

UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

Filling in the Gaps: Computational Tools, Transformer-Based Frameworks, and Integrative Analyses to Advance Complex Disease Risk Prediction

Permalink

<https://escholarship.org/uc/item/2rp6z7j6>

Author

Talwar, James Vishnu

Publication Date

2024

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA SAN DIEGO

**Filling in the Gaps: Computational Tools, Transformer-Based Frameworks, and
Integrative Analyses to Advance Complex Disease Risk Prediction**

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

in

Bioinformatics and Systems Biology

by

James Vishnu Talwar

Committee in charge:

Hannah Carter, Chair
Melissa Gymrek, Co-Chair
Garrison Cottrell
Kathleen Curtius
Maurizio Zanetti

2024

Copyright

James Vishnu Talwar, 2024

All rights reserved.

The dissertation of James Vishnu Talwar is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2024

DEDICATION

To the midnight riders and late-night risers,
The castaway fighters and extreme outliers;

To my family and friends, whose support made this pursuit possible;

And to Dada and Dadi, I hope I would have made you proud.

EPIGRAPH

To see a World in a Grain of Sand
And a Heaven in a Wild Flower,
Hold Infinity in the palm of your hand
And Eternity in an hour.

- *William Blake*

TABLE OF CONTENTS

Dissertation Approval Page	iii
Dedication	iv
Epigraph	v
Table of Contents	vi
List of Figures	ix
List of Tables	xi
Acknowledgements	xii
Vita	xvii
Abstract of the Dissertation	xix
Introduction	1
Chapter 1 GRIEVOUS: your command-line general for resolving cross-dataset genotype inconsistencies	3
1.1 Abstract	3
1.1.1 Summary	3
1.1.2 Availability and Implementation	3
1.2 Introduction	4
1.3 Material and Methods	5
1.3.1 Datasets	5
1.3.2 Identifying Valid Variants	6
1.3.3 GRIEVOUS: design and framework	6
1.4 Results	8
1.5 Conclusion	11
1.6 Ancillary Details	11
1.6.1 Author Contributions	11
1.6.2 Data and Code Availability	11
1.6.3 Conflicts of Interest	11
1.6.4 Acknowledgements	12
Chapter 2 VADeR's On That SNP: A New Hope For Transforming Polygenic Risk Scores And Revealing Genetic Heterogeneity	14
2.1 Abstract	14
2.2 Introduction	15
2.3 Material and Methods	18
2.3.1 Datasets	18
2.3.2 Formulating Feature Sets	19
2.3.3 VADeR Architecture	21

2.3.4	Training VADeR	22
2.3.5	Baselines.....	23
2.3.6	DARTH Patchtribution.....	25
2.3.7	Statistical Analyses	27
2.4	Results.....	27
2.4.1	Evaluating VADeR's Prognostic Potential	27
2.4.2	Linking VADeR Attributable Risk to Loci with DARTH Scores	30
2.4.3	Empirical Validation of DARTH Scores and Regional Risk Attribution...	40
2.5	Discussion	44
2.6	Ancillary Details	47
2.6.1	Author Contributions.....	47
2.6.2	Conflicts of Interest	47
2.6.3	Acknowledgements	48
Chapter 3	Autoimmune alleles at the major histocompatibility locus modify melanoma susceptibility	49
3.1	Abstract	49
3.2	Introduction	50
3.3	Material and Methods	52
3.3.1	Datasets	52
3.3.2	Quantifying immune infiltration	54
3.3.3	Identifying driver mutations	54
3.3.4	Identification and differential expression of conserved antigens	55
3.3.5	Predicting binding affinities	56
3.3.6	HLA population allele representations	57
3.3.7	PRS implementation	58
3.3.8	Multistage carcinogenesis model for melanoma.....	58
3.3.9	Personalizing sojourn time estimates from clock-like mutational signatures	59
3.3.10	Calculating HLA loss of heterozygosity	60
3.3.11	Statistical analyses	61
3.4	Results.....	61
3.4.1	MHC-I autoimmune alleles associate with a later age of diagnosis in melanoma.....	61
3.4.2	MHC-I autoimmune alleles modify melanoma risk	66
3.4.3	MHC-I autoimmune alleles associate with delayed onset of melanoma....	70
3.4.4	Investigating mechanisms of MHC-I autoimmune allele protection	73
3.5	Discussion	81
3.6	Ancillary Details	85
3.6.1	Author Contributions.....	85
3.6.2	Data and Code availability	85
3.6.3	Conflicts of Interest	86
3.6.4	Acknowledgements	86
Epilogue		88
4.1	GRIEVOUS: Conclusions, Limitations, and Future Directions	88
4.2	VADeR: Conclusions, Limitations, and Future Directions	89

