the MHC. Loss of MHC integrity may occur pre-treatment and is frequently observed post-treatment (*5*). In addition to somatic MHC alterations, mounting evidence points to additional pathways for impacting MHC cell surface expression. Decreased transcription of *B2M* or *HLA* genes can decrease the levels of overall MHC, or specific HLA alleles, having varying impacts on ability to present antigen. Germline variation may influence overall expression of certain HLA alleles (*6*). Presence of BRAF V600E mutations

has been reported to cause internal sequestration of MHC-I, also contributing to immune escape (*7*). In addition to antigen presentation function, MHC-I molecule presence on the cell surface provides an inhibitory signal to natural killer (NK) cell mediated effector functions (*8*). Thus, clinical characterization of MHC integrity and, by extension, cell surface presentation, is critical for effective anti-tumor immunity.

My research in Aim 1.2 characterizes the extent of somatic mutations to the MHC-I in the TCGA, which is comprised of non-immunotherapy treated patients. Here, I describe the impacts of somatic MHC-I mutations to the mutation and putative neoantigen landscape and immune infiltration.

1.2.2 Abstract

The major histocompatibility complex class I (MHC-I) molecule is a protein complex that displays intracellular peptides to T cells, allowing the immune system to recognize and destroy infected or cancerous cells. MHC-I is composed of a highly polymorphic HLA-encoded alpha chain that binds the peptide and a Beta-2-microglobulin (B2M) protein that acts as a stabilizing scaffold. HLA mutations have been implicated as a mechanism of immune evasion during tumorigenesis, and B2M is considered a tumor suppressor gene. However, the implications of somatic HLA and B2M mutations have not been fully explored in the context of antigen presentation via the MHC-I molecule during tumor development. To understand the effect that B2M and HLA MHC-I molecule mutations have on mutagenesis, we analyzed the accumulation of mutations in patients from The Cancer Genome Atlas according to their MHC-I molecule mutation status. Somatic B2M and HLA mutations in microsatellite stable tumors were associated with higher overall mutation burden and a larger fraction of HLA-binding neoantigens when

compared to B2M and HLA wild type tumors. B2M and HLA mutations were highly enriched in patients with microsatellite instability. B2M mutations tended to occur relatively early during patients' respective tumor development, whereas HLA mutations were either early or late events. In addition, B2M and HLA mutated patients had higher levels of immune infiltration by natural killer and CD8+ T cells and higher levels of cytotoxicity. Our findings add to a growing body of evidence that somatic B2M and HLA mutations are a mechanism of immune evasion by demonstrating that such mutations are associated with a higher load of neoantigens that should be presented via MHC-I.

1.2.3 Introduction

Immune evasion is one of the hallmark traits characteristic of cancer cells (*9*). The near universal requirement for tumor cells to evade immune elimination implicates the immune system as a major selective force acting on developing tumors. When a tumor cell successfully evades the immune system, the mutations harbored within can persist and propagate as the cell divides.

In humans, the HLA-A, HLA-B, and HLA-C genes encode major histocompatibility complex class I (MHC-I) molecules, which display intracellular peptides on the cell surface for inspection by CD8+ T cells. These T cells have the potential to recognize the MHC-I-peptide complex and become activated cytotoxic T cells (CTLs). Cancer immunotherapies that target CTL activation rely on clinical selection of appropriate neoantigens, mutated peptides specific to tumor cells, to stimulate a response (*10*). Although these cancer immunotherapies are of high interest to patients and clinicians, they have not yet shown widespread clinical success (*11*).

The MHC-I molecule is composed of a highly polymorphic HLA encoded alpha chain and a beta-2-microglobulin (B2M) protein that acts as a stabilizing scaffold. B2M is essential for MHC-I complex formation and peptide presentation. B2M mutations and loss of heterozygosity

(LOH) are linked to decreased MHC class I expression and decreased patient survival (*1*, *2*). In addition to HLA-A, HLA-B, and HLA-C, B2M binds to other immune proteins including CD1, FCGRT, HFE, HLA-E, HLA-G LILRB, and MR1. Somatic mutations in HLA-A and HLA-B have also been shown to be under positive selection during tumorigenesis and are more frequent when tumor immune cell infiltration and cytotoxicity are high (*3*). Importantly, MHC-I molecule presence on the cell surface can provide an inhibitory signal to natural killer (NK) cell mediated effector functions (*8*). In addition to classical HLA molecules HLA-A, HLA-B, HLA-C, and HLA-G, nonclassical HLA-E acts as a ligand to inhibitory receptors on NK cells. Thus, both presence and antigen presentation function of MHC-I molecules contribute to anti-tumor immunity.

A recent study found that an individual's HLA genotype can facilitate immune evasion and shape the landscape of a patient's acquired mutations (*12*). Somatic mutations generating peptides with low affinity for an individual's respective HLA alleles were likely to evade immune detection and persist in the tumor. Somatic LOH in the human leukocyte antigen (HLA) locus is thought to impair immune surveillance and was reported to occur in 40% of non-small-cell lung cancers. The authors found that HLA LOH was significantly associated with a high mutational burden and cancer-specific neoantigens generated from these mutations were biased to bind to the lost HLA allele (*4*). Thus, immune evasion may depend on an individual's unique HLA genotype and the specificity of neoantigens for particular HLA alleles.

We previously observed an uncharacteristic enrichment of somatic mutations at B2M interaction partner binding interfaces (*13*). Whereas for most cancer genes, mutations occurred preferentially on the cancer gene itself, genes encoding B2M binding partners showed almost as many somatic mutations as B2M (**Figure 1.2.1A**). Given B2M's role as a central component of MHC-I, we hypothesized that mutations affecting B2M's interaction with HLA-A, HLA-B, and

HLA-C could facilitate immune evasion by altering the availability of MHC-I molecules with distinct specificities, thus affecting presentation of specific peptides to the immune system (**Figure 1.2.1B**). To gain a better understanding of the role of somatic mutations affecting B2M and its partners in immune evasion, we examined their effect on mutation burden, antigen binding affinity, immune infiltration, and cytotoxicity in tumors sequenced by The Cancer Genome Atlas (TCGA).

1.2.4 Results

HLA and B2M mutations in TCGA. B2M mutation calls were obtained directly from the MAF files provided by TCGA. Because the HLA locus is highly polymorphic, mutation calls against the reference genome are unreliable. Instead, we ran Polysolver (*3*) to simultaneously call patient-specific HLA types and detect somatic mutations affecting a patient's HLA alleles. Out of 10,428 TCGA patients that had the necessary whole exome sequencing data, only 579 patients had an HLA mutation and 125 patients had B2M mutations. Most of these mutations were nonsynonymous (**Figure 1.2.2A**). To determine whether nonsynonymous mutations occurred at amino acid residues with the potential to interfere with formation of the MHC-I molecule, experimental 3D structures for the B2M-HLA complex were obtained from the Protein Data Bank (*14*) and used to annotate amino acid residue location at protein core, surface or at the physical interface between B2M and HLA encoded proteins (**Methods**).

Mutations on HLA proteins, particularly HLA-A, showed a biased distribution with several recurrent hotspots (**Figure 1.2.2B**). Mutations were most concentrated in the α3 domain that mediates interaction with the T cell receptor (TCR) (206 mutations, 40.63% of total; OR = 2.04, $p < 2.58e-09$), and included multiple recurrent hotspots. Fifty-one mutations (10.06%) were observed in the transmembrane domain including additional hotspots. Although mutations were

observed throughout the α1 and α2 domains that form the peptide binding groove, they tended to be less recurrent (88 mutations, 17.36% for both α1 and α2). This may reflect the much larger heterogeneity of this region across HLA alleles.

Recurrent hotspot mutations often targeted interface and core regions on HLA-A, while they targeted core and surface regions on HLA-B, and surface regions on HLA-C (**Figure 1.2.2B**). Since there are many alleles for each HLA protein, we used the consensus of residue annotations across different alleles to annotate each HLA protein (**Figure S1.2.1**). Even though the annotations for most frequently mutated residues agreed between different HLA alleles, there were some exceptions, including residue 231 on HLA-A. Although residue 231 (R231) on HLA-A was annotated as surface based on the consensus across HLA-A alleles, the residue is located very close to the interface region (**Figure 1.2.2B**) and in fact was predicted as an interface residue on 2 of the 6 HLA-A allele structures analyzed. Additionally, although residue 209 (R209) on HLA-A and HLA-B proteins was annotated as ambiguous due to its intermediate value of relative solvent accessible surface area (RSA) for most HLA-A/B structures analyzed, the average RSA across structures is close to the threshold for core annotation (7.17), and R209 was indeed annotated as core in some of them. Overall, the distribution of HLA mutations for the three proteins was consistent with the previous report by Shukla et al. (*3*), though the current analysis incorporates an overall larger number of samples. Mutations in B2M were largely loss of function (**Figure 1.2.2A**) and more broadly distributed (**Figure 1.2.2C**), as expected for a tumor suppressor gene, though several positions were also recurrently mutated.

Expected effects of B2M versus HLA mutations on MHC-I composition. Since B2M is an essential component of all MHC-I molecules, loss of B2M should equally impact MHC-I molecules derived from different HLA alleles. The B2M interface with HLA alleles is shared

across the different alleles (**Figure S1.2.2**), so mutations at this interface are also likely to affect all variants of an individual's MHC-I molecule, although complexes involving B2M and binding partners that use alternative interfaces should not be affected. In contrast, loss of function or interface mutations affecting a specific HLA allele would only affect the MHC-I molecules derived from that allele. Thus, we speculate that B2M mutations are likely to reduce the total amount of MHC-I molecules presenting antigens on the tumor cell surface, while HLA mutations would impact which mutations could be presented as neoantigens.

Mutations in MHC-I proteins are associated with increased mutation burden. We hypothesized that both B2M and HLA mutations would affect MHC-I presentation of mutations. Mice with total lack of B2M express little if any cell surface MHC-I and lack cytotoxic CD8+ T Cells (*15*, *16*). In human lung cancers, an association was found between higher somatic mutation burden and HLA loss of heterozygosity (*4*). If somatic mutations to HLA and B2M similarly impair antigen presentation, we would expect to see an increased mutation burden when comparing to unmutated patients.

We first analyzed 9055 TCGA patients across 31 solid tumor types that had both exome and RNA sequencing data (**Figure 1.2.3A**), removing patients that had synonymous B2M or HLA mutations. We then performed a cancer-specific analysis of 3514 patients across 8 solid tumor types with at least 5 somatic B2M and HLA mutations (**Figure 1.2.3B**). To determine whether somatic mutations to B2M and HLA were associated with an overall higher mutation burden, we compared the total number of expressed nonsynonymous mutations in patients with and without nonsynonymous somatic B2M or HLA mutations. Overall, we observed that both patients with a B2M and an HLA mutation had significantly higher tumor mutation burdens (Mann Whitney test, B2M $p < 1.1e\text{-}20$ and HLA $p < 1.1e\text{-}30$) than patients without (**Figure 1.2.3A**). Pan-cancer, B2M

mutated patients also had significantly higher mutation burdens than HLA mutated patients (Mann Whitney test, $p < 0.0028$). There were approximately equal numbers of early stage (I & II) and late stage (III & IV) tumors in these three groups. We repeated the pan-cancer mutational burden analysis with Cancer Cell Line Encyclopedia (CCLE) data for 25 B2M-mutated cell lines, 114 HLA cell lines, and 1381 non-mutated cell lines, and observed the same trend: cell lines with B2M and HLA mutations had significantly higher overall mutational burden than cell lines without (**Figure S1.2.3**). When we analyzed tumors by tissue type, we observed that certain cancers (stomach adenocarcinoma, endometrial cancer, colorectal cancer, lung adenocarcinoma, and cervical cancer) also had significantly higher mutational burden in mutated patients (**Figure 1.2.3B**). Stomach, uterine and colorectal cancers have documented high rates of microsatellite instability (MSI), thus we evaluated whether B2M and HLA mutations were biased to occur in high MSI tumors. Using MSI annotations available for 10,415 patients from Kautto et al. (*17*), we found a significant bias for B2M and HLA mutations to occur in patients with MSI (Fisher's exact test; B2M OR = 14.66, $p < 8.7e\text{-}24$; HLA OR = 6.28, $p < 2.0e\text{-}36$). To rule out the possibility that MSI was solely driving our results, we reanalyzed the mutational burden between B2M and HLA mutated and unmutated patients, this time retaining only 8668 microsatellite stable (MSS) patients. Interestingly, we found similar trends in elevated mutational burden associated with B2M and HLA mutation (**Figure 1.2.3C-D**), and consequently focused on MSS patients only in the subsequent analyses. Thus, even in MSS tumors, B2M and HLA mutations are associated with an increased nonsynonymous mutational burden.

Mutations in MHC-I proteins are associated with increased binding neoantigen counts.

To obtain more evidence as to whether the elevated mutation counts observed in HLA and B2M mutated patients were a result of the mutation, or vice versa, we compared the fraction of mutations

likely to generate neoantigens across MSS patients with and without B2M and HLA mutations. We speculated that if B2M and HLA mutations are an artifact of higher mutation rates, the proportion of mutations that generate neoantigens should not differ relative to patients without such mutations. However, if these mutations truly facilitate immune escape, neoantigens should be enriched among the observed mutations.

Using HLA allele genotypes called by Polysolver (*3*), we calculated patient-specific MHC-I presentation scores for all expressed mutations observed in each patient's tumor (*12*, *18*). We previously demonstrated that these affinity-based presentation scores, called PHBR-I scores, can distinguish peptides found in complex with MHC-I on the cell surface in mass spectrometry experiments from random peptides simulated from the human proteome, supporting that affinity is a reasonable proxy for cell surface presentation (*12*). Indeed, when we looked at the fraction of expressed mutations considered to be neoantigens at various PHBR-I cutoffs, we found that at any given cutoff, a higher fraction of mutations represented neoantigens in both B2M and HLA mutated patients (**Figure 1.2.4A-B**). This corresponded to overall higher numbers of neoantigens in B2M and HLA mutant tumors (**Figure S.1.2.4**). The higher overall proportion of neoantigens is consistent with both somatic B2M and HLA mutations impairing presentation of neoantigens for immune surveillance.

Assessing bias in neoantigen affinities in patients with mutant HLA alleles. McGranahan *et al*. reported that in lung cancer, subclones that had lost a particular HLA allele tended to accumulate mutations with higher affinity for the lost allele, suggesting that such mutations were no longer subject to immunoediting (*4*). We therefore sought to assess whether mutations accumulating in tumors with HLA mutations showed a bias in affinity toward the affected HLA allele. We first evaluated whether the number of mutant-allele specific mutations in these patients

was higher than the average number of mutations specific to each of the other alleles (**Figure 1.2.4C**). We observed several patients for which the number of mutant-allele specific mutations was indeed higher (**Figure 1.2.4C; red lines**). We note that the current study design differs from the study by McGranahan *et al.* in that we do not have subclone-specific sequencing data, and thus cannot determine which mutations occurred in the same cell population as the mutated HLA allele. We also did not consider allele-specific deletion events, and thus the assumption that the other 5 HLA alleles are intact may be incorrect for some patients.

Timing of somatic mutations in MHC-I proteins. To better understand B2M and HLA mutation timelines, we analyzed the tumor allelic fraction of expressed mutations for all patients. Early clonal mutations are present in a larger fraction of cancer cells than later subclonal mutations and are, therefore, expected to be present in a higher fraction of the reads generated from that site during tumor sequencing. Although this assumption can be complicated by sampling bias and genomic instability of tumors, we nonetheless expect that somatic point mutations with higher read support will in general have occurred at earlier time points than those with lower read support. Since each individual's tumor is unique, we quantified B2M and HLA mutations in terms of their allelic fraction percentile relative to other mutations observed in the same tumor (**Figure 1.2.4D**). Interestingly, B2M mutations tended to be present at higher percentiles than most HLA mutations, suggesting that B2M mutations might occur earlier in tumor development and affect a higher proportion of tumor cells. Most HLA mutations had low percentiles, suggesting these were late, subclonal events, while a subset had high percentiles and likely occurred early during tumor development in those individuals. This observation agrees with the previous report by McGranahan *et al.* that found HLA loss in lung cancer to be predominantly subclonal with a few observations of clonal loss noted. Patients with MSI tended to have HLA mutations with higher

variant allele fraction (VAF) (Fisher's exact test, OR = 73.3, p < 8.1e-16). These findings remained even when we considered only mutations in regions unaffected by copy number changes which can confound VAF estimates (**Figure S1.2.5**). Interestingly, we found that tumors with early HLA mutations had significantly higher levels of neoantigens predicted to specifically bind to the mutated allele than tumors with late HLA mutations (**Figure 1.2.4E**). When we evaluated the bias in specificity of neoantigens for the mutated allele in patients with early HLA loss, we found a significant difference in the number of binding neoantigens between the mutated HLA allele and average of unmutated HLA alleles (**Figure 1.2.4F**). We conclude that somatic B2M and HLA mutations are associated with an overall higher burden of neoantigens, supporting the notion that these mutations facilitate tumor immune escape.

Correlation of B2M versus HLA mutation with immune cell infiltration and cytotoxicity.

Effective antigen presentation via MHC-I is associated with CD8+ T cell driven cytotoxicity. Furthermore, cell surface MHC-I molecules deliver an inhibitory signal to natural killer (NK) cells. Thus, changes to cell surface presentation of neoantigens by MHC-I due to mutations in B2M and HLA may be reflected in immune cell infiltration levels and levels of cytotoxicity. We quantified immune cell infiltration from tumor RNA sequencing data using Cibersort (*19*) and levels of cytotoxicity using the score proposed by Rooney et al. (*20*). While Shukla *et al.* previously evaluated immune infiltrates and cytotoxicity in the context of somatic HLA mutations, to our knowledge B2M mutations have not previously been analyzed in this context (*3*).

CD8+ T cell levels were elevated in tumors with HLA mutations, both pan-cancer (**Figure 1.2.5A**) and in several tumor types (**Figure 1.2.5B**). A possible explanation is that CD8+ T cells are primed in secondary lymphoid organs and travel to the tumor where they accumulate due to the lack of the corresponding MHC-I molecule/peptide complex. NK cell levels were elevated in

tumors with B2M mutations pan-cancer (**Figure 1.2.5C**), however the levels were not significantly different in any given tumor type (**Figure 1.2.5D**). Loss of B2M resulting in reduced cell surface MHC-I molecules should reduce the ability of tumor cells to inhibit NK cell driven cytotoxicity, however it is unclear whether this would affect NK cell levels in the tumor. Cytotoxicity was elevated in both HLA and B2M mutant tumors pan-cancer (**Figure 1.2.5E**) and in several tumor types (**Figure 1.2.5F**). These trends are consistent with the idea that mutations are a mechanism of escape from immune surveillance, as previously suggested by Shukla *et al.* for HLA mutations.

1.2.5 Discussion

Many immunotherapies, such as immune checkpoint inhibitors, rely on the integrity of a patient's immune system to eliminate tumors. Tumors use a variety of strategies to evade the immune system, raising important questions about how different mechanisms of immune evasion could impact response to particular immunotherapies. We found that somatic point mutations in proteins comprising the MHC-I, B2M and HLA, showed signs of positive selection in tumors. This observation motivated our study of the effects of somatic B2M and HLA mutations on accumulation of putative neoantigens in tumors.

Our analysis builds on work by Shukla *et al*. that first applied Polysolver to evaluate patterns of HLA mutation across tumors and showed that such mutations occurred preferentially in tumors with high mutation burden and under strong pressure by the immune system as evidenced by high levels of CD8+ T cell infiltration. Here we further analyze patterns of mutation in tumors with HLA mutations, incorporating information about which mutations are likely to be presented by MHC-I molecules derived from patient-specific HLA alleles, and comparing to tumors with B2M mutations or with unaltered MHC-I. Our analysis supports a model where B2M mutations

reduce the overall levels of cell surface MHC-I molecules while HLA mutations perturb the overall composition of the MHC-I complex landscape, both providing escape from immune surveillance. Our findings are consistent with those of McGranahan *et al.* who reported that somatic loss of heterozygosity in the HLA locus was a common mechanism of immune evasion, and that loss of a specific HLA allele could render a subset of neoantigens within the tumor ineffective at generating an immune response upon checkpoint inhibition. While both B2M and HLA mutated patients showed elevated mutation rates, we observed differences in how neoantigens accumulated in these tumors, with B2M mutant tumors harboring the most neoantigens and tumors with intact MHC-I molecules harboring the least.

Notably, B2M mutations were highly enriched in tumors with microsatellite instability, a phenomenon that has been previously observed in the context of colorectal cancer (*21*) and is now confirmed for other tumor types with high MSI. MSI tumors were associated with higher immune cell infiltration and robust immune responses in this disease (*22*). Previous studies have also linked B2M mutations to increased levels of local immune cytolytic activity in uterine, stomach, colorectal and breast cancer (*20*). It remains unclear to what extent high mutation burden precedes immune infiltration, cytotoxicity and escape via B2M or HLA mutation, or whether the rate and affinity characteristics of the mutations that occur after the event differ from those before. Grasso et al. (*23*) showed that MSI-H colorectal tumors disrupt B2M and HLA genes independent of mutational load with direct effect on T cell infiltration. We conclude that mutations to either component of MHC-I will provide effective escape in the setting of a robust anti-tumor immune response, however mutations to B2M may be more beneficial in settings such as MSI when the number of neoantigens generated is highest.

We note that the current analysis has several limitations. First, our analysis only considered mutations in HLA alleles, whereas other types of variation, including loss of heterozygosity or lack of expression could confer similar effects. In the current analysis, patients with such effects would be grouped with non-mutated tumors, which would reduce the statistical power of the analyses that we performed. In addition, we did not have information about the subclonal membership of particular mutations within the tumor, and thus could not distinguish mutations occurring in the subset of tumor cells with HLA mutation from other mutations in the tumor. Knowledge of the subclonal architecture of the tumor would be helpful to fully investigate the affinity bias of new mutations for the mutated HLA allele. Future studies should address these shortcomings.

Here we show that somatic mutations affecting B2M and HLA genes interact with the accumulation of somatic mutations that generate neoantigens during tumor development. Mutations in both genes relieve pressure by the immune system, allowing the tumor to evade an active immune response. A better understanding of how these mutations differ in shaping the oncogenic landscape may provide insights as to how these factors could contribute to resistance to therapies that induce strong local anti-tumor immunity.

1.2.6 Materials and Methods

Data. All available whole exome sequencing (WXS) data as of 5/3/2018 was downloaded from The Cancer Genome Atlas (TCGA) database via their Genomic Data Commons (GDC) client. Both .bam files and auxiliary .bai files were downloaded. All available somatic mutation data as of 5/21/2018 was downloaded from the TCGA database in the form of TCGA project

mutation annotation files (MAF). Clinical data were also obtained from the GDC (downloaded on 4/25/2017).

Protein structure analysis. Experimental 3D X-ray protein structures for the B2M and HLA-A/B/C complexes were obtained from the Protein Data Bank (PDB) (*14*). Amino acid residues of each PDB structure were annotated based on their 3D location in the protein as core and surface according to their relative solvent accessible surface area (RSA) calculated using Naccess (*24*). Residues with RSA higher than 15 were annotated as surface and residues with RSA lower than 5 were annotated as core, while residues with RSA values between 5 and 15 were annotated as ambiguous. Residues involved in the physical interaction between B2M and HLA proteins are predicted using KFC2 (*25*) and annotated as interface. PDB residue positions were mapped onto the UniProt residue positions using the PDBSWS server (*26*). UniProt residues are numbered based on their position in the protein sequence of the full-length protein, starting from 1. If multiple PDB structures were available for the same protein, we took consensus as the final annotation; and in the case of a tie, the residue was labeled as ambiguous. The residues without known 3D structure are also labeled as ambiguous. We had structures for 6 alleles of HLA-A protein, 15 alleles of HLA-B protein, and 3 alleles of HLA-C protein. We took consensus of residue annotations of different HLA alleles to annotate each HLA protein. VMD (*27*) is used to visualize protein 3D structures (**Figure 1.2.2B**). Exon information for HLA proteins is obtained from the IMGT/HLA database (v3.34; https://www.ebi.ac.uk/ipd/imgt/hla/) (*28*).

HLA typing and mutation calling. HLA genotyping and mutation calling was performed for HLA-A, HLA-B, and HLA-C genes, which encode the human MHC-I complex. We extracted scripts from the Broad Institute's Polysolver Docker container (https://software.broadinstitute.org/cancer/cga/polysolver_run). We verified that the majority of

Polysolver's HLA calls were consistent with that of xHLA (*29*) and therefore used all available Polysolver results. B2M mutations were taken directly from the TCGA MAF files. Patients with somatic B2M or HLA mutations were grouped for subsequent analysis and compared to patients that had neither. We found that only 13 patients had both B2M and HLA mutations.

Microsatellite instability. Microsatellite instability scores for all TCGA patients were obtained from Kautto et al., 2017. Patient MANTIS scores from the paper were binarized to microsatellite instable (MSI-H) and stable (MSS) according to the recommended MANTIS score threshold of 0.4 (*17*).

Determining expressed mutations. We used the bam-readcount tool (https://github.com/genome/bam-readcount) to determine how many RNAseq reads covered a mutated position. To count a mutation as being expressed, we used a read count threshold of 5.

Determining regions with CNVs. Regions affected by copy number variants were determined from TCGA affymetrix SNP6 data by using 0.1 thresholds as the cutoff in either direction. Thus, any region that has a log2 fold change larger than 0.1 or smaller than -0.1 is defined as a position with copy number variation (*30*). For Figure S1.2.7 we excluded any mutations that occurred in regions with copy number variation.

Mutation burden. Mutation counts were obtained from TCGA MAF files for all patients. To obtain nonsynonymous counts, we filtered out mutations outside of coding regions as well as silent mutations and tallied the remaining mutations for each patient. We retained only expressed mutations, and added a pseudocount of 1 for all patients, for all mutation burden analyses. For cancer-type-specific analysis, patients from TCGA tumor types COAD and READ were merged under the name CRC (colon and rectal cancer).

Antigen affinity. We used the netMHCpan4.0 tool (*18*, *31*) to obtain mutation affinity scores for all patient HLA alleles. To determine whether a mutation would be effectively bound as a neoantigen to the MHC-I complex, we binarized affinity scores: mutations with scores <= 2 we considered binding, and mutations with scores > 2 we considered non-binding (*12*, *18*). We then took the harmonic mean of the best ranking neoantigen to calculate the Patient Harmonic-mean Best Rank (PHBR) score (*12*). To evaluate differences in fraction of binding neoantigens at various presentation score (PHBR-I score) cutoffs, we plotted the empirical cumulative distribution function (ECDF) using the median fraction of neoantigens generated from expressed mutations across patients. The Kolmogorov-Smirnov test was used to determine whether the distribution of neoantigen fractions was significantly different for each group (**Figure 1.2.4A**). To determine if the number of neoantigens was significantly different between mutated and control patients at a particular PHBR-I score threshold, we calculated p-values using an unpaired Mann Whitney test for pan-cancer comparisons (**Figure 1.2.4B**). To test the significance of the number of neoantigens between mutated and unmutated HLA alleles, we used a paired Wilcoxon test (**Figure 1.2.4C, Figure 1.2.4F**). The Kolmogorov-Smirnov, Mann Whitney, and Wilcoxon tests implemented in the scipy.stats Python package were used for these analyses.

Allelic fraction analysis. For Polysolver-determined HLA mutations, we obtained the tumor allelic fraction ("tumor_f") from the Mutect output files generated by Polysolver. For all other mutations we calculated tumor allelic fraction from tumor alternate allele reads ("t_alt_count") and tumor read depth ("t_depth") from TCGA MAF files. B2M and HLA mutations were further annotated according to their percentile within the ranked list of mutations in the tumor where they were observed. To determine if the distributions of patients with B2M and

HLA mutations were significantly different than patients without these mutations, we used an unpaired Mann Whitney statistical test from the scipy.stats Python package.

Immune infiltration and cytotoxicity. Immune cell infiltration levels for CD8+ T cells and natural killer cells were obtained by running Cibersort with default parameters and without quantile normalization, on log2 TPM values obtained by reprocessing the TCGA RNAseq data through Sailfish V0.7.6 (*32*). Cytotoxicity was estimated as described in (*20*), by summing the z-scored log2 TPM expression values of granzyme A (*GZMA*) and perforin (*PRF1*). For cancer-type-specific analysis, patients from TCGA tumor types COAD and READ were merged under the name CRC (colon and rectal cancer).

Other statistical considerations. Where appropriate, p-values were adjusted for multiple comparisons using the Benjamini-Hochberg method (*33*).

1.2.7 Figures

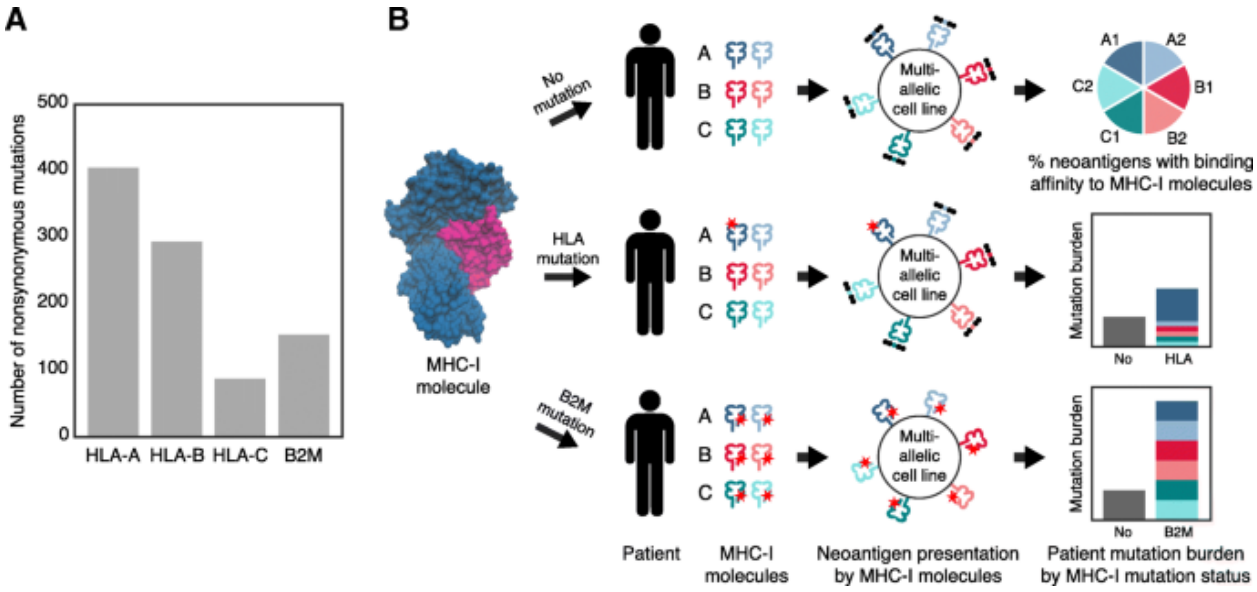
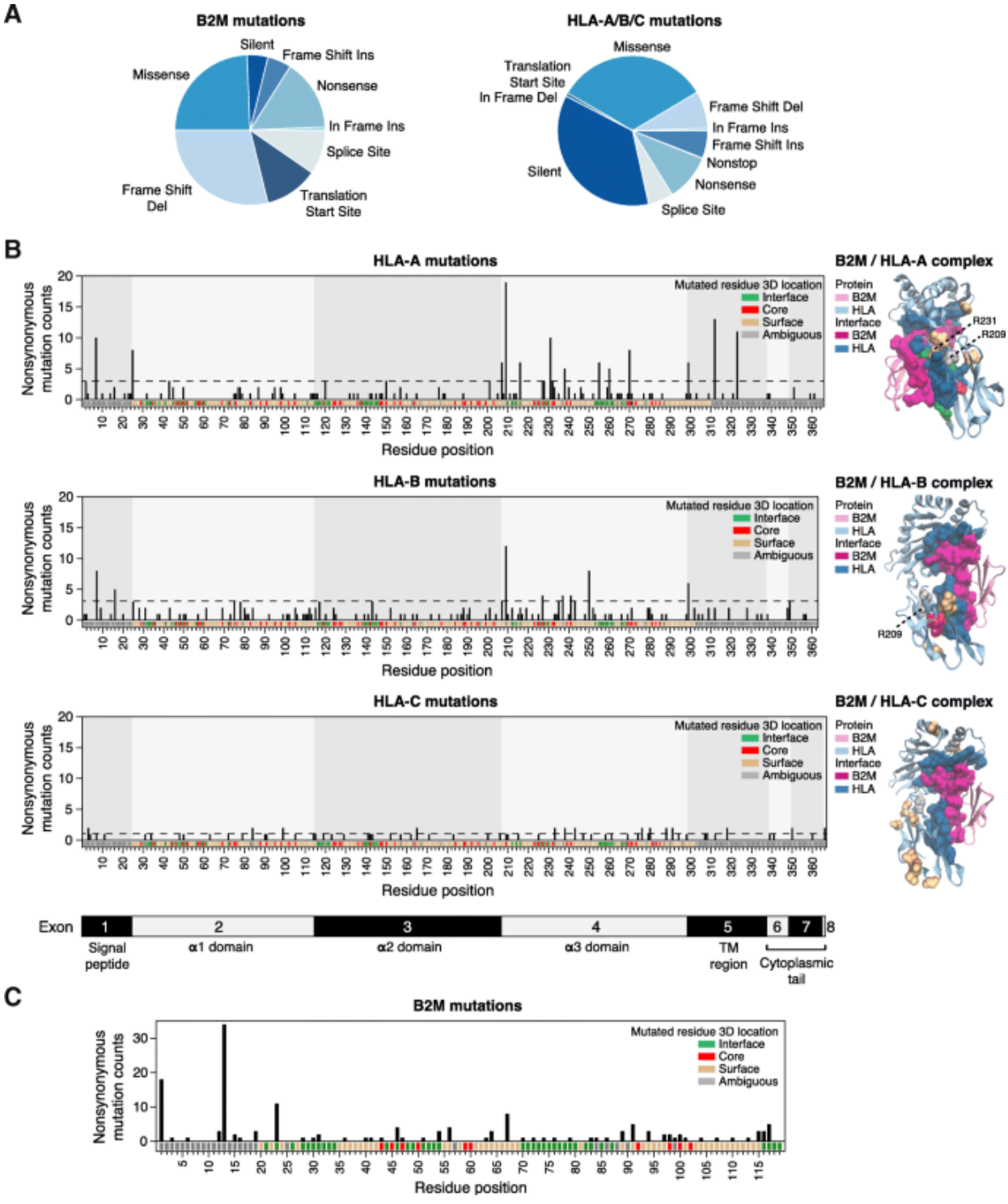


Figure 1.2.1. Somatic mutations affecting components of the MHC-I molecule. (**A**) The total number of nonsynonymous mutations targeting the genes encoding the components of the MHC-I complex, B2M and HLA-A, HLA-B or HLA-C proteins, across all TCGA patients. The HLA mutation counts were obtained via Polysolver. (**B**) Schematic representation of the effects of mutations that alter the cell surface composition of MHC-I. An HLA mutation will affect a specific MHC-I molecule, whereas a B2M mutation will affect all MHC-I molecules; both mutations can increase the mutation burden of the patient. In the case of an HLA mutation, the patient mutation burden should include more neoantigens with affinity for the mutated HLA allele. The MHC-I molecule displayed is composed of B2M (pink) and HLA-A (blue) proteins (PDB: 3bo8)

Figure 1.2.2. Mutational analysis of MHC-I complex. (**A**) Pie charts displaying percentages of types of mutation for the B2M protein; and for the combined HLA-A, HLA-B and HLA-C proteins, respectively, across all TCGA patients. (**B**) Distribution of nonsynonymous mutation counts, obtained from Polysolver, for HLA-A, HLA-B, and HLA-C proteins, across functional domains. The corresponding functional domains of HLA proteins are shown at the bottom. The UniProt sequential residue numbering scheme is used for residue numbering, which requires subtraction of the signal peptide (24 residues) for mapping to the IMGT/HLA residue numbering scheme. On the right, 3D crystal structures of MHC-I complex are displayed as B2M and HLA-A complex (PDB: 3bo8), as B2M and HLA-B complex (PDB: 3b3i), and as B2M and HLA-C complex (PDB: 4 nt6). Purple ribbons indicate B2M protein, while blue ribbons indicate the HLA proteins. The highlighted purple and blue residues correspond to the interface regions of B2M and HLA proteins, respectively. Hotspot mutations for HLA proteins (frequency > 3 for HLA-A, frequency > 3 for HLA-B, and frequency > 1 for HLA-C) are highlighted as green, red, tan and gray indicating interface, core, surface, and ambiguous residues, respectively. (**C**) Distribution of nonsynonymous mutation counts across the entire B2M protein. On the bottom of the plot, all amino acid residues of B2M protein are colored based on their 3D location: interface, core, surface, or ambiguous.


A
B2M mutations
Silent
Frame Shift Ins
Missense
Nonsense
In Frame Ins
Splice Site
Frame Shift Del
Translation Start Site
HLA-A/B/C mutations
Missense
Translation Start Site
In Frame Del
Frame Shift Del
In Frame Ins
Frame Shift Ins
Nonstop
Silent
Nonsense
Splice Site
B
HLA-A mutations
HLA-B mutations
HLA-C mutations
Nonsynonymous mutation counts
Mutated residue 3D location
Interface
Core
Surface
Ambiguous
Residue position
B2M / HLA-A complex
B2M / HLA-B complex
B2M / HLA-C complex
Protein
B2M
HLA
Interface
B2M
HLA
R231
R209
Exon
1
2
3
4
5
6
7
8
Signal peptide
α1 domain
α2 domain
α3 domain
TM region
Cytoplasmic tail
C
B2M mutations

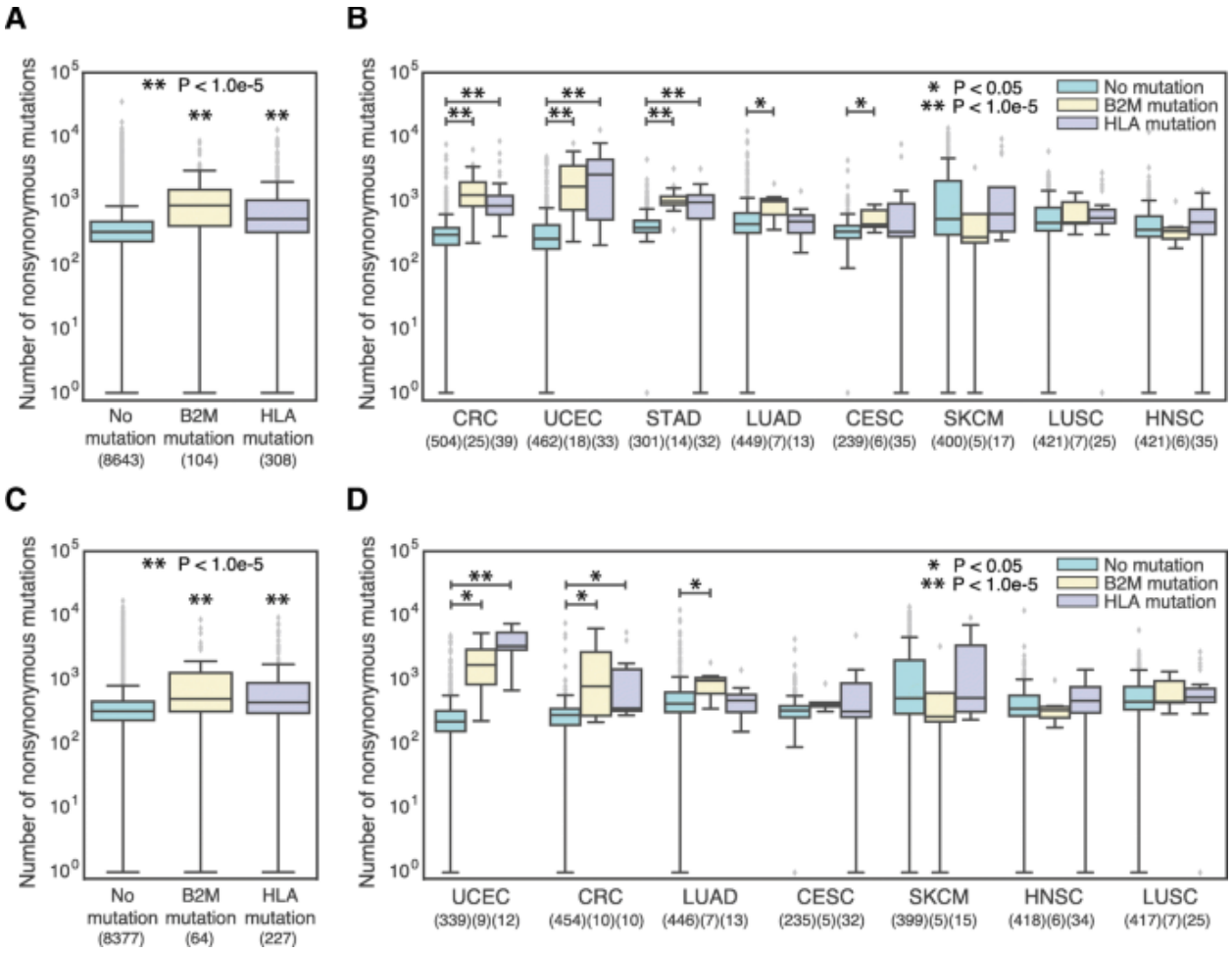


Figure 1.2.3. Increased mutational burden is related to mutations in MHC-I. (A) and **(C)** Boxplots showing the total number of expressed nonsynonymous mutations of TCGA patients who acquired a mutation in their B2M protein or in one of their HLA alleles versus patients who did not acquire any B2M or HLA mutation, (**A**) for all patients, and (**c**) for only MSS patients. Sample sizes for each patient group are written under their name. **(B)** and **(D)** Boxplots showing total number of expressed nonsynonymous mutations for TCGA patients with or without B2M or HLA mutations, (**C**) for all patients, and (**D**) for only MSS patients. Patients are divided by tumor type and only the tumor types with at least 5 mutated patients are shown. *P*-values are adjusted for multiple comparisons using the Benjamini–Hochberg procedure. Sample sizes for each patient group are written under the tissue name.

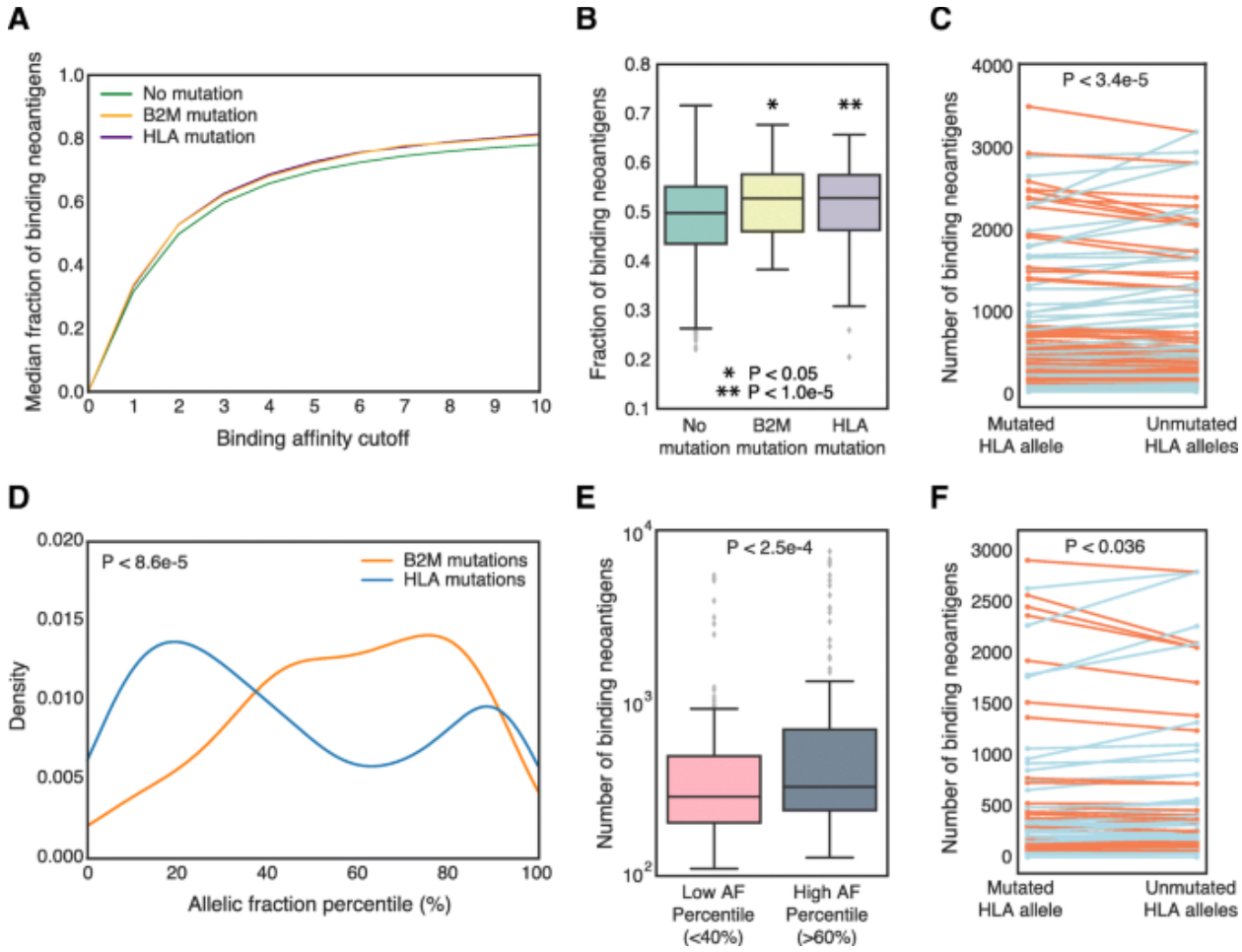


Figure 1.2.4. Analysis of binding neoantigens to patient HLA alleles. (A) Empirical cumulative distribution function showing the proportion of expressed missense and indel mutations labeled as binding neoantigens at different PHBR-I score cutoffs in MSS patients. **(B)** Boxplots comparing the fraction of binding neoantigens in tumors with no B2M or HLA mutation (teal) versus patients with a B2M mutation (yellow) or HLA mutation (purple). A PHBR-I score cutoff of 2 was used to designate a binding neoantigen for this comparison. **(C)** Total number of neoantigens that bind to a patient's mutated HLA allele versus the average number of neoantigens across the unmutated HLA alleles across all cancer types for MSS patients. A red line indicates that there are more neoantigens with binding affinity to the mutated HLA allele than the average across the unmutated HLA alleles; and a blue line depicts the opposite trend. **(D)** Allelic fraction percentile distribution for expressed mutations in MSS patients with B2M and HLA mutations. We used the Kolmogorov-Smirnov statistic to determine whether the two distributions were significantly different. **(E)** Comparison of the number of expressed neoantigens with binding affinity to the patient-specific mutated allele between the low AF percentile (< 40%) and the high AF percentile (> 60%) HLA mutated patients. Patients with MSI and with mutations in both B2M and HLA genes were excluded. **(F)** Comparison of the total number of neoantigens that bind to a patient-specific mutated HLA allele versus the average number of neoantigens with binding affinity to the five unmutated HLA alleles in patients with high allelic fraction percentile HLA mutations.

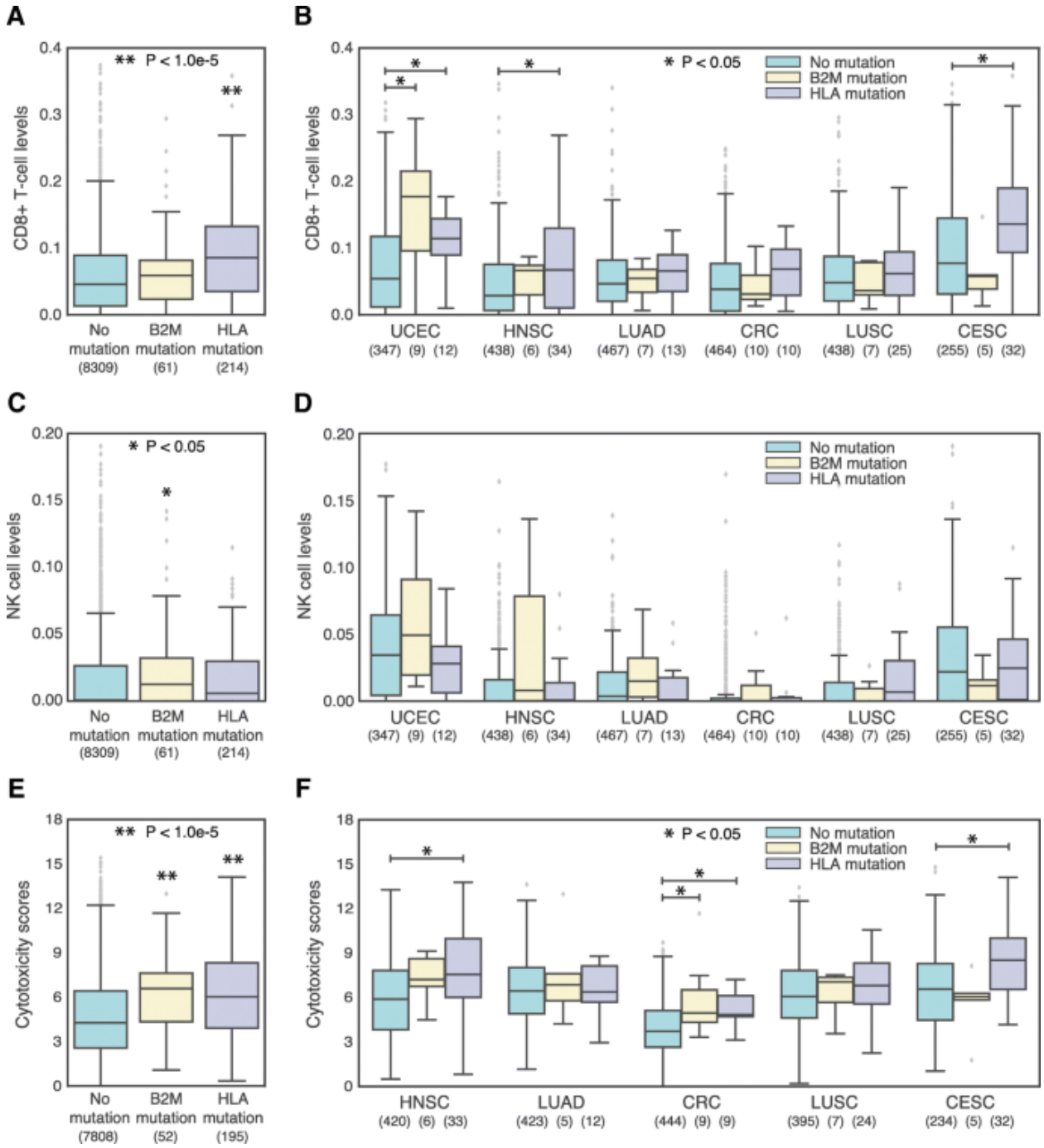


Figure 1.2.5. Increased NK, CD8+ T-cell and cytotoxicity levels are associated with mutations in MHC-I. (A) and **(C)** and **(E)** Boxplots comparing MSS TCGA patients with or without B2M or HLA mutations, in terms of their (**A**) CD8+ T cell levels, (**c**) natural killer (NK) cell levels, and (**E**) cytotoxicity scores. Sample sizes for each patient group are written under their name. **B** and **D** and **F**) Boxplots comparing MSS TCGA patients with or without B2M or HLA mutations, in terms of their (**B**) CD8+ T cell levels, (**D**) natural killer (NK) cell levels, and (**F**) cytotoxicity scores. Patients are divided by tumor type and only the tumor types with at least 5 mutated patients are shown. *P*-values are adjusted for multiple comparisons. Sample sizes for each patient group are written under the tissue name.

1.2.8 Supplemental Data, Tables and Figures

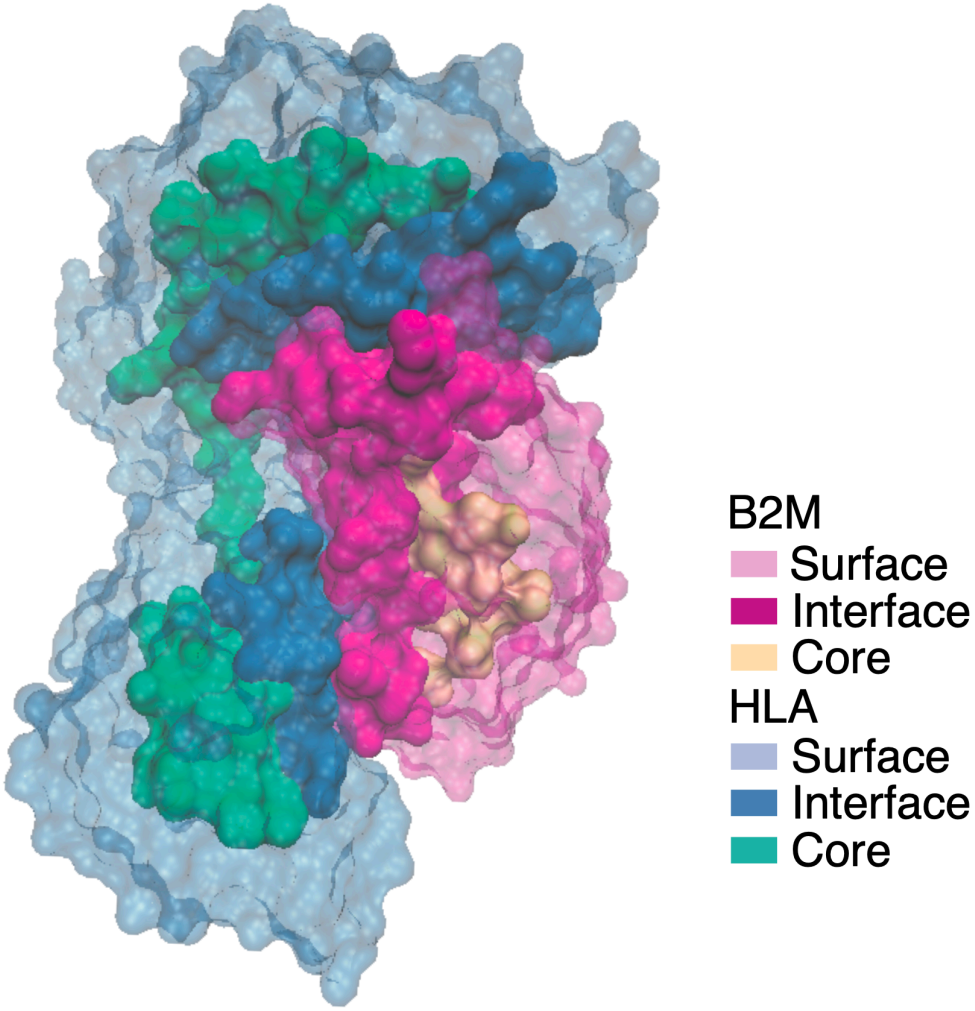


Figure S1.2.1 MHC-I complex 3D structure. 3D crystal structure of MHC-I complex is displayed as B2M/HLA-A complex (PDB: 3bo8). Interface (blue and violet) and core (green and orange) regions of B2M and HLA-A proteins are highlighted, respectively. Transparent blue and violet regions correspond to the surface regions of B2M and HLA-A proteins, respectively.

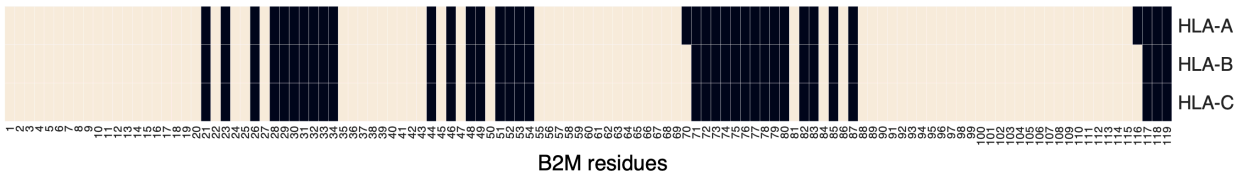


Figure S1.2.2 B2M interface residue positions for HLA alleles. Residues on B2M that interact with HLA-A, HLA-B, HLA-C proteins are highlighted black.

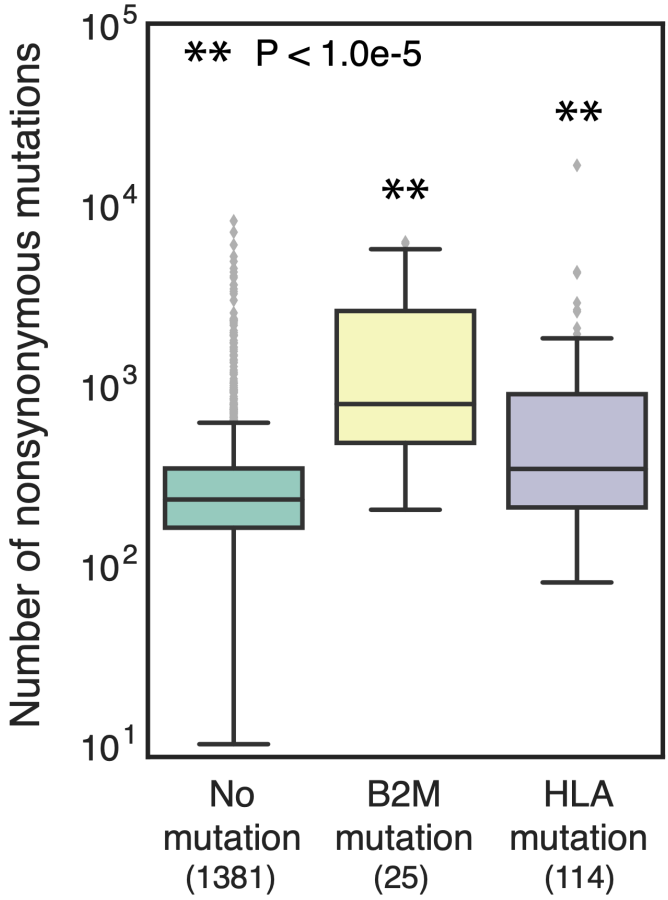


Figure S1.2.3. Mutation burden in CCLE. Boxplots showing the total number of nonsynonymous mutations for CCLE cell lines who acquired a B2M or HLA versus cell lines that did not acquire any B2M or HLA mutation. Sample sizes for each group are written under their name.

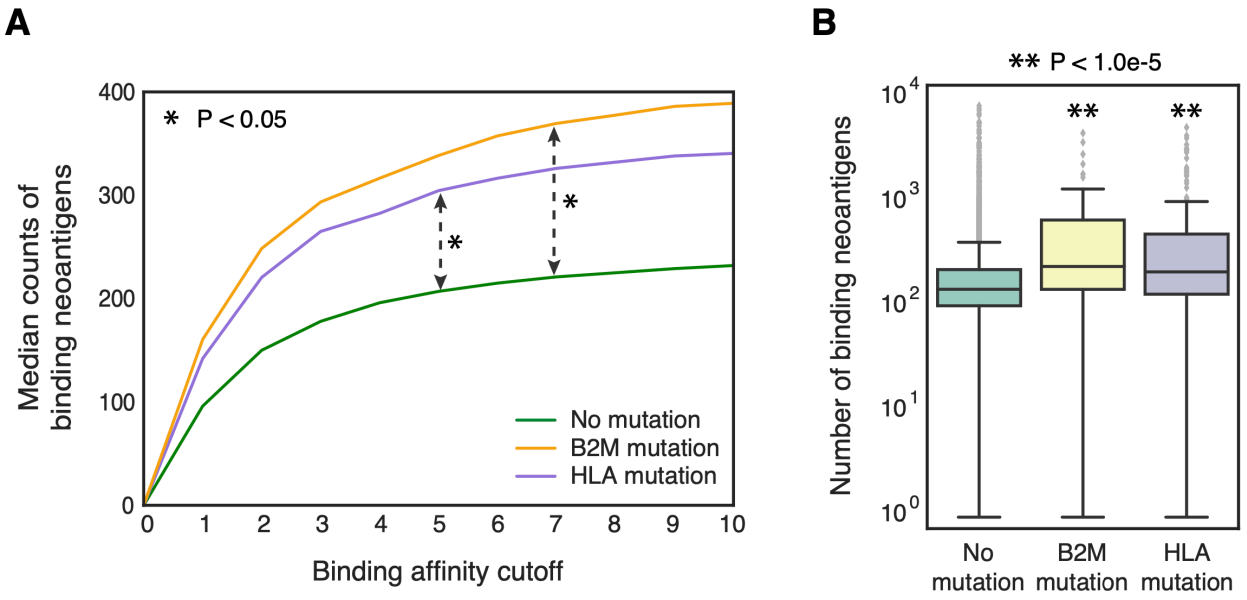


Figure S1.2.4. Total number of binding neoantigens to patient HLA alleles. (A) Distribution of median total counts of binding neoantigens at different PHBR-I score cutoffs for MSS patients. (B) Boxplots comparing the number of neoantigens in MSS patients with no B2M or HLA mutation (teal) versus MSS patients with a B2M mutation (yellow) or an HLA mutation (purple). A PHBR-I score cutoff of 2 was used to designate a binding neoantigen for this comparison.

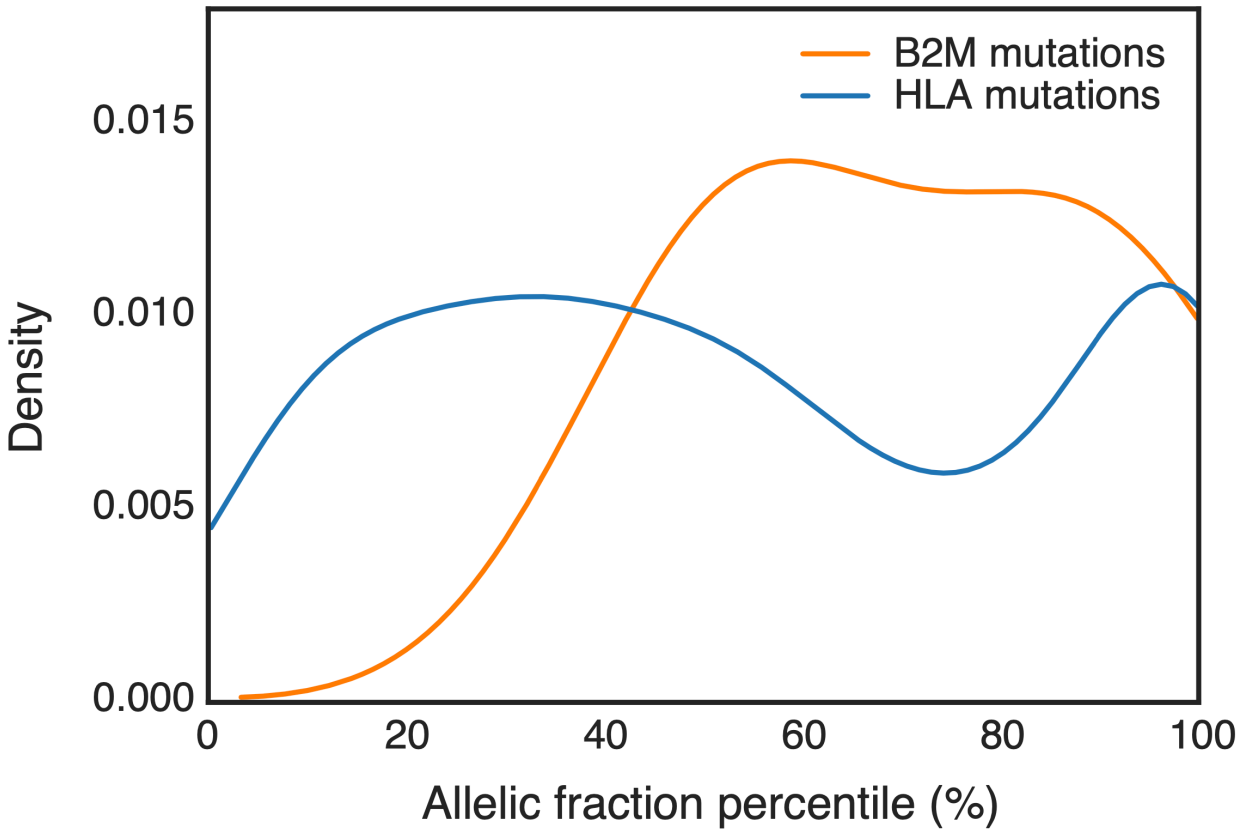


Figure S1.2.5. Allelic fraction percentile distribution for patients with B2M and HLA mutations accounting for aneuploidy. Allelic fraction percentile distribution for expressed mutations in MSS patients with B2M and HLA mutations, excluding all mutations occurring in regions affected by CNVs. Patients that have both B2M and HLA mutations are excluded.

1.2.9 Author Contributions

Original concept: Hannah Carter

Project supervisor: Hannah Carter

Data analysis: Andrea Castro and Kivilcim Ozturk

Data analysis and manuscript draft: Andrea Castro and Kivilcim Ozturk

Data analysis assistance: Rachel M Pyke

Immune cell infiltration analysis: Su Xian

Figure design: Andrea Castro, Kivilcim Ozturk, Hannah Carter

Manuscript writing: Andrea Castro, Kivilcim Ozturk, Maurizio Zanetti and Hannah Carter

1.2.10 Acknowledgements

This work was supported by NIH grants DP5-OD017937 and a CIFAR fellowship to H.C., the SDCSB/CCMI Systems Biology training grant (GM085764 and CA209891) to K.O., and the NIH National Library of Medicine training grant (T15LM011271) to A.C. The results published here are based upon data generated by The Cancer Genome Atlas managed by the NCI and NHGRI. Information about TCGA can be found at https://cancergenome.nih.gov.

All computing was done using the National Resource for Network Biology (NRNB) P41 GM103504.

Chapter 1.2, in full, is a reformatted reprint of the material as it appears in "Elevated neoantigen levels in tumors with somatic mutations in the HLA-A, HLA-B, HLA-C and B2M genes" in *BMC Medical Genomics,* 2019 by Andrea Castro, Kivilcim Ozturk, Rachel Marty Pyke, Su Xian, Maurizio Zanetti, and Hannah Carter. The dissertation author was a primary investigator and author of this paper.

1.2.11 References

1. M. Sade-Feldman, Y. J. Jiao, J. H. Chen, M. S. Rooney, M. Barzily-Rokni, J.-P. Eliane, S. L. Bjorgaard, M. R. Hammond, H. Vitzthum, S. M. Blackmon, D. T. Frederick, M. Hazar-Rethinam, B. A. Nadres, E. E. Van Seventer, S. A. Shukla, K. Yizhak, J. P. Ray, D. Rosebrock, D. Livitz, V. Adalsteinsson, G. Getz, L. M. Duncan, B. Li, R. B. Corcoran, D. P. Lawrence, A. Stemmer-Rachamimov, G. M. Boland, D. A. Landau, K. T. Flaherty, R. J. Sullivan, N. Hacohen, Resistance to checkpoint blockade therapy through inactivation of antigen presentation. *Nat. Commun.* **8**, 1136 (2017).

2. A. B. del Campo, J. A. Kyte, J. Carretero, S. Zinchencko, R. Méndez, G. González-Aseguinolaza, F. Ruiz-Cabello, S. Aamdal, G. Gaudernack, F. Garrido, Others, Immune escape of cancer cells with beta2-microglobulin loss over the course of metastatic melanoma. *International journal of cancer*. **134**, 102–113 (2014).

3. S. A. Shukla, M. S. Rooney, M. Rajasagi, G. Tiao, P. M. Dixon, M. S. Lawrence, J. Stevens, W. J. Lane, J. L. Dellagatta, S. Steelman, C. Sougnez, K. Cibulskis, A. Kiezun, N. Hacohen, V. Brusic, C. J. Wu, G. Getz, Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes. *Nat. Biotechnol.* **33**, 1152–1158 (2015).

4. N. McGranahan, R. Rosenthal, C. T. Hiley, A. J. Rowan, T. B. K. Watkins, G. A. Wilson, N. J. Birkbak, S. Veeriah, P. Van Loo, J. Herrero, C. Swanton, TRACERx Consortium, Allele-Specific HLA Loss and Immune Escape in Lung Cancer Evolution. *Cell*. **171**, 1259–1271.e11 (2017).

5. M. J. Christopher, A. A. Petti, M. P. Rettig, C. A. Miller, E. Chendamarai, E. J. Duncavage, J. M. Klco, N. M. Helton, M. O'Laughlin, C. C. Fronick, R. S. Fulton, R. K. Wilson, L. D. Wartman, J. S. Welch, S. E. Heath, J. D. Baty, J. E. Payton, T. A. Graubert, D. C. Link, M. J. Walter, P. Westervelt, T. J. Ley, J. F. DiPersio, Immune Escape of Relapsed AML Cells after Allogeneic Transplantation. *N. Engl. J. Med.* **379**, 2330–2341 (2018).

6. M. Pagadala, V. H. Wu, E. Pérez-Guijarro, H. Kim, A. Castro, J. Talwar, C. Gonzalez-Colin, S. Cao, B. J. Schmiedel, R. M. Salem, G. P. Morris, O. Harismendy, S. P. Patel, J. P. Mesirov, M. Zanetti, C.-P. Day, C. C. Fan, W. K. Thompson, G. Merlino, J. Silvio Gutkind, P. Vijayanand, H. Carter, Germline variants that influence the tumor immune microenvironment also drive response to immunotherapy. *bioRxiv* (2021), p. 2021.04.14.436660.

7. S. D. Bradley, Z. Chen, B. Melendez, A. Talukder, J. S. Khalili, T. Rodriguez-Cruz, S. Liu, M. Whittington, W. Deng, F. Li, C. Bernatchez, L. G. Radvanyi, M. A. Davies, P. Hwu, G. Lizée, BRAFV600E Co-opts a Conserved MHC Class I Internalization Pathway to Diminish Antigen Presentation and CD8+ T-cell Recognition of Melanoma. *Cancer Immunol Res*. **3**, 602–609 (2015).

8. T. Kambayashi, J. Michaëlsson, L. Fahlén, B. J. Chambers, C. L. Sentman, K. Kärre, H. G. Ljunggren, Purified MHC class I molecules inhibit activated NK cells in a cell-free system in vitro. *Eur. J. Immunol.* **31**, 869–875 (2001).

9. D. Hanahan, R. A. Weinberg, Hallmarks of cancer: the next generation. *Cell*. **144**, 646–674 (2011).

10. T. N. Schumacher, R. D. Schreiber, Neoantigens in cancer immunotherapy. *Science*. **348**, 69–74 (2015).

11. S. A. Rosenberg, Raising the bar: the curative potential of human cancer immunotherapy. *Sci. Transl. Med.* **4**, 127ps8 (2012).

12. R. Marty, S. Kaabinejadian, D. Rossell, M. J. Slifker, J. van de Haar, H. B. Engin, N. de Prisco, T. Ideker, W. H. Hildebrand, J. Font-Burgada, H. Carter, MHC-I Genotype Restricts the Oncogenic Mutational Landscape. *Cell*. **171**, 1272–1283.e15 (2017).

13. H. B. Engin, J. F. Kreisberg, H. Carter, Structure-Based Analysis Reveals Cancer Missense Mutations Target Protein Interaction Interfaces. *PLoS One*. **11**, e0152929 (2016).

14. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne, The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).

15. M. Zijlstra, M. Bix, N. E. Simister, J. M. Loring, D. H. Raulet, R. Jaenisch, Beta 2-microglobulin deficient mice lack CD4-8+ cytolytic T cells. *Nature*. **344**, 742–746 (1990).

16. B. H. Koller, P. Marrack, J. W. Kappler, O. Smithies, Normal development of mice deficient in beta 2M, MHC class I proteins, and CD8+ T cells. *Science*. **248**, 1227–1230 (1990).

17. E. A. Kautto, R. Bonneville, J. Miya, L. Yu, M. A. Krook, J. W. Reeser, S. Roychowdhury, Performance evaluation for rapid detection of pan-cancer microsatellite instability with MANTIS. *Oncotarget*. **8**, 7452–7463 (2017).

18. I. Hoof, B. Peters, J. Sidney, L. E. Pedersen, A. Sette, O. Lund, S. Buus, M. Nielsen, NetMHCpan, a method for MHC class I binding prediction beyond humans. *Immunogenetics*. **61**, 1–13 (2009).

19. A. M. Newman, C. L. Liu, M. R. Green, A. J. Gentles, W. Feng, Y. Xu, C. D. Hoang, M. Diehn, A. A. Alizadeh, Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods*. **12**, 453–457 (2015).

20. M. S. Rooney, S. A. Shukla, C. J. Wu, G. Getz, N. Hacohen, Molecular and genetic properties of tumors associated with local immune cytolytic activity. *Cell*. **160**, 48–61 (2015).

21. M. Kloor, S. Michel, M. von Knebel Doeberitz, Immune evasion of microsatellite unstable colorectal cancers. *Int. J. Cancer*. **127**, 1001–1010 (2010).

22. M. Kloor, M. von Knebel Doeberitz, The Immune Biology of Microsatellite-Unstable Cancer. *Trends Cancer Res.* **2**, 121–133 (2016).

23. C. S. Grasso, M. Giannakis, D. K. Wells, T. Hamada, X. J. Mu, M. Quist, J. A. Nowak, R. Nishihara, Z. R. Qian, K. Inamura, T. Morikawa, K. Nosho, G. Abril-Rodriguez, C. Connolly,

H. Escuin-Ordinas, M. S. Geybels, W. M. Grady, L. Hsu, S. Hu-Lieskovan, J. R. Huyghe, Y. J. Kim, P. Krystofinski, M. D. M. Leiserson, D. J. Montoya, B. B. Nadel, M. Pellegrini, C. C. Pritchard, C. Puig-Saus, E. H. Quist, B. J. Raphael, S. J. Salipante, D. S. Shin, E. Shinbrot, B. Shirts, S. Shukla, J. L. Stanford, W. Sun, J. Tsoi, A. Upfill-Brown, D. A. Wheeler, C. J. Wu, M. Yu, S. H. Zaidi, J. M. Zaretsky, S. B. Gabriel, E. S. Lander, L. A. Garraway, T. J. Hudson, C. S. Fuchs, A. Ribas, S. Ogino, U. Peters, Genetic Mechanisms of Immune Evasion in Colorectal Cancer. *Cancer Discov.* **8**, 730–749 (2018).

24. S. J. Hubbard, J. M. Thornton, NACCESS: Department of Biochemistry and Molecular Biology, University College London. *Software available at http://www. bioinf. manchester. ac. uk/naccess/nacdownload. html* (1993).

25. X. Zhu, J. C. Mitchell, KFC2: a knowledge-based hot spot prediction method based on interface solvation, atomic density, and plasticity features. *Proteins*. **79**, 2671–2683 (2011).

26. A. C. R. Martin, Mapping PDB chains to UniProtKB entries. *Bioinformatics*. **21**, 4297–4301 (2005).

27. W. Humphrey, A. Dalke, K. Schulten, VMD: visual molecular dynamics. *J. Mol. Graph.* **14**, 33–8, 27–8 (1996).

28. J. Robinson, J. A. Halliwell, J. D. Hayhurst, P. Flicek, P. Parham, S. G. E. Marsh, The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Res.* **43**, D423–31 (2015).

29. C. Xie, Z. X. Yeo, M. Wong, J. Piper, T. Long, E. F. Kirkness, W. H. Biggs, K. Bloom, S. Spellman, C. Vierra-Green, C. Brady, R. H. Scheuermann, A. Telenti, S. Howard, S. Brewerton, Y. Turpaz, J. C. Venter, Fast and accurate HLA typing from short-read next-generation sequence data with xHLA. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 8059–8064 (2017).

30. R. Beroukhim, C. H. Mermel, D. Porter, G. Wei, S. Raychaudhuri, J. Donovan, J. Barretina, J. S. Boehm, J. Dobson, M. Urashima, K. T. Mc Henry, R. M. Pinchback, A. H. Ligon, Y.-J. Cho, L. Haery, H. Greulich, M. Reich, W. Winckler, M. S. Lawrence, B. A. Weir, K. E. Tanaka, D. Y. Chiang, A. J. Bass, A. Loo, C. Hoffman, J. Prensner, T. Liefeld, Q. Gao, D. Yecies, S. Signoretti, E. Maher, F. J. Kaye, H. Sasaki, J. E. Tepper, J. A. Fletcher, J. Tabernero, J. Baselga, M.-S. Tsao, F. Demichelis, M. A. Rubin, P. A. Janne, M. J. Daly, C. Nucera, R. L. Levine, B. L. Ebert, S. Gabriel, A. K. Rustgi, C. R. Antonescu, M. Ladanyi, A. Letai, L. A. Garraway, M. Loda, D. G. Beer, L. D. True, A. Okamoto, S. L. Pomeroy, S. Singer, T. R. Golub, E. S. Lander, G. Getz, W. R. Sellers, M. Meyerson, The landscape of somatic copy-number alteration across human cancers. *Nature*. **463**, 899–905 (2010).

31. V. Jurtz, S. Paul, M. Andreatta, P. Marcatili, B. Peters, M. Nielsen, NetMHCpan-4.0: Improved Peptide-MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. *J. Immunol.* **199**, 3360–3368 (2017).

32. R. Patro, S. M. Mount, C. Kingsford, Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nat. Biotechnol.* **32**, 462–464 (2014).

33. Y. Benjamini, Y. Hochberg, Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Series B Stat. Methodol.* **57**, 289–300 (1995).

1.3.1 Foreword

Up until this point, my research has largely focused on the role of the MHC in cancer. However, as gatekeepers of adaptive immunity, MHC class I and II are relevant in many other diseases and disorders, including infection and autoimmunity. MHC-I and MHC-II, though MHC-II is usually only present on professional antigen-presenting cells such as macrophages, B cells, and dendritic cells, also influence autoimmune susceptibility (1) and infection. Historically, certain alleles have been identified as protective or predisposing against autoimmune disorders including Type-1 diabetes, Rheumatoid arthritis, Celiac disease, Ankylosing spondylitis, and Multiple sclerosis (2, 3). This protection or predisposition stems from clonal elimination or retention of T cells during thymic development (4). MHC presentation of antigen during thymic development is required for T cell selection (5); T cells that recognize self-peptides are clonally eliminated, leaving behind clones that should be able to recognize a multitude of foreign antigen if presented via MHC.

In infection, MHC-I can present intracellular foreign peptides for recognition and elimination by CD8+ T cells (6), while MHC-II presents foreign peptides obtained via mechanisms ranging from macropinocytosis, endocytosis, phagocytosis and autophagy (7) to CD4+ T cells. Stimulated CD4+ T cells can help activate macrophages, CD8+ T cells, and B cells to fight infection (8, 9). Despite the fact that the majority of presented foreign peptides on MHC-I are immunogenic (10), MHC presentation can be a potential bottleneck for adaptive immunity. Like some tumor cells, viruses can downregulate MHC-I cell surface expression to avoid immune detection (11) in addition to targeting along the antigen presentation pathway. Here, I describe another potential mechanism of immune evasion by SARS-CoV-2, where overall poor MHC-II presentation of the critical ACE2 receptor binding motif (RBM) may affect subsequent CD4+ T cell stimulation and maturation of neutralizing antibody generating B cells.

1.3.2 Abstract

SARS-CoV-2 antibodies develop within two weeks of infection, but wane relatively rapidly post-infection, raising concerns about whether antibody responses will provide protection upon re-exposure. Here we revisit T-B cooperation as a prerequisite for effective and durable neutralizing antibody responses centered on a mutationally constrained RBM B cell epitope. T-B cooperation requires co-processing of B and T cell epitopes by the same B cell and is subject to MHC-II restriction. We evaluated MHC-II constraints relevant to the neutralizing antibody response to a mutationally-constrained B cell epitope in the receptor binding motif (RBM) of the spike protein. Examining common MHC-II alleles, we found that peptides surrounding this key B cell epitope are predicted to bind poorly, suggesting a lack MHC-II support in T-B cooperation, impacting generation of high-potency neutralizing antibodies in the general population. Additionally, we found that multiple microbial peptides had potential for RBM cross-reactivity, supporting previous exposures as a possible source of T cell memory.

1.3.3 Introduction

Upon infection with SARS-CoV-2 the individual undergoes seroconversion. In mildly symptomatic patients, seroconversion occurs between day 7 and 14, includes IgM and IgG, and outlasts virus detection with generally higher IgG levels in symptomatic than asymptomatic groups in the early convalescent phase (12). Alarmingly, the IgG levels in both asymptomatic and symptomatic patients decline during the early convalescent phase, with a median decrease of ~75% within 2–3 months after infection (13). This suggests that the systemic antibody response which follows natural infection with SARS-CoV-2 is short-lived, with the possibility of no residual immunity after 6-12 months (14) affecting primarily neutralizing antibodies in plasma (15). Early

activated B cells produce antibodies in quasi-germline configuration and are likely 'innate-like B cells' (16–19) that have not undergone somatic hypermutation and maturation. Consistent with the above argument, a lack of germinal center formation but robust activation of non-germinal type B cells has been reported in cases of severe COVID-19 infection, impairing production of long-lived memory or high affinity B cells (20).

The generation of an antibody response requires cooperation between a B cell producing specific antibody molecules and a CD4 T cell (helper cell) activated by an epitope on the same antigen as that recognized by the B cell (T-B cooperation) (21). This reaction occurs in the germinal center (22, 23). Excluded from this rule are responses against carbohydrates and antigens with repeating motifs that alone cross-link the B cell antigen receptor leading to B cell activation (24). Discovered over 50 years ago (25–27), it also became apparent that T-B cooperation is restricted by Major Histocompatibility Complex class II (MHC-II) molecules (28–30). T-B cooperation plays a key role in the facilitation and strength of the antibody response (26, 31) and the size of the antibody response is proportional to the number of Th cells activated by the B cell during T-B cooperation (29, 30, 32). The importance of T cell help during the activation of antigen specific B cells to protein antigens driving B cell selection is emphasized by recent experiments where the injection of a conjugate of antigen (OVA) linked with an anti-DEC205 antibody induced a greater proliferation of DEC205+ relative to DEC205- B cells consistent with a T helper effect on B cell activation (33).

T-B cooperation requires that the epitopes recognized by the B and T cell be on the same portion of the antigen (27, 34, 35) leading to a model requiring the contextual internalization and co-processing of T and B cell epitopes (21) which is consistent with the principle of linked (e.g. associative) recognition of antigen (36). Studies in vitro using human T and B lymphocytes showed

that an antigen specific B cell can present antigen to CD4 T cells even if antigen is present at very low concentration (10^{-11} – 10^{-12} M) (37). Presentation of antigen by the B cell also facilitates the cooperation between CD4 T cells of different specificities resulting in enhanced generation of memory CD4 T cells (38). However, T-B cooperation is not the only form of cooperative interaction among lymphocytes as cooperation exists between CD4 T and CD8 T cells (39) and between two CD4 T cells responding to distinct epitopes on the same antigen (40).

A model based on coprocessing of T and B epitopes also led to the suggestion that preferential T-B pairing could be based on topological proximity (41–45) so that during BCR-mediated internalization the T cell epitope is protected by the paratope of the BCR. Indeed, a more recent study showed that not only is CD4 T cell help a limiting factor in the development of antibodies to smallpox (vaccinia virus), but that there also exists a deterministic epitope linkage of specificities in T-B cooperation against this viral pathogen (46). Collectively, it appears that T-B pairing and MHC-II restriction are key events in the selection of the antibody response to pathogens and that operationally T-B cooperation and MHC-II restriction are key events in the generation of an adaptive antibody response, suggesting that lack of or defective T-B preferential pairing could result in an antibody response that is suboptimal, short-lived, or both.

The relevance of T-B cooperation in protective antiviral responses has been documented in numerous systems. In the influenza A virus (PR8) system it was shown that while Th1 CD4 T cell responses on their own are ineffective at promoting recovery from infection, antibodies generated through T-B cooperation were indispensable in the protective response against the virus (47). In a different influenza A strain, it was shown that T-B cooperation and CD4 T cells represent a limiting factor in the kinetics and early magnitude of the primary B cell response to virus challenge and provide help in a preferential way (i.e. intra-molecular but nor inter-molecular) (48).

Additionally, CD40-CD40L (costimulatory molecules found on B cells and CD4 T cells, respectively) interaction is required for the generation of antibody responses, isotype switching and memory responses in non-viral model systems (49). In LCMV (lymphocytic choriomeningitis virus) and VSV (vesicular stomatitis virus) abrogation of CD40-CD40L interaction prevented T-B cooperation and thus inhibited antiviral protection (50). Interestingly, this study also showed that the activation of CD4 T cells (e.g., inflammatory CD4 T cells) not associated with the activation of B cells was not compromised (50). These data demonstrate the relevance of T-B cooperation in the antibody response in protection against viral infection.

In SARS-CoV-2, neutralizing antibodies (NAbs) are a key defense mechanism against infection and transmission. NAbs generated by single memory B cell VH/VL cloning from convalescent COVID-19 patients have been extremely useful in defining the fine epitope specificity of the antibody response in COVID-19 individuals. At present, SARS-CoV-2 NAbs can be distinguished into three large categories. 1) Repurposed antibodies, that is, NAbs discovered and characterized in the context of SARS-CoV and subsequently found to neutralize SARS-CoV-2 via cross-reactivity. These antibodies map away from the receptor binding domain (RBD) of the spike protein (51–53). 2) Non-RBD neutralizing antibodies discovered in SARS-CoV-2 patients whose paratope is specific for sites outside the RBD (54). 3) RBD antibodies, including NAbs, derived from SARS-CoV-2 patients that map to a restricted site in the RBD (18, 55–60). Cryo-EM of this third antibody category shows that they bind to residues in or around the four amino acids Phe-Asp-Cys-Tyr (FNCY) in the receptor binding motif (RBM) (residues 437-508) which is inside the larger RBD (residues 319-541) at the virus:ACE2 interface (56). Although the RBD has been shown to be an immunodominant target of serum antibodies in COVID-19 patients (61), high potency NAbs are directed against a conserved portion of the RBM on or around

the FNCY patch, a sequence only found in the RBD of SARS-CoV-2 and not in other coronaviruses. NAbs that make contact with the FNCY patch outperform other NAbs that do not in competition binding assays, highlighting the importance of the region in neutralizing ACE2 binding (54). Indeed while the RBD is mutationally tolerant, the RBM is constrained to the wild-type amino acids (62), implying that the B cell epitope included in this region of the virus:ACE2 interface is resistant to antigenic drift. Thus, we may refer to this site as a key RBM B cell epitope in the generation of potent NAbs.

Antibody responses against SARS-CoV-2 depend on CD4 T cell help. Spike-specific CD4 T cell responses have been found to correlate with the magnitude of the anti-RBD IgG response whereas non-spike CD4 T cell responses do not(63). However, in unexposed patients, spike-specific CD4 T cells reactive with MHC-II peptides proximal to the central B cell epitope represent a minority (~10%) of the total CD4 T cell responses, which are dominated by responses against either the distal portion of the spike protein or other structural antigens (64). Surprisingly, these CD4 T cell responses are largely cross-reactive and originate from previous coronavirus infections (65).

As mounting evidence suggests that the NAb response in COVID-19 patients is relatively short-lived, we decided to test the hypothesis that associative recognition of a key RBM B cell epitope (in and around the FNCY patch) and proximal MHC-II-restricted epitopes may be defective with detrimental effects on preferential T-B pairing. Specifically, we hypothesize that the inability to present SARS-CoV-2 peptide sequences near putative B cell epitopes may impair memory cell generation and consequently reduce the strength and longevity of overall and neutralizing antibody responses. To quantify the potential effects of T-B cooperation in vivo, we analyzed all 15mer putative MHC-II epitopes (+/- 50 amino acid residues) relative to the key RBM

B cell epitope for coverage by all known 5,620 human MHC-II alleles and predicted binding affinity. The analysis shows that there exists in general less availability of effective T cell epitopes in close proximity to the key RBM B cell epitope in the human population.

1.3.4 Results

Topology of a key RBM B cell epitope. Within the 222 amino acid long RBD of the spike protein (residues 319-541), the RBM (residues 437-508) is the portion of the spike protein that establishes contact with the ACE2 receptor (**Figure 1.3.1A**). The contact residues span a relatively large surface involving approximately 17 residues (56), among them residues F486, N487, Y489 form a loop, which we term the FNCY patch, which is surface exposed and protrudes up towards the ACE2 receptor from the bulge of the RBD (**Figure 1.3.1B-C**). F486 forms hydrophobic interactions with three ACE2 residues (L79, M82, W83). N487 forms hydrogen bonds with Q24 and W83, and Y489 is linked with K31 via a hydrophobic interaction. This makes the amino acid residues in or around the FNCY patch a logical B cell epitope target for antibodies blocking the virus:receptor interaction. In addition, these core residues are mutationally constrained by the ACE2 contact surface (62). Not surprisingly, a set of recently reported potently neutralizing antibodies generated by single B cell VH/VL cloning from convalescent COVID-19 patients all bear paratopes that include the FNCY patch in their recognition site (54, 58–60, 66) (**Figure 1.3.1D**). While other residues (Q493, N501, and Y505) are also shared between ACE2 and the paratope of these antibodies, they are not as protruding and are on a β-sheet unlike the FNCY patch which is organized in a short loop as a result of the C480:C488 disulfide bond. Thus, blockade of the RBM:ACE2 interaction (neutralization) depends at least in part on a B cell epitope in the RBM

that is structurally and functionally critical to the interaction, virus internalization, and cell infectivity.

Prediction of MHC-II affinity for 15mer peptides proximal to the RBM B cell epitope. In the T-B cooperation model, B cell activation and production of NAbs is dependent on CD4 T cell responses to MHC-II restricted peptides. To test the hypothesis that the generation of NAbs against a mutationally constrained B cell epitope in the RBM reflects the efficiency of processing and presentation of MHC-II peptides proximal to the FNCY patch, we evaluated the landscape of MHC-II peptide restriction across the entire SARS-CoV-2 spike protein with respect to common MHC-II alleles in the human population. To assess the potential for effective restriction by MHC-II molecules in a reasonable proportion of the population, we devised a position-based score that assigns each amino acid residue the median affinity of the best overlapping peptide, where median affinity is calculated across the 1911 most common MHC-II alleles (**Figure 1.3.2A**), which was highly correlated with scores across all 5620 MHC-II alleles (**Figure 1.3.2B**; Pearson rho=0.99, p<2.2e-308). While a number of sites along the spike protein are predicted to generate high affinity peptides for most common MHC-II alleles, the region around the FNCY patch was depleted for generally effective binders (**Figure 1.3.2C**, Fisher's exact OR=0.21, p=0.015, **Methods**). Interestingly, the RBM region containing the FNCY patch was free of glycans that could potentially mask the epitope (**Figure 1.3.2D**). We further evaluated the distributions of binding affinities for the 20 best-ranked peptides across all sites in the spike protein (**Figure 1.3.2E**), and in comparison, the distributions for the best 20 peptides overlapping positions within +/- 50 residues of the FNCY patch (**Figure 1.3.2F**). In the best case, less than half of the considered MHC-II alleles bound a shared peptide close to the FNCY patch, whereas at other sites there were multiple peptides that could be bound by nearly all of the MHC-II alleles (**Figure 1.3.2E**). This

suggested overall less availability of effective T cell epitopes in close proximity to the FNCY B cell epitope, which could limit the availability of T cell help during an epitope-specific T-B cooperative interaction in the germinal center.

To further assess whether population variation in MHC-II alleles might contribute to heterogeneity in potential to generate neutralizing antibodies, we also evaluated the potential of MHC-II supertypes to restrict peptides from neighboring the FNCY patch. Greenbaum et al. previously defined 7 supertypes that group MHC-II alleles based on shared binding repertoire. These 7 supertypes account for between 46%-77% of haplotypes and cover over 98% of individuals when all four loci are considered together (67). We revisited our analysis of peptide restriction proximal to the FNCY patch treating each supertype separately. There was considerable variability in potential to effectively present FNCY patch proximal sequences across supertypes (**Figure 1.3.3A-B**, X^2=175, p=3.75e-35). Only 3 supertypes (DP2, main DP and DR4) commonly presented peptides overlapping the FNCY patch (**Figure 1.3.3B**). We were able to obtain population allele frequencies for four populations from the Be The Match registry (68) and Du et al. (69). These data show that DR4 is relatively infrequent across the populations evaluated, whereas main DR, main DP, and DP2 are more common (**Figure 1.3.3C**), and thus could be more important for MHC-II restriction supportive of neutralizing antibodies. While there were some large population-specific differences in main DP and DP2 supertype frequencies, these frequency estimates are based on a limited population sample and may provide only a rough approximation. In general, DP and DR haplotypes were able to restrict more FNCY patch proximal sequences (**Figure 1.3.3D**).

Cross-reactivity to a non-coronavirus MHC-II binding peptide as a potential driver of T cell responses helping antibody response to the RBM B cell epitope. Interestingly, Mateus et al.

reported pre-existing CD4 T cell responses to peptides derived from the spike protein using T cells from unexposed individuals, suggesting previous exposures to other human coronaviruses could potentially generate protective immunity toward SARS-CoV-2. Indeed, regions of higher coronavirus homology were associated with more T cell responses in their data (65). This represents the most comprehensive interrogation of the spike protein with response to CD4 T cell responses to date. They screened all 15mers of the spike protein in pooled format and further evaluated 66 predicted MHC-II peptides that generated CD4 T cell responses. Visualizing the landscape of the CD4 T cell responses described in their work by percent positive response (**Figure 1.3.4A**) or spot forming cells (**Figure 1.3.4B**), we noted relatively few responses proximal to the FNCY patch in the RBM. Accordingly, few other coronaviruses had limited homology to the FNCY region, and none fully included the FNCY patch (**Figure 1.3.5A**).

A notable exception in Mateus' results is peptide 486FNCYFPLQSYGFQPT500, which was reported to induce a CD4 T cell response in an unexposed individual. In this case, the peptide was restricted by HLA-DRB1*0101 or HLA-DQA1*0101/DQB1*0501. We found that the peptide sequence had greater in silico predicted affinity to HLA-DRB1*0101. To explain the conundrum, we blasted this peptide against the "refseq_protein" database excluding SARS-CoV-2 (**Methods**). Surprisingly, the sequences with the best homology for this query were not from coronaviruses but rather from common pathogens, first among them parasites of the Cryptosporidium genus of apicomplexan parasitic alveolates. These sequences included conserved anchor positions for the HLA-DRB*0101 allele making it plausible that a prior exposure could account for the formation of a memory CD4 T cell response (**Figure 1.3.5B-C**). To further assess the potential for other prior exposures in generating immune memory for sequences proximal to the FNCY patch we blasted all 15mers within +/-30 amino acids of the FNCY patch and filtered

the resulting sequences based on restriction by consensus MHC-II supertypes (67). We found peptides associated with multiple microbial organisms that may meet the criteria to potentially generate CD4 T cell memory relevant to the RBM of SARS-CoV-2 (**Figure 1.3.5D**).