4.3 MHC-I Autoimmune Utility from Liability (MAUL): Conclusions, Limitations, and Future Directions	93
Appendix A Supplemental Material for Chapter 1	95
A.1 Supplementary Figures	95
A.2 Supplementary Tables	97
Appendix B Supplemental Material for Chapter 2	98
B.1 Supplementary Figures	98
Appendix C Supplemental Material for Chapter 3	101
C.1 Supplementary Figures	101
Bibliography	121

LIST OF FIGURES

Figure 1.1.	GRIEVOUS: An overview	7
Figure 2.1.	VADeR model overview	28
Figure 2.2.	Comparing average attributions alignments between classical PRS and VADeR with DARTH scores	36
Figure 2.3.	Identifying genetic substructure driving predicted risk from DARTH scores	39
Figure 2.4.	Quantifying patch contributions to VADeR predicted disease risk through patch genotype shuffling	42
Figure 3.1.	Effect of MHC-I AI alleles on melanoma age at diagnosis and immune infiltration	65
Figure 3.2.	Contextualization and quantification of MHC-I AI alleles' effect on melanoma risk	68
Figure 3.3.	Identification and coverage of melanoma driver mutations including BRAF p.Val600Glu and its impact on age at diagnosis and immune infiltration ...	75
Figure 3.4.	Conserved antigen coverage characterization including HLA LOH effects ...	78
Figure A.1.	GRIEVOUS <i>realign</i> : The three-step workflow	96
Figure B.1.	Across the (UMAP) dimensions: Projecting personalized patch-specific DARTH scores	99
Figure C.1.	Distributions of stably expressed gene scores in GTEx	102
Figure C.2.	Identifying regions of conserved antigens for which autoimmune alleles exhibit improved predicted binding	103
Figure C.3.	PRS SNP minor allele frequencies across melanoma datasets	104
Figure C.4.	Multistage carcinogenesis melanoma modeling	105
Figure C.5.	Covariates and associations with melanoma in the TCGA	106
Figure C.6.	Effect of omitting each of the 7 individual autoimmune alleles from carrier status in TCGA	107
Figure C.7.	Effect of HLA supertype on age at diagnosis in TCGA	109
Figure C.8.	Regression modeling of autoimmune allele effects on melanoma age of diagnosis in TCGA	110

Figure C.9.	Melanoma external validation set exhibits significant sex-specific differences in age of diagnosis	111
Figure C.10.	HLA-B*27:02 associates with earlier age of diagnosis across datasets	112
Figure C.11.	PRS and MHC-I autoimmune alleles exhibit independent effects with age ..	113
Figure C.12.	Quantifying melanoma sojourn time in TCGA	114
Figure C.13.	Driver mutation and conserved antigen coverage for HLA population allele representations	116
Figure C.14.	Predicted median driver mutation rank scores for both common and autoimmune MHC-I alleles	117
Figure C.15.	CD8+ T-cell infiltration levels in TCGA primary melanoma tumors	118
Figure C.16.	Distributions of the predicted binding repertoire of conserved antigens between MHC-I autoimmune and common alleles	119
Figure C.17.	Extended: Identifying regions of conserved antigens for which autoimmune alleles exhibit improved predicted binding	120

LIST OF TABLES

Table 1.1.	Summary of datasets and GRIEVOUS realignment results	10
Table 2.1.	Summary of VADeR model hyperparameters.	22
Table 2.2.	Test set performances of VADeR and baseline models across multiple metrics	31
Table A.1.	GRIEVOUS rescues polygenic risk score AUC performances in prostate cancer datasets	97