1.3.5 Discussion

SARS-CoV-2 uses the RBD of the spike protein to bind to the ACE2 receptor on target cells. The actual contact with ACE2 is mediated by a discrete number of amino acids that have been visualized by cryo-EM (56, 70). Although several SARS-related coronaviruses share 75% homology and interact with ACE2 on target cells (71, 72) the RBM in SARS-CoV-2 is unique to this virus. In vitro binding measurements show that SARS-CoV-2 RBD binds to ACE2 with an affinity in the low nanomolar range (73). Mutations in this motif could be detrimental to the virus's ability to infect ACE2 positive human cells. Since the RBD is an immunodominant site in the antibody response in humans (61) it is not surprising that the paratope of some antibodies isolated from convalescent individuals via single B cell VH/VL cloning, and selected on the basis of high neutralization potency, all seem to bind a surface encompassing the FNCY patch in the RBM (18, 19, 55, 57–60). Arguably, this motif corresponds to a relevant B cell epitope in the spike protein of SARS-CoV-2 and is a logical target of potent neutralizing antibodies.

Although antibodies directed to this site have been isolated by different groups, little is known about their contribution to the pool of antibodies in serum of SARS-CoV-2 infected individuals, but evidence suggests they are likely to be rare. In one study they were found to represent a subdominant fraction of the anti-RBD response (60) while the estimated frequency of antigen-specific B cells ranges from 0.07 to 0.005% of all the total B cells in COVID-19 convalescent individuals (74). In a second study, the identification of two ultra-potent NAbs

having a paratope involving the FNCY patch required screening of 800 clones from twelve individuals (19). This suggests that a potent NAb response to a mutationally constrained RBM epitope is a rare component of the total anti-virus response consistent, with the observation that there is no correlation between RBM site-specific neutralizing antibodies and serum half-maximal neutralization titer (NT50) (74). Here we show that the core RBM B cell epitope is apparently uncoupled from preferential T-B pairing, a prerequisite for a coordinated activation of B cells against the pathogen. We analyzed MHC-II binding of 15mer peptides in the spike protein upstream (-50 aa) or downstream (+50 aa) of the central RBM B cell epitope and found both low coverage by 1911 common MHC-II alleles and a depletion of binding 15mers proximal to the FNCY patch versus other exposed areas on the spike protein. This could be due to the fact that a sizeable proportion (40%) of CD4 T cells responding to the spike protein are memory responses found in SARS-CoV-2 unexposed individuals (63, 75) or other structural protein of SARS-CoV-2 such as the N protein (64). Thus, it is possible that these conserved responses are used as a decoy mechanism to polarize the response away from the RBM. However, this does not rule out the contribution of a bias in frequency of specific B cells in the available repertoire.

Corroboration to our hypothesis also comes from Mateus et al. (65) who tested sixty-six 15mer peptides of the spike protein in SARS-CoV-2 unexposed individuals and found that CD4 T cell responses against this narrow RBM site account for only 2/110 (1.8%) of the total CD4 T cell response to 15mer peptides of the spike protein. Surprisingly, a CD4 T cell response against peptide FNCYFPLQSYGFQPT was by CD4 T cells of an unexposed individual. Since this peptide has low homology with previous human coronaviruses, we reasoned that this could either represent a case of TCR cross-reactivity since a single TCR can engage large numbers of unique MHC/peptide combinations without requiring degeneracy in their recognition (76, 77).

Remarkably, however, a BLAST analysis revealed a 10 amino acid sequence match with proteins from pathogens including those from the Cryptosporidium genus, with identity in binding motif and anchor residues (agretope) for the restricting MHC-II allele strongly suggesting peptide cross-reactivity. Cryptosporidium hominis is a parasite that causes watery diarrhea that can last up to 3 weeks in immunocompetent patients (78). Additional possibilities for cross-reactivity to the RBM, albeit of a lesser stringency, involve antigens from Micromonospora, Pseudomonas, Blastococcus, Lactobacillus, and Bacteroides (**Figure 1.3.5D**). Thus, it appears as if memory CD4 T cells reactive with peptides in the RBM may reflect the immunological history of the individual that, as evidenced by this case, can be unrelated to infection by other coronaviruses. Interestingly, the great majority (64-88%) of COVID-19 positive individuals in homeless shelters in Los Angeles and Boston were found to be asymptomatic (79). This suggests that the status of the immune system, which itself reflects past antigenic exposure, may be a determining factor in the generation of a protective immune response after SARS-CoV-2 infection.

The findings reported herein have considerable implications for natural immunity to SARS-CoV-2. The fact that there seems to be an overall suboptimal T-B preferential pairing suggests that B cells that respond to the RBM B cell epitope may receive inadequate T cell help. This is consistent with the observation that in general potent neutralizing antibodies to the RBM undergo very limited somatic mutation (19, 57) and are by and large in quasi-germline configuration (80). Since T cell help is also necessary to initiate somatic hypermutation in B cell through CD40 or CD38 signaling in the germinal center (81), it follows that one important implication of our study is that defective T-B pairing may negatively influence the normal process of germinal center maturation of the B cell response in response to SARS-CoV-2 infection in a critical way.

Which antigens can generate T cell responses depends on the binding specificities of MHC-II molecules, which are highly polymorphic in the human population. We noted a general trend for MHC-II alleles to less effectively present peptides from the RBM region, but also observed some variability across MHC-II supertypes. The main DP and DP2 haplotypes were both common and had the highest potential to present peptides, suggesting that most individuals should carry at least one allele capable of presenting peptides in this region. Which of the two DP haplotypes was more common varied by ancestral population, thus it is possible that differences in the haplotypes could translate to differences in T-B cooperativity levels within groups, though binding affinities for epitopes near the FNCY patch were similar for both. DQ and DR supertypes were less able to present peptides near FNCY, with the exception of DR4, which is among the less common supertypes. Importantly, our analysis was limited to predicted affinity of peptides to MHC-II, and other characteristics such as expression levels, stability or differences in interactions with molecular chaperones likely also contribute to whether FNCY proximal peptides are available to support T-B cooperation (82).

The present study assesses the probability of SARS-CoV-2 peptides of the Spike protein to bind and be presented by MHC-II molecules. Our study is limited by the following: results are an estimate based on an algorithm that encompasses many biophysical variables for MHC-II presentation but certainly not all. In addition, while we believe the epitope containing the FNCY patch is promising for inducing a protective neutralizing response, it is not the sole determinant of a protective antibody response to SARS-CoV-2; as neutralizing antibodies against other portions of the spike and other non-structural proteins have been reported (52, 53, 83–86).

In light of our findings, it can be predicted that, in general, a specific RBM antibody response may be short-lived and that residual immunity from a primary infection may not be

sufficient to prevent reinfection after 6-9 months. Sporadic cases of re-infection have been reported by the media in Hong Kong and Nevada (87). A third case has been reported in a care-home resident who after the second infection produced only low levels of antibodies (88). Finally, silent re-infections in young workers in a COVID-19 ward who tested positive for the new coronavirus and became reinfected several months later with no symptoms in either instance have been reported (89). It is tempting to speculate that waning antibody levels or a poorly developed specific NAb antibody response to SARS-CoV-2 can potentially put people at risk of reinfection. Other factors to consider are a bias in the available B cell repertoire in the population and the extent to which a defective T-B cooperation influences the longevity of terminally differentiated plasma cells in the bone marrow (90).

In summary, we provide evidence that MHC-II constrains the CD4 T cell response for epitopes that are best positioned to facilitate T-B pairing in generating and sustaining a potent neutralizing antibody response against a mutationally constrained RBM B cell epitope. Furthermore, we show that the immunological history of the individual, not necessarily related to infection by other coronaviruses, may confer immunologic advantage. Finally, these findings may have implications for the quality and persistence of a protective, neutralizing antibody response to RBM induced by current SARS-CoV-2 vaccines.

1.3.6. Materials and Methods

Affinity analysis. NetMHCIIpan version 4.0 was used to predict peptide-MHC-II affinity (*91*) for generated 15mers along the SARS-CoV-2 spike protein.

Spike protein analyses. SARS-CoV-2 spike protein sequence and protein regions were obtained from https://www. uniprot.org/uniprot/P0DTC2. Glycan data were obtained from (*92*) and true-positive sites were aggregated across 3 replicates. To assess depletion of effective binders near the FNCY patch, we performed a Fisher's exact test for binding (median affinity across common alleles <10) versus proximity (+/- 50 amino acids) to FNCY for positions free of glycans. We excluded positions within 10 amino acids of a glycan using the data obtained from Watanabe et al. and added a pseudocount of 1.

The SARS1, MERS1, HCoV-229E, HCoV-NL63, HCoV-OC43, and HCoV-HKU1 spike protein sequences were also downloaded from UniProt (P59594, K9N5Q8, P15423, Q6Q1S2, P36334, Q0ZME7, respectively). Multiple sequence alignment was performed on the EMBL-EBI Clustal Omega web server using default parameters (*93*).

Structure analysis. The 6M0J 3D X-ray structure for the protein complex containing the SARS-CoV-2 spike protein RBD (P0DTC2) interaction with ACE2 (Q9BYF1) from (*56*). The structure figures were prepared using VMD (*94*).

Supertype analysis. Supertypes were obtained from (*67*). All alpha/beta combinations spanning any of these types were included, resulting in 279 alleles. US supertype frequencies for alleles in DRB1 and DQB1 were obtained from the Be the Match registry (*68*), US frequencies for alleles in DPB1 were obtained from (*69*) as DPB1 was not available from the Be the Match registry. Available allele frequencies within each supertype were summed for Fig 3C.

Motif analysis. All 13-20mer peptides adhering to the following parameters were downloaded from the IEDB (*95*): MHC-II assay, positive only, DRB1*01:01 allele, linear peptides; and any peptides with post-translational modifications or noncanonical amino acids were removed. The remaining 10,117 peptides were input into Gibbs cluster v2.0 using the default MHC-II ligand parameters.

BLAST analysis. 15mers were generated along a sliding window +/-30 amino acids from the FNCY patch start and end (455–518, 0-index) and input into NCBI BLAST (*96*) using the 'refseq_protein' database and excluding SARS-CoV-2 (taxid:2697049). Identified peptides were then evaluated for binding affinity and any peptide binding to at least one allele was retained.

1.3.7 Figures

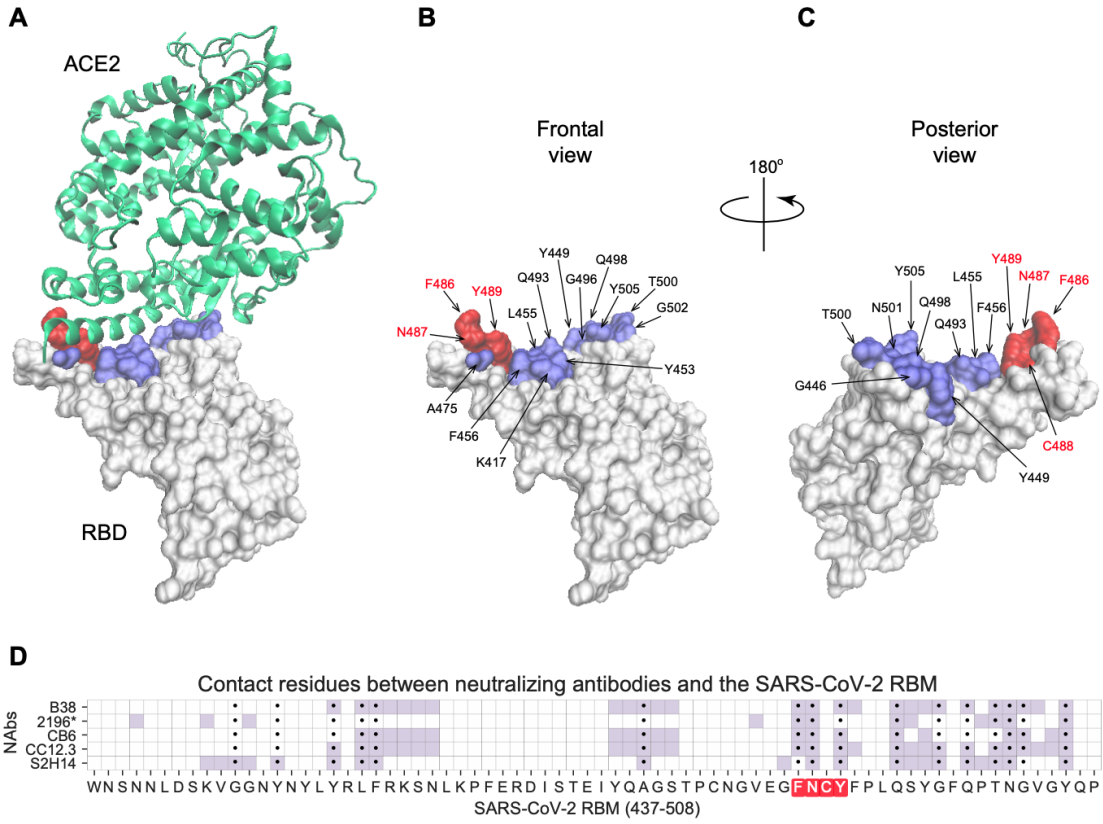


Figure 1.3.1. Visualization of the FNCY core of the RBM B cell epitope on the SARS-CoV-2 spike protein RBD. (A) 3D structure of the SARS-CoV-2 spike protein RBD (white) binding the ACE2 receptor (green) (PDB: 6M0J) with contact residues highlighted in blue and the FNCY patch highlighted in red. (B-C) Spike protein RBD with ACE2 contact residues and FNCY patch residues labeled in two orientations (front and back). (D) Heatmap of neutralizing antibody contact residues (purple) on the spike protein RBM region (positions 437-508). Black dots indicate ACE2 contact residues and the FNCY patch is highlighted in red.

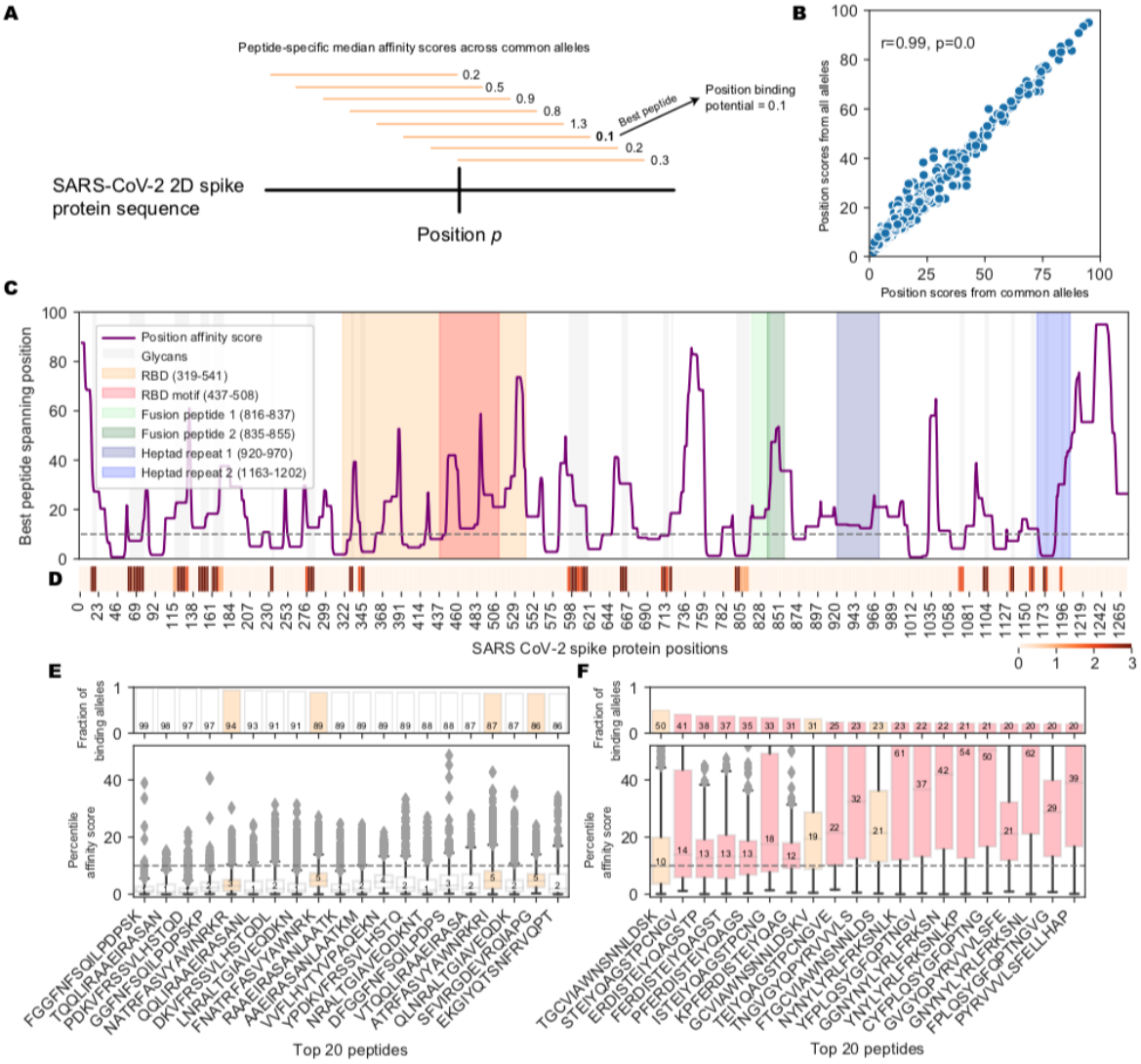


Figure 1.3.2. Landscape of MHC-II binding affinity across spike protein 2D sequence. (A) Overview of the position affinity score. (B) Scatterplot showing position affinity scores estimated using only common (>10% frequency) MHC-II alleles (x-axis) versus across all MHC-II alleles (y-axis). (C) Lineplot showing the position affinity scores across common MHC-II alleles (**Methods**). Annotated domains from UniProt are highlighted. (D) Heatmap showing amino acid positions that are glycosylated(57). (E) Barplots (top) and boxplots (bottom) describing the fraction of binding MHC-II alleles and corresponding affinity percentile rank distributions respectively for the top 20 peptides with the highest fraction of common binding alleles. The binding threshold of 10 is shown as a dotted line, with values less than 10 indicating binding. Colors correspond to the regions listed in C. (F) Barplots (top) and boxplots (bottom) describing the fraction of binding MHC-II alleles and corresponding affinity percentile rank distributions respectively for the top 20 peptides within +/-50 amino acids of the FNCY B cell epitope. Colors correspond to the regions listed in C.

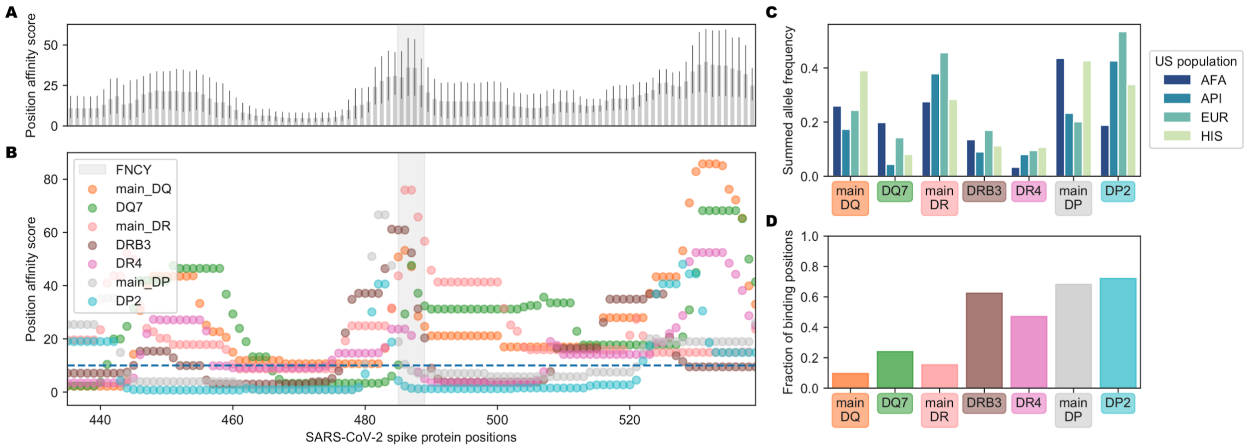


Figure 1.3.3. Population variation affecting availability of FNCY proximal T cell epitopes. (A) Barplot showing the aggregated supertype position affinity scores for each position +/- 50 amino acids from the FNCY patch (grey zone). (B) Scatterplot showing the specific supertype position scores for each position +/- 50 amino acids from the FNCY patch (grey zone). The binding threshold of 10 is shown as a dashed blue line, with points below the threshold indicating binding. (C) Barplot showing United States population frequencies, summed across the available alleles in each supertype. (D) Fraction of positions falling below the binding threshold within the region of interest for each supertype.

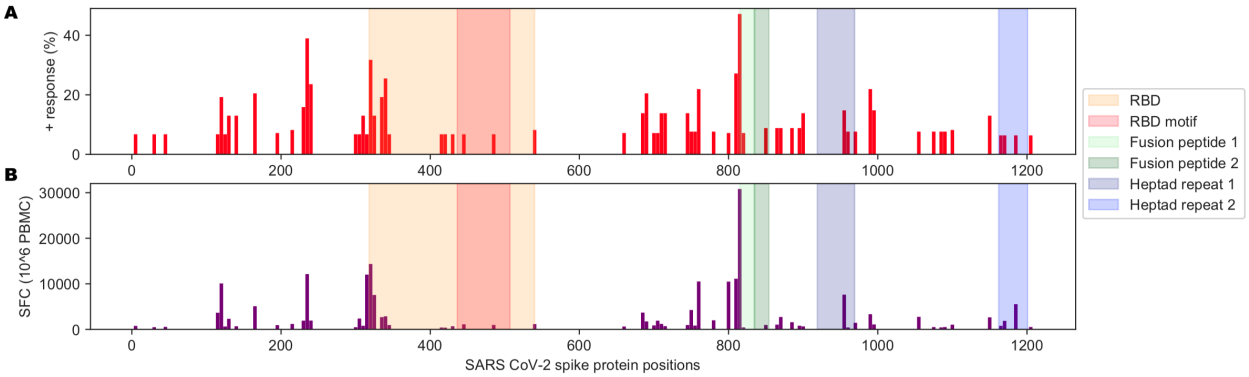
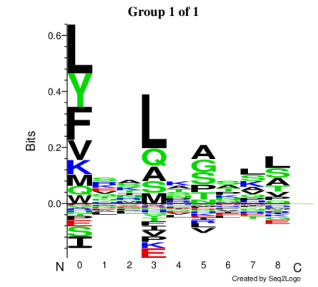


Figure 1.3.4. Immunological history of relevance to SARS-CoV-2. (A) Barplot showing the percentage of positive responses toward SARS-CoV-2 peptides from unexposed individuals. (B) Barplot showing the number of spot-forming cells (SFC) for tested SARS-CoV-2 peptides against PBMCs from unexposed individuals. Data from Table S1 from Mateus et al.

A

```
P15423|SPIKE_HCoV-229E      GKVN--------------------NFVKFGSVCFSLKD--------IPGGCAM------  398
Q6Q1S2|SPIKE_HCoV-NL63      SKLN--------------------NFQKFKTICFSTVE--------VPGSCNF------  579
P36334|SPIKE_HCoV-OC43      STWNKRFGFIEDSVFKPRPAGVLTNHDVVYAQHCFKAPKNFCPCK--LNGSCVGSGPGKN  507
Q0ZME7|SPIKE_HCoV-HKU1      SSWNRRYGFGSF---------NLSSYDVVYSDHCFSVNSDFCPCA--DPSV-VNSCAKSK  489
K9N5Q8|SPIKE_MERS1          SYINKCSRFLSD-------------DRTEVPQLVNAN--QYSPCVSIVPST-VW------  535
P0DTC2|SPIKE_SARS2          LYRLFRKSNLKP-------------FERDISTEIYQA--GSTPCNGVEGFNCYF------  490
P59594|SPIKE_SARS1          KYRYLRHGKLRP-------------FERDISNVPFSP--DGKPCTP-PALNCYW------  476
```

B



C

BLAST alignment	Organism
Query YFPLQSYGFQ YFPLQSYGFQ Sbjct YFPLQSYGFQ	Cryptosporidium parvum Iowa II
Query YFPLQSYGFQ YFPLQSYGFQ Sbjct YFPLQSYGFQ	Cryptosporidium hominis TU502
Query YFPLQSYGFQ YFP QSYGFQ Sbjct YFPFQSYGFQ	Cryptosporidium ubiquitum
Query LQSYGFQPT LQSYGFQPT Sbjct LQSYGFQPT	Vanderwaltozyma polyspora DSM 70294

D

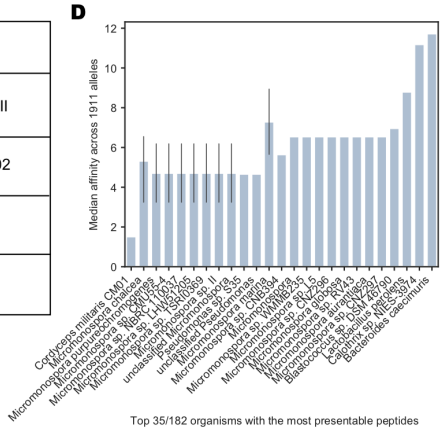


Figure 1.3.5. Learned immunity to other targets that could support T cell responses to SARS-CoV-2. (A) Multiple sequence alignment between SARS-CoV-2, SARS1, MERS, and other human coronaviruses, focusing on the region surrounding the FNCY B cell epitope. (B) SeqLogo plot obtained by clustering IEDB peptides reported to bind to DRB1*01:01. (C) Top results after blasting the FNCYFPLQSYGFQPT peptide against all reference proteins. (D) Barplot describing best peptide affinities across MHC-II alleles of the top 35 unique organisms with one or more peptides matching a peptide with high similarity to 15mers +/-30aa from the FNCY binding epitope based on BLAST analysis. The closer to 0, the greater the binding potential.

1.3.8 Author Contributions

Original concept: Maurizio Zanetti and Hannah Carter

Data curation and formal analysis: Andrea Castro and Kivilcim Ozturk

Funding acquisition: Maurizio Zanetti and Hannah Carter

Manuscript writing: Andrea Castro, Kivilcim Ozturk, Maurizio Zanetti and Hannah Carter

1.3.9 Acknowledgements

This work was supported by an NIH National Library of Medicine Training Grant T15LM011271 to A.C., an Emerging Leader Award #18-022-ELA from The Mark Foundation for Cancer Research (https://themarkfoundation.org), a Canadian Institute For Advanced Research (CIFAR) (https://www.cifar.ca) fellowship #FL-000655 to H.C. and NIH NCI RO1 CA220009 to M. Z. and H.C.

Chapter 1.3, in full, is a reformatted reprint of the material as it appears in "In silico analysis suggests less effective MHC-II presentation of SARS-CoV-2 RBM peptides: Implication for neutralizing antibody responses" in *Plos one,* 2021 by Andrea Castro, Kivilcim Ozturk, Maurizio Zanetti, and Hannah Carter. The dissertation author was a primary investigator and author of this paper.

1.3.10 References

1. G. T. Nepom, H. Erlich, MHC Class-II Molecules and Autoimmunity. *Annu. Rev. Immunol.* **9**, 493–525 (1991).

2. V. Matzaraki, V. Kumar, C. Wijmenga, A. Zhernakova, The MHC locus and genetic susceptibility to autoimmune and infectious diseases. *Genome Biol.* **18**, 76 (2017).

3. S. Tsai, P. Santamaria, MHC Class II Polymorphisms, Autoreactive T-Cells, and Autoimmunity. *Front. Immunol.* **4**, 321 (2013).

4. J. W. Kappler, N. Roehm, P. Marrack, T cell tolerance by clonal elimination in the thymus. *Cell*. **49**, 273–280 (1987).

5. L. Klein, B. Kyewski, P. M. Allen, K. A. Hogquist, Positive and negative selection of the T cell repertoire: what thymocytes see (and don't see). *Nat. Rev. Immunol.* **14**, 377–391 (2014).

6. E. W. Hewitt, The MHC class I antigen presentation pathway: strategies for viral immune evasion. *Immunology*. **110**, 163–169 (2003).

7. P. A. Roche, K. Furuta, The ins and outs of MHC class II-mediated antigen processing and presentation. *Nat. Rev. Immunol.* **15**, 203–216 (2015).

8. R. V. Luckheeram, R. Zhou, A. D. Verma, B. Xia, $CD4^+$ T cells: differentiation and functions. *Clin. Dev. Immunol.* **2012**, 925135 (2012).

9. S. L. Swain, K. K. McKinstry, T. M. Strutt, Expanding roles for $CD4^+$ T cells in immunity to viruses. *Nat. Rev. Immunol.* **12**, 136–148 (2012).

10. N. P. Croft, S. A. Smith, J. Pickering, J. Sidney, B. Peters, P. Faridi, M. J. Witney, P. Sebastian, I. E. A. Flesch, S. L. Heading, A. Sette, N. L. La Gruta, A. W. Purcell, D. C. Tscharke, Most viral peptides displayed by class I MHC on infected cells are immunogenic. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 3112–3117 (2019).

11. M. Koutsakos, H. E. G. McWilliam, T. E. Aktepe, S. Fritzlar, P. T. Illing, N. A. Mifsud, A. W. Purcell, S. Rockman, P. C. Reading, J. P. Vivian, J. Rossjohn, A. G. Brooks, J. M. Mackenzie, J. D. Mintern, J. A. Villadangos, T. H. O. Nguyen, K. Kedzierska, Downregulation of MHC Class I Expression by Influenza A and B Viruses. *Front. Immunol.* **10**, 1158 (2019).

12. R. Wölfel, V. M. Corman, W. Guggemos, M. Seilmaier, S. Zange, M. A. Müller, D. Niemeyer, T. C. Jones, P. Vollmar, C. Rothe, M. Hoelscher, T. Bleicker, S. Brünink, J. Schneider, R. Ehmann, K. Zwirglmaier, C. Drosten, C. Wendtner, Virological assessment of hospitalized patients with COVID-2019. *Nature*. **581**, 465–469 (2020).

13. Q.-X. Long, X.-J. Tang, Q.-L. Shi, Q. Li, H.-J. Deng, J. Yuan, J.-L. Hu, W. Xu, Y. Zhang, F.-J. Lv, K. Su, F. Zhang, J. Gong, B. Wu, X.-M. Liu, J.-J. Li, J.-F. Qiu, J. Chen, A.-L. Huang,

Clinical and immunological assessment of asymptomatic SARS-CoV-2 infections. *Nat. Med.* **26**, 1200–1204 (2020).

14. C. Rydyznski Moderbacher, S. I. Ramirez, J. M. Dan, A. Grifoni, K. M. Hastie, D. Weiskopf, S. Belanger, R. K. Abbott, C. Kim, J. Choi, Y. Kato, E. G. Crotty, C. Kim, S. A. Rawlings, J. Mateus, L. P. V. Tse, A. Frazier, R. Baric, B. Peters, J. Greenbaum, E. Ollmann Saphire, D. M. Smith, A. Sette, S. Crotty, Antigen-Specific Adaptive Immunity to SARS-CoV-2 in Acute COVID-19 and Associations with Age and Disease Severity. *Cell* (2020), doi:10.1016/j.cell.2020.09.038.

15. J. Prévost, R. Gasser, G. Beaudoin-Bussières, J. Richard, R. Duerr, A. Laumaea, S. P. Anand, G. Goyette, M. Benlarbi, S. Ding, H. Medjahed, A. Lewin, J. Perreault, T. Tremblay, G. Gendron-Lepage, N. Gauthier, M. Carrier, D. Marcoux, A. Piché, M. Lavoie, A. Benoit, V. Loungnarath, G. Brochu, E. Haddad, H. D. Stacey, M. S. Miller, M. Desforges, P. J. Talbot, G. T. Gould Maule, M. Côté, C. Therrien, B. Serhir, R. Bazin, M. Roger, A. Finzi, Cross-sectional evaluation of humoral responses against SARS-CoV-2 Spike. *Cell Rep Med*, 100126 (2020).

16. W. Wen, W. Su, H. Tang, W. Le, X. Zhang, Y. Zheng, X. Liu, L. Xie, J. Li, J. Ye, L. Dong, X. Cui, Y. Miao, D. Wang, J. Dong, C. Xiao, W. Chen, H. Wang, Immune cell profiling of COVID-19 patients in the recovery stage by single-cell sequencing. *Cell Discov*. **6**, 31 (2020).

17. B. Ju, Q. Zhang, J. Ge, R. Wang, J. Sun, X. Ge, J. Yu, S. Shan, B. Zhou, S. Song, X. Tang, J. Yu, J. Lan, J. Yuan, H. Wang, J. Zhao, S. Zhang, Y. Wang, X. Shi, L. Liu, J. Zhao, X. Wang, Z. Zhang, L. Zhang, Human neutralizing antibodies elicited by SARS-CoV-2 infection. *Nature*. **584**, 115–119 (2020).

18. L. Liu, P. Wang, M. S. Nair, J. Yu, M. Rapp, Q. Wang, Y. Luo, J. F.-W. Chan, V. Sahi, A. Figueroa, X. V. Guo, G. Cerutti, J. Bimela, J. Gorman, T. Zhou, Z. Chen, K.-Y. Yuen, P. D. Kwong, J. G. Sodroski, M. T. Yin, Z. Sheng, Y. Huang, L. Shapiro, D. D. Ho, Potent neutralizing antibodies against multiple epitopes on SARS-CoV-2 spike. *Nature*. **584**, 450–456 (2020).

19. M. A. Tortorici, M. Beltramello, F. A. Lempp, D. Pinto, H. V. Dang, L. E. Rosen, M. McCallum, J. Bowen, A. Minola, S. Jaconi, F. Zatta, A. De Marco, B. Guarino, S. Bianchi, E. J. Lauron, H. Tucker, J. Zhou, A. Peter, C. Havenar-Daughton, J. A. Wojcechowskyj, J. B. Case, R. E. Chen, H. Kaiser, M. Montiel-Ruiz, M. Meury, N. Czudnochowski, R. Spreafico, J. Dillen, C. Ng, N. Sprugasci, K. Culap, F. Benigni, R. Abdelnabi, S.-Y. C. Foo, M. A. Schmid, E. Cameroni, A. Riva, A. Gabrieli, M. Galli, M. S. Pizzuto, J. Neyts, M. S. Diamond, H. W. Virgin, G. Snell, D. Corti, K. Fink, D. Veesler, Ultrapotent human antibodies protect against SARS-CoV-2 challenge via multiple mechanisms. *Science* (2020), doi:10.1126/science.abe3354.

20. N. Kaneko, H.-H. Kuo, J. Boucau, J. R. Farmer, H. Allard-Chamard, V. S. Mahajan, A. Piechocka-Trocha, K. Lefteri, M. Osborn, J. Bals, Y. C. Bartsch, N. Bonheur, T. M. Caradonna, J. Chevalier, F. Chowdhury, T. J. Diefenbach, K. Einkauf, J. Fallon, J. Feldman, K. K. Finn, P. Garcia-Broncano, C. A. Hartana, B. M. Hauser, C. Jiang, P. Kaplonek, M.

Karpell, E. C. Koscher, X. Lian, H. Liu, J. Liu, N. L. Ly, A. R. Michell, Y. Rassadkina, K. Seiger, L. Sessa, S. Shin, N. Singh, W. Sun, X. Sun, H. J. Ticheli, M. T. Waring, A. L. Zhu, G. Alter, J. Z. Li, D. Lingwood, A. G. Schmidt, M. Lichterfeld, B. D. Walker, X. G. Yu, R. F. Padera Jr, S. Pillai, Massachusetts Consortium on Pathogen Readiness Specimen Working Group, Loss of Bcl-6-Expressing T Follicular Helper Cells and Germinal Centers in COVID-19. *Cell*. **183**, 143–157.e13 (2020).

21. N. A. Mitchison, T-cell-B-cell cooperation. *Nat. Rev. Immunol.* **4**, 308–312 (2004).

22. J. Jacob, G. Kelsoe, K. Rajewsky, U. Weiss, Intraclonal generation of antibody mutants in germinal centres. *Nature*. **354**, 389–392 (1991).

23. C. Berek, A. Berger, M. Apel, Maturation of the immune response in germinal centers. *Cell*. **67**, 1121–1129 (1991).

24. M. Zanetti, D. Glotz, Considerations on thymus-dependent and -independent antigens in acquired and natural immunity. *Ann. Inst. Pasteur Immunol.* **139**, 192–193 (1988).

25. H. N. Claman, E. A. Chaperon, R. F. Triplett, Thymus-marrow cell combinations. Synergism in antibody production. *Proc. Soc. Exp. Biol. Med.* **122**, 1167–1171 (1966).

26. N. A. Mitchison, The carrier effect in the secondary response to hapten-protein conjugates. I. Measurement of the effect with transferred cells and objections to the local environment hypothesis. *Eur. J. Immunol.* **1**, 10–17 (1971).

27. K. Rajewsky, E. Rottländer, G. Peltre, B. Müller, The immune response to a hybrid protein molecule; specificity of secondary stimulation and of tolerance induction. *J. Exp. Med.* **126**, 581–606 (1967).

28. D. H. Katz, T. Hamaoka, M. E. Dorf, B. Benacerraf, Cell interactions between histoincompatible T and B lymphocytes. The H-2 gene complex determines successful physiologic lymphocyte interactions. *Proc. Natl. Acad. Sci. U. S. A.* **70**, 2624–2628 (1973).

29. J. Sprent, Restricted helper function of F1 hybrid T cells positively selected to heterologous erythrocytes in irradiated parental strain mice. II. Evidence for restrictions affecting helper cell induction and T-B collaboration, both mapping to the K-end of the H-2 complex. *J. Exp. Med.* **147**, 1159–1174 (1978).

30. B. Jones, C. A. Janeway Jr, Cooperative interaction of B lymphocytes with antigen-specific helper T lymphocytes is MHC restricted. *Nature*. **292**, 547–549 (1981).

31. N. A. Mitchison, The carrier effect in the secondary response to hapten-protein conjugates. II. Cellular cooperation. *Eur. J. Immunol.* **1**, 18–27 (1971).

32. C. A. Janeway Jr, Cellular cooperation during in vivo anti-hapten antibody responses. I. The effect of cell number on the response. *J. Immunol.* **114**, 1394–1401 (1975).

33. Z. Shulman, A. D. Gitlin, S. Targ, M. Jankovic, G. Pasqual, M. C. Nussenzweig, G. D. Victora, T follicular helper cell dynamics in germinal centers. *Science*. **341**, 673–677 (2013).

34. F. Celada, E. E. Sercarz, Preferential pairing of T-B specificities in the same antigen: the concept of directional help. *Vaccine*. **6**, 94–98 (1988).

35. F. Manca, A. Kunkl, D. Fenoglio, A. Fowler, E. Sercarz, F. Celada, Constraints in T-B cooperation related to epitope topology on E. coli β-galactosidase. I. The fine specificity of T cells dictates the fine specificity of antibodies directed to conformation-dependent determinants. *Eur. J. Immunol.* **15**, 345–350 (1985).

36. P. Bretscher, M. Cohn, A theory of self-nonself discrimination. *Science*. **169**, 1042–1049 (1970).

37. A. Lanzavecchia, Antigen-specific interaction between T and B cells. *Nature*. **314** (1985), pp. 537–539.

38. D. R. Kroeger, C. D. Rudulier, P. A. Bretscher, Antigen presenting B cells facilitate CD4 T cell cooperation resulting in enhanced generation of effector and memory CD4 T cells. *PLoS One*. **8**, e77346 (2013).

39. D. Cassell, J. Forman, Linked recognition of helper and cytotoxic antigenic determinants for the generation of cytotoxic T lymphocytes. *Ann. N. Y. Acad. Sci.* **532**, 51–60 (1988).

40. M. Gerloni, S. Xiong, S. Mukerjee, S. P. Schoenberger, M. Croft, M. Zanetti, Functional cooperation between T helper cell determinants. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 13269–13274 (2000).

41. J. A. Berzofsky, L. K. Richman, D. J. Killion, Distinct H-2-linked Ir genes control both antibody and T cell responses to different determinants on the same antigen, myoglobin. *Proc. Natl. Acad. Sci. U. S. A.* **76**, 4046–4050 (1979).

42. J. A. Berzofsky, A. N. Schechter, G. M. Shearer, D. H. Sachs, Genetic control of the immune response to staphylococcal nuclease. III. Time-course and correlation between the response to native nuclease and the response to its polypeptide fragments. *J. Exp. Med.* **145**, 111–122 (1977).

43. J. A. Berzofsky, A. N. Schechter, G. M. Shearer, D. H. Sachs, Genetic control of the immune response to staphylococcal nuclease. IV. H-2-linked control of the relative proportions of antibodies produced to different determinants of native nuclease. *J. Exp. Med.* **145**, 123–135 (1977).

44. M. Zanetti, E. Sercarz, J. Salk, The immunology of new generation vaccines. *Immunol. Today*. **8**, 18–25 (1987).

45. F. Celada, A. Kunkl, F. Manca, D. Fenoglio, A. Fowler, U. Krzych, E. Sercarz, Preferential pairings in T-B encounters utilizing Th cells directed against discrete portions of b-

galactosidase and B cells primed with the native enzyme or a hapten epitope. *Regulation of the Immune System*, 637–646 (1984).

46. A. Sette, M. Moutaftsi, J. Moyron-Quiroz, M. M. McCausland, D. H. Davies, R. J. Johnston, B. Peters, M. Rafii-El-Idrissi Benhnia, J. Hoffmann, H.-P. Su, K. Singh, D. N. Garboczi, S. Head, H. Grey, P. L. Felgner, S. Crotty, Selective CD4+ T cell help for antibody responses to a large viral pathogen: deterministic linkage of specificities. *Immunity*. **28**, 847–858 (2008).

47. K. Mozdzanowska, M. Furchner, K. Maiese, W. Gerhard, CD4+ T cells are ineffective in clearing a pulmonary infection with influenza type A virus in the absence of B cells. *Virology*. **239**, 217–225 (1997).

48. S. Alam, Z. A. G. Knowlden, M. Y. Sangster, A. J. Sant, CD4 T cell help is limiting and selective during the primary B cell response to influenza virus infection. *J. Virol.* **88**, 314–324 (2014).

49. D. C. Parker, The functions of antigen recognition in T cell-dependent B cell activation. *Semin. Immunol.* **5**, 413–420 (1993).

50. A. Oxenius, K. A. Campbell, C. R. Maliszewski, T. Kishimoto, H. Kikutani, H. Hengartner, R. M. Zinkernagel, M. F. Bachmann, CD40-CD40 ligand interactions are critical in T-B cooperation but not for other anti-viral CD4+ T cell functions. *J. Exp. Med.* **183**, 2209–2218 (1996).

51. Z. Lv, Y.-Q. Deng, Q. Ye, L. Cao, C.-Y. Sun, C. Fan, W. Huang, S. Sun, Y. Sun, L. Zhu, Q. Chen, N. Wang, J. Nie, Z. Cui, D. Zhu, N. Shaw, X.-F. Li, Q. Li, L. Xie, Y. Wang, Z. Rao, C.-F. Qin, X. Wang, Structural basis for neutralization of SARS-CoV-2 and SARS-CoV by a potent therapeutic antibody. *Science*. **369**, 1505–1509 (2020).

52. D. Pinto, Y.-J. Park, M. Beltramello, A. C. Walls, M. A. Tortorici, S. Bianchi, S. Jaconi, K. Culap, F. Zatta, A. De Marco, A. Peter, B. Guarino, R. Spreafico, E. Cameroni, J. B. Case, R. E. Chen, C. Havenar-Daughton, G. Snell, A. Telenti, H. W. Virgin, A. Lanzavecchia, M. S. Diamond, K. Fink, D. Veesler, D. Corti, Cross-neutralization of SARS-CoV-2 by a human monoclonal SARS-CoV antibody. *Nature*. **583**, 290–295 (2020).

53. M. Yuan, N. C. Wu, X. Zhu, C.-C. D. Lee, R. T. Y. So, H. Lv, C. K. P. Mok, I. A. Wilson, A highly conserved cryptic epitope in the receptor binding domains of SARS-CoV-2 and SARS-CoV. *Science*. **368**, 630–633 (2020).

54. L. Piccoli, Y.-J. Park, M. A. Tortorici, N. Czudnochowski, A. C. Walls, M. Beltramello, C. Silacci-Fregni, D. Pinto, L. E. Rosen, J. E. Bowen, O. J. Acton, S. Jaconi, B. Guarino, A. Minola, F. Zatta, N. Sprugasci, J. Bassi, A. Peter, A. De Marco, J. C. Nix, F. Mele, S. Jovic, B. F. Rodriguez, S. V. Gupta, F. Jin, G. Piumatti, G. Lo Presti, A. F. Pellanda, M. Biggiogero, M. Tarkowski, M. S. Pizzuto, E. Cameroni, C. Havenar-Daughton, M. Smithey, D. Hong, V. Lepori, E. Albanese, A. Ceschi, E. Bernasconi, L. Elzi, P. Ferrari, C. Garzoni, A. Riva, G. Snell, F. Sallusto, K. Fink, H. W. Virgin, A. Lanzavecchia, D. Corti, D. Veesler, Mapping Neutralizing and Immunodominant Sites on the SARS-CoV-2 Spike Receptor-Binding

Domain by Structure-Guided High-Resolution Serology. *Cell* (2020), doi:10.1016/j.cell.2020.09.037.

55. C. O. Barnes, A. P. West Jr, K. E. Huey-Tubman, M. A. G. Hoffmann, N. G. Sharaf, P. R. Hoffman, N. Koranda, H. B. Gristick, C. Gaebler, F. Muecksch, J. C. C. Lorenzi, S. Finkin, T. Hägglöf, A. Hurley, K. G. Millard, Y. Weisblum, F. Schmidt, T. Hatziioannou, P. D. Bieniasz, M. Caskey, D. F. Robbiani, M. C. Nussenzweig, P. J. Bjorkman, Structures of Human Antibodies Bound to SARS-CoV-2 Spike Reveal Common Epitopes and Recurrent Features of Antibodies. *Cell*. **182**, 828–842.e16 (2020).

56. J. Lan, J. Ge, J. Yu, S. Shan, H. Zhou, S. Fan, Q. Zhang, X. Shi, Q. Wang, L. Zhang, X. Wang, Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature*. **581**, 215–220 (2020).

57. T. F. Rogers, F. Zhao, D. Huang, N. Beutler, A. Burns, W.-T. He, O. Limbo, C. Smith, G. Song, J. Woehl, L. Yang, R. K. Abbott, S. Callaghan, E. Garcia, J. Hurtado, M. Parren, L. Peng, S. Ramirez, J. Ricketts, M. J. Ricciardi, S. A. Rawlings, N. C. Wu, M. Yuan, D. M. Smith, D. Nemazee, J. R. Teijaro, J. E. Voss, I. A. Wilson, R. Andrabi, B. Briney, E. Landais, D. Sok, J. G. Jardine, D. R. Burton, Isolation of potent SARS-CoV-2 neutralizing antibodies and protection from disease in a small animal model. *Science*. **369**, 956–963 (2020).

58. R. Shi, C. Shan, X. Duan, Z. Chen, P. Liu, J. Song, T. Song, X. Bi, C. Han, L. Wu, G. Gao, X. Hu, Y. Zhang, Z. Tong, W. Huang, W. J. Liu, G. Wu, B. Zhang, L. Wang, J. Qi, H. Feng, F.-S. Wang, Q. Wang, G. F. Gao, Z. Yuan, J. Yan, A human neutralizing antibody targets the receptor-binding site of SARS-CoV-2. *Nature*. **584**, 120–124 (2020).

59. Y. Wu, F. Wang, C. Shen, W. Peng, D. Li, C. Zhao, Z. Li, S. Li, Y. Bi, Y. Yang, Y. Gong, H. Xiao, Z. Fan, S. Tan, G. Wu, W. Tan, X. Lu, C. Fan, Q. Wang, Y. Liu, C. Zhang, J. Qi, G. F. Gao, F. Gao, L. Liu, A noncompeting pair of human neutralizing antibodies block COVID-19 virus binding to its receptor ACE2. *Science*. **368**, 1274–1278 (2020).

60. S. J. Zost, P. Gilchuk, J. B. Case, E. Binshtein, R. E. Chen, J. P. Nkolola, A. Schäfer, J. X. Reidy, A. Trivette, R. S. Nargi, R. E. Sutton, N. Suryadevara, D. R. Martinez, L. E. Williamson, E. C. Chen, T. Jones, S. Day, L. Myers, A. O. Hassan, N. M. Kafai, E. S. Winkler, J. M. Fox, S. Shrihari, B. K. Mueller, J. Meiler, A. Chandrashekar, N. B. Mercado, J. J. Steinhardt, K. Ren, Y.-M. Loo, N. L. Kallewaard, B. T. McCune, S. P. Keeler, M. J. Holtzman, D. H. Barouch, L. E. Gralinski, R. S. Baric, L. B. Thackray, M. S. Diamond, R. H. Carnahan, J. E. Crowe Jr, Potently neutralizing and protective human antibodies against SARS-CoV-2. *Nature*. **584**, 443–449 (2020).

61. L. Premkumar, B. Segovia-Chumbez, R. Jadi, D. R. Martinez, R. Raut, A. Markmann, C. Cornaby, L. Bartelt, S. Weiss, Y. Park, C. E. Edwards, E. Weimer, E. M. Scherer, N. Rouphael, S. Edupuganti, D. Weiskopf, L. V. Tse, Y. J. Hou, D. Margolis, A. Sette, M. H. Collins, J. Schmitz, R. S. Baric, A. M. de Silva, The receptor binding domain of the viral spike protein is an immunodominant and highly specific target of antibodies in SARS-CoV-2 patients. *Sci Immunol*. **5** (2020), doi:10.1126/sciimmunol.abc8413.

62. T. N. Starr, A. J. Greaney, S. K. Hilton, D. Ellis, K. H. D. Crawford, A. S. Dingens, M. J. Navarro, J. E. Bowen, M. A. Tortorici, A. C. Walls, N. P. King, D. Veesler, J. D. Bloom, Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. *Cell*. **182**, 1295–1310.e20 (2020).

63. A. Grifoni, D. Weiskopf, S. I. Ramirez, J. Mateus, J. M. Dan, C. R. Moderbacher, S. A. Rawlings, A. Sutherland, L. Premkumar, R. S. Jadi, D. Marrama, A. M. de Silva, A. Frazier, A. F. Carlin, J. A. Greenbaum, B. Peters, F. Krammer, D. M. Smith, S. Crotty, A. Sette, Targets of T Cell Responses to SARS-CoV-2 Coronavirus in Humans with COVID-19 Disease and Unexposed Individuals. *Cell*. **181**, 1489–1501.e15 (2020).

64. N. Le Bert, A. T. Tan, K. Kunasegaran, C. Y. L. Tham, M. Hafezi, A. Chia, M. H. Y. Chng, M. Lin, N. Tan, M. Linster, W. N. Chia, M. I.-C. Chen, L.-F. Wang, E. E. Ooi, S. Kalimuddin, P. A. Tambyah, J. G.-H. Low, Y.-J. Tan, A. Bertoletti, SARS-CoV-2-specific T cell immunity in cases of COVID-19 and SARS, and uninfected controls. *Nature*. **584**, 457–462 (2020).

65. J. Mateus, A. Grifoni, A. Tarke, J. Sidney, S. I. Ramirez, J. M. Dan, Z. C. Burger, S. A. Rawlings, D. M. Smith, E. Phillips, S. Mallal, M. Lammers, P. Rubiro, L. Quiambao, A. Sutherland, E. D. Yu, R. da Silva Antunes, J. Greenbaum, A. Frazier, A. J. Markmann, L. Premkumar, A. de Silva, B. Peters, S. Crotty, A. Sette, D. Weiskopf, Selective and cross-reactive SARS-CoV-2 T cell epitopes in unexposed humans. *Science* (2020), doi:10.1126/science.abd3871.

66. M. Yuan, H. Liu, N. C. Wu, C.-C. D. Lee, X. Zhu, F. Zhao, D. Huang, W. Yu, Y. Hua, H. Tien, T. F. Rogers, E. Landais, D. Sok, J. G. Jardine, D. R. Burton, I. A. Wilson, Structural basis of a shared antibody response to SARS-CoV-2. *Science*. **369**, 1119–1123 (2020).

67. J. Greenbaum, J. Sidney, J. Chung, C. Brander, B. Peters, A. Sette, Functional classification of class II human leukocyte antigen (HLA) molecules reveals seven different supertypes and a surprising degree of repertoire sharing across supertypes. *Immunogenetics*. **63**, 325–335 (2011).

68. M. Maiers, L. Gragert, W. Klitz, High-resolution HLA alleles and haplotypes in the United States population. *Hum. Immunol.* **68**, 779–788 (2007).

69. Z. Du, HLA-DPA1 and HLA-DPB1 Frequencies in the US Populations (2017), (available at https://atcmeetingabstracts.com/abstract/hla-dpa1-and-hla-dpb1-frequencies-in-the-us-populations/).

70. J. Shang, G. Ye, K. Shi, Y. Wan, C. Luo, H. Aihara, Q. Geng, A. Auerbach, F. Li, Structural basis of receptor recognition by SARS-CoV-2. *Nature*. **581**, 221–224 (2020).

71. X.-Y. Ge, J.-L. Li, X.-L. Yang, A. A. Chmura, G. Zhu, J. H. Epstein, J. K. Mazet, B. Hu, W. Zhang, C. Peng, Y.-J. Zhang, C.-M. Luo, B. Tan, N. Wang, Y. Zhu, G. Crameri, S.-Y. Zhang, L.-F. Wang, P. Daszak, Z.-L. Shi, Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor. *Nature*. **503**, 535–538 (2013).

72. W. Ren, X. Qu, W. Li, Z. Han, M. Yu, P. Zhou, S.-Y. Zhang, L.-F. Wang, H. Deng, Z. Shi, Difference in receptor usage between severe acute respiratory syndrome (SARS) coronavirus and SARS-like coronavirus of bat origin. *J. Virol.* **82**, 1899–1907 (2008).

73. A. C. Walls, Y.-J. Park, M. A. Tortorici, A. Wall, A. T. McGuire, D. Veesler, Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell*. **181**, 281–292.e6 (2020).

74. D. F. Robbiani, C. Gaebler, F. Muecksch, J. C. C. Lorenzi, Z. Wang, A. Cho, M. Agudelo, C. O. Barnes, A. Gazumyan, S. Finkin, T. Hägglöf, T. Y. Oliveira, C. Viant, A. Hurley, H.-H. Hoffmann, K. G. Millard, R. G. Kost, M. Cipolla, K. Gordon, F. Bianchini, S. T. Chen, V. Ramos, R. Patel, J. Dizon, I. Shimeliovich, P. Mendoza, H. Hartweger, L. Nogueira, M. Pack, J. Horowitz, F. Schmidt, Y. Weisblum, E. Michailidis, A. W. Ashbrook, E. Waltari, J. E. Pak, K. E. Huey-Tubman, N. Koranda, P. R. Hoffman, A. P. West Jr, C. M. Rice, T. Hatziioannou, P. J. Bjorkman, P. D. Bieniasz, M. Caskey, M. C. Nussenzweig, Convergent antibody responses to SARS-CoV-2 in convalescent individuals. *Nature*. **584**, 437–442 (2020).

75. J. Braun, L. Loyal, M. Frentsch, D. Wendisch, P. Georg, F. Kurth, S. Hippenstiel, M. Dingeldey, B. Kruse, F. Fauchere, E. Baysal, M. Mangold, L. Henze, R. Lauster, M. A. Mall, K. Beyer, J. Röhmel, S. Voigt, J. Schmitz, S. Miltenyi, I. Demuth, M. A. Müller, A. Hocke, M. Witzenrath, N. Suttorp, F. Kern, U. Reimer, H. Wenschuh, C. Drosten, V. M. Corman, C. Giesecke-Thiel, L. E. Sander, A. Thiel, SARS-CoV-2-reactive T cells in healthy donors and patients with COVID-19. *Nature* (2020), doi:10.1038/s41586-020-2598-9.

76. M. E. Birnbaum, J. L. Mendoza, D. K. Sethi, S. Dong, J. Glanville, J. Dobbins, E. Ozkan, M. M. Davis, K. W. Wucherpfennig, K. C. Garcia, Deconstructing the peptide-MHC specificity of T cell recognition. *Cell*. **157**, 1073–1087 (2014).

77. L. K. Selin, M. Cornberg, M. A. Brehm, S.-K. Kim, C. Calcagno, D. Ghersi, R. Puzone, F. Celada, R. M. Welsh, CD8 memory T cells: cross-reactivity and heterologous immunity. *Semin. Immunol.* **16**, 335–347 (2004).

78. R. Gharpure, A. Perez, A. D. Miller, M. E. Wikswo, R. Silver, M. C. Hlavsa, Cryptosporidiosis Outbreaks - United States, 2009-2017. *MMWR Morb. Mortal. Wkly. Rep.* **68**, 568–572 (2019).

79. D. P. Oran, E. J. Topol, Prevalence of Asymptomatic SARS-CoV-2 Infection : A Narrative Review. *Ann. Intern. Med.* **173**, 362–367 (2020).

80. C. Kreer, M. Zehner, T. Weber, M. S. Ercanoglu, L. Gieselmann, C. Rohde, S. Halwe, M. Korenkov, P. Schommers, K. Vanshylla, V. Di Cristanziano, H. Janicki, R. Brinker, A. Ashurov, V. Krähling, A. Kupke, H. Cohen-Dvashi, M. Koch, J. M. Eckert, S. Lederer, N. Pfeifer, T. Wolf, M. J. G. T. Vehreschild, C. Wendtner, R. Diskin, H. Gruell, S. Becker, F. Klein, Longitudinal Isolation of Potent Near-Germline SARS-CoV-2-Neutralizing Antibodies from COVID-19 Patients. *Cell*. **182**, 1663–1673 (2020).

81. S. Bergthorsdottir, A. Gallagher, S. Jainandunsing, D. Cockayne, J. Sutton, T. Leanderson, D. Gray, Signals that initiate somatic hypermutation of B cells in vitro. *J. Immunol.* **166**, 2228–2234 (2001).

82. M. Anczurowski, N. Hirano, Mechanisms of HLA-DP Antigen Processing and Presentation Revisited. *Trends Immunol.* **39**, 960–964 (2018).

83. C. Wang, W. Li, D. Drabek, N. M. A. Okba, R. van Haperen, A. D. M. E. Osterhaus, F. J. M. van Kuppeveld, B. L. Haagmans, F. Grosveld, B.-J. Bosch, A human monoclonal antibody blocking SARS-CoV-2 infection. *Nat. Commun.* **11**, 2251 (2020).

84. K. M. McAndrews, D. P. Dowlatshahi, J. Dai, L. M. Becker, J. Hensel, L. M. Snowden, J. M. Leveille, M. R. Brunner, K. W. Holden, N. S. Hopkins, A. M. Harris, J. Kumpati, M. A. Whitt, J. J. Lee, L. L. Ostrosky-Zeichner, R. Papanna, V. S. LeBleu, J. P. Allison, R. Kalluri, Heterogeneous antibodies against SARS-CoV-2 spike receptor binding domain and nucleocapsid with implications for COVID-19 immunity. *JCI Insight*. **5** (2020), doi:10.1172/jci.insight.142386.

85. N. M. A. Okba, M. A. Müller, W. Li, C. Wang, C. H. GeurtsvanKessel, V. M. Corman, M. M. Lamers, R. S. Sikkema, E. de Bruin, F. D. Chandler, Y. Yazdanpanah, Q. Le Hingrat, D. Descamps, N. Houhou-Fidouh, C. B. E. M. Reusken, B.-J. Bosch, C. Drosten, M. P. G. Koopmans, B. L. Haagmans, Severe Acute Respiratory Syndrome Coronavirus 2-Specific Antibody Responses in Coronavirus Disease Patients. *Emerg. Infect. Dis.* **26**, 1478–1488 (2020).

86. C. Fenwick, A. Croxatto, A. T. Coste, F. Pojer, C. André, C. Pellaton, A. Farina, J. Campos, D. Hacker, K. Lau, B.-J. Bosch, S. Gonseth Nussle, M. Bochud, V. D'Acremont, D. Trono, G. Greub, G. Pantaleo, Changes in SARS-CoV-2 Spike versus Nucleoprotein Antibody Responses Impact the Estimates of Infections in Population-Based Seroprevalence Studies. *J. Virol.* **95** (2021), doi:10.1128/JVI.01828-20.

87. R. L. Tillett, J. R. Sevinsky, P. D. Hartley, H. Kerwin, N. Crawford, A. Gorzalski, C. Laverdure, S. C. Verma, C. C. Rossetto, D. Jackson, M. J. Farrell, S. Van Hooser, M. Pandori, Genomic evidence for reinfection with SARS-CoV-2: a case study. *Lancet Infect. Dis.* (2020), doi:10.1016/S1473-3099(20)30764-7.

88. J. D. Goldman, K. Wang, K. Roltgen, S. C. A. Nielsen, J. C. Roach, S. N. Naccache, F. Yang, O. F. Wirz, K. E. Yost, J.-Y. Lee, K. Chun, T. Wrin, C. J. Petropoulos, I. Lee, S. Fallen, P. M. Manner, J. A. Wallick, H. A. Algren, K. M. Murray, Y. Su, J. Hadlock, J. Jeharajah, W. R. Berrington, G. P. Pappas, S. T. Nyatsatsang, A. L. Greninger, A. T. Satpathy, J. S. Pauk, S. D. Boyd, J. R. Heath, Reinfection with SARS-CoV-2 and Failure of Humoral Immunity: a case report. *medRxiv* (2020), doi:10.1101/2020.09.22.20192443.

89. V. Gupta, R. C. Bhoyar, A. Jain, S. Srivastava, R. Upadhayay, M. Imran, B. Jolly, M. K. Divakar, D. Sharma, P. Sehgal, G. Ranjan, R. Gupta, V. Scaria, S. Sivasubbu, Asymptomatic reinfection in two healthcare workers from India with genetically distinct SARS-CoV-2. *Clin. Infect. Dis.* (2020), doi:10.1093/cid/ciaa1451.

90. M. K. Slifka, M. Matloubian, R. Ahmed, Bone marrow is a major site of long-term antibody production after acute viral infection. *J. Virol.* **69**, 1895–1902 (1995).

91. B. Reynisson, B. Alvarez, S. Paul, B. Peters, M. Nielsen, NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Res.* (2020) (available at https://academic.oup.com/nar/advance-article-abstract/doi/10.1093/nar/gkaa379/5837056).

92. Y. Watanabe, J. D. Allen, D. Wrapp, J. S. McLellan, M. Crispin, Site-specific glycan analysis of the SARS-CoV-2 spike. *Science*. **369**, 330–333 (2020).

93. F. Madeira, Y. M. Park, J. Lee, N. Buso, T. Gur, N. Madhusoodanan, P. Basutkar, A. R. N. Tivey, S. C. Potter, R. D. Finn, R. Lopez, The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res.* **47**, W636–W641 (2019).

94. W. Humphrey, A. Dalke, K. Schulten, VMD: visual molecular dynamics. *J. Mol. Graph.* **14**, 33–8, 27–8 (1996).

95. R. Vita, S. Mahajan, J. A. Overton, S. K. Dhanda, S. Martini, J. R. Cantrell, D. K. Wheeler, A. Sette, B. Peters, The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Res.* **47**, D339–D343 (2019).

96. E. W. Sayers, T. Barrett, D. A. Benson, E. Bolton, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, S. Federhen, M. Feolo, I. M. Fingerman, L. Y. Geer, W. Helmberg, Y. Kapustin, D. Landsman, D. J. Lipman, Z. Lu, T. L. Madden, T. Madej, D. R. Maglott, A. Marchler-Bauer, V. Miller, I. Mizrachi, J. Ostell, A. Panchenko, L. Phan, K. D. Pruitt, G. D. Schuler, E. Sequeira, S. T. Sherry, M. Shumway, K. Sirotkin, D. Slotta, A. Souvorov, G. Starchenko, T. A. Tatusova, L. Wagner, Y. Wang, W. J. Wilbur, E. Yaschenko, J. Ye, Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **39**, D38–51 (2011).

CHAPTER 2: Sex and age can influence tumor-immune interaction

2.1 Foreword

It is well established that sex is a biological variable that affects both innate and adaptive immune responses across multiple species (*1–4*). Age in and of itself results in a decline in immune function (*5*), and can also modify sex differences in immune responses (*1*). Differences in sex (including reproductive status) and age are associated with varying susceptibility and incidence of autoimmunity, infection, and cancer. Generally, females tend to have higher rates of autoimmune diseases such as Graves disease, Hashimoto thyroiditis, Multiple sclerosis, Rheumatoid arthritis, Lupus, and Type-1 diabetes while males have nearly a two-fold greater mortality rate from bladder, larynx, esophagus, and lung cancer (*1*, *6*, *7*). Investigations into the mechanistic root of these differences highlighted sex-chromosome-linked immune genes (*8*), Y-chromosome-linked tumor suppressors (*9*), and tumor suppressors that escape X-inactivation (*10*). There are also differential concentrations of sex steroids in men and women, which bind to immune cells and can influence signalling pathways as well as cytokine production (*11*, *12*). Furthermore, males accumulate somatic mutations earlier than females, with earlier risk of developing truncal mutations (*13*, *14*). Despite these facts, sex and age are generally not considered in cancer care and treatment.

The growing promise of immunotherapy has resulted in an increase in immune checkpoint blockade clinical trials for multiple tumor types (*15*). Existing biomarkers are poor at predicting response to treatment, and researchers are attempting to pinpoint what other variables are important for long-term clinical benefit. Earlier meta-analyses of immunotherapy studies provide conflicting evidence of sex differences in response (*16*, *17*). However, a more recent meta-analysis reported sex-associated biases in immune cell abundances and immune checkpoints in some tumor types, as well as a differential mutational/neoantigen load in a few clinical trials (*18*). Even more

recently, a preprint identified differential expression of mitochondrial genes as a likely driver of autoimmune incidence and multiple cancers (*19*). Altogether, in addition to general association of sex and age to cancer incidence and immunotherapy response, it is crucial to understand why such differences exist. At a glance, the stronger overall immune responses observed in younger and female individuals might lead to the paradoxical assumption that younger and female patients should generally respond well to immunotherapy. However, as I show in Aim 2, the interaction of sex and age as biological variables results in differential tumor-immune interaction and thus varying availability of visible driver mutations. My work provides some mechanistic insight into the cause of immunotherapy-response discrepancies.

2.2 Abstract

Individual MHC genotype constrains the mutational landscape during tumorigenesis. Immune checkpoint inhibition reactivates immunity against tumors that escaped immune surveillance in approximately 30% of cases. Recent studies demonstrated poorer response rates in female and younger patients. Although immune responses differ with sex and age, the role of MHC- based immune selection in this context is unknown. We find that tumors in younger and female individuals accumulate more poorly presented driver mutations than those in older and male patients, despite no differences in MHC genotype. Younger patients show the strongest effects of MHC-based driver mutation selection, with younger females showing compounded effects and nearly twice as much MHC-II based selection. This study presents evidence that strength of immune selection during tumor development varies with sex and age, and may influence the availability of mutant peptides capable of driving effective response to immune checkpoint inhibitor therapy.

2.3 Introduction

The major histocompatibility complex (MHC) exposes protein content on the cell surface to allow detection of antigens by the immune system. This applies to non-self-antigens such as viral proteins, and self-proteins that include tumor antigens. Tumor cells harbor oncogenic alterations that can be presented to the immune system by the MHC, causing immune recognition and elimination (immune surveillance) (*20*). However, in order to grow, invade, and spread, tumors must evade immune surveillance. Common mechanisms of immune evasion include loss of the MHC molecules and the upregulation of immune checkpoint molecules on cell surfaces that normally regulate the amplitude and duration of a T-cell response (*21*). Immune checkpoint blockade (ICB) uses antibodies to block these immune checkpoint molecules, and can invigorate inactive and/or exhausted T cells to produce antitumor effects that confer long-term survival benefits in certain types of cancer (*22*). However, ICB is effective in only 10-40% of patients for reasons that remain unclear. Meta-analyses of clinical trials in multiple cancer types treated with ICB suggest that young and female patients are characterized by low response rates (*18*, *23–26*). The reason(s) for the poor response of these two populations remains elusive.

An accumulating body of literature points to sexual dimorphism in immune responses (*1*). Moderated by genetic and hormonal factors, females have twice the antibody response to influenza vaccines (*27*) and higher CD4+ T-cell counts than males (*28*). Moreover, females are far more susceptible to autoimmune diseases (*7*), demonstrating a stark imbalance in the way the immune response causes diseases in the two sexes. Immuno-sequencing of over 800 individuals revealed sex-associated differences in the extent to which HLA molecules propagate selection and expansion of CD8+ T cells (*29*). Interestingly, a stronger immune response in females has been observed across several species (*3*, *4*, *30*), and sexual dimorphism has been demonstrated in

immune selection and restriction of intratumor genetic heterogeneity in a mouse model of B-cell lymphoma (*31*). In addition, a recent study has found sex-based differences in molecular biomarkers and immune checkpoint expression in multiple tumor types treated with ICB (*18*). Altogether, these studies suggest that these differences are sex-specific and not lifestyle dependent.

Studies have demonstrated age-related changes in immune response as well. As humans age, there is a decrease of general immune function including production of IL-2, a pivotal growth factor for T cells (*32*). Reduced thymic output, lower numbers of naïve T cells, and overall reshaping of the size and specificity of the T-cell repertoire by microbial pathogens may explain why, for example, about 90% of excess deaths during flu season occur in patients greater than 65 years of age (*33*). In addition, elderly people have reduced phagocytic function and HLA-II expression on antigen presenting cells (*34*). Collectively, these factors render elderly individuals less able to mount a T-cell response to new antigens and respond to vaccination.

Recently, we developed the Patient Harmonic-mean Best Rank (PHBR) score that quantifies patients' ability to present somatic mutations in their tumor by their specific MHC-I and MHC-II haplotypes (*35*, *36*). PHBR-I and PHBR-II scores aggregate predicted peptide-MHC molecule binding affinities from established tools (*37*, *38*) to produce a mass spectrometry-validated, residue-centric, and patient-specific presentation score that captures a mutant peptide's visibility to the immune system. In previous publications we used PHBR scores to assess the role of MHC genotype in shaping mutation accumulation during tumorigenesis (*35*, *36*). We found that patients tend to accumulate driver mutations that cannot be effectively presented by their own MHC molecules, likely a consequence of immune-based elimination of tumor cells harboring well-presented driver mutations, a selective process referred to as immunoediting (*39*). This analysis revealed that thyroid carcinoma and low-grade glioma patients experience the highest MHC-based

selective pressure on driver mutations (*35*, *36*). Interestingly, these tumor types also had the youngest average age at diagnosis compared to all studied tumor types. In light of these observations, we reasoned that younger and female patients may experience stronger immunoediting early in their tumor history, accumulating mutations that are less favorably presented by their MHC, i.e., mutations more invisible to their immune system, at the time of diagnosis. Predictably, a depletion of potentially immunogenic mutant peptides would cause ICB to be ineffective. At first approximation we ruled out an effect due to sex-specific (MHC-I Pearson R = 0.99, MHC-II Pearson R = 0.99) or age-specific (MHC-I Pearson R = 0.98, MHC-II Pearson R=0.99) imbalances in MHC genotype frequencies. Therefore, we sought to test the hypothesis that sex- and age-specific differences in driver mutation presentation are the result of differential immunoediting.

Here, we find that female and younger patients exhibit stronger immune selection in their tumors, measured by the affinity of their observed, expressed driver mutations compared to male and older patients. MHC-II appears to have a stronger effect compared to MHC-I. Our findings, based on TCGA samples, are validated in an independent validation cohort.

2.4 Results

Fewer presentable drivers in female and younger patients. We focused on a set of 1018 driver mutations, defined in (*35*), as driver mutations are more prevalent in the clonal architecture of an individual's cancer and confer a selective growth advantage. We assigned MHC-I and MHC-II types using PolySolver and HLA-HD, two exome-based calling methods (*40*, *41*) and considered only microsatellite-stable TCGA tumors. After excluding 515 patients from class I and 1064 patients from class II analyses due to HLA genotype incompatibility with NetMHCpan affinity prediction software, 9913 patients with MHC-I calls and 7174 patients with MHC-II calls

remained. These patients were diverse in sex, with more males than females, and a broad distribution of age at diagnosis. PHBR-I and -II scores were calculated for all patients across the 1018 driver events by taking the harmonic mean of each allele's best NetMHCpan percentile rank affinity score, providing an estimate of each patient's potential to present each mutation via MHC-I and MHC-II, respectively. Importantly, the PHBR-I and PHBR-II scores aggregate percentile rank scores of mutated peptides relative to large numbers of random peptide provided by NetMHCpan-4.0 and NetMHCIIpan3.2. For single peptide-HLA pairs, percentile rank scores of 0.5% and 2% for MHC-I and 2% and 10% for MHC-II have been used to represent strong and weak binding cutoffs respectively (*42*, *43*).

To rule out other covariates, we performed a series of control analyses. We categorized patients into subgroups according to sex (male versus female) and age (younger versus older based on pan- cancer 30th and 70th percentiles at age of diagnosis for categorical analyses). For sex-specific analyses, we further excluded seven sex- specific tumor types (breast, cervical, ovarian, uterine, prostate, and testicular cancer). First, we established that there were similar average numbers of driver mutations across sex and age patient groups. We previously found that TCGA patients with somatic MHC-I mutations had altered mutational landscapes, with a higher fraction of binding mutant peptides than patients without MHC-I mutations (*44*). To ensure that somatic MHC-I mutations would not skew the driver mutation PHBR-I score distributions, we compared scores for patients with and without MHC-I mutations grouped by sex and age and found no significant differences (**Figure S2.1**). We then compared the distributions of patient PHBR-I and PHBR-II scores across the 1018 driver mutations (**Figure S2.2A-D**) and found significant p values, but very small effect sizes between groups. To ensure that the potential to present driver mutations was consistent across sex and age, we compared the fraction of presented drivers at various score

thresholds, and found no significant differences (**Figure S2.2E-F**). The overall similarity of MHC presentation suggests that patients of both sexes and various ages at diagnosis present driver mutations with roughly equivalent efficacy, implying that specificity of MHC presentation resulting from specific allele combinations is not a mechanism causing differences in ICB response rate.

We therefore reasoned that the discrepancy might be due to differences in the strength of immune selection, e.g., tumors with stronger immunoediting should retain fewer driver mutations that are presentable to T cells by the patient's own MHC molecules. For sex- and age-specific groups in each cohort, we compared the PHBR-I and PHBR-II score distributions for observed, RNA-expressed driver mutations observed in patient tumors, excluding 4782 patients with no drivers from the list of 1018. While the number of observed drivers was not significantly different between sex and age groups, younger female patients were overrepresented in the group with no observed driver mutations (Fisher's exact test: class I: OR = 1.12, $p < 0.12$; class II: OR = 1.28, $p < 0.015$). We note this group had an overrepresentation of thyroid cancer cases, a disease associated with low mutational burden and that typically only has a single driver mutation (*45*). We therefore performed sex-specific analysis for unique 2900 patients and age-specific analysis for 3928 unique patients.

Across pan-cancer cohorts, females were at a significant disadvantage (higher PHBR scores) in presenting their driver mutations by both their MHC-I and MHC-II molecules (**Figure 2.1A-B**, $p < 2.6e{-}04$ and $p < 1.2e{-}07$, respectively). Younger patients also tended to have worse presentation of driver mutations by both MHC-I and MHC-II molecules (**Figure 2.1C-D,** $p < 2.4e{-}5$ and $p < 7.3e{-}04$, respectively). Notably, the shift in PHBR score distributions between groups occurs near the threshold for weak binding. Given that a limited number of somatic

mutations generate mutant peptides and not all of these are immunogenic, this small shift may translate to significantly less opportunity to generate a host antitumor response upon ICB. Importantly, we found that these observed between-group differences in PHBR scores were far greater (falling outside the 99% confidence interval) than differences when we randomly reassigned mutations across patients and recalculated patient-specific PHBR scores (**Methods; Figure S2.3**), and were an order of magnitude greater than the effect sizes observed when comparing score distributions independent of mutation occurrence (**Figure S2.2**). We also found differences in affinity independent of the PHBR score, using median NetMHCpan affinity scores across all alleles (**Figure S2.4**). Altogether this suggests that score differences do indeed result from the interaction of inherited MHC genotype with the observed mutations. Interestingly, the mutation-specific fraction of RNA reads mapping to these driver mutations was significantly lower for females and younger patients (**Figure S2.5**), further supporting sex- and age-based differential strength in immune selection.

We next examined evidence for sex and age differences in specific tumor types, adjusting age thresholds according to tumor type. There was a general trend for female and younger patients' tumors to have higher median PHBR-I and II scores across tumor types, although the difference was only statistically significant in melanoma (**Figure S2.6A**). We observed more variability in the trends across tumor types by age. Younger individuals trended toward higher median PHBR-I and II scores in tumors where the 30th/70th percentile was associated with a large age gap and the younger age threshold was under 55, with some notable exceptions that included rectal cancer, thyroid cancer, stomach cancer, and liver (**Figure S2.6B**). Overall these trends suggest that stronger pan-cancer immune selection in younger and female patients results broadly from effects observed across multiple tumor types.

Next, we explored the effect of age and sex in the context of the immune system's ability to eliminate effectively-presented mutations by modeling the relationship between mutation occurrence and immune visibility as modeled by PHBR-I and II scores. We constructed sex- and age-specific generalized additive models with random effects to account for variation in mutation rate across individuals, and examined the coefficients corresponding to independent and interaction effects for PHBR-I, PHBR-II, and sex or age to assess their contribution to immune selection for expressed mutations observed ≥2 times in the cohort, excluding patients with no observed, expressed driver mutations. To control for the fact that some driver mutations occurred in the same tumor, and thus are not completely independent events, we included patient ID as a random effect in our linear model. In both models, we found that PHBR-I and PHBR-II scores alone had significant effects on the probability of a mutation to be a target of immune selection (**Table 2.1**). Positive coefficients for both PHBR scores indicate that the higher the PHBR score (i.e., poorer presentation), the higher the probability of mutation. Furthermore, when we quantified the influence of both scores on probability of mutation using odds ratios between respective 25th and 75th percentiles, we found that PHBR-II (OR: 3.4, CI [3.19, 3.6]) has a much larger impact on probability of mutation than PHBR-I (OR: 1.27, CI [1.26, 1.29]), echoing the larger effect sizes seen in Figure 2.1. As expected, sex and age alone did not influence the probability of mutation; however, of particular interest are the interaction terms that indicate the influence of PHBR scores on probability of mutation within the context of sex and age. Both the PHBR-I:sex and PHBR-I:age interactions as well as the PHBR-II:sex and PHBR-II:age interactions were significant. The negative PHBR:age estimates indicate stronger effects of PHBR-I as well as PHBR-II contribution to the probability of mutation in younger patients. On the other hand, positive PHBR:sex estimates indicate stronger effects of PHBR-I and PHBR-II contributing to probability of mutation in

females according to the model formulation (**Methods**). Collectively, these results suggest stronger immune selection in females and younger patients.

As females and younger patients both demonstrated stronger immune selection compared to males and older patients, we further partitioned the cohorts simultaneously by sex and age, and investigated the distribution of PHBR-I and -II scores for these groups. We found that sex and age effects are cumulative, with tumors in younger females exhibiting significantly higher selective pressure by MHC than those in the other three groups (**Figure 2.2**). We noticed a profound difference between PHBR score distributions between younger females and older males. Because younger males had worse presentation of their driver mutations compared to older females (**Figure 2.2**), we sought to ensure that sex had an effect on immune selection independent of age. In two models incorporating sex, age, and PHBR-I and PHBR-II scores, respectively, both PHBR:sex and PHBR:age were independently significant for both class I and class II (**Table S2.1**). These results demonstrate that more aggressive immune selection in younger females selects for tumors with driver mutations that are less visible to the immune system.

Mutational signatures do not explain differential selection. We next explored whether sex- and age-specific effects could be driven by differences in environmental exposure rather than the strength of immune selection. Mutational signatures assign specific mutations to different mutagenic processes, allowing the exploration of differences in environmental exposure across sex and age. We compared the sex-specific occurrence of mutational signatures in each tumor type and found only a minority of instances where signature strength was weakly but significantly associated with sex (**Figure 2.3A**). Importantly, only three of the signatures (01, 02, and 05) where we observed significant sex-specific differences contribute to the set of driver mutations used for this analysis (**Figure 2.3B**). Since signatures 01 and 05 are endogenous rather than exposure

associated signatures, this suggests a very low impact of environmental exposures on sex-specific effects of immune selection on drivers. Furthermore, when we excluded the tumor types with significant signature differences (glioblastoma multiforme, GBM and liver hepatocellular carcinoma, LIHC), we still observed sex- and age-related differences (**Table S2.2**). In addition, only two signatures correlated with age, both of which have known association with aging (*46*). We examined C>T and T>C mutations, which are hallmarks of signature 01 and 05, respectively, and found that observed driver mutations in these categories were broadly distributed across age at diagnosis. To explain weaker immune selection in older individuals, age-related mutations would have to be better presented (have lower PHBR scores) than other mutations. Instead, we found that C>T and T>C mutations were significantly more poorly presented (had slightly higher PHBR scores) than other mutations across all possible MHC-I and MHC-II alleles, suggesting that these mutations, and by extension, signatures 01 and 05, could not drive the apparent age-associated difference in immune selection (**Figure 2.3C**). Thus, we conclude that the sex- and age-specific effects on immune selection are not likely due to environmental exposure differences (*46*, *47*).

Validation in an independent non-TCGA cohort. We sought validation of our findings in a cohort of 342 patients (309 with compatible MHC-I type calls and 277 with MHC-II type calls) compiled from published dbGaP studies and non-TCGA samples in the International Cancer Genome Consortium (ICGC) database (*48*) and filtered to exclude tumor types not represented in TCGA. While fewer tumor types were represented relative to the discovery cohort, these patients were diverse with respect to sex and age at diagnosis, with slightly more males than females, and similar average numbers of driver mutations. As in the discovery cohort, we found some significant differences in patient PHBR score distributions across the 1018 driver mutations, also

with very small effect sizes between groups. Likewise, there was no difference in the fraction of presented drivers at various score thresholds (**Figure S2.7**). The majority of our validation cohort did not have expression data, so we predicted RNA expression using a logistic regression classifier trained on the TCGA cohort (**Methods**).

We found, as in the discovery cohort, that effectively-presented driver mutations were significantly depleted in younger and female patients compared to older and male patients (**Figure 2.4A-D**). These differences were an order of magnitude greater than the effect sizes observed when comparing score distributions independent of mutation occurrence (**Figure S2.7E-H**).

When we examined the simultaneous effects of sex and age (**Figure 2.4E-F**), younger females once again had significantly worse presentation of their driver mutations than older males across both MHC-I and MHC-II ($p < 0.001$, $p < 0.007$). We repeated the sex- and age-specific analyses using the generalized additive models and found that, for both sex and age, PHBR-II scores alone significantly influenced the probability of mutation, with higher PHBR scores (i.e., worse presentation) leading to higher probability of mutation (**Table S2.3**). While PHBR-II:sex and PHBR-II:age coefficients trended in the same direction, with stronger effects in females and younger patients, they did not reach significance, likely due to sample size.

2.5 Discussion

Here, we present evidence that both sex and age impact the driver mutations that arise and persist during tumorigenesis. We found that younger and female patients accumulate driver mutations in their tumors that are less readily presented by their MHC molecules (**Figure 2.5**), suggesting a stronger toll by immune selection early in tumorigenesis. This finding is consistent with recent meta-analyses across multiple tumors showing sex- and age-dependent differences in

response to ICB (*23–26*). We also observed the strongest effects in MHC-II based selection, in agreement with the fact that females have higher CD4+ T-cell counts than males (*49*). A prevalent role of MHC-II driven immune selection can be explained by the fact that CD4+ T cells, besides direct effector function comparable to that of CD8+ T cells, also play a deep-rooted regulatory role in cooperating with CD8+ T cells via associative recognition of antigen (*50*, *51*). Their function in orchestrating T-cell immunity, in general terms, makes them privileged actors, hence targets of immune selection as revealed herein. In older individuals, immune selection effects by MHC-II presentation of driver mutations are mitigated by a reduced CD4+/CD8+ ratio (*52*) and greater telomere attrition in CD4+ T cells than in CD8+ T cells (*53*) leading to accelerated senescence. Taken together, the evidence suggests that tumors developing in younger and female patients are prone to stronger immunoediting than those in older and male patients.

Our findings based on the TCGA were reproduced in the smaller validation cohort where we once again observed poorer MHC-based presentation of driver mutations in females versus males and younger versus older patients, with presentation being worse in younger and female patients. When modeling the influence of MHC genotype on the probability of observing driver mutations, the estimated effect sizes are modest, although relatively large compared to effects detected by genome wide association studies where odds ratios are often <1.240. Several sources of uncertainty, including errors in patient genotyping, prediction of the peptide-HLA binding affinities used to calculate the PHBR score, and errors in somatic mutation calling could obscure the true effects (*35*). More accurate estimates will likely require larger sample sizes, and ideally availability of expression data as non- expressed mutations should not reflect the effects of immune selection.

In this analysis, we focused on a set of recurrent missense and indel mutations in established driver genes developed in our previous work. This is motivated by the assumption that these are more likely to occur early during tumorigenesis, and may thus provide a view of immune selection before various mechanisms of immune evasion occur (*36*). However, it is unlikely that immune selection operates differently on different categories of mutation, and nondriver mutation-derived neoantigens should be equally capable of triggering a T-cell response. Whether tumor cells can evade T cell responses more easily when they are targeted against nonessential nondriver mutations remains an important question. It has been suggested that ICB responses are most effective when a clonal driver neoantigen is present (*54*). While we did not observe large sex or age bias in the mutational signatures associated with the 1018 driver mutations, we speculate that it is possible non-driver mutations could show differences in their potential to serve as neoantigens if the underlying mutational processes are active at different times or are biased to generate mutations in expressed protein coding sequences with characteristics that bias their presentation.

Notwithstanding some limitations, our analysis provides a compelling case for the paradigm that immune selection exerts its toll differently with respect to sex and age, with a greater effect in younger females. Of note, the younger female cohort had the poorest driver mutation presentation across both the discovery and validation cohorts, suggesting that these effects are strong and complementary. Although our analysis suggests that younger age is associated with stronger antitumor immune responses, we strongly suggest caution in considering whether this trend could generalize to pediatric tumors. The genomic landscape of pediatric tumors is distinct from that of adulthood tumors, with lower mutation burdens, different driver events and more germline factors and the characteristics of the pediatric immune system differ greatly from those of an adult (*55*). Furthermore, we are unable to control for other sex- and age-related factors

beyond predicted MHC presentation of driver mutation-derived peptides. These possibilities may include (a) differences in the antigen processing machinery preceding surface exposure of MHC-peptide complexes, and (b) genetic and epigenetic factors causing preferential mutation accumulation in the cohorts for reasons other than immunoediting.

In conclusion, this study indicates that immune selection exerts its toll differently with respect to sex and age, with a greater effect in younger females. As such, the response rate to ICB may be dependent on the strength of immune selection occurring early in tumorigenesis. Methods to accurately predict the impact of immunoediting on a patient-specific basis may lead to better predictive algorithms for response to therapy. As a corollary, we posit that ICB treatment is likely to have a reduced effect in younger female patients since this treatment will attempt to reactivate T cells for immunologically invisible neoantigens. Rather, adaptive T-cell therapy against patient-validated neoantigens or therapeutic vaccination against conserved antigens will likely be more beneficial in these patients. Notably prior to treatment with ICB, male sex (and less consistently older age) are associated with higher risk of recurrence and death in melanoma and may stand to benefit more from ICB (*56*, *57*), thus it is also possible that overall stronger immune surveillance could prove advantageous in the context of ICB despite differences in the quality of neoantigens. Finally, these findings shed light on the role of immune surveillance in cancer progression.

2.6 Materials and Methods

Data acquisition. Data were obtained from publicly available sources including The Cancer Genome Atlas (TCGA) Research Network (http://cancergenome.nih.gov/). TCGA normal exome sequences and TCGA clinical data were downloaded from the GDC on June 23-26th, 2018

and April 25th, 2017 respectively. Furthermore, TCGA somatic mutations were accessed from the NCI Genomic Data Commons (https://portal.gdc.cancer.gov/) on May 14th, 2017.

Validation cohort. dbGaP studies (accession numbers: phs001493.v1.p1.c2, phs001041.v1.p1.c1, phs001425.v1.p1.c1, phs001493.v1.p1.c1, phs000980.v1.p1.c1, phs001469.v1.p1.c1, phs000452.v2.p1.c1, phs001451.v1.p1.c1, phs001519.v1.p1.c1, phs001565.v1.p1.c1) were obtained from the dbGaP database and WXS/WGS data obtained from the Sequence Read Archive (SRA) (*57*). Somatic mutation files were obtained from the respective papers associated with each study. Additional non-TCGA patients' WXS/WGS data was obtained from the ICGC and somatic mutation data from the ICGC DCC Data Release on (April 2, 2019 (PCAWG), March 18, 2019 (THCA-SA)). The validation cohort's MHC-I and -II genotypes were typed using HLA-HD (*41*) and PHBR scores calculated using the method described in "Presentation score assignment".

HLA typing. HLA genotyping was performed for class I genes HLA-A, HLA-B, HLA-C, and class II genes HLA-DRB1, HLA-DPA1, HLA-DPB1, HLA-DQA1, and HLA-DQB1, which encode three protein determinants of MHC-I peptide binding specificity, HLA-DR, HLA-DP, and HLA-DQ. TCGA samples were typed with Polysolver (*40*), with default parameters, for class I and typed with HLA-HD27, using default parameters, for class II. Both tools require germline (whole blood or tissue matched) whole exome sequenced samples. Samples with very low coverage on specific genes are left untyped by HLA-HD. Patients were assigned an HLA-DR type if they were successfully typed for HLA-DRB1. Patients were assigned HLA-DP and -DQ types if they had successful typing for HLA-DPA1/HLA-DPB1 and HLA-DQA1/HLA-DQB1, respectively. Class I and class II types were validated by xHLA (*58*), run with default parameters, and only patients where all alleles agreed in both classes were included in the analysis.

Presentation score assignment. We used patient presentation scores, as defined in (*35*), to represent a particular patient's ability to present a residue given their distinct set of HLA types. For class I, 6 HLA alleles were considered (HLA-A, HLA- B, and HLA-C). For class II, 12 HLA-encoded MHC-II molecules (4 combinations of HLA-DPA1/DPB1 and HLA-DQA1/DQB1; 2 alleles of HLA-DRB1 considered twice each, since HLA-DRA1 is invariant, for consistency between resulting molecules). NetMHCpan4.0 (*42*) and NetMHCIIpan3.2 (*43*) were used to calculate binding affinities. The PHBR score was assigned as the harmonic mean of the best residue presentation scores for each group of MHC-I and MHC-II molecules. A lower patient presentation score indicates that the patient's MHC molecules are more likely to present a residue on the cell surface.

Set of driver mutations. Somatic mutations were considered to be recurrent and oncogenic if they occurred in one of the 100 most highly ranked oncogenes or tumor suppressors described by Davoli et al. (*59*) and were observed in at least three TCGA samples. Among these, we retained only mutations that would result in predictable protein sequence changes that could generate neoantigens, including missense mutations and inframe indels. A total of 1018 mutations (512 missense mutations from oncogenes, 488 missense mutations from tumor suppressors, 11 indels from oncogenes and 7 indels from tumor suppressors) were obtained (*35*).

Modeling the effects of PHBR score on mutation probability. We built two matrices, for PHBR-I scores and PHBR-II scores, from the 1018 mutations and the 1912 patients with both PHBR-I and -II calls. Next, we built a binary mutation matrix $y_{ij} \in \{0,1\}$ indicating whether patient i has a specific mutation j. We evaluated the relationship between this binary matrix, the matched 1912×1018 matrices with log PHBR-I and -II scores, $x1_{ij}$ and $x2_{ij}$, respectively, and the variable of interest (sex or age) for patient i and mutation j. We fit a generalized additive model for the

centered log PHBR-I, centered log PHBR-II scores, centered sex (coded 0/1 for males/females) or centered age, and mutation probability with the GAM function in the MGCV R package (*60*). To estimate the effects of PHBR and sex or age on probability of mutation, we considered the following random effects models:

(1) $\text{Logit}\,(P(y_{ij} = 1)) = \beta_1 x1_{ij} + \beta_2 x2_{ij} + \beta_3 \text{Sex}_i + \beta_4 (x1_{ij} * \text{Sex}_i) + \beta_5 (x2_{ij} * \text{Sex}_i) + \eta_i$

(2) $\text{Logit}\,(P(y_{ij} = 1)) = \beta_1 x1_{ij} + \beta_2 x2_{ij} + \beta_3 \text{Age}_i + \beta_4 (x1_{ij} * \text{Age}_i) + \beta_5 (x2_{ij} * \text{Age}_i) + \eta_i$

And PHBR-I and PHBR-II specific models (results in **Table S2.1**):

(3) $\text{Logit}\,(P(y_{ij} = 1)) = \beta_1 x1_{ij} + \beta_2 \text{Age}_i + \beta_3 \text{Sex}_i + \beta_4 (x1_{ij} * \text{Sex}_i) + \beta_5 (x1_{ij} * \text{Age}_i) + \eta_i$

(4) $\text{Logit}\,(P(y_{ij} = 1)) = \beta_1 x2_{ij} + \beta_2 \text{Age}_i + \beta_3 \text{Sex}_i + \beta_4 (x2_{ij} * \text{Sex}_i) + \beta_5 (x2_{ij} * \text{Age}_i) + \eta_i$

where $\eta_i \sim N(0, \theta_\eta)$ are random effects capturing different mutation propensities among patients, using patient IDs. In these models βn measures the effect of the log-PHBR-I, log-PHBR-II, and sex or age. This analysis was repeated for the validation cohort.