ACKNOWLEDGEMENTS

The end of my Ph.D. feels like crossing the finish line of a marathon for which I didn't even know I was training. Defined by countless late nights that almost always blurred into early mornings, filled with frequent setbacks and those rare but cherished moments of success, this journey was unlike any prior academic pursuit. It required a fundamental shift in my learning paradigm, one that meant abandoning well-trodden paths of certainty and embracing the challenges of navigating ambiguity, exploration, and persistence. Success was no longer measured by immediate results but by the willingness to push through obstacles, adapt, and grow. Yet, this journey was far from solitary. At every step, I was supported, challenged, and inspired by those around me: mentors who shared their wisdom, peers who provided camaraderie, and loved ones who offered unwavering encouragement. To them, I owe a debt of gratitude that words can only begin to express.

First and foremost, I would like to express my deepest gratitude to my advisor, Hannah Carter. Five years ago, when faced with the constrained optimization problem of selecting Ph.D. students, you chose to take a chance on me over many other exceptional aspiring researchers. Thank you not only for the opportunity to embark on this journey, but also for your invaluable advice, guidance, patience, flexibility, generosity, and (hopefully) unwavering belief, all of which have been instrumental in helping me reach this point. Like Athena guiding Telemachus, you have skillfully steered me through uncharted waters, helping me navigate the complexities of this academic odyssey. I am profoundly grateful for the chance you gave me to chart a new research direction in the lab, exploring the intersection of deep learning and genomics at the cutting edge of science. It has been incredibly rewarding to witness how this work has grown since its InSNPtion, and I'm excited for its future. Stylistically, chronotypically, methodologically, and perhaps idiosyncratically, I suspect I've stood out among your students, though I hope in a way that has left an indelibly positive and unique mark. It has truly been an honor to be a part of your research family.

I would also like to extend my heartfelt thanks to the rest of my committee. To Gary Cottrell: as the fearless leader in deep learning, you ignited my passion for the field and provided guidance that launched critical avenues of my research. Your mentorship not only strengthened my

foundation in the field but also sparked the curiosity and skill set that continues to fuel my pursuit of innovative side projects, a fire that shows no signs of burning out. To Melissa Gymrek: your unwavering dedication as a committee member, combined with your enthusiasm and insightful suggestions, has been instrumental in shaping my research. Your Personal Genomics class remains one of the best at UCSD, and, like Gary, you played a central role in deepening my passion for this field. To Kit Curtius: I am deeply grateful for your exemplary feedback, positivity, and thoughtful input. Your collaboration, especially in incorporating custom mathematical models into our investigation of the MHC-I autoimmune cancer interplay, significantly strengthened our work and made it far richer. To Maurizio Zanetti: thank you for generously sharing your vast and unparalleled knowledge of the immune system. Your eager enthusiasm and engagement with my research, and even my broader interests, has been truly inspiring. And, of course, thank you for your invaluable restaurant recommendations; I'm sad to report I haven't had the chance to try them yet, but they are at the top of my post-graduation list.

Beyond my academic mentors, my family has been an unwavering source of strength and balance throughout this journey. I owe an immeasurable debt of gratitude to my parents, Keshav and Claire Talwar. Thank you for instilling in me a relentless drive, an insatiable curiosity, a competitive spirit, and a foundation of academic excellence. Beyond that, I am deeply grateful for the oasis of reprieve you created; coming home always helped me reset and recharge, ready to face the next challenge in what often felt like an interminable journey. To my sister, Anna-Maria: thank you for always keeping me grounded, never letting me get too high or too low, and for knowing exactly when a well-timed caffeinated gift was in order. Your unwavering sense of responsibility and your remarkable (visual-spatial) talents continue to inspire me. To my brothers, John-David and Cai-Matthew: thank you for embracing our never-ending scoreboard of athletics, sports, and general game shenanigans. To an outsider, these moments might seem trivial, but they offered me invaluable escapes to reconnect with a sense of fun and lightheartedness, bringing much-needed balance to my Ph.D. journey.