2.7 Figures

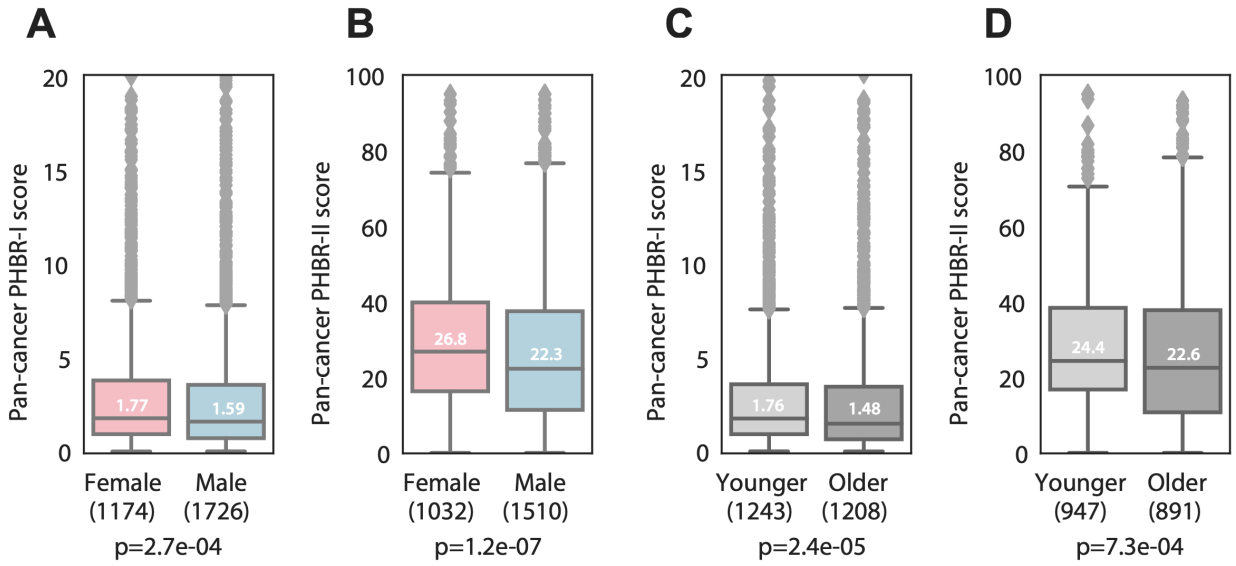


Figure 2.1. Sex- and age-specific MHC presentation of observed, RNA-expressed driver mutations. Box plots denoting the distribution of (**A**) PHBR-I and (**B**) PHBR-II scores for expressed driver mutations in female and male pan-cancer patients. c, d Box plots denoting the distribution of (**C**) PHBR-I and (**D**) PHBR-II scores for expressed driver mutations in younger and older pan-cancer patients. P values were calculated using the one-tailed Mann–Whitney U test. Median values are shown in each boxplot. All box plots include the median line, the box denotes the interquartile range (IQR), whiskers denote the rest of the data distribution and outliers are denoted by points greater than ±1.5 × IQR. The following effect sizes were calculated using Cliff's d: (A) r = −0.0654, (B) −0.104, (C) −0.081, (D) −0.0734.

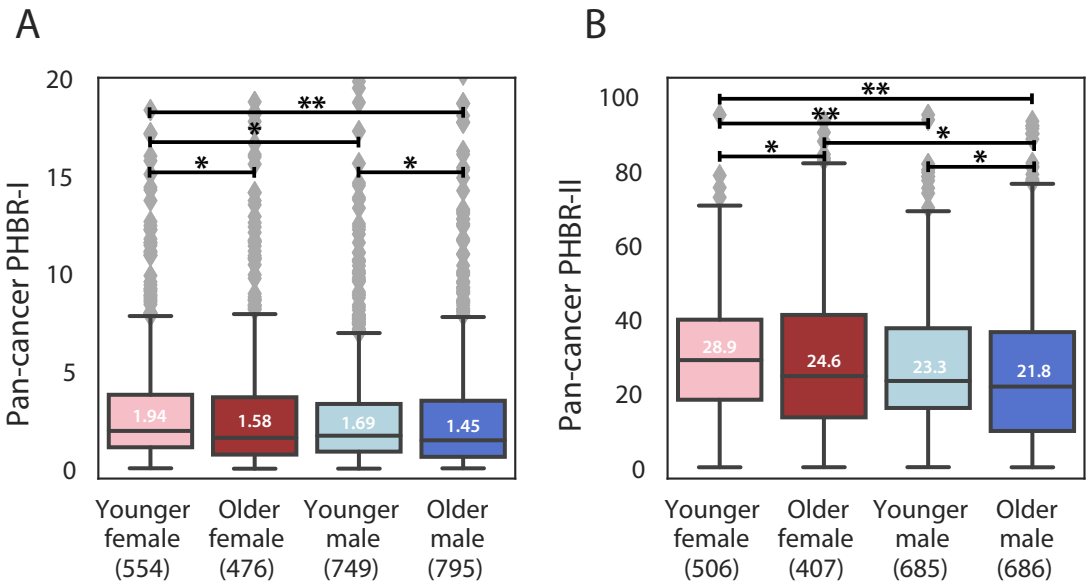


Figure 2.2. Integrated sex- and age-specific analysis. (**A**) PHBR-I and (**B**) PHBR-II scores for the observed driver mutations in pan-cancer integrated sex- and age- specific patient cohorts. One asterisk indicates p values < 0.05 and two asterisks indicates p values < 0.001. All p-values were calculated using a one-tailed Mann–Whitney U test. The Benjamini–Hochberg method was used to adjust for multiple comparisons. Median values are shown in each boxplot. Exact p-values for (A) include: YF, OM: 0.7e−05; YF, OF: 0.005; YF, YM: 0.008; YM, OM: 0.008; OF, OM: 0.08; OF, YM: 0.22. Exact p values for (B) include: YF, OM: 5.51e−07; YF, YM: 0.0003; YM, OM: 0.035; YF, OF: 0.038; OF, YM: 0.17. Y = younger, O = older, F = female, M = male. All box plots include the median line, the box denotes the interquartile range (IQR), whiskers denote the rest of the data distribution and outliers are denoted by points greater than ±1.5 × IQR.

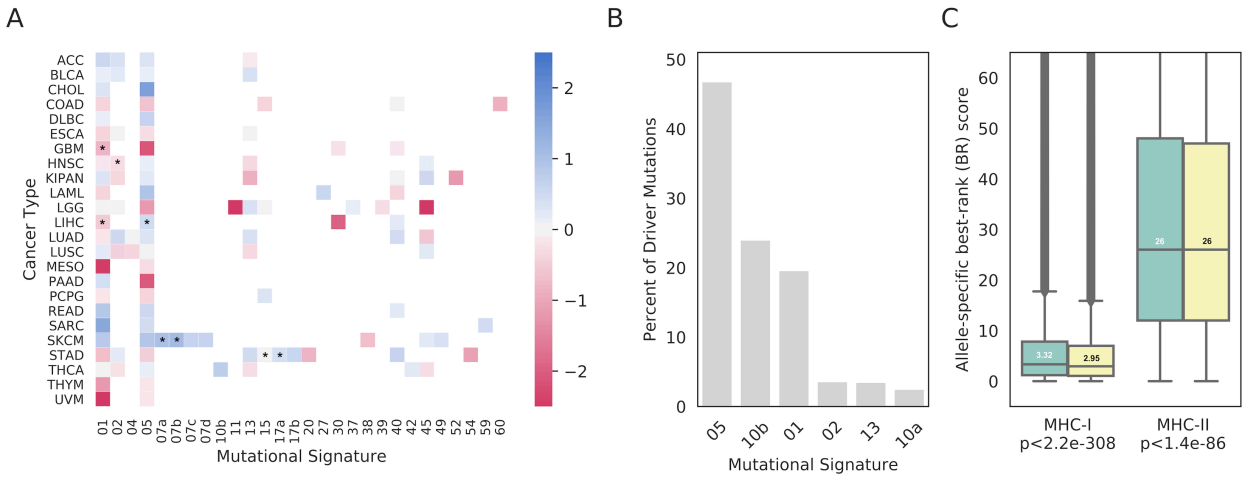


Figure 2.3. Sex-specific exposure analysis with mutational signatures. (**A**) Heatmap of log2 male (blue) to female (pink) ratios of mutational signatures for each tumor type with asterisks denoting a significantly different ratio between male and female sexes. (**B**) The percentage of mutations in the set of driver mutations that are part of each mutational signature. (**C**) Boxplot comparing MHC-I and MHC-II presentation scores across all possible alleles for C>T or T>C driver mutations (green) versus driver mutations resulting from other base substitutions (yellow); 1,063,975 and 2,051,300 affinity scores were evaluated for C>T or T>C mutations for class I and II, respectively; and 1,851,025 and 3,568,700 affinity scores were evaluated for other mutations for class I and II, respectively. Exact p-values were calculated using a one-tailed Mann–Whitney U test: (C) 2.2e−308 and 1.4e−86. Median values are denoted in each boxplot. All box plots include the median line, the box denotes the interquartile range (IQR), whiskers denote the rest of the data distribution and outliers are denoted by points greater than ±1.5 × IQR.

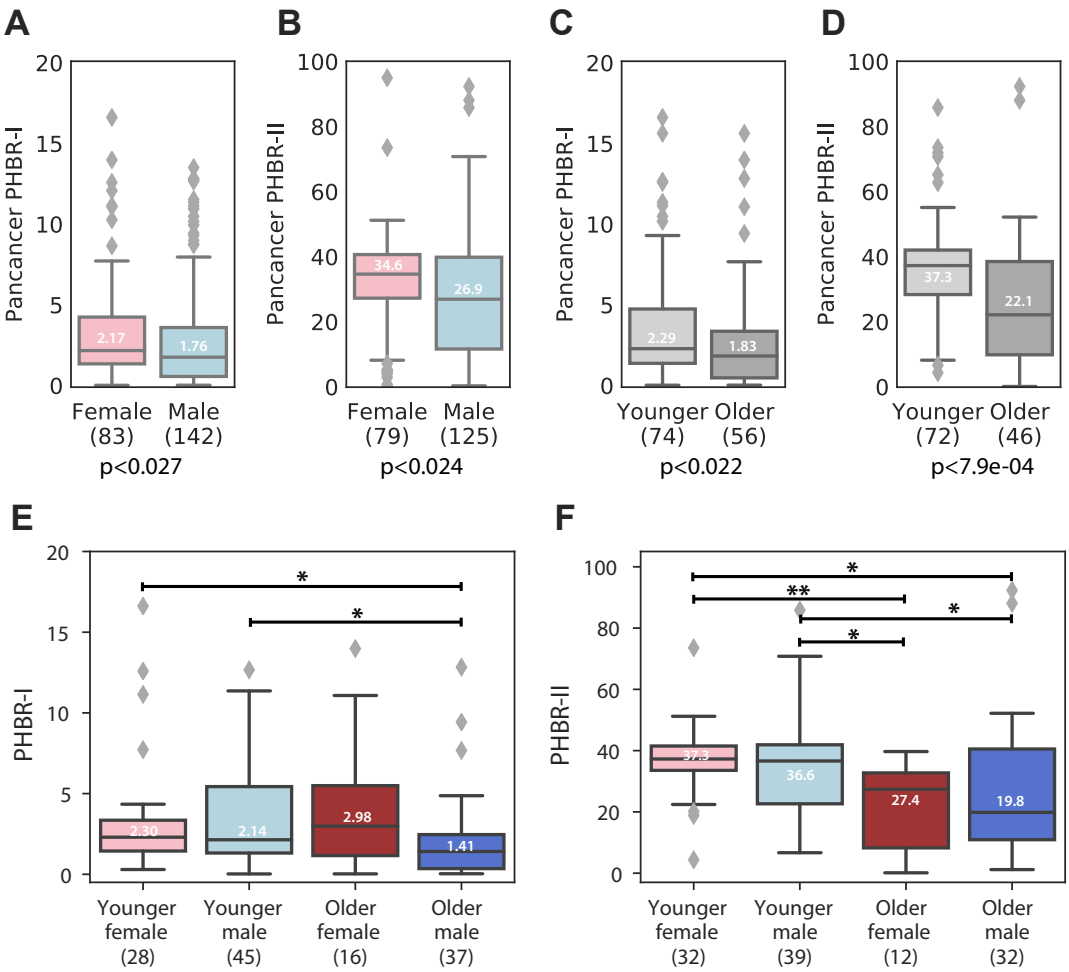


Figure 2.4. Sex- and age-specific MHC presentation of observed driver mutations in the validation cohort. Box plots denoting the distribution of (**A**) PHBR- I and (**B**) PHBR-II scores for driver mutations in female and male pan-cancer patients. Exact p-values were calculated using a one-tailed Mann–Whitney U test: (A) 0.027 and (B) 0.024, and effect sizes were calculated using Cliff's d: (A) r = −0.154, (B) r = −0.164. Box plots denoting the distribution of (**C**) PHBR-I and (**D**) PHBR-II scores for driver mutations in younger and older pan-cancer patients. Exact p-values were calculated using a one-tailed Mann–Whitney U test: (C) 0.022 and (D) 7.9e−04, and effect sizes were calculated using Cliff's d: (C) r = −0.207, (D) −0.346. Box plots denoting the distribution of (**E**) PHBR-I and (**F**) PHBR-II scores for driver mutations among integrated sex- and age-specific pan-cancer patient cohorts. One asterisk indicates p-values < 0.05 and two asterisks indicate p values < 0.001. P-values were calculated using a one-tailed Mann–Whitney U test. The Benjamini–Hochberg method was used to adjust for multiple comparisons for (E, F). Median values are shown in each boxplot. Exact p-values for (E) include: YM, OM: 0.024; YF, OM: 0.028; OF, OM: 0.070; YF, OF: 0.56; YF, YM: 0.49; OF, YM: 0.50. Exact p values for (F) include: YF, OF: 0.0083; YF, OM: 0.013; OF, YM: 0.023; YM, OM: 0.045; YF, YM: 0.24; OF, OM: 0.34. Y = younger, O = older, F = female, M = male. All box plots include the median line, the box denotes the interquartile range (IQR), whiskers denote the rest of the data distribution and outliers are denoted by points greater than ±1.5 × IQR.

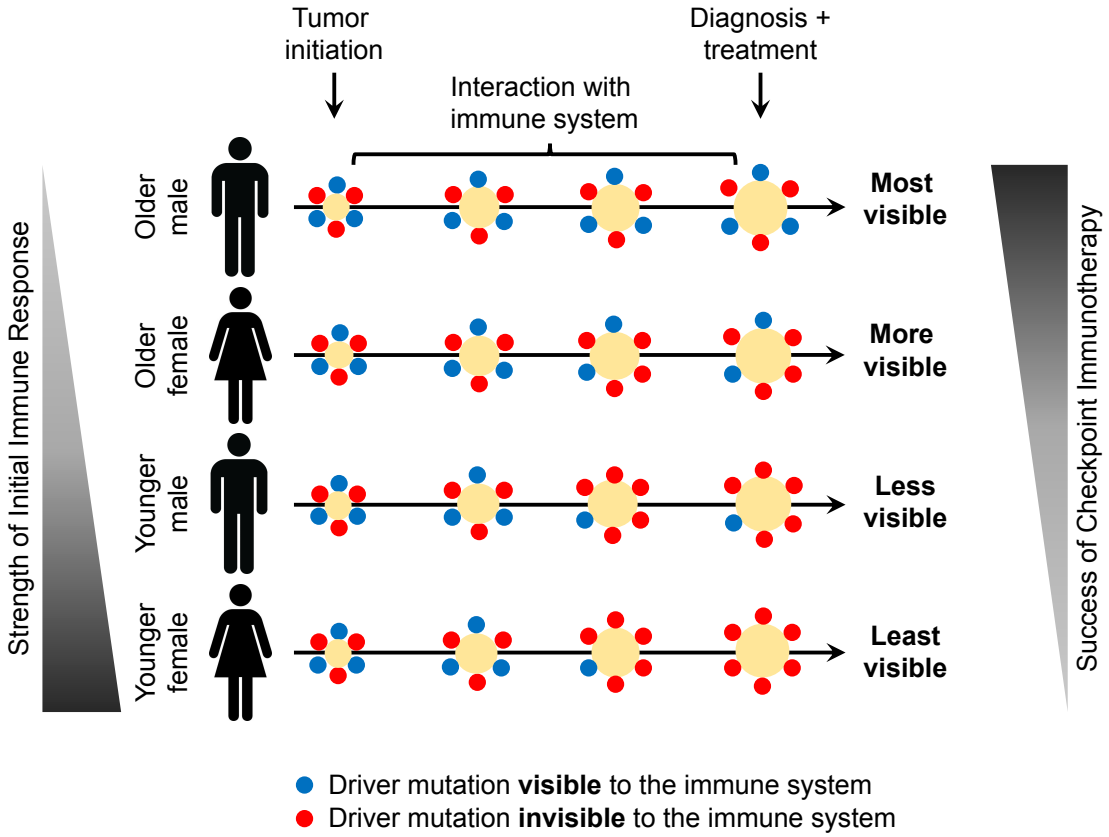


Figure 2.5. Proposed model of the relationship between immune selection and immunotherapy in cancer patients. Young females experience the strongest immune response, rendering their diagnosed tumors more invisible to the immune system and difficult to treat with ICB. On the other extreme, old males experience the weakest immune response, leaving their diagnosed tumors more visible to the immune system and open to attack when stimulated with ICB. Blue dots indicate immunologically visible driver mutations while red dots indicate immunologically invisible driver mutations at various time points.

2.8 Tables

Table 2.1. Quantitative estimate of the association between PHBR score and mutation occurrence in sex- and age-specific cohorts. Estimates and p-values are shown for a generalized additive model with random effects relating PHBR scores to the set of expressed driver mutations observed ≥2 times in this cohort. P-values were calculated via Wald tests using the Bayesian covariance matrix for the coefficients. Variables and their respective estimates and p-values have been bolded if significant ($p<0.05$).

	Parametric coefficients	**Estimate**	**Pr(>|z|)**
Sex analysis	**PHBR-I**	**0.048**	**0.0035**
	PHBR-II	**0.31**	**1.66e-56**
	Sex	-0.02	0.59
	PHBR-I:Sex	**0.07**	**0.02**
	PHBR-II:Sex	**0.15**	**0.00035**
Age analysis	**PHBR-I**	**0.043**	**0.0078**
	PHBR-II	**0.31**	**1.01e-54**
	Age	-0.0025	0.06
	PHBR-I:Age	**-0.0029**	**0.005**
	PHBR-II:Age	**-0.0035**	**0.007**

2.9 Supplemental Data, Tables and Figures

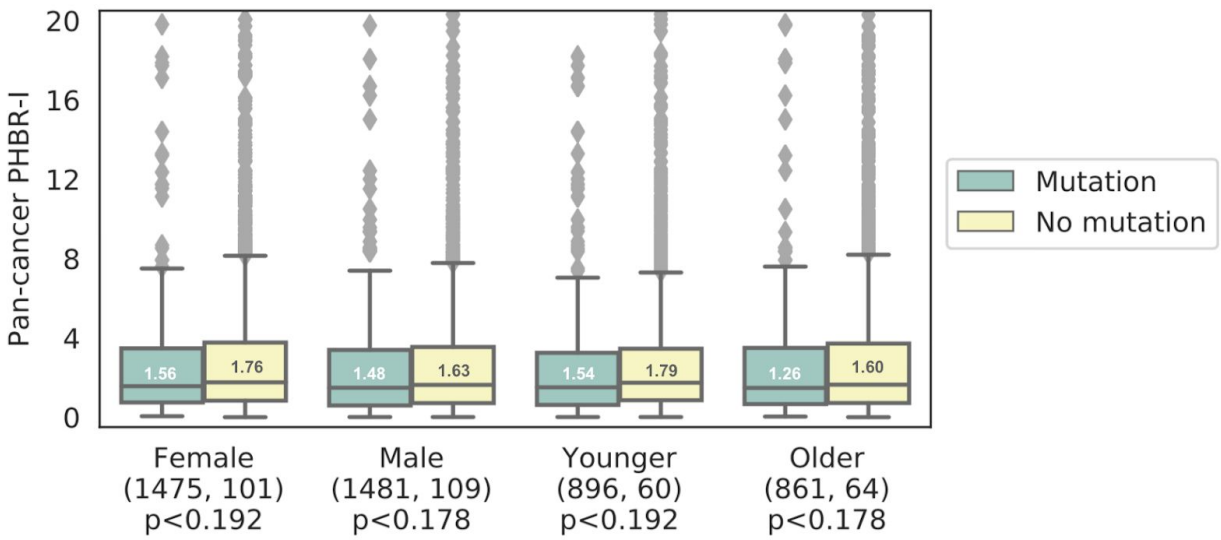


Figure S2.1. MHC-I mutation control analysis. Sex- and age-specific MHC presentation of common driver mutations for patients with and without MHC-I mutations. Box plots denote the distribution of PHBR-I scores for expressed driver mutations in female, male, younger, and older pan-cancer patients with and without MHC-I mutations. P-values were calculated using the one-tailed Mann Whitney U test. Exact p-values are: (A) 0.192, (B) 0.178, (C) 0.192, (D) 0.178. Median values are indicated in each boxplot. All boxplots include the median line, the box denotes the interquartile range (IQR), whiskers denote the rest of the data distribution and outliers are denoted by points greater than +/- 1.5 * IQR.

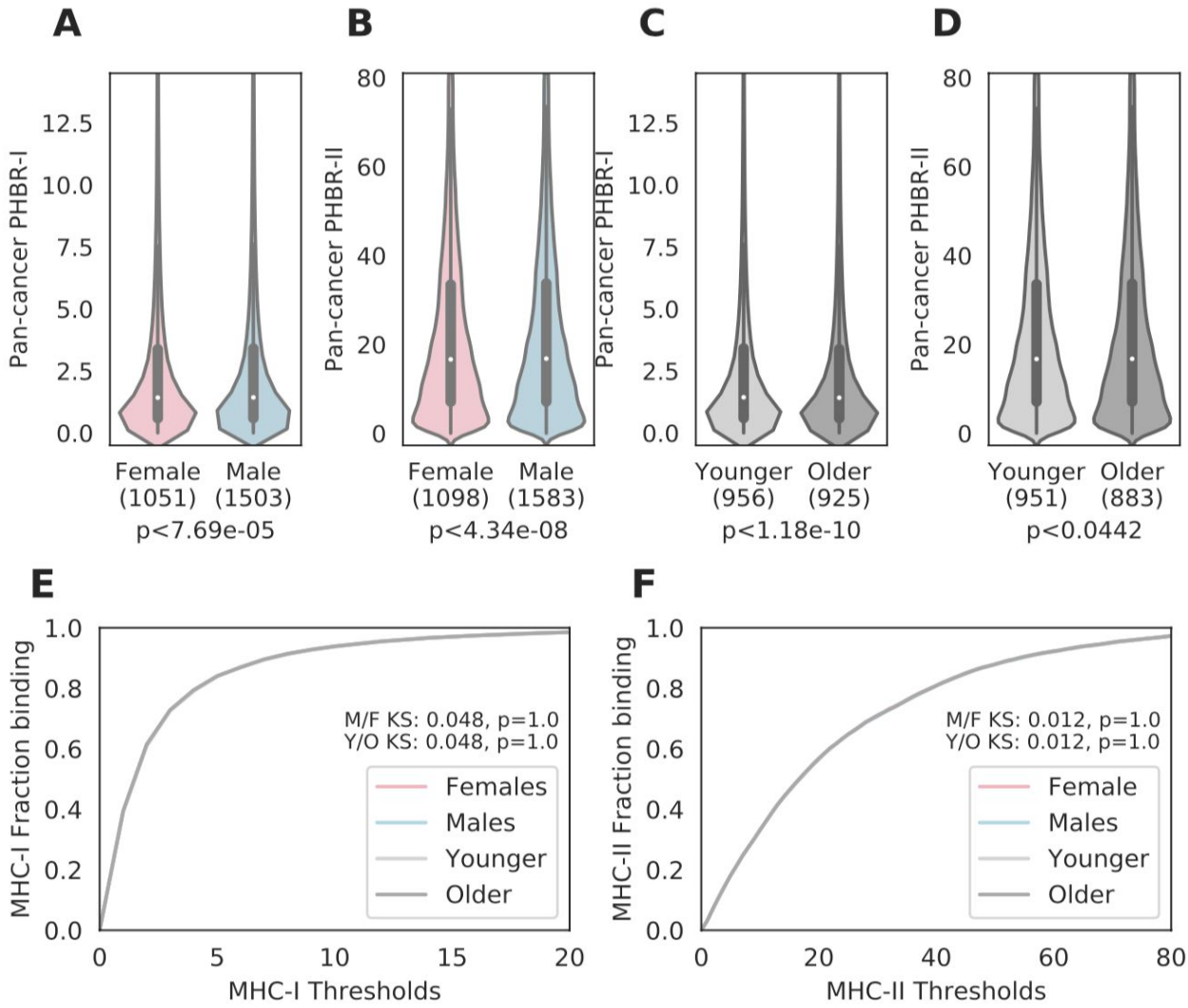


Figure S2.2. Sex- and age-specific MHC presentation of common driver mutations. (**A-D**) Violin plots denoting the sex and age stratified distribution of (**A, C**) PHBR-I and (**B, D**) PHBR-II scores across all common cancer driving mutations. P-values were calculated using the one-tailed Mann Whitney U test. Exact p-values are: (A) 7.69e-05, (B) 4.34e-08, (C) 1.18e-10, (D) 0.0442. Effect sizes were calculated using Cliff's d: (A) r=-0.00276, (B) r=-0.00381, (C) r=-0.00529, (D) r=-0.00144. Median PHBR scores are: (A) 1.42 F, 1.43 M, (B) 16.65 F, 16.81 M (C) 1.44 Y, 1.42 O (D) 16.72 Y, 16.75 O. (E, F) Empirical cumulative distribution functions showing the fraction of driver mutations predicted to bind to (E) MHC-I and (F) MHC-II at different PHBR score thresholds. Distributions were compared using the Kolmogorov-Smirnov two-sample test.

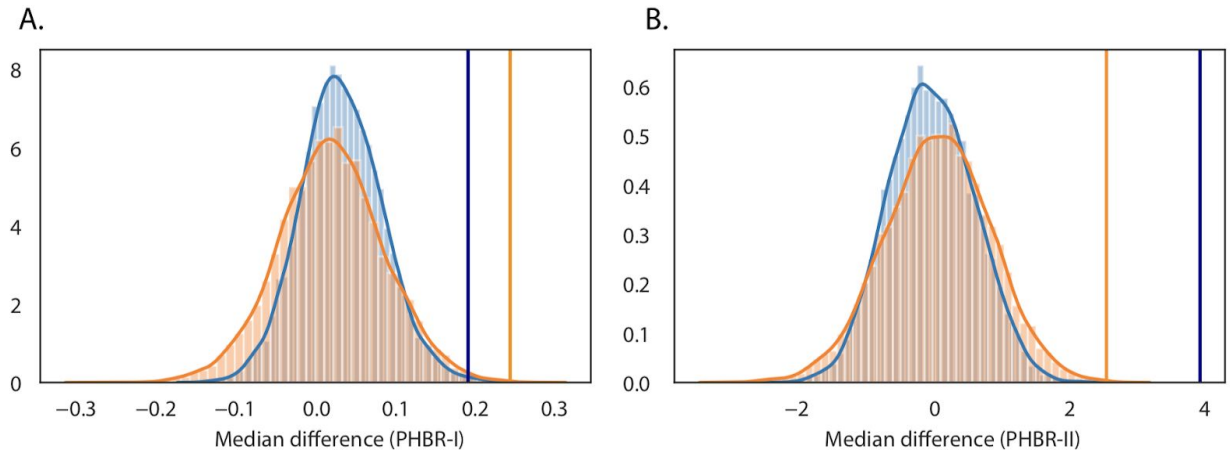


Figure S2.3. Shuffling driver mutations control analysis. Distributions of the median difference in (**A**) PHBR-I and (**B**) PHBR-II scores between sex-specific (blue) and age-specific (orange) patient groups estimated by shuffling driver mutations between patients 10,000 times while maintaining constant column and row counts. Solid lines indicate the observed score differences between cohorts.

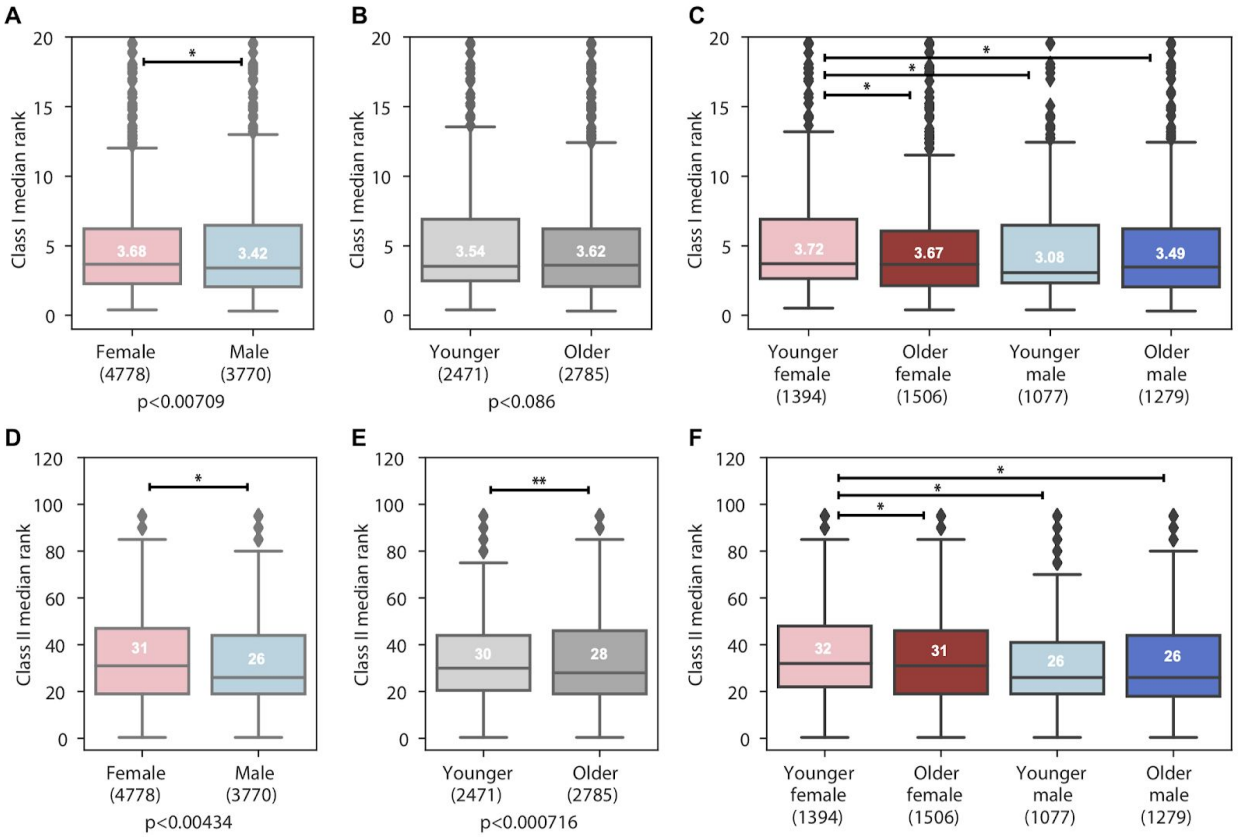


Figure S2.4. NetMHCpan alternate affinity analysis. Boxplots showing the distribution of median NetMHCpan (**A-C**) class I and (**D-F**) class II affinity scores for observed, expressed driver mutations across sex and age in our cohort. P-values were calculated using the one-tailed Mann Whitney U test. The Benjamini-Hochberg method was used to adjust for multiple comparisons for (C) and (F). Exact p-values are: (A) 0.00709, (B) 0.086, (C) YF, YM: 0.016; YF, OM: 0.019; YF, OF: 0.043; OF, YM: 0.27; OF, OM: 0.33; YM, OM: 0.42, (D) 0.00434, (E) 7.16e-04, (F) YF, OM: 4.64e-05; YF, YM: 0.0005; YF, OF: 0.0008; YM, OM: 0.24; OF, OM: 0.21; OF, YM: 0.43. Y=younger, O=older, F=female, M=male. One asterisk indicates p-values <0.05 and two asterisks indicates p-values <0.001. Median values are indicated in each boxplot. All boxplots include the median line, the box denotes the interquartile range (IQR), whiskers denote the rest of the data distribution and outliers are denoted by points greater than +/- 1.5 * IQR.

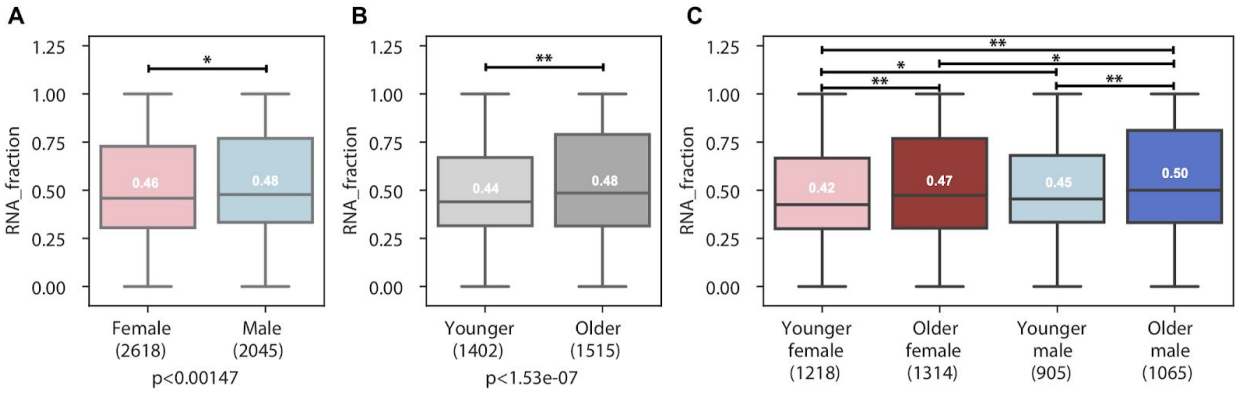


Figure S2.5. Sex- and age-specific analysis of mutation RNA fraction. Box plots showing the distribution of fraction of RNA reads supporting the mutated allele in (**A**) female and male patients, (**B**) younger and older patients, and (**C**) integrated sex- and age-specific patient cohorts. P-values were calculated using the one-tailed Mann Whitney U test. The Benjamini-Hochberg method was used to adjust for multiple comparisons for (C). Exact p-values are: (A) 0.000147, (B) 1,53e-07, (C) YF, OM: 1.19e-08; YF, OF: 0.0003; YM, OM: 0.0006; YF, YM: 0.008; OF, OM: 0.015; OF, YM: 0.11. Y=younger, O=older, F=female, M=male. One asterisk indicates p-values <0.05 and two asterisks indicates p-values <0.001. Median values are indicated in each boxplot. All boxplots include the median line, the box denotes the interquartile range (IQR), whiskers denote the rest of the data distribution and outliers are denoted by points greater than +/- 1.5 * IQR.

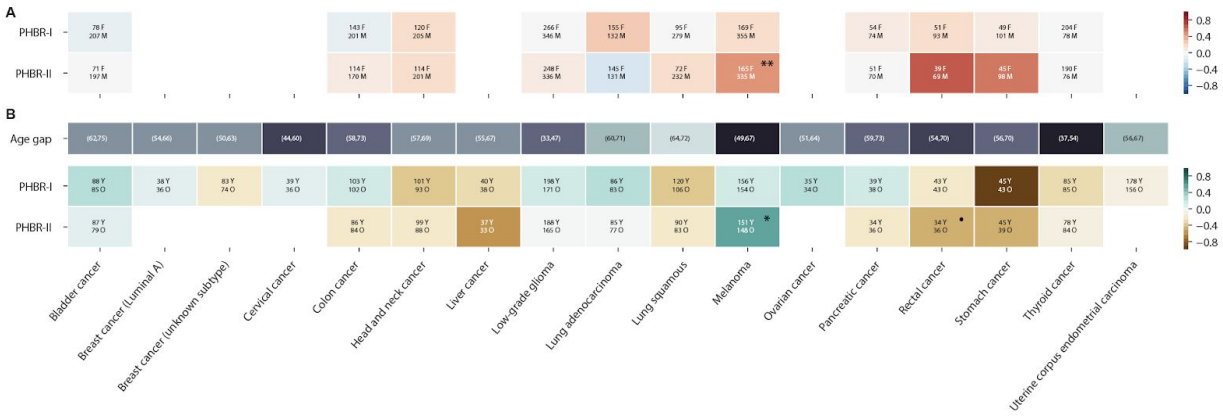


Figure S2.6. Disease-specific sex- and age-specific analysis of driver affinities. Heatmaps showing disease-specific PHBR-I and PHBR-II median ratios for (**A**) females vs. males where red coloring indicates higher median female PHBR distributions, and blue coloring indicates higher median male PHBR distributions. (**B**) Younger vs. older patients where green coloring indicates higher median younger PHBR distributions and yellow indicates higher median older PHBR distributions. "Age gap" shows disease-specific age thresholds (30th and 70th percentile), with darker coloring indicating a wider age gap and vice versa. A minimum of 30 patients in each group was required; blanks indicate that fewer than 30 samples were available in one of the categories. P-values were calculated using the one-tailed Mann Whitney U test. The Benjamini-Hochberg method was used to adjust for multiple comparisons for (A) and (B). Exact p-values for significant comparisons are: (A) 0.00098 for PHBR-II in melanoma, (B) 0.04 for PHBR-II in melanoma, and 0.07 for PHBR-II in rectal cancer. One asterisk indicates p-values <0.05, two asterisks indicates p-values <0.001, and a dot indicates p-values <0.1.

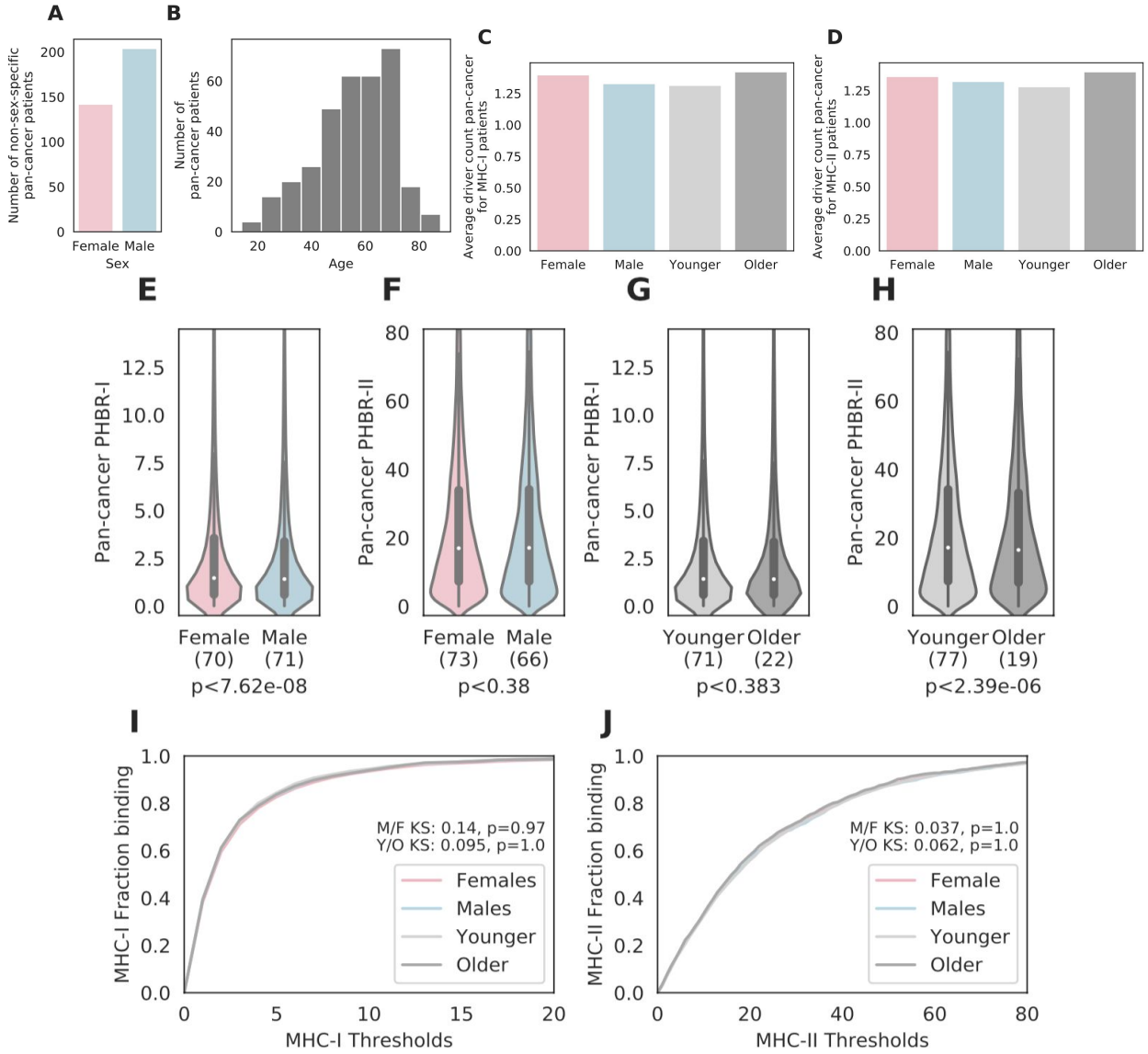


Figure S2.7. Overview of the validation cohort. (**A**) A bar plot comparing 346 male and female patients in the pan-cancer validation cohort. (**B**) A histogram denoting the distribution of ages when patients were diagnosed with cancer for 335 patients in the pan-cancer validation cohort. Bar plots denoting the average number of driver mutations for 346 patients in each sex- and age-specific cohort for (**C**) patients with MHC-I calls, and (**D**) patients with MHC-II calls. Median PHBR scores are: (A) 1.46 F, 1.41 M, (B) 16.97 F, 17.07 M (C) 1.41 Y, 1.40 O (D) 17.15 Y, 16.50 O. (E-H) Violin plots denoting the distribution of (E, G) PHBR-I and (F, H) PHBR-II scores across all common cancer driving mutations. P-values were calculated using the one-tailed Mann Whitney U test. Exact p-values are: (E) 7.62e-08, (F) 0.38, (G) 0.383, (H) 2.39e-06. Effect sizes were calculated using Cliff's d: (A) r=-0.016, (B) r=-0.000938, (C) r=-0.00132, (D) r=-0.0212. Empirical cumulative distribution functions showing the fraction of driver mutations predicted to bind to (**I**) MHC-I and (**J**) MHC-II at different PHBR score thresholds. Distributions were compared using the Kolmogorov-Smirnov two-sample test.

Table S2.1. Quantitative estimate of the association between PHBR score and mutation occurrence in sex- and age-specific TCGA cohorts. Estimates and p-values are shown for a generalized additive model with random effects relating PHBR-I and PHBR-II scores to the occurrence of driver mutations observed ≥2 times in the TCGA cohort. P-values were calculated via Wald tests using the Bayesian covariance matrix for the coefficients. Variables and their respective estimates and p-values have been bolded if significant ($p<0.05$).

	Parametric coefficients	Estimate	Pr(>|z|)
PHBR-I analysis	**PHBR-I**	**0.13**	**7.54e-22**
	Sex	-0.003	0.93
	Age	**-0.0026**	**0.02**
	PHBR-I:Sex	**0.08**	**0.003**
	PHBR-I:Age	**-0.0028**	**0.001**
PHBR-II analysis	**PHBR-II**	**0.33**	**2.05e-72**
	Sex	-0.05	0.22
	Age	-0.002	0.17
	PHBR-II:Sex	**0.16**	**2.73e-05**
	PHBR-II:Age	**-0.004**	**0.000518**

Table S2.2. Quantitative estimate of the association between PHBR score and mutation occurrence in sex- and age-specific TCGA cohorts, without tumor types significantly associated with sex-specific mutational signature ratios. Estimates and p-values are shown for a generalized additive model with random effects relating PHBR scores to occurence of driver mutations observed ≥2 times in the TCGA cohort. P-values were calculated via Wald tests using the Bayesian covariance matrix for the coefficients. Variables and their respective estimates and p-values have been bolded if significant ($p<0.05$).

	Parametric coefficients	Estimate	Pr(>|z|)
Sex analysis	**PHBR-I**	**0.06**	**0.002**
	PHBR-II	**0.33**	**1.9e-41**
	Sex	-0.03	0.54
	PHBR-I:Sex	0.075	0.07
	PHBR-II:Sex	**0.12**	**0.01**
Age analysis	**PHBR-I**	**0.06**	**0.004**
	PHBR-II	**0.33**	**2.2e-40**
	Age	-0.0021	0.17
	PHBR-I:Age	**-0.0044**	**0.0005**
	PHBR-II:Age	**-0.004**	**0.01**

Table S2.3. Quantitative estimate of the association between PHBR score and mutation occurrence in sex- and age-specific validation cohorts. Estimates and p-values are shown for a generalized additive model with random effects relating PHBR scores to occurence of driver mutations observed in the validation cohort. P-values were calculated via Wald tests using the Bayesian covariance matrix for the coefficients. Variables and their respective estimates and p-values have been bolded if significant ($p<0.05$).

	Parametric coefficients	Estimate	Pr(>|z|)
Sex analysis	PHBR-I	-0.17303	0.0576
	PHBR-II	**0.26725**	**0.0464**
	Sex	0.21348	0.2901
	PHBR-I:Sex	-0.09819	0.5917
	PHBR-II:Sex	0.30335	0.2647
Age analysis	PHBR-I	-0.153878	0.087
	PHBR-II	**0.287683**	**0.0302**
	Age	-0.009375	0.1771
	PHBR-I:Age	-0.007862	0.2371
	PHBR-II:Age	-0.009064	0.3329

2.10 Author Contributions

Original concept: Rachel M Pyke and Hannah Carter

Project supervisor: Maurizio Zanetti and Hannah Carter

Project planning and experimental design: Andrea Castro, Rachel M Pyke, Chi-Ping Day, Maurizio Zanetti, and Hannah Carter

Data acquisition, processing and analysis: Andrea Castro and Rachel M Pyke

Statistical advising: Xinlian Zhang and Wesley K Thompson

Mutational signature analysis: Ludmil Alexandrov

Manuscript writing: Andrea Castro, Rachel M Pyke, Maurizio Zanetti, and Hannah Carter

2.11 Acknowledgements

We would like to thank T. Cameron Waller, Tina Wang, and Trey Ideker for scientific discussion. This work was supported by an NIH National Library of Medicine Training Grant T15LM011271 to A.C., a NSF graduate fellowship #2015205295 to R.M.P., NIH grants DP5-OD017937, an Emerging Leader Award from The Mark Foundation for Cancer Research, grant #18-022-ELA and a CIFAR fellowship to H.C. and RO1 CA220009 to M.Z. and H.C., P41-GM103504 for computing resources provided by the National Resource for Network Biology (NRNB). We would like to thank the TCGA research network for providing data used in the analyses, the ICGC database, as well as the following studies used in the validation cohort. phs001493.v1.p1.c2 and phs001451.v1.p1.c1 We would also like to thank the Blavatnik Family Foundation, grants from the Broad Institute SPARC program, the National Institutes of Health (NCI-5R01CA155010-02, NHLBI-5R01HL103532-03, NCI-SPORE-2P50CA101942-11A1, NCI-R50-RCA211482A), the Francis and Adele Kittredge Family Immuno-Oncology and Melanoma Research Fund, the Faircloth Family Research Fund, and the DFCI Center for Cancer Immunotherapy Research fellowship and Leukemia and Lymphoma Society. phs001041.v1.p1.c1. We thank Martin Miller at Memorial Sloan Kettering Cancer Center (MSKCC) for his assistance with the NetMHC server, Agnes Viale and Kety Huberman at the MSKCC Genomics Core, Annamalai Selvakumar and Alice Yeh at the MSKCC HLA typing laboratory for their technical assistance, and John Khoury for assistance in chart review. phs001425.v1.p1.c1 Christine N. Spencer, Pei-Ling Chen, Michael T. Tetzlaff, Michael A. Davies, Jeffrey E. Gershenwald, Sapna P. Patel, Adi Diab, Isabella C. Glitza, Hussein Tawbi, Alexander J. Lazar, Patrick Hwu, Wen-Jen Hwu, Scott E. Woodman, Rodabe N. Amaria, Victor G. Prieto, and Jennifer A. Wargo enrolled subjects and contributed samples. phs001493.v1.p1.c1 This study was supported by an AACR

KureIt grant. phs000980.v1.p1.c1. We thank the members of the Thoracic Oncology Service and the Chan and Wolchok labs at MSKCC for helpful discussions, as well as the Immune Monitoring Core at MSKCC, including L. Caro, R. Ramsawak, and Z. Mu, for exceptional support with processing and banking peripheral blood lymphocytes. We thank P. Worrell and E. Brzostowski for help in identifying tumor specimens for analysis. We thank A. Viale for superb technical assistance. We thank D. Philips, M. van Buuren, and M. Toebes for help performing the combinatorial coding screens. This work was supported by the Geoffrey Beene Cancer Research Center (MDH, NAR, TAC, JDW, AS), the Society for Memorial Sloan Kettering Cancer Center (MDH), Lung Cancer Research Foundation (WL), Frederick Adler Chair Fund (TAC), The One Ball Matt Memorial Golf Tournament (EBG), Queen Wilhelmina Cancer Research Award (TNS), The STARR Foundation (TAC, JDW), the Ludwig Trust (JDW), and a Stand Up To Cancer-Cancer Research Institute Cancer Immunology Translational Cancer Research Grant (JDW, TNS, TAC). Stand Up To Cancer is a program of the Entertainment Industry Foundation administered by the American Association for Cancer Research. phs001469.v1.p1.c1. This work was supported by NIH grants R35CA197633, P01CA168585, 5P50CA168536, and GM08042. A comprehensive description of the dataset can be found at PMID:29320474. phs001519.v1.p1.c1. We thank the Ben and Catherine Ivy Foundation, the Blavatnik Family Foundation, the Broad Institute SPARC program, and NIH (NCI-1RO1CA155010-02 (to C.J.W.)), NHLBI-5R01HL103532-03 (to C.J.W.), Francis and Adele Kittredge Family Immuno-Oncology and Melanoma Research Fund (to P.A.O.), Faircloth Family Research Fund (to P.A.O.), NIH/ NCI R21 CA216772-01A1 (to D.B.K.), NCI-SPORE-2P50CA101942-11A1 (to D.B.K.); NHLBI-T32HL007627 (to J.B.I.); NCI (R50CA211482) (to S.A.S.), Zuckerman STEM Leadership Program (to I.T.); Benoziyo Endowment Fund for the Advancement of Science (to I.T.); P50 CA165962 (SPORE) and P01

CA163205 (to K.L.L.); DFCI Center for Cancer Immunotherapy Research fellowship (to Z.H.); Howard Hughes Medical Institute Medical Research Fellows Program (to A.J.A.); and American Cancer Society PF-17-042-01–LIB (to N.D.M.). C.J.W. is a scholar of the Leukemia and Lymphoma Society. We thank the Center for Neuro-Oncology, J. Russell and Dana-Farber Cancer Institute (DFCI) Center for Immuno-Oncology (CIO) staff; B. Meyers, C. Harvey and S. Bartel (Clinical Pharmacy); M. Severgnini, K. Kleinsteuber, and E. McWilliams, (CIO laboratory); M. Copersino (Regulatory Affairs); T. Bowman (DFHCC Specialized Histopathology Core Laboratory); A. Lako (CIO); M. Seaman and D. H. Barouch (BIDMC); the Broad Institute's Biological Samples, Genetic Analysis and Genome Sequencing Platforms; J. Petricciani and M. Krane for regulatory advice; B. McDonough (CSBio), I. Javeri and K. Nellaiappan (CuriRx) for peptide development. phs001565.v1.p1.c1 The research reported in this article was supported by BroadIgnite, BroadNext10, NIH K08CA188615, the Howard Hughes Medical Institute, and Stand Up To Cancer—American Cancer Society Lung Cancer Dream Team Translational Research Grant (grant number: SU2C-AACR-DT17-15). Stand Up To Cancer is a program of the Entertainment Industry Foundation. Research grants are administered by the American Association for Cancer Research, the scientific partner of SU2C.

Chapter 2, in full, is a reformatted reprint of the material as it appears as "Strength of immune selection in tumors varies with sex and age" in *Nature Communications*, 2020 by Andrea Castro, Rachel Marty Pyke, Xinlian Zhang, Wesley Kurt Thompson, Chi-Ping Day, Ludmil B. Alexandrov, Maurizio Zanetti and Hannah Carter. The dissertation author was a primary investigator and author of this paper.

2.12 References

1. S. L. Klein, K. L. Flanagan, Sex differences in immune responses. *Nat. Rev. Immunol.* **16**, 626–638 (2016).

2. J. A. Fargallo, J. Martínez-Padilla, A. Toledano-Díaz, J. Santiago-Moreno, J. A. Dávila, Sex and testosterone effects on growth, immunity and melanin coloration of nestling Eurasian kestrels. *J. Anim. Ecol.* **76**, 201–209 (2007).

3. P. L. Pap, G. A. Czirják, C. I. Vágási, Z. Barta, D. Hasselquist, Sexual dimorphism in immune function changes during the annual cycle in house sparrows. *Naturwissenschaften*. **97**, 891–901 (2010).

4. S. Mondal, U. Rai, Sexual Dimorphism in Phagocytic Activity of Wall Lizard's Splenic Macrophages and Its Control by Sex Steroids. *Gen. Comp. Endocrinol.* **116**, 291–298 (1999).

5. E. Montecino-Rodriguez, B. Berent-Maoz, K. Dorshkind, Causes, consequences, and reversal of immune system aging. *J. Clin. Invest.* **123**, 958–965 (2013).

6. R. Voskuhl, Sex differences in autoimmune diseases. *Biol. Sex Differ.* **2**, 1 (2011).

7. D. L. Jacobson, S. J. Gange, N. R. Rose, N. M. Graham, Epidemiology and estimated population burden of selected autoimmune diseases in the United States. *Clin. Immunol. Immunopathol.* **84**, 223–243 (1997).

8. B. Stamova, Y. Tian, G. Jickling, C. Bushnell, X. Zhan, D. Liu, B. P. Ander, P. Verro, V. Patel, W. C. Pevec, N. Hedayati, D. L. Dawson, E. C. Jauch, A. Pancioli, J. P. Broderick, F. R. Sharp, The X-chromosome has a different pattern of gene expression in women compared with men with ischemic stroke. *Stroke*. **43**, 326–334 (2012).

9. A. Cáceres, A. Jene, T. Esko, L. A. Pérez-Jurado, J. R. González, Extreme Downregulation of Chromosome Y and Cancer Risk in Men. *J. Natl. Cancer Inst.* **112**, 913–920 (2020).

10. A. Dunford, D. M. Weinstock, V. Savova, S. E. Schumacher, J. P. Cleary, A. Yoda, T. J. Sullivan, J. M. Hess, A. A. Gimelbrant, R. Beroukhim, M. S. Lawrence, G. Getz, A. A. Lane, Tumor-suppressor genes that escape from X-inactivation contribute to cancer sex bias. *Nat. Genet.* **49**, 10–16 (2017).

11. L. G. vom Steeg, S. L. Klein, SeXX Matters in Infectious Disease Pathogenesis. *PLoS Pathog.* **12**, e1005374 (2016).

12. S. L. Klein, C. Roberts, Eds., *Sex Hormones and Immunity to Infection* (Springer, Berlin, Heidelberg, 2010).

13. D. I. Podolskiy, A. V. Lobanov, G. V. Kryukov, V. N. Gladyshev, Analysis of cancer genomes reveals basic features of human aging and its role in cancer development. *Nat. Commun.* **7**, 12157 (2016).

14. A. J. Levine, N. A. Jenkins, N. G. Copeland, The Roles of Initiating Truncal Mutations in Human Cancers: The Order of Mutations and Tumor Cell Type Matters. *Cancer Cell*. **35**, 10–15 (2019).

15. S. Bagchi, R. Yuan, E. G. Engleman, Immune Checkpoint Inhibitors for the Treatment of Cancer: Clinical Impact and Mechanisms of Response and Resistance. *Annu. Rev. Pathol.* **16**, 223–249 (2021).

16. F. Conforti, L. Pala, V. Bagnardi, T. De Pas, M. Martinetti, G. Viale, R. D. Gelber, A. Goldhirsch, Cancer immunotherapy efficacy and patients' sex: a systematic review and meta-analysis. *Lancet Oncol.* **19**, 737–746 (2018).

17. T. Zhang, H. Jia, Z. Wu, Sex as a predictor of response to cancer immunotherapy. *Lancet Oncol.* **19** (2018), p. e374.

18. Y. Ye, Y. Jing, L. Li, G. B. Mills, L. Diao, H. Liu, L. Han, Sex-associated molecular differences for cancer immunotherapy. *Nat. Commun.* **11**, 1779 (2020).

19. D. R. Crawford, S. Sinha, N. U. Nair, B. M. Ryan, J. S. Barnholtz-Sloan, S. M. Mount, A. Erez, K. Aldape, P. E. Castle, P. S. Rajagopal, C.-P. Day, A. A. Schäffer, E. Ruppin, Sex biases in cancer and autoimmune disease incidence are strongly positively correlated with mitochondrial gene expression across human tissues. *bioRxiv* (2022), p. 2021.09.07.459207.

20. F. M. Burnet, The concept of immunological surveillance. *Prog. Exp. Tumor Res.* **13**, 1–27 (1970).

21. R. D. Schreiber, L. J. Old, M. J. Smyth, Cancer immunoediting: integrating immunity's roles in cancer suppression and promotion. *Science*. **331**, 1565–1570 (2011).

22. A. Ribas, J. D. Wolchok, Cancer immunotherapy using checkpoint blockade. *Science*. **359**, 1350–1355 (2018).

23. A. Nosrati, K. K. Tsai, S. M. Goldinger, P. Tumeh, B. Grimes, K. Loo, A. P. Algazi, T. D. L. Nguyen-Kim, M. Levesque, R. Dummer, O. Hamid, A. Daud, Evaluation of clinicopathological factors in PD-1 response: derivation and validation of a prediction scale for response to PD-1 monotherapy. *Br. J. Cancer*. **116**, 1141–1147 (2017).

24. Y. Wu, Q. Ju, K. Jia, J. Yu, H. Shi, H. Wu, M. Jiang, Correlation between sex and efficacy of immune checkpoint inhibitors (PD-1 and CTLA-4 inhibitors). *Int. J. Cancer* (2018), doi:10.1002/ijc.31301.

25. A. Botticelli, C. E. Onesti, I. Zizzari, B. Cerbelli, P. Sciattella, M. Occhipinti, M. Roberto, F. Di Pietro, A. Bonifacino, M. Ghidini, P. Vici, L. Pizzuti, C. Napoletano, L. Strigari, G. D'Amati, F. Mazzuca, M. Nuti, P. Marchetti, The sexist behaviour of immune checkpoint inhibitors in cancer therapy? *Oncotarget*. **8**, 99336–99346 (2017).

26. C. H. Kugel 3rd, S. M. Douglass, M. R. Webster, A. Kaur, Q. Liu, X. Yin, S. A. Weiss, F. Darvishian, R. N. Al-Rohil, A. Ndoye, R. Behera, G. M. Alicea, B. L. Ecker, M. Fane, M. J.

Allegrezza, N. Svoronos, V. Kumar, D. Y. Wang, R. Somasundaram, S. Hu-Lieskovan, A. Ozgun, M. Herlyn, J. R. Conejo-Garcia, D. Gabrilovich, E. L. Stone, T. S. Nowicki, J. Sosman, R. Rai, M. S. Carlino, G. V. Long, R. Marais, A. Ribas, Z. Eroglu, M. A. Davies, B. Schilling, D. Schadendorf, W. Xu, R. K. Amaravadi, A. M. Menzies, J. L. McQuade, D. B. Johnson, I. Osman, A. T. Weeraratna, Age Correlates with Response to Anti-PD1, Reflecting Age-Related Differences in Intratumoral Effector and Regulatory T-Cell Populations. *Clin. Cancer Res.* (2018), doi:10.1158/1078-0432.CCR-18-1116.

27. R. J. M. Engler, Half- vs Full-Dose Trivalent Inactivated Influenza Vaccine (2004-2005). *Arch. Intern. Med.* **168**, 2405 (2008).

28. M. Abdullah, P.-S. Chai, M.-Y. Chong, E. R. M. Tohit, R. Ramasamy, C. P. Pei, S. Vidyadaran, Gender effect on in vitro lymphocyte subset levels of healthy individuals. *Cell. Immunol.* **272**, 214–219 (2012).

29. T. Schneider-Hohendorf, D. Görlich, P. Savola, T. Kelkka, S. Mustjoki, C. C. Gross, G. C. Owens, L. Klotz, K. Dornmair, H. Wiendl, N. Schwab, Sex bias in MHC I-associated shaping of the adaptive immune system. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 2168–2173 (2018).

30. E. M. Hill-Burns, A. G. Clark, X-Linked Variation in Immune Response in Drosophila melanogaster. *Genetics*. **183**, 1477–1491 (2009).

31. I. Milo, M. Bedora-Faure, Z. Garcia, R. Thibaut, L. Périé, G. Shakhar, L. Deriano, P. Bousso, The immune system profoundly restricts intratumor genetic heterogeneity. *Sci Immunol*. **3** (2018), doi:10.1126/sciimmunol.aat1435.

32. A. K. Simon, G. A. Hollander, A. McMichael, Evolution of the immune system in humans from infancy to old age. *Proc. Biol. Sci.* **282**, 20143085 (2015).

33. N. Jiang, J. He, J. A. Weinstein, L. Penland, S. Sasaki, X.-S. He, C. L. Dekker, N.-Y. Zheng, M. Huang, M. Sullivan, P. C. Wilson, H. B. Greenberg, M. M. Davis, D. S. Fisher, S. R. Quake, Lineage structure of the human antibody repertoire in response to influenza vaccination. *Sci. Transl. Med.* **5**, 171ra19 (2013).

34. A. Agrawal, S. Agrawal, S. Gupta, Dendritic cells in human aging. *Exp. Gerontol.* **42**, 421–426 (2007).

35. R. Marty, S. Kaabinejadian, D. Rossell, M. J. Slifker, J. van de Haar, H. B. Engin, N. de Prisco, T. Ideker, W. H. Hildebrand, J. Font-Burgada, H. Carter, MHC-I Genotype Restricts the Oncogenic Mutational Landscape. *Cell*. **171**, 1272–1283.e15 (2017).

36. R. Marty, W. K. Thompson, R. M. Salem, M. Zanetti, H. Carter, Evolutionary Pressure against MHC Class II Binding Cancer Mutations. *Cell* (2018), doi:10.1016/j.cell.2018.08.048.

37. M. Nielsen, M. Andreatta, NetMHCpan-3.0; improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets. *Genome Med.* **8**, 33 (2016).

38. E. Karosiene, M. Rasmussen, T. Blicher, O. Lund, S. Buus, M. Nielsen, NetMHCIIpan-3.0, a common pan-specific MHC class II prediction method including all three human MHC class II isotypes, HLA-DR, HLA-DP and HLA-DQ. *Immunogenetics*. **65**, 711–724 (2013).

39. G. P. Dunn, A. T. Bruce, H. Ikeda, L. J. Old, R. D. Schreiber, Cancer immunoediting: from immunosurveillance to tumor escape. *Nat. Immunol.* **3**, 991–998 (2002).

40. S. A. Shukla, M. S. Rooney, M. Rajasagi, G. Tiao, P. M. Dixon, M. S. Lawrence, J. Stevens, W. J. Lane, J. L. Dellagatta, S. Steelman, C. Sougnez, K. Cibulskis, A. Kiezun, N. Hacohen, V. Brusic, C. J. Wu, G. Getz, Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes. *Nat. Biotechnol.* **33**, 1152–1158 (2015).

41. S. Kawaguchi, K. Higasa, M. Shimizu, R. Yamada, F. Matsuda, HLA-HD: An accurate HLA typing algorithm for next-generation sequencing data. *Hum. Mutat.* **38**, 788–797 (2017).

42. V. Jurtz, S. Paul, M. Andreatta, P. Marcatili, B. Peters, M. Nielsen, NetMHCpan-4.0: Improved Peptide-MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. *J. Immunol.* **199**, 3360–3368 (2017).

43. K. K. Jensen, M. Andreatta, P. Marcatili, S. Buus, J. A. Greenbaum, Z. Yan, A. Sette, B. Peters, M. Nielsen, Improved methods for predicting peptide binding affinity to MHC class II molecules. *Immunology*. **154**, 394–406 (2018).

44. W. C. Wong, D. Kim, H. Carter, M. Diekhans, M. C. Ryan, R. Karchin, CHASM and SNVBox: toolkit for detecting biologically important single nucleotide mutations in cancer. *Bioinformatics*. **27**, 2147–2148 (2011).

45. Cancer Genome Atlas Research Network, Integrated genomic characterization of papillary thyroid carcinoma. *Cell*. **159**, 676–690 (2014).

46. L. B. Alexandrov, P. H. Jones, D. C. Wedge, J. E. Sale, P. J. Campbell, S. Nik-Zainal, M. R. Stratton, Clock-like mutational processes in human somatic cells. *Nat. Genet.* **47**, 1402–1407 (2015).

47. L. B. Alexandrov, S. Nik-Zainal, D. C. Wedge, S. A. J. R. Aparicio, S. Behjati, A. V. Biankin, G. R. Bignell, N. Bolli, A. Borg, A.-L. Børresen-Dale, S. Boyault, B. Burkhardt, A. P. Butler, C. Caldas, H. R. Davies, C. Desmedt, R. Eils, J. E. Eyfjörd, J. A. Foekens, M. Greaves, F. Hosoda, B. Hutter, T. Ilicic, S. Imbeaud, M. Imielinski, N. Jäger, D. T. W. Jones, D. Jones, S. Knappskog, M. Kool, S. R. Lakhani, C. López-Otín, S. Martin, N. C. Munshi, H. Nakamura, P. A. Northcott, M. Pajic, E. Papaemmanuil, A. Paradiso, J. V. Pearson, X. S. Puente, K. Raine, M. Ramakrishna, A. L. Richardson, J. Richter, P. Rosenstiel, M. Schlesner, T. N. Schumacher, P. N. Span, J. W. Teague, Y. Totoki, A. N. J. Tutt, R. Valdés-Mas, M. M. van Buuren, L. van 't Veer, A. Vincent-Salomon, N. Waddell, L. R. Yates, Australian Pancreatic Cancer Genome Initiative, ICGC Breast Cancer Consortium, ICGC MMML-Seq Consortium, ICGC PedBrain, J. Zucman-Rossi, P. A. Futreal, U. McDermott, P. Lichter, M. Meyerson, S. M. Grimmond, R. Siebert, E. Campo, T. Shibata, S. M. Pfister, P. J. Campbell, M. R. Stratton, Signatures of mutational processes in human cancer. *Nature*. **500**, 415–421 (2013).

48. J. Zhang, J. Baran, A. Cros, J. M. Guberman, S. Haider, J. Hsu, Y. Liang, E. Rivkin, J. Wang, B. Whitty, M. Wong-Erasmus, L. Yao, A. Kasprzyk, International Cancer Genome Consortium Data Portal--a one-stop shop for cancer genomics data. *Database*. **2011** (2011), pp. bar026–bar026.

49. A. Amadori, R. Zamarchi, G. De Silvestro, G. Forza, G. Cavatton, G. A. Danieli, M. Clementi, L. Chieco-Bianchi, Genetic control of the CD4/CD8 T-cell ratio in humans. *Nat. Med.* **1**, 1279–1283 (1995).

50. J. A. Keene, J. Forman, Helper activity is required for the in vivo generation of cytotoxic T lymphocytes. *J. Exp. Med.* **155**, 768–782 (1982).

51. M. Gerloni, S. Xiong, S. Mukerjee, S. P. Schoenberger, M. Croft, M. Zanetti, Functional cooperation between T helper cell determinants. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 13269–13274 (2000).

52. J. J. Goronzy, F. Fang, M. M. Cavanagh, Q. Qi, C. M. Weyand, Naive T cell maintenance and function in human aging. *J. Immunol.* **194**, 4073–4080 (2015).

53. N. H. Son, S. Murray, J. Yanovski, R. J. Hodes, N. Weng, Lineage-specific telomere shortening and unaltered capacity for telomerase expression in human T and B lymphocytes with age. *J. Immunol.* **165**, 1191–1196 (2000).

54. R. Marty Pyke, W. K. Thompson, R. M. Salem, J. Font-Burgada, M. Zanetti, H. Carter, Evolutionary Pressure against MHC Class II Binding Cancer Mutations. *Cell*. **175**, 1991 (2018).

55. N. McGranahan, A. J. S. Furness, R. Rosenthal, S. Ramskov, R. Lyngaa, S. K. Saini, M. Jamal-Hanjani, G. A. Wilson, N. J. Birkbak, C. T. Hiley, T. B. K. Watkins, S. Shafi, N. Murugaesu, R. Mitter, A. U. Akarca, J. Linares, T. Marafioti, J. Y. Henry, E. M. Van Allen, D. Miao, B. Schilling, D. Schadendorf, L. A. Garraway, V. Makarov, N. A. Rizvi, A. Snyder, M. D. Hellmann, T. Merghoub, J. D. Wolchok, S. A. Shukla, C. J. Wu, K. S. Peggs, T. A. Chan, S. R. Hadrup, S. A. Quezada, C. Swanton, Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. *Science*. **351**, 1463–1469 (2016).

56. S. N. Gröbner, B. C. Worst, J. Weischenfeldt, I. Buchhalter, K. Kleinheinz, V. A. Rudneva, P. D. Johann, G. P. Balasubramanian, M. Segura-Wang, S. Brabetz, S. Bender, B. Hutter, D. Sturm, E. Pfaff, D. Hübschmann, G. Zipprich, M. Heinold, J. Eils, C. Lawerenz, S. Erkek, S. Lambo, S. Waszak, C. Blattmann, A. Borkhardt, M. Kuhlen, A. Eggert, S. Fulda, M. Gessler, J. Wegert, R. Kappler, D. Baumhoer, S. Burdach, R. Kirschner-Schwabe, U. Kontny, A. E. Kulozik, D. Lohmann, S. Hettmer, C. Eckert, S. Bielack, M. Nathrath, C. Niemeyer, G. H. Richter, J. Schulte, R. Siebert, F. Westermann, J. J. Molenaar, G. Vassal, H. Witt, ICGC PedBrain-Seq Project, ICGC MMML-Seq Project, B. Burkhardt, C. P. Kratz, O. Witt, C. M. van Tilburg, C. M. Kramm, G. Fleischhack, U. Dirksen, S. Rutkowski, M. Frühwald, K. von Hoff, S. Wolf, T. Klingebiel, E. Koscielniak, P. Landgraf, J. Koster, A. C. Resnick, J. Zhang, Y. Liu, X. Zhou, A. J. Waanders, D. A. Zwijnenburg, P. Raman, B. Brors, U. D. Weber, P. A. Northcott, K. W. Pajtler, M. Kool, R. M. Piro, J. O. Korbel, M. Schlesner, R. Eils, D. T.

W. Jones, P. Lichter, L. Chavez, M. Zapatka, S. M. Pfister, The landscape of genomic alterations across childhood cancers. *Nature*. **555**, 321–327 (2018).

57. R. Leinonen, H. Sugawara, M. Shumway, I. N. S. D. Collaboration, The sequence read archive. *Nucleic Acids Res.* **39**, D19–D21 (2010).

58. C. Xie, Z. X. Yeo, M. Wong, J. Piper, T. Long, E. F. Kirkness, W. H. Biggs, K. Bloom, S. Spellman, C. Vierra-Green, C. Brady, R. H. Scheuermann, A. Telenti, S. Howard, S. Brewerton, Y. Turpaz, J. C. Venter, Fast and accurate HLA typing from short-read next-generation sequence data with xHLA. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 8059–8064 (2017).

59. T. Davoli, A. W. Xu, K. E. Mengwasser, L. M. Sack, J. C. Yoon, P. J. Park, S. J. Elledge, Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. *Cell*. **155**, 948–962 (2013).

60. S. N. Wood, mgcv: GAMs and generalized ridge regression for R. *R news*. **1**, 20–25 (2001).

CHAPTER 3: Subcellular location is a novel feature that improves immunogenicity prediction

3.1 Foreword

Over the past decade, there has been surge of interest in researching and exploiting neoantigens as targets. Thanks to expanded availability of datasets and progress in machine learning, *in silico* prediction of MHC presentation has improved greatly (*1–4*). However, cell surface presentation of antigen alone does not guarantee T cell recognition and response, also known as immunogenicity. Other factors such as peptide-MHC stability, affinity relative to the wildtype peptide, dissimilarity to the proteome, and presence of certain amino acids have all been correlated with immunogenicity, but often fail to validate experimentally (*5, 6*). Among other factors, these attempt to quantify the likelihood that 1) the mutant peptide can bind to MHC (affinity), 2) the peptide-MHC complex is stable enough to remain on the cell surface for some time (stability), and 3) T cells able to recognize this peptide-MHC complex have not been eliminated during negative selection in the thymus (agretopicity and foreignness); all of which are required in some degree to enable T cell recognition and response of neopeptide. Currently studied factors focus on neopeptide characteristics that are more relevant in the later portion of the antigen presentation pathway; after proteosomal cleavage and TAP transport into the endoplasmic reticulum. Motivated by initial experimental studies that identified a location bias for eluted MHC-presented peptides, I hypothesized that location could affect ability to even enter the antigen presentation pathway. Aim 3 is the first investigation of its kind to assess the extent to which protein subcellular location can help improve immunogenicity prediction. Even now, this remains

a difficult challenge, one that requires significant strides in improvement in order to realize consistent, effective immunotherapy responses.

3.2 Abstract

Antigen presentation via the major histocompatibility complex (MHC) is essential for anti-tumor immunity, however the rules that determine what tumor-derived peptides will be immunogenic are still incompletely understood. Here we investigate whether constraints on peptide accessibility to the MHC due to protein subcellular location are associated with potential for peptide immunogenicity. Analyzing over 380,000 peptides from studies of MHC presentation and peptide immunogenicity, we find clear spatial biases in both eluted and immunogenic peptides. We find that including parent protein location improves prediction of peptide immunogenicity in multiple datasets. In human immunotherapy cohorts, location was associated with response to a neoantigen vaccine, and immune checkpoint blockade responders generally had a higher burden of neopeptides from accessible locations. We conclude that protein subcellular location adds important information for optimizing immunotherapies.

3.3 Introduction

Accurate prediction of immunogenic neopeptides is crucial for the effective application of neoantigen-based cancer treatments such as neoantigen vaccines, immune checkpoint blockade, and adoptive T cell therapy. These immunotherapies all depend on cell surface display of tumor-derived peptides by the major histocompatibility complex (MHC) for immune surveillance by T cells. Current approaches to predict neopeptide immunogenicity (i.e. which tumor mutations will result in neopeptides that are both displayed and recognized by T cells as foreign) largely focus on peptide-MHC affinity, with non-standardized, and sometimes controversial, incorporation of other peptide characteristics such as peptide-MHC stability, agretopicity (*7*), foreignness (*8*), hydrophobicity, mutation position within the neopeptide (*9*), and neopeptide RNA abundance (*6*). However, the utility of these features for predicting immunogenicity varies across experiments and cohorts and ultimately, current tools still yield many false positive neoantigen predictions (*5*, *10*). Therefore, it follows that there must be other factors that contribute to T cell recognition of peptide-bound MHC that remain to be accounted for.

The canonical pathways by which peptides are added to the MHC for binding differ for class I and class II, with class I peptides requiring transport to the endoplasmic reticulum via TAP transporters (*11*) and class II peptides arriving via endosomes generated through phagocytosis or B cell receptor internalization by antigen presenting cells (*12*). It stands to reason that these distinct pathways could result in peptides from different proteins being more accessible to MHC class I versus class II molecules. Indeed, studies profiling eluted peptide-MHC complexes noted enrichment for peptide origin from intracellular compartments for MHC-I eluted peptides, and for secreted, cell membrane, and extracellular proteins for MHC-II eluted peptides (*13–16*). Interestingly, a more recent work found a bias for MHC-I presented peptides to come from proteins

with certain molecular functions such as intracellular structural proteins while MHC-II was biased to present membrane transport proteins (*17*).

These observations imply that proteins in different cellular contexts (location or molecular function, which are correlated (*18*) can have varying levels of access to the MHC-I or MHC-II presentation pathway. This could further constrain the landscape of peptides that are presented to T cells during cancer development and during T cell selection in the thymus. In the first case, a tumor mutation that would otherwise be effectively bound and displayed by the MHC may not be presented because peptides from the source protein never reach the MHC. For the second, effective presentation of self-antigen during thymic development is required for clonal deletion of corresponding T cells. This pruning of the T cell repertoire is essential to prevent inappropriate activity against self; mutations to the AIRE gene that promotes tissue-specific self-antigen expression during selection result in widespread, multi-organ autoimmunity (*19*). However, if the peptide repertoire available for T cell selection is constrained by cellular context, it is conceivable that the remaining T cell repertoire would be more capable of mounting a response against peptides from those hidden cellular contexts.

Based on this reasoning, we hypothesized that subcellular location of a source protein could influence the immunogenic potential of its derivative peptides. We collected and analyzed data on peptides presented by the MHC and peptides documented to generate immune responses. To determine the extent to which cellular location can predict neopeptide immunogenicity, we trained and evaluated machine learning models on datasets of experimentally tested neopeptides. FInally, we evaluated source protein subcellular location in the context of immunotherapy response and tumor remodeling by immunotherapy. Together, these analyses support the view that protein subcellular location is a determinant of neopeptide immunogenicity.

3.4 Results

Certain cellular components are enriched for immunogenic peptides. We performed gene ontology cellular component enrichment analysis on over 380,000 MHC class I and II eluted peptides from a diverse set of normal tissues and tumor cell lines (**Figure S3.1**). We confirmed enrichment for MHC-I presented peptides in the cytosol, nucleoplasm, and extracellular exosome components and for MHC-II presented peptides in the extracellular exosome, region, and space (**Figure S3.1**) as previously described (*13–16*). Expression analysis of genes from enriched locations revealed increased expression compared to genes from depleted locations (**Figure S3.2**), though 62% of highly expressed genes were not significantly enriched in eluted pMHC, reaffirming previous findings that gene expression alone does not drive peptide elution (*16*). We also evaluated protein turnover rates from 4 human cell types including B cells, natural killer cells, hepatocytes, and monocytes (*20*) as peptides presented by MHC-I are derived from degraded peptides in the cell (*21*). Thus, a high turnover rate could result in more peptides being available for presentation. Instead, we found that overall, proteins from enriched location categories tended to have longer predicted half lives (**Figure S3.3**), though the cell types evaluated were limited. To understand the implications of these findings for peptide-directed T cell responses, we next sought to correlate peptide immunogenicity with subcellular origin.

We hypothesized that the effects of protein subcellular localization bias on peptide availability for presentation and T cell selection would influence the immunogenicity of tumor neoepitopes. Therefore we sought to test whether incorporation of peptide parent protein location would improve prediction of neopeptide immunogenicity. We began by querying the immune epitope database (IEDB) (*22*) for neoepitopes that were assayed for immunogenicity. After filtering (**Methods**), 2,943 neoepitopes remained, of which 813 (27.6%) were reported to elicit a

positive T cell assay result. Cellular component enrichment analysis of parent proteins (n=325) of immunogenic peptides did not reveal any significantly enriched locations, likely due to limited sample size. However, these peptides did tend to come from locations where more eluted peptides were observed on average (**Figure S3.4**). Analysis of unique nonimmunogenic parent proteins (n=772) whose peptides passed the minimum MHC-I binding threshold (<2 netMHCpan percentile rank) showed enrichment for locations including integral component of plasma membrane, plasma membrane, which are depleted for eluted peptide-MHC (**Figure S3.1**) but also some other membrane locations that were enriched for MHC-bound peptides, such as the endoplasmic and sarcoplasmic reticulum membranes.

As proteins can localize to multiple locations and have multiple cellular components, we sought a representation capable of capturing complex location patterns. We first used pretrained, 200-dimensional gene ontology embeddings (*23*) to represent each protein's gene ontology (GO) cellular component annotations, summing embedding vectors if a protein was associated with multiple GO terms. For visualization and use in a machine learning model, we used UMAP dimensionality reduction of these summed embeddings (**Methods**). As the embeddings and UMAP reduced features are in a non-Euclidean space, proteins with similar locations (e.g. integral component of membrane vs integral component of membrane + plasma membrane) are not necessarily near each other in 2D space (**Figure S3.5**).

We next evaluated locations of immunogenic or non-immunogenic neopeptides from the IEDB, using unique proteins evaluated in respective studies (Methods). As expected, immunogenic and non-immunogenic peptide source proteins had many overlapping locations (**Figure 3.1A,B**). However, we still observed certain combinations of locations with more immunogenic peptides than not (**Figure 3.1C**) and vice versa. For example, proteins localizing to

both the mitochondrion and mitochondrial matrix, or the mitochondrial matrix alone had fewer immunogenic peptides while more immunogenic peptides were found in proteins localizing to the centrosome, cytoplasm, nucleoplasm, and nucleus regions, though small sample sizes and use of several different assays to determine immunogenicity may reduce power.

Parent protein location improves peptide immunogenicity prediction in multiple datasets.

Next, we sought to test whether incorporating location as a feature would improve immunogenicity prediction. We performed 10-fold cross validation using a random forest classifier on the IEDB dataset (Methods) with and without adding location to a feature set that comprised peptide-MHC binding affinity (nM) (*24*), peptide-MHC stability (*25*), and foreignness (*6*, *8*). This dataset did not include many MHC-II peptides, and did not provide enough information about MHC-II alleles to calculate pMHC affinities, therefore we focused on MHC-I peptides. We found that adding location as a feature improved both the area under the receiver operating characteristic (ROC) curve (**Figure 3.1D**) and precision-recall (PR) curve (**Figure 3.1E**), and contributed to 38% of the model's predictive power (**Figure 3.1F**). We examined the predicted differences between the two models by using the median Youden index to classify peptides as immunogenic or not for each model. 254 peptides were differentially classified between the two models, with 134 now classified as immunogenic and 120 not immunogenic in the location model. The re-classified peptides were enriched for true positives and negatives (Fisher's exact OR: 1.74, p=0.04), and true positives included peptides from the cytosol while the true negatives included peptides from the nucleus (**Figure S3.6**). Interestingly, among these newly classified peptides, immunogenic versus non-immunogenic peptides had significantly higher median GTEx gene expression but similar affinity, stability, and foreignness scores (**Figure S3.6**), suggesting that location may help predict gene expression. As gene expression was correlated with peptide-MHC elution (**Figure S3.2**) (*26*), we

repeated the analysis, this time including median gene expression obtained from GTEx as a feature. We found that the benefit of including location as a feature persisted, suggesting that location provides distinct information than expression for immunogenicity prediction (**Figure S3.7**).

Next, we tested our model on unseen datasets not included in the IEDB database. First, we analyzed around 900 peptides from Wells *et al*. as this dataset represented the largest collection of immunologically tested peptides we could identify (*6*). As these data were designed to benchmark neoantigen prediction algorithms, they were partitioned into a ~600 peptide discovery set and a ~300 peptide validation set. Initial analysis revealed differences in the distribution of MHC affinity and stability of immunogenic peptides between the IEDB and Wells datasets (**Figure S3.8**) and the parent proteins of immunogenic peptides identified in Wells *et al*., originated from locations that were infrequently observed in the IEDB. Interestingly, while a model trained solely on IEDB did not perform as well as a model trained on the discovery partition provided in the Wells study in ROC analysis (AUROC of 89% vs 93%), it significantly improved the precision-recall curve (AUPRC of 64% vs 9.2%) (**Figure S3.8**). This suggests that filtering candidate neoantigens based on location may significantly reduce false positive predictions. To address systematic differences in the feature sets, we trained a new model combining the IEDB with the Wells discovery set, and were able to achieve a higher recall on the test set (AUROC of 92%) while retaining the benefit for reducing false positives (69% AUPRC) (**Figure 3.2A-C**).

Several peptide features including tumor abundance (expression) and agretopicity that were identified in the Wells dataset as being predictive of immunogenicity but were not available in the IEDB dataset. Therefore, we trained a separate model on the Wells discovery set alone incorporating these additional features and tested it on the independent Wells test set. We found that incorporating location improved both the AUROC (89% vs 67%) and AUPRC (9.6% vs

7.5%). In this model, the location features contributed 22% of feature importance, just below affinity (**Figure S3.9**). These experiments suggest that location improves prediction of immunogenic peptides, with the greatest benefit likely coming from reduction in the number of false positive predictions. The large improvement in precision and recall when incorporating the IEDB underscores the benefit of a large training set for capturing the information provided by parent protein location.

We evaluated performance on a second independent dataset of 43 assayed MHC-I neoepitopes from advanced ovarian cancer patients (*27*) not seen in the IEDB cohort. Of these, 3 (6.9%) were validated as immunogenic. The tested neopeptides once again had significantly different affinity and stability than the IEDB dataset, and while 16 locations were shared between datasets, these did not include the parent proteins for the 3 immunogenic peptides (**Figure S3.10**). We ran the model trained on IEDB alone with and without location features, as well as the model trained on both the IEDB and Wells datasets. We observed improved performance as with the addition of datasets and incorporation of location (**Figure S3.11**), 45% vs 65% AUROC and 6.2% vs 9% AUPRC. Taken together, these findings suggest that immunogenicity prediction benefits from incorporating parent protein subcellular location and can be improved through aggregation of independent datasets across cancer types.

Immunotherapy response reflects neoepitope parent protein subcellular location. Most T cell based immunotherapies, such as anti-cancer vaccines and immune checkpoint blockade (ICB) depend on the availability of immunogenic peptides to drive effector T cell responses. We speculated that if parent protein subcellular location constrains the set of mutations in a tumor that could potentially be immunogenic, then we should find associations between location and immunotherapy responses. We evaluated three ways in which location might be apparent in human

immunotherapy studies. First, we sought to determine whether location was a determinant of T cell response in a neoantigen vaccine study, then we asked whether neopeptides from locations that were more immunogenic were more likely to be depleted by immunotherapy (immunoediting), and finally investigated whether location could improve estimation of the effective neoantigen burden and consequently stratification of responders and nonresponders to ICB.

We first investigated the association of location with immune response in a neoantigen vaccine study (*28*) and found that parent proteins of neopeptides able to induce a post-vaccination response (75/125 tested; 120 distinct parent proteins) were enriched for previously observed immunogenic locations (Fisher's exact 3.49, p=0.029). Because neoantigen vaccine studies have reported predominantly CD4+ T cell responses (*28–31*), we also investigated whether the vaccine neopeptides associated with response came from locations from which more MHC-I or MHC-II peptides were eluted by the HLA ligand atlas. The majority of neopeptides tested (92/125, 73.6%) had parent proteins from which peptides had been found eluted from both MHC-I and MHC-II. Nineteen neopeptides' parent proteins were only observed to be eluted from MHC-I and 7 were exclusive to MHC-II (**Figure S3.12A**). Unlike having parent proteins in a location previously associated with immunogenic peptides, the number of MHC eluted peptides from neopeptide parent proteins alone did not correlate with response, although there was a weak trend for neopeptide parent proteins exclusive to MHC-II to have higher numbers of eluted peptides observed (**Figure S3.12B**). Thus, although we observed correlation between number of eluted peptides and immunogenicity in general (**Figure S3.4**), subcellular location may be a more nuanced determinant of immunogenic potential.

Since we found an association between parent protein subcellular location and post-vaccine response, we speculated that tumor clones cleared during treatment would be more likely to harbor

mutations in proteins from immunogenic locations. To further explore this possibility, we evaluated 73 melanoma patients with paired pre- and on-treatment samples to see if there were notable differences between eliminated and persistent neopeptides before and after treatment (*32*). We focused on responders (n=38, partial/complete response, or >6 months of stable disease) as these patients should have a relatively intact immune response compared to non-responders. While responders had better overall presentation of evaluated neopeptides, eliminated neopeptides did not have significantly better overall MHC allele-specific presentation compared to retained neopeptides in both responders and nonresponders (**Figure S3.13**), suggesting that neopeptide elimination is not driven solely by affinity or stability in this dataset. However, we note that this analysis is complicated by the non-independence of mutations that coexist within the same subclones. To investigate further, we examined 12,915 retained neopeptides from responders that were predicted to be presented by MHC-I (NetMHCpan rank < 2) and, therefore, should have been eliminated by the patient's immune system. We found that eliminated neopeptides tended to be enriched for locations where immunogenic peptides were previously observed (combining the immunogenic peptides from the IEDB, Wells et al., Liu et al.) (Fisher's exact OR: 1.08, p=0.09).

Next, we studied the potential for parent protein location to improve ICB response stratification. We evaluated cohorts with whole exome sequencing data, as well as one profiled using a deep sequenced gene panel. We began by looking for association of previously immunogenic locations with response status. We found that in a large cohort of melanoma patients (n=122) (*33*) where tumor mutation burden associated with response (**Figure S3.14A**), responders compared to nonresponders had a higher burden of proteins from locations where immunogenic responses have been previously observed (Fisher's exact OR: 1.06, p=0.0014). Considering only mutations from immunogenic locations as putative neoantigens, responders still had a significantly

higher burden of neoantigens relative to non-responders (**Figure 3.3A**). We repeated this analysis in another melanoma cohort (n=110) (*34*), where higher tumor mutation burden (TMB) was also associated with response (**Figure S3.14B**) and this time found no significant association between neopeptide parent protein location and response (Fisher's exact OR: 0.95, p=0.24). This could be due in part to ignoring other major determinants of immunogenicity such as affinity for the MHC. Therefore, we used a model trained on all datasets with immunogenicity information (IEDB + Wells + Liu (ovarian)) to classify neopeptides in both ICB cohorts as immunogenic or not based on the Youden index of the trained model (Methods**)**. In the Van Allen cohort, the predicted burden of immunogenic peptides in responders was significantly higher than in non-responders (**Figure 3.3B**). Comparing predicted neoantigens from models with and without location, to the burden of neopeptides with MHC-I affinity <500nM, we found that filtering out neopeptides predicted not to be immunogenic widened the gap between responders and non-responders in both cohorts, with the largest difference between responders and non-responders obtained with the model including location (**Figure 3.3C**), supporting the potential of location to improve stratification of patient groups pre-treatment.

Finally, we analyzed a cohort of 83 diverse tumors treated with immune checkpoint monotherapy that were profiled pre-treatment with the Foundation Medicine gene panel. Of 325 genes on this panel, 40 (12.3%) encoded proteins with subcellular locations from which immunogenic peptides had previously been observed, including *ABL1, ALK, APC, ARAF, C11ORF30, EMSY, CCND3, CDK4, CDKN1A, CDKN2A, CREBBP, EGFR, EZH2, FAM46C, FGF19, FGF3, FGF4, FUBP1, GATA3, ID3, INPP4B, JAK1, KDM5C, KMT2D,MLL2, KRAS, MAP3K1, MDM4, MET, MYCL, MYCL1, NPM1, NT5C2, PALB2, PBRM1, PTPRO, RARA, SMO, TBX3, TET2, TIPARP* and *TP53.* Thirty-seven of these 40 were from regions where only MHC-I

peptides had been previously observed, while 2 genes (*BCORL1* and *EPHA3*) were associated with locations from which only MHC-II peptides had been observed.

First, we asked whether the burden of somatic mutations in the 40 genes was informative for stratifying patient outcomes. Focusing on mutation burden in proteins from immunogenic locations reduced the total number of mutations under consideration while preserving the potential to distinguish responders from non-responders and those with stable disease (SD) (**Figure 3.4A-B**). Second, we asked whether effective presentation of one or more neopeptides from these 40 proteins was a better determinant of outcome than presentation of one or more neopeptides across all proteins in the panel. For this analysis, we focused on the 71 out of 83 patients that carried at least one mutation in these 40 genes. While patient MHC genotype specific presentation scores (PHBR scores, (*35*)) were able to stratify responders from non-responders when all proteins were considered (**Figure 3.4C**), the stratification improved when we focused on only the 40 proteins from immunogenic locations overall (**Figure 3.4D**) and in high TMB patients (**Figure S3.15**). We revisited this analysis using a Cox Proportional Hazards model with covariates as described previously (*36*), and found that when we focused on the 40 panel genes encoding proteins from immunogenic locations, presentation (PHBR score) was more significantly associated with outcome in high TMB patients, and the model had an improved (lower) Akaike information criterion score (**Table S3.1**). Altogether these results support that subcellular location of parent proteins is a determinant of the effective neoantigen burden in the setting of immunotherapy.

3.5 Discussion

While immunotherapy has generated more durable responses than targeted therapies (*37*), the fraction of patients that respond is lower. Notably, immunotherapy tends to have higher

response rates in tumor types with a high burden of somatic mutations which is thought to be a proxy for having a large number of immunogenic mutations. Mapping the mutations in a tumor genome to the subset that are likely to create immunogenic neoantigens is therefore important for realistically assessing the potential for immunotherapy response as well as for designing effective cancer vaccines. Consequently, a variety of metrics have been developed to reveal putative neoantigens in tumor genomes, with the most common being peptide-MHC binding affinity, peptide-MHC complex stability, peptide agretopicity, foreignness, and mutation expression. Here we analyzed peptides from eluted peptide-MHC and found that the subcellular location of proteins also influences which peptides are presented by the MHC. Using a high-dimensional cellular location embedding that captured multi-localization mapped to a 2 dimensional representation, we analyzed the implications of parent protein location relative to peptide immunogenicity and immunotherapy response. Immunogenic peptides were biased toward specific subcellular locations and a higher burden of mutations from these regions was associated with more benefit from immunotherapy in multiple cohorts. These findings provide the first evidence that parent protein locations influence both neopeptide presentation and T cell recognition and elimination.

We evaluated both the subcellular locations of proteins from which MHC-I and MHC-II bound peptides originate as well as those associated with peptides labeled as immunogenic based on experimental assays. We note that these locations may not fully overlap. While stable presentation by the MHC is a prerequisite for immunogenicity, it's possible that not all locations from which peptides are sourced will generate immunogenic peptides. This is dependent in part on the extent of thymic selection. Furthermore, we note that not all experimental assays used to profile immunogenicity will fully recapitulate the dependence on protein location, which could lead to the appearance of some immunogenic peptides coming from regions with no HLA

presented peptides. Furthermore, there was substantially less information about MHC-II peptides than MHC-I, leading to more limited assessment of immunogenicity in locations where peptides were predominantly eluted from MHC-II. Nonetheless, more MHC-II presented peptides than MHC-I presented peptides were associated with vaccine response in multiple vaccine studies (*28*–*31*).

In general, we speculate that the location constraint could more strongly affect peptide availability for MHC-I. Peptides from different compartments within the cell may have more variable access to the ER, which depends largely on transport from the cytoplasm by TAP family transporters (*38*). Peptides displayed by MHC-II come mostly from proteins internalized by antigen presenting cells, however MHC-I and MHC-II has been found abundantly in extracellular exosomes derived from B cells, which may explain the significant enrichment in eluted peptides (*39*, *40*). In addition, the diversity and availability of such proteins could change drastically in the presence of apoptotic or necrotic cells in the tumor immune microenvironment, making proteins from previously unavailable locations more accessible. Cross-priming may allow some exceptions to location constraints as well (*41*).

These considerations are particularly important in the context of cancer vaccines. Effective vaccine design depends on selecting peptides that will induce robust immune responses. Inclusion of peptides that stimulate T cell expansion but are not effectively displayed by the MHC at the tumor site creates the risk of generating immunodominance toward ineffective targets (*38*). The resulting T cell expansions could be dominated by clones incapable of suppressing the tumor, while more relevant clones are outcompeted in competition for antigen on the APCs (*42*), nutrient starved (*43*) and may become more easily exhausted (*44*). Thus, it may be important to avoid including peptides from parent proteins that are less accessible to the MHC. More stringent

constraints on peptide accessibility to MHC-I might make selection of effective peptides for MHC-I more challenging than for MHC-II.

Another possible consideration is whether biases in protein location during thymic tolerance render the T cell repertoire more sensitive to proteins from certain locations. Including peptides from these locations could be beneficial. This also leads to the speculation that protein localization changes in tumor cells could alter accessibility to the MHC. If these proteins were less subject to thymic tolerance, they could potentially be more potently immunogenic. One study found that an inverted form of melanoma antigen with altered localization, Melan-A, was recognized by T cells while the native orientation and a variant expressed in the cytosol were not (*45*). Although alterations in localization signals are reportedly rare (*46*, *47*), differences in trafficking could be more common (*48*). For example, we found that some mitochondrial regions were depleted for immunogenic proteins, however mitochondrial derived vesicles may provide a pathway for proteins from these regions to the MHC (*49*).

We note several limitations to our study. The pretrained location embeddings were based on characteristics of normal cells, and will reflect any biases or gaps present in the Gene Ontology (*50*). Furthermore, many proteins map to multiple locations (*51*) and have multiple associated cellular component terms. In this study we weighted each component equally, but it is likely that some locations may be predominant or transient. Immunogenicity is based on experimental assays in the IEDB performed on 325 proteins by various groups using various assays. These proteins could reflect selection bias. Similarly, locations associated with MHC eluted peptides may reflect the specific alleles that were profiled. In addition, MHC-II datasets may be biased toward B cells, whereas differences in internalization mechanisms among antigen presenting cell types such as dendritic cells or macrophages could create differences in which proteins are more accessible.

Despite these limitations, we found that incorporating protein location into analysis of immunotherapy cohorts was helpful in several ways. We used location to revise the effective neoantigen burden in tumors and better stratify potential for immunotherapy response. While expression and location provided independent benefit for inferring immunogenicity, correlation between these measures suggests that location could serve as a generic proxy for expression in studies where expression was not directly measured. Studying the effects of location in the context of tumor immunoediting is further made difficult by patterns of co-segregating mutation and subclone-specific mechanisms of immune evasion can confound the association with neoantigen characteristics. More insight may be gained from future single cell studies where it is possible to define the clonal architecture of tumors and determine which mutations coexist within the same clones. Indeed, Mehrabadi *et al* found that location bias of mutated proteins correlated with immunoediting of specific tumor subclones in a murine model of melanoma (*52*). Location information was also beneficial in a cohort that was profiled with a gene panel however, suggesting that this information could still be relevant for the more limited data commonly generated in clinical settings. Thus, we conclude that protein subcellular location contributes to shaping the tumor-immune interface and can potentially be leveraged to improve the effective application of immunotherapies.

3.6 Materials and Methods

GO analysis. Gene ontology enrichment analysis was performed using GOATOOLS (https://github.com/tanghaibao/goatools) (*53*) using the standard parameters, and retaining enriched or depleted results if the Benjamini-Hochberg corrected p-value was less than 0.05. (**Table S3.2**)

Abelin 2019 peptides. Peptides were obtained from the published Supplementary Data S1B. Peptides were mapped to parent UniProt sequences and filtered out if they mapped to multiple parent proteins. 69653/76561 (90.7%) peptides uniquely mapped to 1 parent protein sequence.

Isolation and purification of HLA-DR bound peptides. The human B cell lymphoblastoid cell lines 721.221, JThom (9004), OLL (9100), and SPACHECO (9072) were grown in complete RPMI 1640 medium (Gibco) supplemented with 10% fetal bovine serum (FBS; Gibco/Invitrogen Corp). The HeLa cell line was grown in DMEM/F12K (Gibco) supplemented with 10% fetal bovine serum (FBS; Gibco/Invitrogen Corp). The cells were grown in large scale cultures in roller bottles and the cell viability was maintained at >90% throughout the experiments. To induce HLA-Class II surface expression, HeLa cells were treated with IFNɣ (500 U/mL) for 72 hours after which the cells were harvested, washed twice with ice cold PBS and spun down at 2500xg for 10 minutes. The cell pellets were snap frozen in LN_2 and stored at -80 until downstream processing. All cell lines were subjected to high-resolution sequence-based HLA typing (HLA-A, -B, -C, DR, DP and DQ) for authentication prior to large scale culture and data collection.

HLA-DR molecules were purified from the cells by affinity chromatography using the anti-human HLA-DR antibody (clone L243) coupled to CNBr-activated Sepharose 4 Fast Flow (Amersham Pharmacia Biotech, Orsay, France) as described previously (*54*). Briefly, frozen cell pellets were pulverized using Retsch Mixer Mill MM400, resuspended in lysis buffer comprised of Tris pH 8.0 (50 mM), Igepal, 0.5%, NaCl (150 mM) and complete protease inhibitor cocktail (Roche, Mannheim, Germany). Lysates were centrifuged in an Optima XPN-80 ultracentrifuge (Beckman Coulter, IN, USA) and filtered supernatants were loaded on immunoaffinity columns. After a minimum of 3 passages, columns were washed sequentially with a series of wash buffers (*55*) and were eluted with 0.2 N acetic acid. The HLA was denatured, and the peptides were

isolated by adding glacial acetic acid and heat. The mixture of peptides and HLA-DR was subjected to reverse phase high performance liquid chromatography (RP-HPLC).

Fractionation of the HLA/Peptide Mixture by RP-HPLC. RP-HPLC was used to reduce the complexity of the peptide mixture eluted from the affinity column. First, the eluate was dried under vacuum using a CentriVap concentrator (Labconco, Kansas City, Missouri, USA). The solid residue was dissolved in 10% acetic acid and fractionated using a Paradigm MG4 instrument (Michrom BioResources, Auburn, California, USA). An acetonitrile (ACN) gradient was run at pH 2 using a two-solvent system. Solvent A contained 2% ACN in water, and solvent B contained 5% water in ACN. Both solvent A and Solvent B contained 0.1% trifluoroacetic acid (TFA). The column was pre-equilibrated at 2% solvent B. Then the sample was loaded at a flow rate of 120 µl/min and a two-segment gradient was run at 160 µl/min flow rate as described in detail in (*54*). Fractions were collected in 2 min intervals using a Gilson FC 203B fraction collector (Gilson, Middleton, Wisconsin, USA), and the ultra-violet (UV) absorption profile of the eluate was recorded at 215 nm wavelength.

Nano LC-MS/MS Analysis. Peptide-containing HPLC fractions were dried, resuspended in a solvent composed of 10% acetic acid, 2% ACN and iRT peptides (Biognosys, Schlieren, Switzerland) as internal standards. Fractions were applied individually to an Eksigent nanoLC 415 nanoscale RP-HPLC (AB Sciex, Framingham, Massachusetts, USA), including a 5-mm long, 350 µm internal diameter Chrom XP C18 trap column with 3 µm particles and 120 Å pores, and a 15-cm-long ChromXP C18 separation column (75 µm internal diameter) packed with the same medium (AB Sciex, Framingham, Massachusetts, USA). An ACN gradient was run at pH 2.5 using a two-solvent system. Solvent A was 0.1% formic acid in water, and solvent B was 0.1% formic acid in 95% ACN in water. The column was pre-equilibrated at 2% solvent B. Samples were loaded

at 5 μL/min flow rate onto the trap column and run through the separation column at 300 nL/min with two linear gradients: 10% to 40% B for 70 minutes, followed by 40% to 80% B for 7 minutes.

The column effluent was ionized using the nanospray III ion source of an AB Sciex TripleTOF 5600 quadrupole time-of-flight mass spectrometer (AB Sciex, Framingham, MA, USA) with the source voltage set to 2,400 V. Information-dependent analysis (IDA) method was used for data acquisition (*54*). PeakView Software version 1.2.0.3 (AB Sciex, Framingham, MA, USA) was used for data visualization.

Peptide Identification and Source Protein Information. Peptide sequences were identified using PEAKS Studio 10.5 software (Bioinformatics Solutions, Waterloo, Canada). A database composed of SwissProt Homo sapiens (taxon identifier 9606) and iRT peptide sequences was used as the reference for database search. Variable post-translational modifications (PTM) including acetylation, deamination, pyroglutamate formation, oxidation, sodium adducts, phosphorylation, and cysteinylation were included in database search. Identified peptides were further filtered at a false discovery rate (FDR) of 1% using PEAKS decoy-fusion algorithm

Cellular component location embedding. Gene Ontology (GO) cellular component (CC) annotations for all UniProt protein IDs was obtained from uniprot.org. Pretrained 200 dimensional vectors for 64,649 GO terms were obtained from (*23*). Vectors were mapped to UniProt IDs and summed if a UniProt ID had more than 1 associated GO CC term. UMAP dimensionality reduction (*56*) using the "hyperboloid" metric was applied, then mapped to the Poincare disk model. The resulting 2 values for each protein were used as features.

IEDB Data. Peptides were selected from the Immune Epitope Database and Analysis Resource (www.iedb.org) (*22*) on July 16, 2021 using filters "Epitope Structure: Linear

Sequence", "Included Related Structures: neo-epitope", "No B cell Assays", "No MHC assays", "MHC Restriction Type: Class I", "Host: Homo sapiens (human)" and "Include Positive Assays", "Include Negative Assays" (T cell assays). This resulted in 3754 peptides. Peptides whose "Assay Antigen Antigen Description" sequence did not match "Epitope Description" sequence were filtered, resulting in 3521 peptides. Peptides were also filtered out if they did not have an associated UniProt ID, found in the "Related Object Parent Protein IRI" field, resulting in 3367 peptides. Peptides were additionally removed if they were not in the UniProt CC file (downloaded on April 17, 2020 from uniprot.org by selecting all Human peptides and choosing "Gene ontology (cellular component)" in the column selection, see *Cellular component location embedding*), resulting in 3125 peptides. Peptides were dropped if they did not have a specific associated HLA allele (e.g. Allele Name = "HLA class I" or Allele Name = "HLA-A2") or if they were not a simple linear sequence (e.g. ILCETCLIV + AIB(C3, C6)). This resulted in 2943 peptides. Affinity, stability, and foreignness features were calculated as described in "Validation Data".

Hex plots. Parent protein locations were plotted for each unique protein and immunogenic state for each study (i.e. if peptides from ProteinA had positive and negative tests, the parent protein would be retained twice, once in each immunogenic category).

Random forest model. The RandomForestClassifier from sklearn v0.24.2 was used using random state 2021. The StratifiedKFold function was used to perform the 10-fold splits, also using random state 2021. The Youden indices for each fold were obtained by taking the threshold that had the greatest TPR-FPR (i.e. greatest area under the curve). The median Youden index was used to classify peptides as immunogenic or not for downstream analyses.

Wells Data. Experimentally validated peptides were obtained from published supplementary tables S4 and S7 in (*6*). Peptides were mapped to parent proteins by iterating

through all UniProt proteins and looking for a match to any peptide with a wildcard in the given mutated position (e.g. Python code to find the wildtype peptide corresponding to "FLCEILRSMSI" with mutated position 10: re.findall(r'(?=("FLCEILRSM.I"))', protein_sequence). 10 peptides without a mutated position were excluded. 584 (97.6%) of neopeptide sequences mapped to 1 unique parent wildtype sequence. Peptides with matched wildtype sequences mapping to multiple UniProt IDs were dropped. Missing foreginness or agretopicity scores were recalculated using the methods described in Wells *et al.* The resulting 558 peptides from the discovery set was used to train a Random Forest classifier using sklearn (v0.24.2).

Wells Validation Data. The trained model was tested on the 310 peptide validation dataset from Wells et al., as well as 43 peptides from ovarian tumors (*27*). As these datasets did not include all features from the discovery dataset, NetMHCstabpan (v1.0) (*25*) was used to predict pMHC stability, NetMHCpan (v4.0) (*24*) was used to predict pMHC binding affinity, and the antigen.garnish package (https://github.com/andrewrech/antigen.garnish) was used to calculate foreignness as described in Luksa, Wells. Finally, agretopicity was calculated by taking the ratio of mutant to wildtype binding affinity.

3.7 Figures

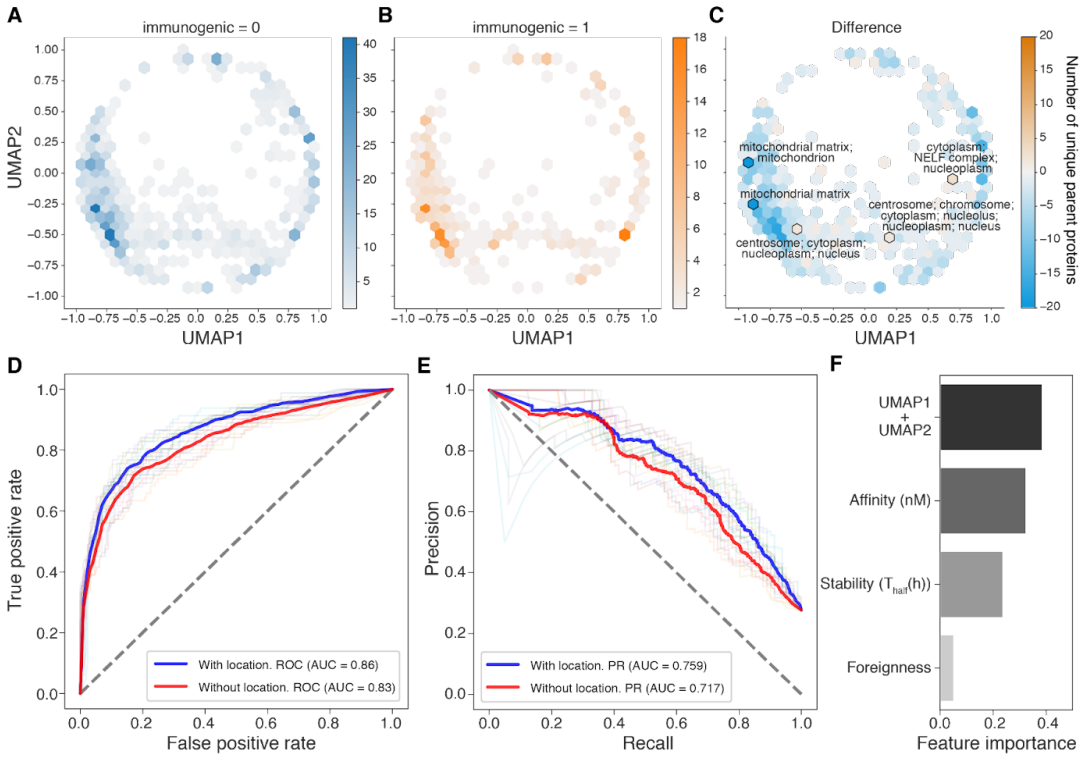


Figure 3.1. Overview of T cell assayed neoepitopes from IEDB. Hexplots of embedded location for (**A**) non-immunogenic (blue) and (**B**) immunogenic (orange) peptides. (**C**) Hexplot depicting the difference between immunogenic and non-immunogenic hexplots in A and B. Orange indicates more immunogenic peptides, and blue indicates more non-immunogenic peptides. (**D**) Area under the receiver operating characteristic curve (AUROC) and (**E**) area under the precision recall curve (AUPRC) for 10-fold cross validation using a Random Forest model incorporating peptide affinity, stability, and foreignness (Methods) with and without parent protein location features. (F) Barplot of model feature importances.

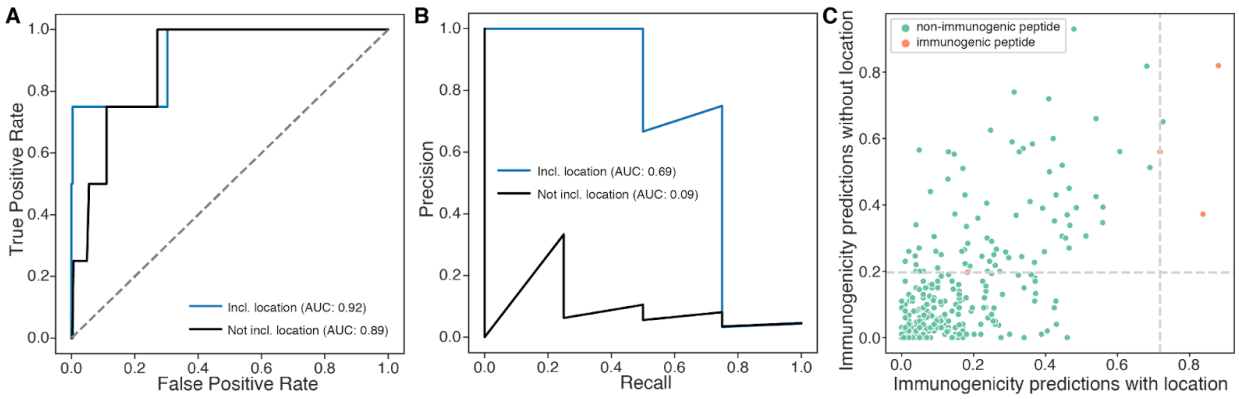


Figure 3.2. Predicting immunogenicity on unseen datasets. (**A**) Area under the receiver operating characteristic curve (AUROC) (**B**) and area under the precision recall curve (AUPRC) for the unseen validation dataset with and without parent protein location features. (**C**) Scatterplot of the predicted probabilities for unseen test neopeptides to be immunogenic with and without location as a feature. Dashed lines indicate the Youden index for each model, used to optimally threshold predictions. False positives are reduced in the model with location.

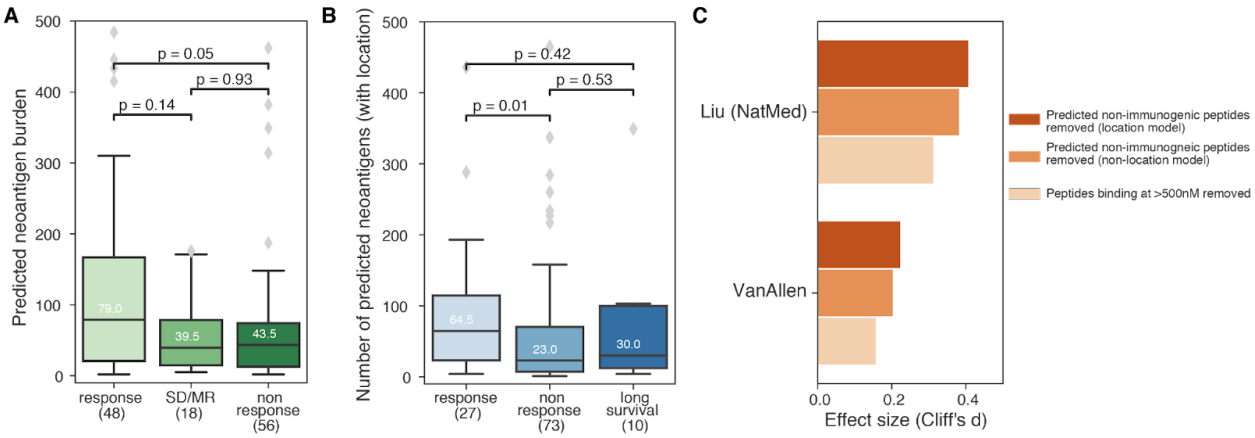


Figure 3.3. ICB responders carry a higher burden of mutations in proteins from immunogenic locations. (**A**) Predicted neoantigen burden versus response category in the Liu cohort when retaining only mutations in proteins from subcellular locations previously observed to source immunogenic peptides. (**B**) Predicted neoantigen burden versus response category in the VanAllen cohort where neoantigen status is predicted using a model trained on 3 sources of immunogenic peptide and features including peptide MHC affinity, stability, agretopicity and location. (**C**) Barplot of effect sizes between responders and nonresponders where neoantigen status is predicted using a model trained on 3 sources of immunogenic peptide and MHC affinity, stability and agretopicity, with and without location.

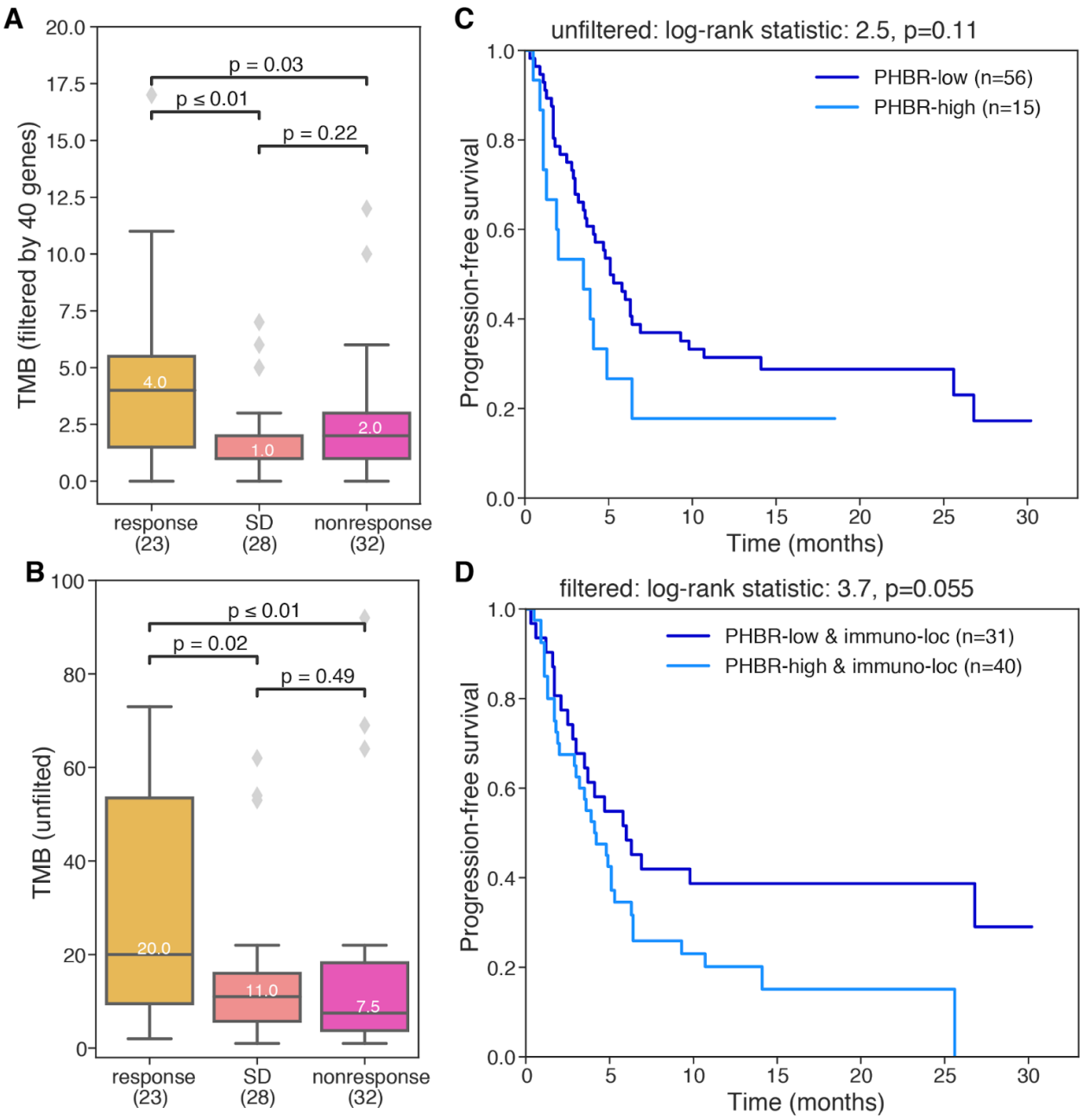


Figure 3.4. Focusing on immunogenic locations improves response prediction in a gene panel profiled cohort. Tumor mutation burden (**A**) focusing on the 40 genes whose proteins localize to previously observed immunogenic subcellular locations and (**B**) all genes in the gene panel. Kaplan Meier curves showing the effect of the best presented mutation on progression-free survival (**C**) using all genes in the panel and (**D**) using only the 40 genes of interest.

3.8 Supplemental Data, Tables and Figures

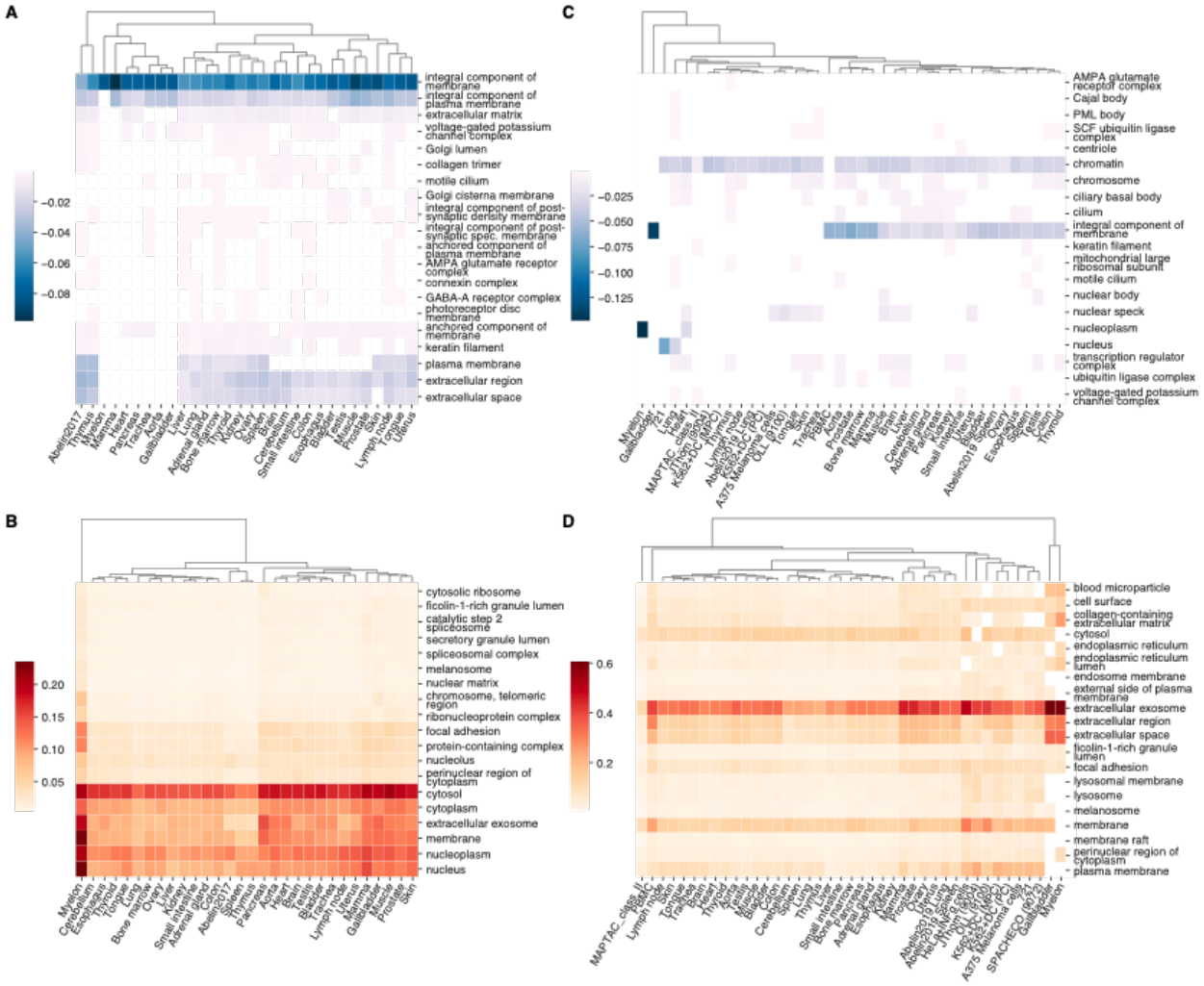


Figure S3.1. Overview of enrichment or depletion of cellular components in multiple datasets. (**A**) Clustermap of the top 20 cellular components depleted in eluted peptide-MHC (pMHC) class I from normal and cell lines. (**B**) Clustermap of 21 cellular components enriched in eluted pMHC-I complexes across all evaluated normal tissues, indicated by "(N)" (Marcu et al., 2021) and evaluated cell lines (Abelin et al., 2017). The color indicates the difference in study vs population enrichment. Clustermaps of (**C**) depleted and (**D**) enriched cellular components for eluted pMHC-II from 721.221, JThom (9004), OLL (9100), and SPACHECO (9072) B cells, HeLa cells stimulated with IFN-γ, (Abelin et al., 2019), and (Abelin et al., 2019; Marcu et al., 2021).

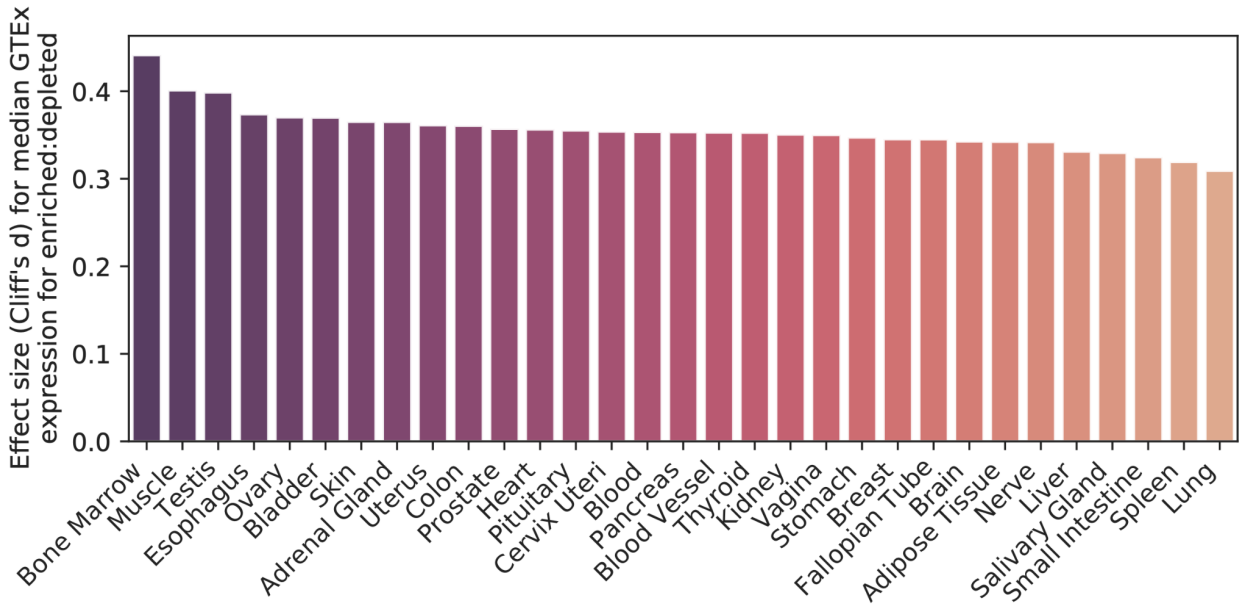


Figure S3.2. Correlation of gene expression and eluted peptide location. Barplot for each tissue showing the effect size comparing the median GTEx gene expression for enriched over depleted genes. All comparisons show that genes in enriched locations have higher median gene expression than genes in depleted locations.

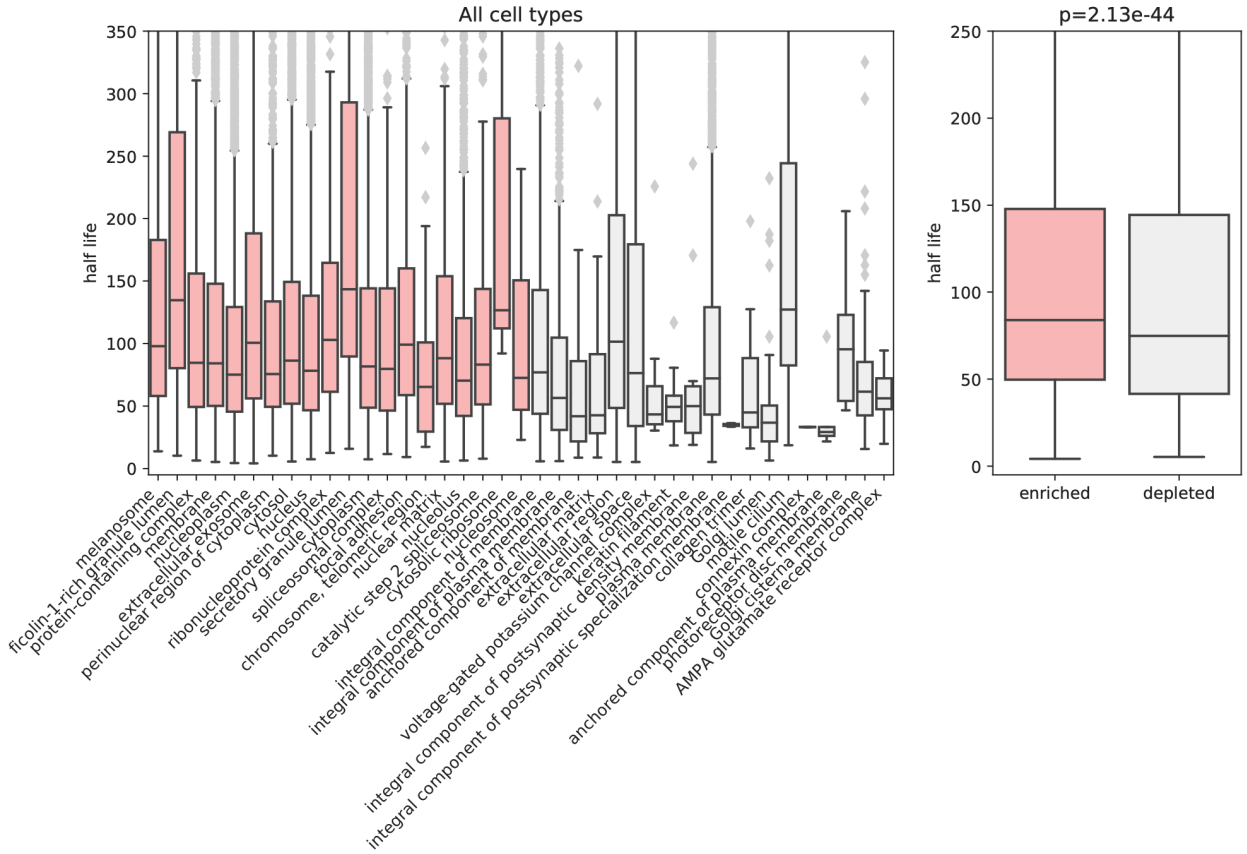


Figure S3.3. Relationship between protein turnover and elution for the top 20 most frequently enriched or depleted cellular components across evaluated tissues or cell lines. The Mann-Whitney U test was used to compare statistical significance.

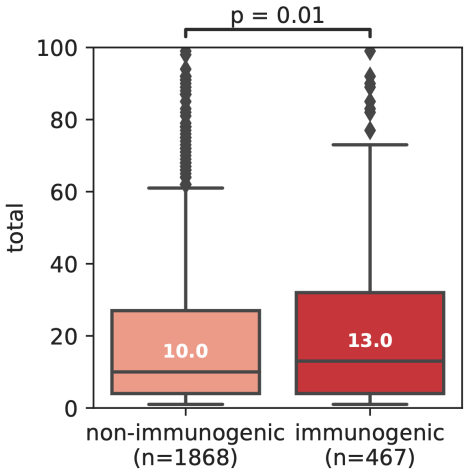


Figure S3.4. Analysis of the relationship between elution and immunogenicity. Boxplot comparing the frequency of eluted peptides versus immunogenicity assay results for proteins that have been evaluated for class I immunogenicity in the IEDB.

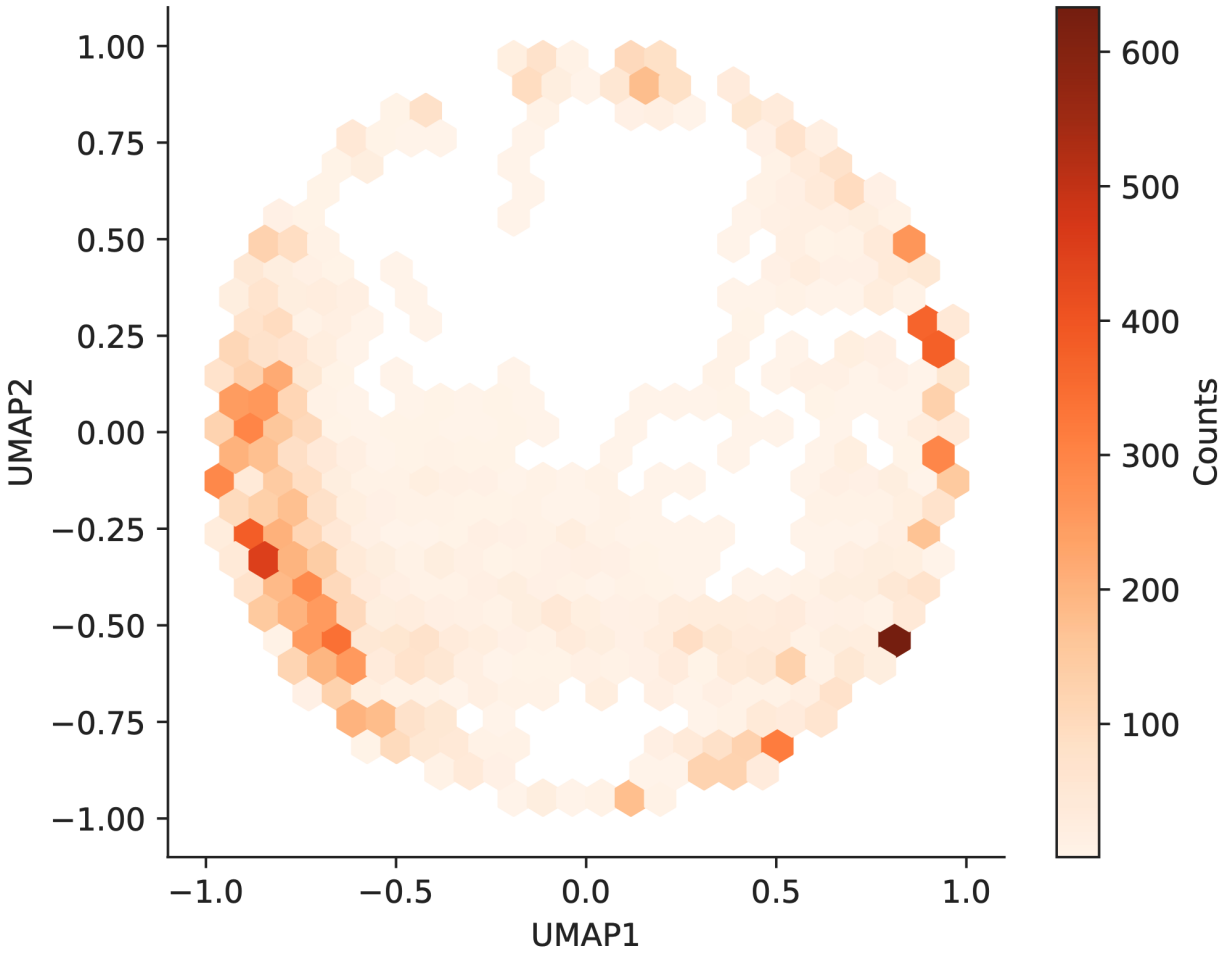


Figure S3.5. Overview of UMAP location embeddings for all unique UniProt proteins. Hexplot of UMAP location embeddings for all unique UniProt proteins with reviewed status and unique gene names.

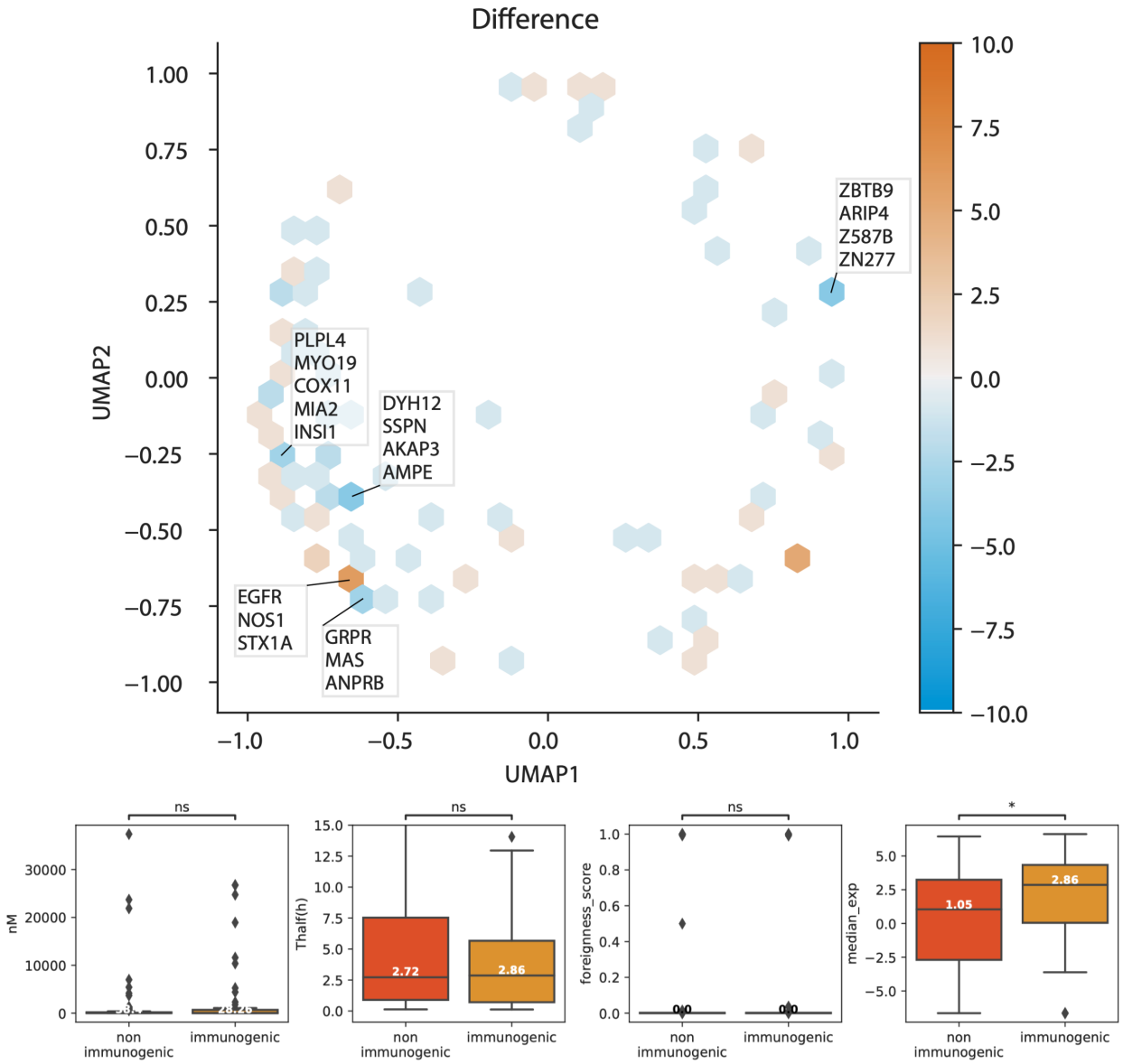


Figure S3.6. Overview of differentially classified peptides between the models with and without location as a feature. Orange indicates locations with more immunogenic peptides compared to non-immunogenic peptides and vice versa. Locations with more than 3 immunogenic or non-immunogenic genes are highlighted.

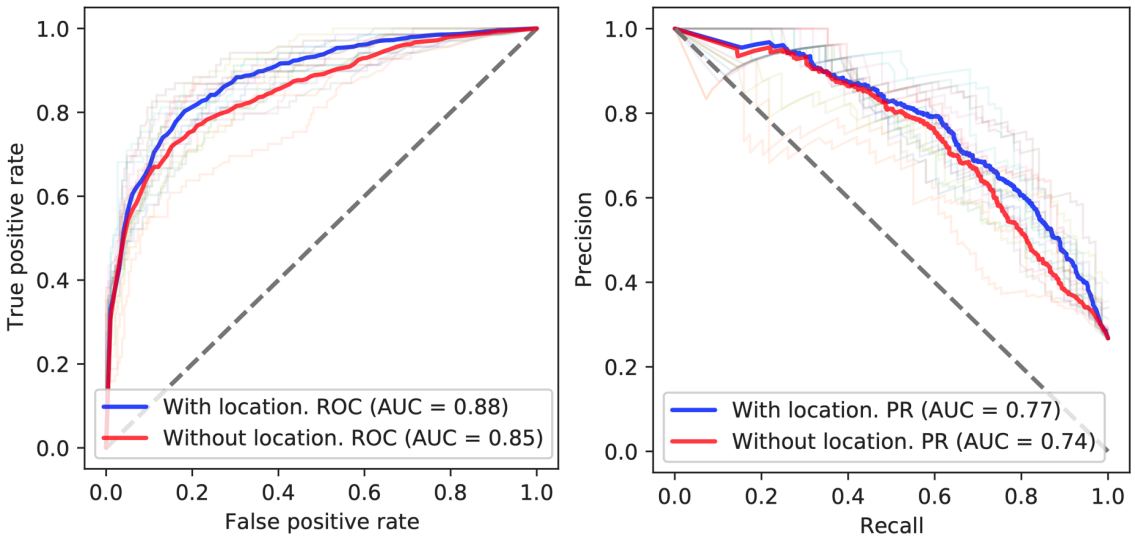


Figure S3.7. Analysis of the effects of incorporating gene expression in the random forest model. (Left) Area under the receiver operating characteristic curve (AUROC) and (right) area under the precision recall curve (AUPRC) for 10-fold cross validation using a Random Forest model incorporating median GTEx gene expression, peptide affinity, stability, and foreignness (Methods) with and without parent protein location features.

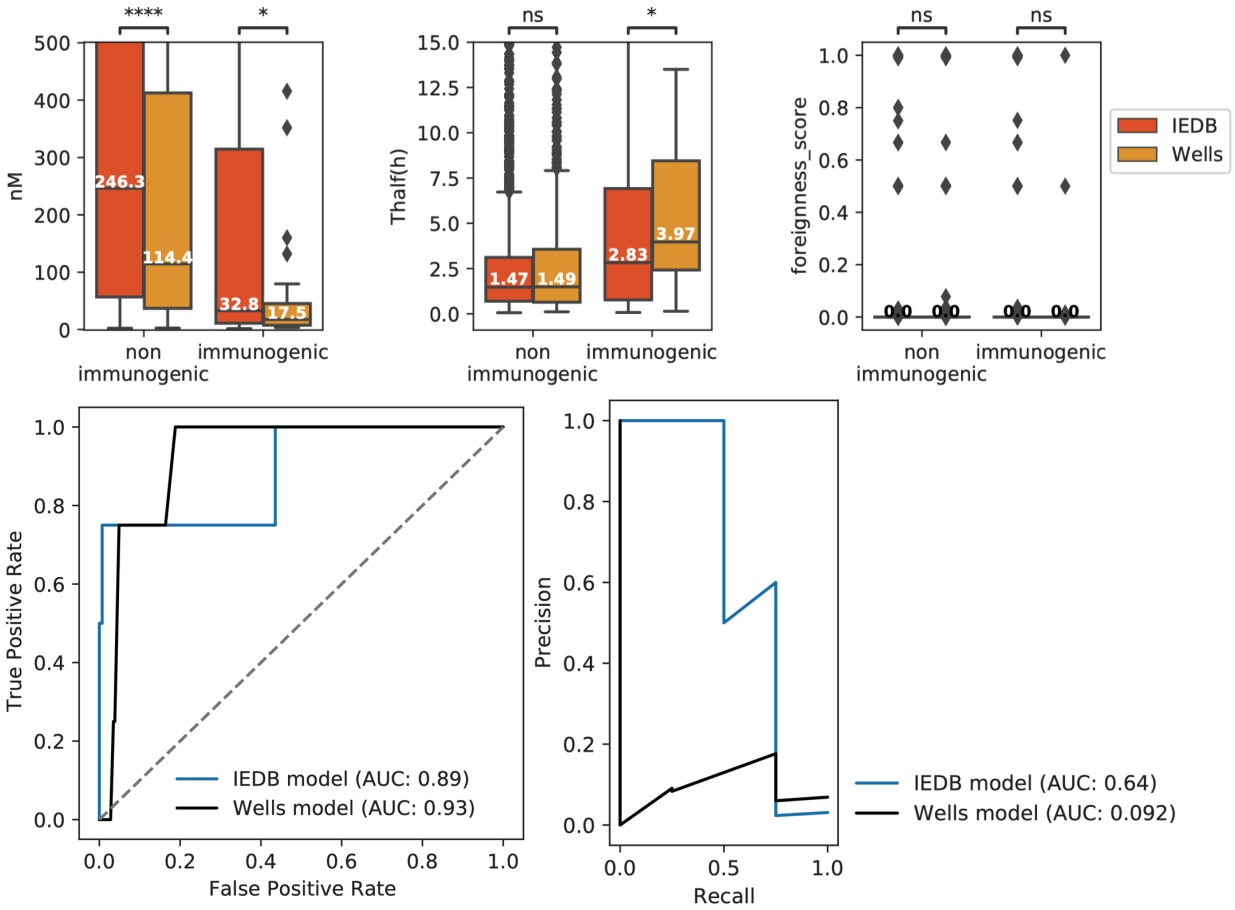


Figure S3.8. Comparison of the IEDB and Wells et al. datasets. (Top) Boxplots comparing affinity (measured in nM), stability (measured by half life), and foreignness stratified by immunogenicity. The Mann-Whitney U test was used to compare statistical significance. (Bottom panel) Area under the receiver operating characteristic and precision recall curves using the random forest model trained on the IEDB dataset and Wells discovery set to test on the Wells test set.

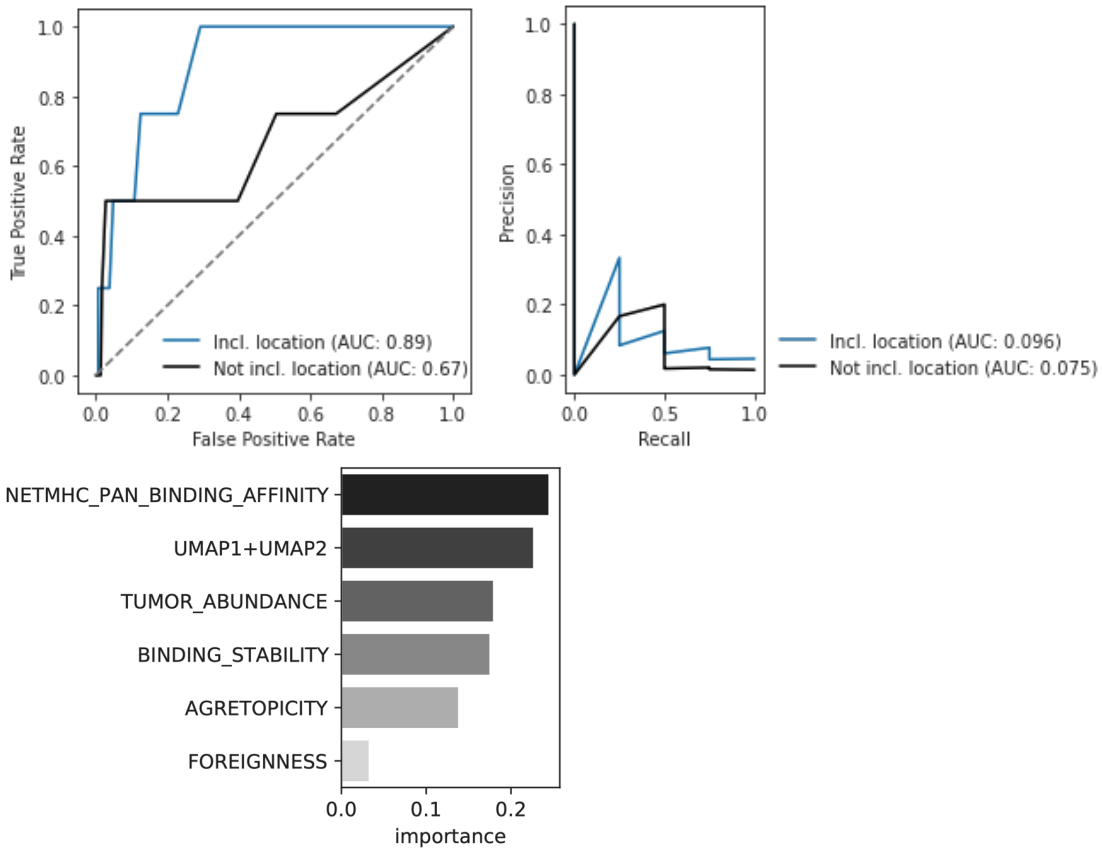


Figure S3.9. AUROC and AUPRC plots for the model trained on the Wells discovery dataset and tested on the Wells test dataset. Features include peptide affinity, stability, tumor abundance, agretopicity, and foreignness. Barplot denoting feature importance for the model.

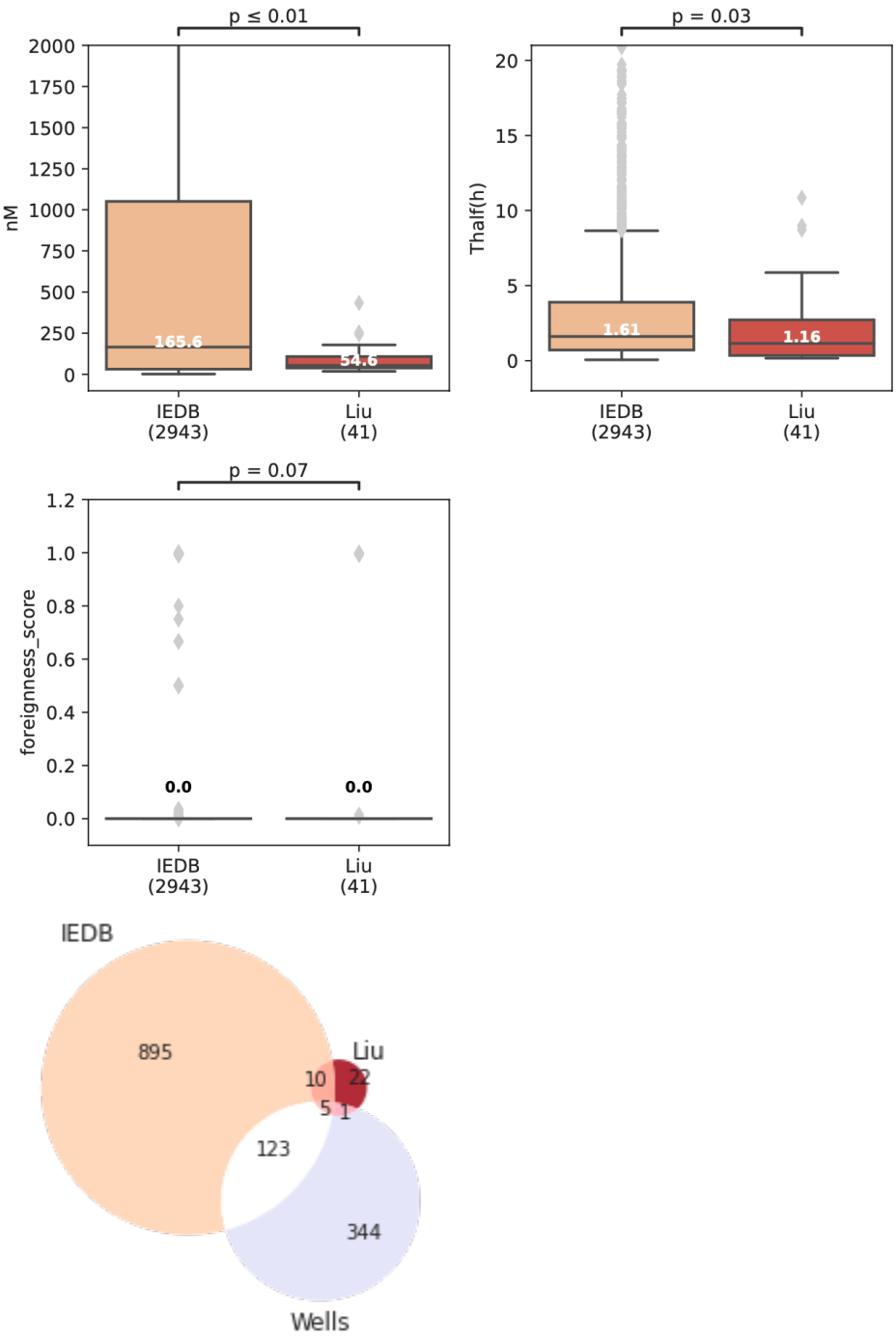


Figure S3.10. Comparison of the IEDB and Liu et al. datasets. Boxplots comparing affinity (measured in nM), stability (measured by half-life), and foreignness. The Mann-Whitney U test was used to compare statistical significance. The Venn diagram shows the overlapping unique locations. All overlapping locations were non-immunogenic in Liu.

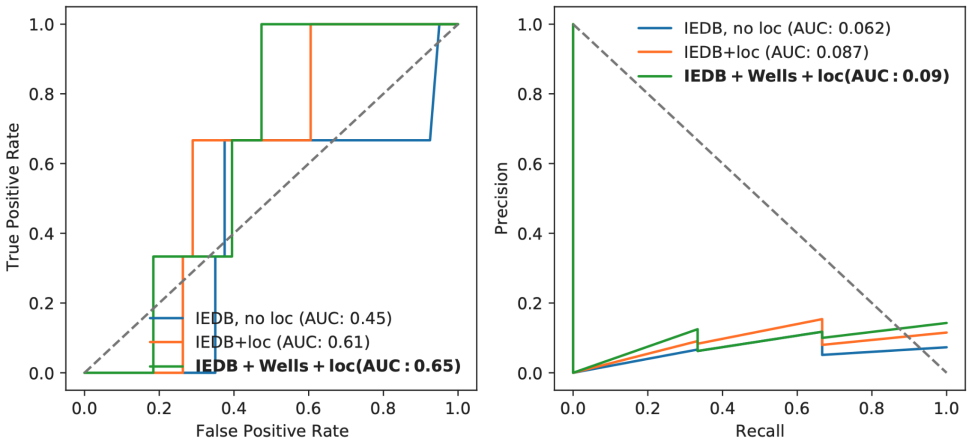


Figure S3.11. Testing pretrained models on the unseen Liu ovarian dataset. (Left) AUROC and (right) AUPRC curves for the IEDB model without location, with location, and aggregated model with IEDB, Wells et al., and location.

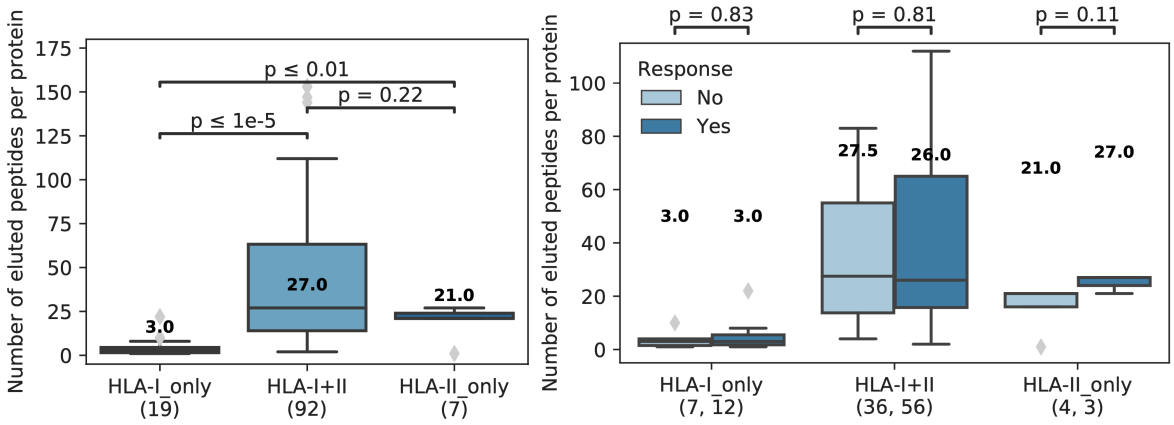


Figure S3.12. Analysis of neopeptide vaccine parent protein MHC elution patterns. (**A**) Boxplots showing the number of eluted peptides in the HLA ligand atlas associated with the parent proteins of the 125 neopeptides evaluated by Sahin *et al*. The majority were from proteins from which peptides were found in both MHC-I and MHC-II eluted complexes. Parent proteins exclusive to MHC-I tended to have lower eluted peptide counts than parent proteins exclusive to MHC-II. (**B**) Boxplots showing the number of eluted peptides as in panel A, but further divided according to whether the number of MHC eluted peptides was not associated with post-vaccination response.

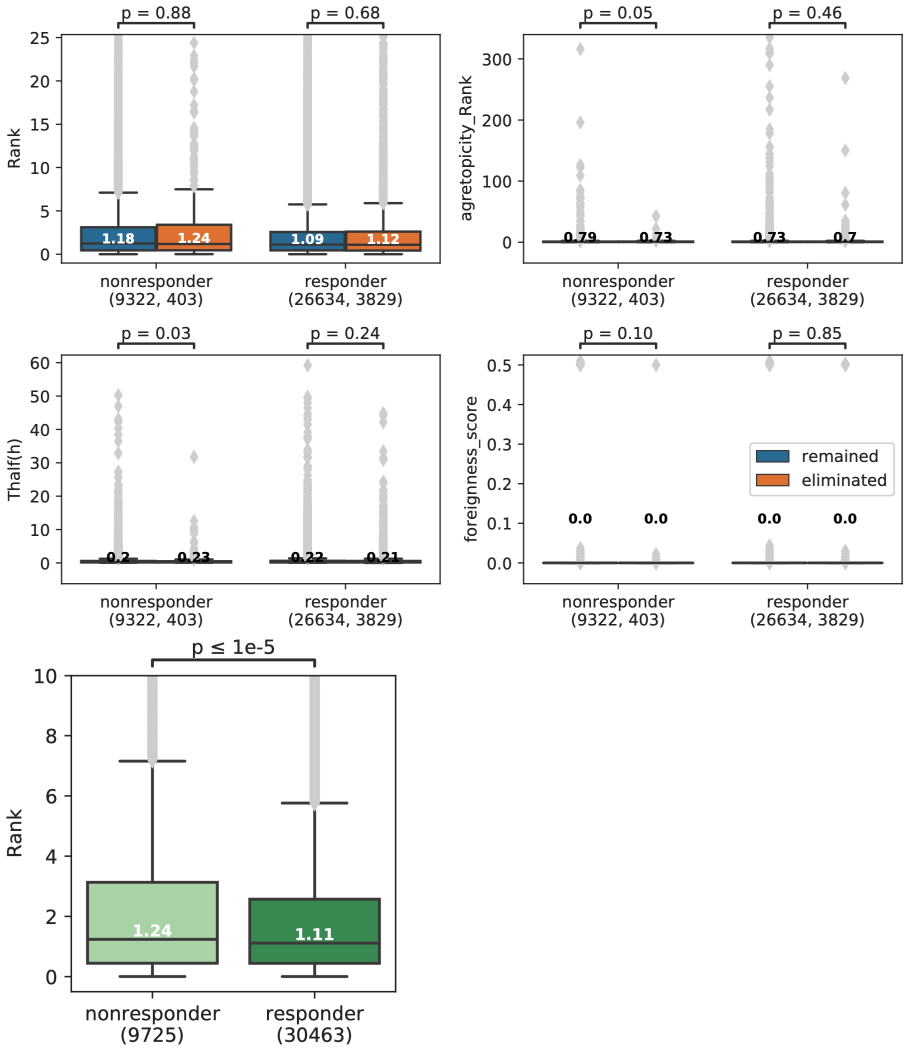


Figure S3.13. Comparison of neopeptide characteristics in the Riaz et al. dataset. (Top) Boxplots comparing affinity, agretopicity, stability, and foreignness between eliminated versus remaining neopeptides for both responders and nonresponders. (Bottom) Comparison of neopeptide affinity between responders and nonresponders.

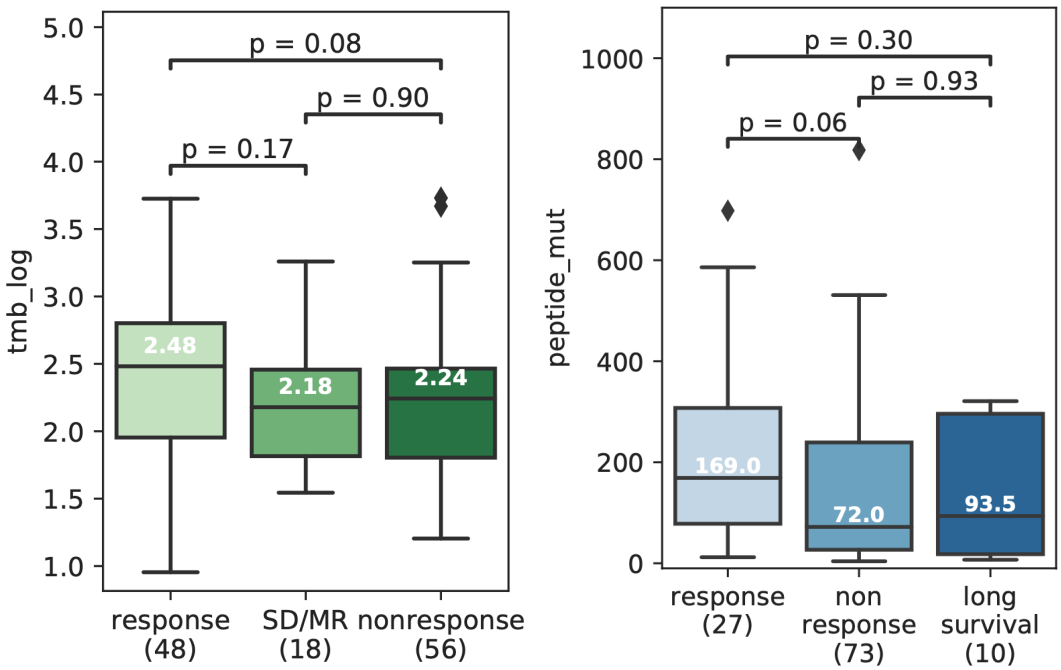


Figure S3.14. Initial association of tumor mutation burden with response. (**A**) Liu *et al.* and (**B**) Van Allen *et al.* Tumor mutation burden was defined in Liu *et al.* as the number of somatic mutations, and in Van Allen *et al.* as the number of neopeptides <500nM.

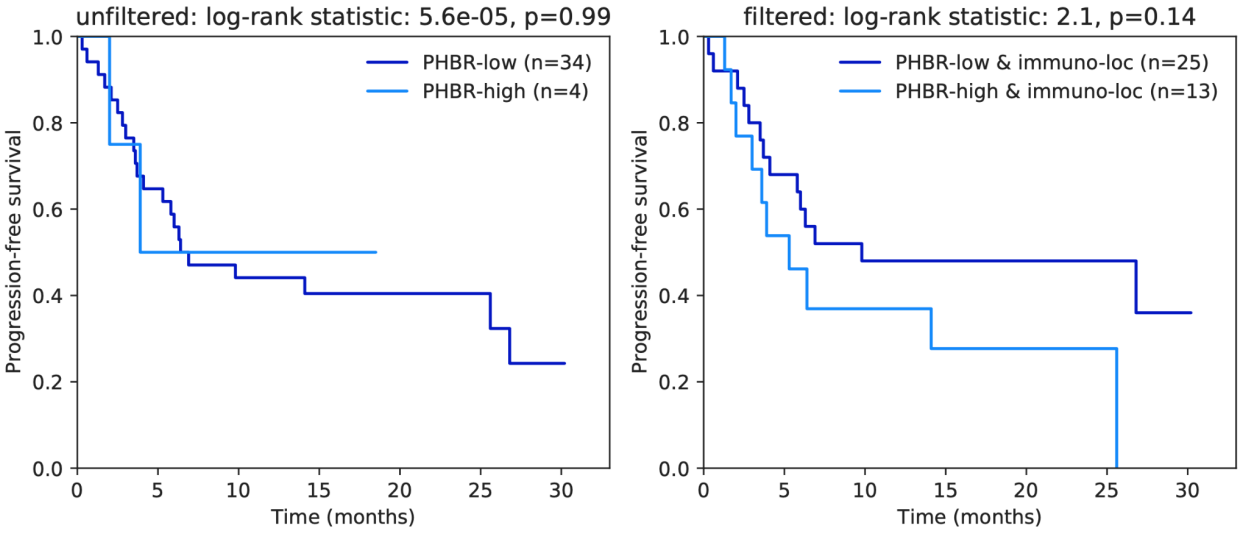


Figure S3.15. Kaplan Meier curves showing the effect of the best presented mutation on progression-free survival. (Left) using all genes in the panel and (Right) using only the 40 genes of interest for high TMB patients.

Table S3.1. Cox proportional hazards results.

		Hazard ratios (exp(coef))	**Standard error (se(coeff))**	**p-value**
Model focusing on the 40 genes that fall in locations seen to be previously immunogenic Partial AIC=147.9	PHBR score	1.27	0.09	**0.01**
	TMB	0.99	0.004	0.57
	Age at treatment with immunotherapy	0.97	0.02	0.13
	Gender	1.55	0.47	0.35
Unfiltered model. Focuses on all genes from the gene panel Partial AIC=151.7	PHBR score	1.35	0.22	0.17
	TMB	0.99	0.004	0.46
	Age at treatment with immunotherapy	0.98	0.02	0.33
	Gender	1.88	0.46	0.17

3.9 Author Contributions

Original concept: Andrea Castro

Project supervisor: Hannah Carter

Data acquisition, processing and analysis: Andrea Castro, Saghar Kabbienejadian, and Hooman Yari

Peptide elution experiments and supervision: Saghar Kabbienejadian, Hooman Yari and William Hildebrand

Figures: Andrea Castro and Hannah Carter

Manuscript writing: Andrea Castro, Maurizio Zanetti, and Hannah Carter

3.10 Acknowledgements

This work was supported by Emerging Leader Award from The Mark Foundation for Cancer Research grant #18-022-ELA and NIH U24CA248138-01A1 grant subaward 20051-01-144-384 to HC and NIH grant R01CA220009 to HC and MZ. Some analyses herein used data from the GTEx project, which was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS.

Chapter 3, in full, is a reformatted reprint of the material currently being prepared for submission for publication as "Source protein subcellular location is a novel feature to improve prediction of neoantigens for immunotherapy" by Andrea Castro, Saghar Kaabinejadian, Hooman Yari, William Hildebrand, Maurizio Zanetti, and Hannah Carter. The dissertation author was a primary investigator and author of this paper.

3.11 References

1. B. Reynisson, B. Alvarez, S. Paul, B. Peters, M. Nielsen, NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Res.* (2020) (available at https://academic.oup.com/nar/advance-article-abstract/doi/10.1093/nar/gkaa379/5837056).

2. T. J. O'Donnell, A. Rubinsteyn, U. Laserson, MHCflurry 2.0: Improved Pan-Allele Prediction of MHC Class I-Presented Peptides by Incorporating Antigen Processing. *Cell Syst*. **11**, 42–48.e7 (2020).

3. M. Nielsen, M. Andreatta, B. Peters, S. Buus, Immunoinformatics: Predicting Peptide–MHC Binding. *Annu. Rev. Biomed. Data Sci.* **3**, 191–215 (2020).

4. X. M. Shao, R. Bhattacharya, J. Huang, I. K. A. Sivakumar, C. Tokheim, L. Zheng, D. Hirsch, B. Kaminow, A. Omdahl, M. Bonsack, A. B. Riemer, V. E. Velculescu, V. Anagnostou, K. A. Pagel, R. Karchin, High-Throughput Prediction of MHC Class I and II Neoantigens with MHCnuggets. *Cancer Immunol Res*. **8**, 396–408 (2020).

5. M. Yadav, S. Jhunjhunwala, Q. T. Phung, P. Lupardus, J. Tanguay, S. Bumbaca, C. Franci, T. K. Cheung, J. Fritsche, T. Weinschenk, Z. Modrusan, I. Mellman, J. R. Lill, L. Delamarre, Predicting immunogenic tumour mutations by combining mass spectrometry and exome sequencing. *Nature*. **515**, 572–576 (2014).

6. D. K. Wells, M. M. van Buuren, K. K. Dang, V. M. Hubbard-Lucey, K. C. F. Sheehan, K. M. Campbell, A. Lamb, J. P. Ward, J. Sidney, A. B. Blazquez, A. J. Rech, J. M. Zaretsky, B. Comin-Anduix, A. H. C. Ng, W. Chour, T. V. Yu, H. Rizvi, J. M. Chen, P. Manning, G. M. Steiner, X. C. Doan, Tumor Neoantigen Selection Alliance, T. Merghoub, J. Guinney, A. Kolom, C. Selinsky, A. Ribas, M. D. Hellmann, N. Hacohen, A. Sette, J. R. Heath, N. Bhardwaj, F. Ramsdell, R. D. Schreiber, T. N. Schumacher, P. Kvistborg, N. A. Defranoux, Key Parameters of Tumor Epitope Immunogenicity Revealed Through a Consortium Approach Improve Neoantigen Prediction. *Cell*. **183**, 818–834.e13 (2020).

7. E. Ghorani, R. Rosenthal, N. McGranahan, J. L. Reading, M. Lynch, K. S. Peggs, C. Swanton, S. A. Quezada, Differential binding affinity of mutated peptides for MHC class I is a predictor of survival in advanced lung cancer and melanoma. *Ann. Oncol.* **29**, 271–279 (2018).

8. M. Łuksza, N. Riaz, V. Makarov, V. P. Balachandran, M. D. Hellmann, A. Solovyov, N. A. Rizvi, T. Merghoub, A. J. Levine, T. A. Chan, J. D. Wolchok, B. D. Greenbaum, A neoantigen fitness model predicts tumour response to checkpoint blockade immunotherapy. *Nature*. **551**, 517–520 (2017).

9. J. Schmidt, A. R. Smith, M. Magnin, J. Racle, J. R. Devlin, S. Bobisse, J. Cesbron, V. Bonnet, S. J. Carmona, F. Huber, G. Ciriello, D. E. Speiser, M. Bassani-Sternberg, G. Coukos, B. M. Baker, A. Harari, D. Gfeller, Prediction of neo-epitope immunogenicity reveals TCR

recognition determinants and provides insight into immunoediting. *Cell Rep Med*. **2**, 100194 (2021).

10. A. Castro, M. Zanetti, H. Carter, Neoantigen Controversies (2021), doi:10.1146/annurev-biodatasci-092820-112713.

11. M. Wieczorek, E. T. Abualrous, J. Sticht, M. Álvaro-Benito, S. Stolzenberg, F. Noé, C. Freund, Major Histocompatibility Complex (MHC) Class I and MHC Class II Proteins: Conformational Plasticity in Antigen Presentation. *Front. Immunol.* **8**, 292 (2017).

12. P. A. Roche, K. Furuta, The ins and outs of MHC class II-mediated antigen processing and presentation. *Nat. Rev. Immunol.* **15**, 203–216 (2015).

13. J. G. Abelin, D. Harjanto, M. Malloy, P. Suri, T. Colson, S. P. Goulding, A. L. Creech, L. R. Serrano, G. Nasir, Y. Nasrullah, C. D. McGann, D. Velez, Y. S. Ting, A. Poran, D. A. Rothenberg, S. Chhangawala, A. Rubinsteyn, J. Hammerbacher, R. B. Gaynor, E. F. Fritsch, J. Greshock, R. C. Oslund, D. Barthelme, T. A. Addona, C. M. Arieta, M. S. Rooney, Defining HLA-II Ligand Processing and Binding Rules with Mass Spectrometry Enhances Cancer Epitope Prediction. *Immunity*. **51**, 766–779.e17 (2019).

14. M. Bassani-Sternberg, S. Pletscher-Frankild, L. J. Jensen, M. Mann, Mass spectrometry of human leukocyte antigen class I peptidomes reveals strong effects of protein abundance and turnover on antigen presentation. *Mol. Cell. Proteomics*. **14**, 658–673 (2015).

15. I. M. M. Schellens, I. Hoof, H. D. Meiring, S. N. M. Spijkers, M. C. M. Poelen, J. A. M. van Gaans-van den Brink, K. van der Poel, A. I. Costa, C. A. C. M. van Els, D. van Baarle, C. Kesmir, Comprehensive Analysis of the Naturally Processed Peptide Repertoire: Differences between HLA-A and B in the Immunopeptidome. *PLoS One*. **10**, e0136417 (2015).

16. H. Pearson, T. Daouda, D. P. Granados, C. Durette, E. Bonneil, M. Courcelles, A. Rodenbrock, J.-P. Laverdure, C. Côté, S. Mader, S. Lemieux, P. Thibault, C. Perreault, MHC class I-associated peptides derive from selective regions of the human genome. *J. Clin. Invest.* **126**, 4690–4701 (2016).

17. V. Karnaukhov, W. Paes, I. B. Woodhouse, T. Partridge, A. Nicastri, S. Brackenridge, D. Scherbinin, D. M. Chudakov, I. V. Zvyagin, N. Ternette, H. Koohy, P. Borrow, M. Shugay, HLA binding of self-peptides is biased towards proteins with specific molecular functions. *bioRxiv* (2021), doi:10.1101/2021.02.16.431395.

18. Z. Lu, L. Hunter, in *Biocomputing 2005* (WORLD SCIENTIFIC, 2004), pp. 151–161.

19. Y. Xing, K. A. Hogquist, T-cell tolerance: central and peripheral. *Cold Spring Harb. Perspect. Biol.* **4** (2012), doi:10.1101/cshperspect.a006957.

20. T. Mathieson, H. Franken, J. Kosinski, N. Kurzawa, N. Zinn, G. Sweetman, D. Poeckel, V. S. Ratnu, M. Schramm, I. Becher, M. Steidel, K.-M. Noh, G. Bergamini, M. Beck, M. Bantscheff, M. M. Savitski, Systematic analysis of protein turnover in primary cells. *Nat. Commun.* **9**, 689 (2018).

21. E. Milner, E. Barnea, I. Beer, A. Admon, The turnover kinetics of major histocompatibility complex peptides of human cancer cells. *Mol. Cell. Proteomics*. **5**, 357–365 (2006).

22. R. Vita, S. Mahajan, J. A. Overton, S. K. Dhanda, S. Martini, J. R. Cantrell, D. K. Wheeler, A. Sette, B. Peters, The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Res.* **47**, D339–D343 (2019).

23. J. Kim, D. Kim, K.-A. Sohn, HiG2Vec: Hierarchical Representations of Gene Ontology and Genes in the Poincaré Ball. *Bioinformatics* (2021), doi:10.1093/bioinformatics/btab193.

24. V. Jurtz, S. Paul, M. Andreatta, P. Marcatili, B. Peters, M. Nielsen, NetMHCpan-4.0: Improved Peptide-MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. *J. Immunol.* **199**, 3360–3368 (2017).

25. M. Rasmussen, E. Fenoy, M. Harndahl, A. B. Kristensen, I. K. Nielsen, M. Nielsen, S. Buus, Pan-Specific Prediction of Peptide–MHC Class I Complex Stability, a Correlate of T Cell Immunogenicity. *The Journal of Immunology*. **197**, 1517–1524 (2016).

26. J. G. Abelin, D. B. Keskin, S. Sarkizova, C. R. Hartigan, W. Zhang, J. Sidney, J. Stevens, W. Lane, G. L. Zhang, T. M. Eisenhaure, K. R. Clauser, N. Hacohen, M. S. Rooney, S. A. Carr, C. J. Wu, Mass Spectrometry Profiling of HLA-Associated Peptidomes in Mono-allelic Cells Enables More Accurate Epitope Prediction. *Immunity*. **46**, 315–326 (2017).

27. S. Liu, J. Matsuzaki, L. Wei, T. Tsuji, S. Battaglia, Q. Hu, E. Cortes, L. Wong, L. Yan, M. Long, A. Miliotto, N. W. Bateman, S. B. Lele, T. Chodon, R. C. Koya, S. Yao, Q. Zhu, T. P. Conrads, J. Wang, G. L. Maxwell, A. A. Lugade, K. Odunsi, Efficient identification of neoantigen-specific T-cell responses in advanced human ovarian cancer. *Journal for ImmunoTherapy of Cancer*. **7**, 1–17 (2019).

28. U. Sahin, E. Derhovanessian, M. Miller, B.-P. Kloke, P. Simon, M. Löwer, V. Bukur, A. D. Tadmor, U. Luxemburger, B. Schrörs, T. Omokoko, M. Vormehr, C. Albrecht, A. Paruzynski, A. N. Kuhn, J. Buck, S. Heesch, K. H. Schreeb, F. Müller, I. Ortseifer, I. Vogler, E. Godehardt, S. Attig, R. Rae, A. Breitkreuz, C. Tolliver, M. Suchan, G. Martic, A. Hohberger, P. Sorn, J. Diekmann, J. Ciesla, O. Waksmann, A.-K. Brück, M. Witt, M. Zillgen, A. Rothermel, B. Kasemann, D. Langer, S. Bolte, M. Diken, S. Kreiter, R. Nemecek, C. Gebhardt, S. Grabbe, C. Höller, J. Utikal, C. Huber, C. Loquai, Ö. Türeci, Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer. *Nature*. **547**, 222–226 (2017).

29. P. A. Ott, Z. Hu, D. B. Keskin, S. A. Shukla, J. Sun, D. J. Bozym, W. Zhang, A. Luoma, A. Giobbie-Hurder, L. Peter, C. Chen, O. Olive, T. A. Carter, S. Li, D. J. Lieb, T. Eisenhaure, E. Gjini, J. Stevens, W. J. Lane, I. Javeri, K. Nellaiappan, A. M. Salazar, H. Daley, M. Seaman, E. I. Buchbinder, C. H. Yoon, M. Harden, N. Lennon, S. Gabriel, S. J. Rodig, D. H. Barouch, J. C. Aster, G. Getz, K. Wucherpfennig, D. Neuberg, J. Ritz, E. S. Lander, E. F. Fritsch, N. Hacohen, C. J. Wu, An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature*. **547**, 217–221 (2017).

30. D. B. Keskin, A. J. Anandappa, J. Sun, I. Tirosh, N. D. Mathewson, S. Li, G. Oliveira, A. Giobbie-Hurder, K. Felt, E. Gjini, S. A. Shukla, Z. Hu, L. Li, P. M. Le, R. L. Allesøe, A. R. Richman, M. S. Kowalczyk, S. Abdelrahman, J. E. Geduldig, S. Charbonneau, K. Pelton, J. B. Iorgulescu, L. Elagina, W. Zhang, O. Olive, C. McCluskey, L. R. Olsen, J. Stevens, W. J. Lane, A. M. Salazar, H. Daley, P. Y. Wen, E. A. Chiocca, M. Harden, N. J. Lennon, S. Gabriel, G. Getz, E. S. Lander, A. Regev, J. Ritz, D. Neuberg, S. J. Rodig, K. L. Ligon, M. L. Suvà, K. W. Wucherpfennig, N. Hacohen, E. F. Fritsch, K. J. Livak, P. A. Ott, C. J. Wu, D. A. Reardon, Neoantigen vaccine generates intratumoral T cell responses in phase Ib glioblastoma trial. *Nature*. **565**, 234–239 (2019).

31. N. Hilf, S. Kuttruff-Coqui, K. Frenzel, V. Bukur, S. Stevanović, C. Gouttefangeas, M. Platten, G. Tabatabai, V. Dutoit, S. H. van der Burg, P. Thor Straten, F. Martínez-Ricarte, B. Ponsati, H. Okada, U. Lassen, A. Admon, C. H. Ottensmeier, A. Ulges, S. Kreiter, A. von Deimling, M. Skardelly, D. Migliorini, J. R. Kroep, M. Idorn, J. Rodon, J. Piró, H. S. Poulsen, B. Shraibman, K. McCann, R. Mendrzyk, M. Löwer, M. Stieglbauer, C. M. Britten, D. Capper, M. J. P. Welters, J. Sahuquillo, K. Kiesel, E. Derhovanessian, E. Rusch, L. Bunse, C. Song, S. Heesch, C. Wagner, A. Kemmer-Brück, J. Ludwig, J. C. Castle, O. Schoor, A. D. Tadmor, E. Green, J. Fritsche, M. Meyer, N. Pawlowski, S. Dorner, F. Hoffgaard, B. Rössler, D. Maurer, T. Weinschenk, C. Reinhardt, C. Huber, H.-G. Rammensee, H. Singh-Jasuja, U. Sahin, P.-Y. Dietrich, W. Wick, Actively personalized vaccination trial for newly diagnosed glioblastoma. *Nature*. **565**, 240–245 (2019).

32. N. Riaz, J. J. Havel, V. Makarov, A. Desrichard, W. J. Urba, J. S. Sims, F. S. Hodi, S. Martín-Algarra, R. Mandal, W. H. Sharfman, S. Bhatia, W.-J. Hwu, T. F. Gajewski, C. L. Slingluff Jr, D. Chowell, S. M. Kendall, H. Chang, R. Shah, F. Kuo, L. G. T. Morris, J.-W. Sidhom, J. P. Schneck, C. E. Horak, N. Weinhold, T. A. Chan, Tumor and Microenvironment Evolution during Immunotherapy with Nivolumab. *Cell*. **171**, 934–949.e16 (2017).

33. D. Liu, B. Schilling, D. Liu, A. Sucker, E. Livingstone, L. Jerby-Arnon, L. Zimmer, R. Gutzmer, I. Satzger, C. Loquai, S. Grabbe, N. Vokes, C. A. Margolis, J. Conway, M. X. He, H. Elmarakeby, F. Dietlein, D. Miao, A. Tracy, H. Gogas, S. M. Goldinger, J. Utikal, C. U. Blank, R. Rauschenberg, D. von Bubnoff, A. Krackhardt, B. Weide, S. Haferkamp, F. Kiecker, B. Izar, L. Garraway, A. Regev, K. Flaherty, A. Paschen, E. M. Van Allen, D. Schadendorf, Integrative molecular and clinical modeling of clinical outcomes to PD1 blockade in patients with metastatic melanoma. *Nat. Med.* **25**, 1916–1927 (2019).

34. E. M. Van Allen, D. Miao, B. Schilling, S. A. Shukla, C. Blank, L. Zimmer, A. Sucker, U. Hillen, M. H. G. Foppen, S. M. Goldinger, J. Utikal, J. C. Hassel, B. Weide, K. C. Kaehler, C. Loquai, P. Mohr, R. Gutzmer, R. Dummer, S. Gabriel, C. J. Wu, D. Schadendorf, L. A. Garraway, Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. *Science*. **350**, 207–211 (2015).

35. R. Marty, S. Kaabinejadian, D. Rossell, M. J. Slifker, J. van de Haar, H. B. Engin, N. de Prisco, T. Ideker, W. H. Hildebrand, J. Font-Burgada, H. Carter, MHC-I Genotype Restricts the Oncogenic Mutational Landscape. *Cell*. **171**, 1272–1283.e15 (2017).

36. A. M. Goodman, A. Castro, R. M. Pyke, R. Okamura, S. Kato, P. Riviere, G. Frampton, E. Sokol, X. Zhang, E. D. Ball, H. Carter, R. Kurzrock, MHC-I genotype and tumor mutational burden predict response to immunotherapy. *Genome Med.* **12**, 45 (2020).

37. H. Ledford, Cocktails for cancer with a measure of immunotherapy. *Nature*. **532**, 162–164 (2016).

38. J. W. Yewdell, J. R. Bennink, Immunodominance in major histocompatibility complex class I-restricted T lymphocyte responses. *Annu. Rev. Immunol.* **17**, 51–88 (1999).

39. R. Wubbolts, R. S. Leckie, P. T. M. Veenhuizen, G. Schwarzmann, W. Möbius, J. Hoernschemeyer, J.-W. Slot, H. J. Geuze, W. Stoorvogel, Proteomic and biochemical analyses of human B cell-derived exosomes. Potential implications for their function and multivesicular body formation. *J. Biol. Chem.* **278**, 10963–10972 (2003).

40. M. Colombo, G. Raposo, C. Théry, Biogenesis, secretion, and intercellular interactions of exosomes and other extracellular vesicles. *Annu. Rev. Cell Dev. Biol.* **30**, 255–289 (2014).

41. C. Kurts, B. W. S. Robinson, P. A. Knolle, Cross-priming in health and disease. *Nat. Rev. Immunol.* **10**, 403–414 (2010).

42. Z. Garcia, E. Pradelli, S. Celli, H. Beuneu, A. Simon, P. Bousso, Competition for antigen determines the stability of T cell-dendritic cell interactions during clonal expansion. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 4553–4558 (2007).

43. N. Kedia-Mehta, D. K. Finlay, Competition for nutrients and its role in controlling immune responses. *Nat. Commun.* **10**, 2123 (2019).

44. N. Malandro, S. Budhu, N. F. Kuhn, C. Liu, J. T. Murphy, C. Cortez, H. Zhong, X. Yang, G. Rizzuto, G. Altan-Bonnet, T. Merghoub, J. D. Wolchok, Clonal Abundance of Tumor-Specific CD4(+) T Cells Potentiates Efficacy and Alters Susceptibility to Exhaustion. *Immunity*. **44**, 179–193 (2016).

45. D. Rimoldi, K. Muehlethaler, S. Salvi, D. Valmori, P. Romero, J. C. Cerottini, F. Levy, Subcellular localization of the melanoma-associated protein Melan-AMART-1 influences the processing of its HLA-A2-restricted epitope. *J. Biol. Chem.* **276**, 43189–43196 (2001).

46. X. Wang, S. Li, Protein mislocalization: mechanisms, functions and clinical applications in cancer. *Biochim. Biophys. Acta*. **1846**, 13–25 (2014).

47. K. Laurila, M. Vihinen, Prediction of disease-related mutations affecting protein localization. *BMC Genomics*. **10**, 122 (2009).

48. H.-T. Tzeng, Y.-C. Wang, Rab-mediated vesicle trafficking in cancer. *J. Biomed. Sci.* **23**, 70 (2016).

49. D. Matheoud, A. Sugiura, A. Bellemare-Pelletier, A. Laplante, C. Rondeau, M. Chemali, A. Fazel, J. J. Bergeron, L.-E. Trudeau, Y. Burelle, E. Gagnon, H. M. McBride, M. Desjardins,

Parkinson's Disease-Related Proteins PINK1 and Parkin Repress Mitochondrial Antigen Presentation. *Cell*. **166**, 314–327 (2016).

50. P. Gaudet, C. Dessimoz, in *The Gene Ontology Handbook*, C. Dessimoz, N. Škunca, Eds. (Springer New York, New York, NY, 2017), pp. 189–205.

51. P. J. Thul, L. Åkesson, M. Wiking, D. Mahdessian, A. Geladaki, H. Ait Blal, T. Alm, A. Asplund, L. Björk, L. M. Breckels, A. Bäckström, F. Danielsson, L. Fagerberg, J. Fall, L. Gatto, C. Gnann, S. Hober, M. Hjelmare, F. Johansson, S. Lee, C. Lindskog, J. Mulder, C. M. Mulvey, P. Nilsson, P. Oksvold, J. Rockberg, R. Schutten, J. M. Schwenk, Å. Sivertsson, E. Sjöstedt, M. Skogs, C. Stadler, D. P. Sullivan, H. Tegel, C. Winsnes, C. Zhang, M. Zwahlen, A. Mardinoglu, F. Pontén, K. von Feilitzen, K. S. Lilley, M. Uhlén, E. Lundberg, A subcellular map of the human proteome. *Science*. **356** (2017), doi:10.1126/science.aal3321.

52. F. R. Mehrabadi, K. L. Marie, E. Pérez-Guijarro, S. Malikić, E. S. Azer, H. H. Yang, C. Kızılkale, C. Gruen, W. Robinson, H. Liu, M. C. Kelly, C. Marcelus, S. Burkett, A. Buluç, F. Ergün, M. P. Lee, G. Merlino, C.-P. Day, S. Cenk Sahinalp, Profiles of expressed mutations in single cells reveal subclonal expansion patterns and therapeutic impact of intratumor heterogeneity. *bioRxiv* (2021), p. 2021.03.26.437185.

53. D. V. Klopfenstein, L. Zhang, B. S. Pedersen, F. Ramírez, A. W. Vesztrocy, A. Naldi, C. J. Mungall, J. M. Yunes, O. Botvinnik, M. Weigel, W. Dampier, C. Dessimoz, P. Flick, H. Tang, GOATOOLS: A Python library for Gene Ontology analyses. *Sci. Rep.* **8**, 1–17 (2018).

54. S. Kaabinejadian, C. Barra, B. Alvarez, H. Yari, W. H. Hildebrand, M. Nielsen, Accurate MHC Motif Deconvolution of Immunopeptidomics Data Reveals a Significant Contribution of DRB3, 4 and 5 to the Total DR Immunopeptidome. *Front. Immunol.* **13** (2022), doi:10.3389/fimmu.2022.835454.

55. A. W. Purcell, S. H. Ramarathinam, N. Ternette, Mass spectrometry-based identification of MHC-bound peptides for immunopeptidomics. *Nat. Protoc.* **14**, 1687–1707 (2019).

56. L. McInnes, J. Healy, J. Melville, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv [stat.ML]* (2018), (available at http://arxiv.org/abs/1802.03426).