I would also like to express my heartfelt gratitude to my uncles, Sanjiv and Kapil Talwar, for your unwavering guidance, immense generosity, and steadfast mentorship throughout my life. Your kindness and constant support have been a source of strength for me, and no words

can fully capture how grateful I am for everything you have done. To the rest of my Aussie family – my aunt Karen Byrne and my cousins, Rishikesh and Indra Talwar - thank you for opening your home to me and allowing me to stay for an extended period during a particularly challenging time before my Ph.D. Your kindness, tireless hospitality, and constant encouragement was invaluable in my recovery from a prolonged and puzzling autoimmune illness, ultimately enabling and motivating me to pursue this journey.

To my Ph.D. family in the Carter Lab: I am deeply grateful for the friendships, mentorship, and camaraderie that have shaped my journey. To Michelle Dow: thank you for your daily positivity during the early days of my Ph.D. Stopping by your office was always a pleasure and never failed to brighten my day, even when you claimed to be grumpy. To Andrea Castro, my BISB buddy and mentor: you were an incredible resource both in research and beyond, always willing to answer my questions, share your insights, and uncover opportunities that I never would have noticed. Your generosity with your time and knowledge made a lasting impact, and I'm grateful we had the chance to be both colleagues and friends. To Meghana Pagadala: thank you for your collaboration and support. Your genuine interest in all aspects of my life, patience with my endless curiosity, and enthusiastic contributions across our projects made the journey not only more productive but also more enjoyable. I'm truly honored to have worked alongside you and to call you a friend. Finally, to Clarence Mah and honorary Carter Lab alum Michelle Ragsac: we started this journey together. Whether chatting about research, science, or logistics in our (windowless) office or catching up outside of UCSD, your friendship helped make this experience immeasurably better.

To my Ph.D. brothers: Adam (Bobby) Klie and Benjamin Rico (III): thank you for making this journey not just manageable but memorable. Bobby, whether in the office, the apartment, or at the gym, thank you for indulging my eternal willingness to “roll the dice.” Whether we were pushing each other in the gym, the pool, the track (or lunge equivalent), or tackling research and classes, every interaction, no matter how punishing, was a highlight of my day, even, or especially when we might have overdone it. Rico, thank you for matching and indulging my random, repetitive quoting, for the late-night gaming sessions, polar bear ocean swims, basketball grenades, and countless workouts. Your positivity and equanimity, even in the

face of adversity, always brightened my day without fail. To both of you, thank you for the blend of ridiculous and mundane moments we shared at the Fleet House, from uproarious laughter over absurd antics to quiet evenings that grounded the chaos of the Ph.D. life. It was a privilege and an honor to share this journey with you, and the bonds we forged in the crucible of our shared Ph.D. experiences will remain among the most treasured parts of my life.

There are so many others without whom this journey would not have been possible. While space prevents me from naming everyone, please know that even if you're not mentioned here, your support and impact have been deeply appreciated. To Gibraan Rahman: from battling rattlesnakes (yes, that really happened) during our first year to tackling classes and research together, thank you for always being a friend and for enjoying my endless recounting of crazy encounters. To Veronica Gong: I've lost count of the number of injuries I sustained throughout my Ph.D., but without fail, you helped me recover fully from every single one. Thank you for everything. To Willy Markuske: your time as our cluster wizard was chronically underappreciated, but your dedication to maintaining, upgrading, and troubleshooting the cluster made this journey so much smoother. I can't thank you enough for your invaluable efforts. To the rest of the Carter Lab members and alumni not named above, and to a host of other professors: thank you for making this experience such a wonderfully enriching journey. Finally, to the NY crew: thank you for your friendship and for being an outlet outside of research. I know I missed more than a few trips and weddings, but I'm looking forward to catching up and reconnecting on the other side.

The content of Chapter 1, in full, is a reformatted reprint of the material as it appears in "GRIEVOUS: your command-line general for resolving cross-dataset genotype inconsistencies" in *Bioinformatics*, 2024 by James V. Talwar, Adam Klie, Meghana S. Pagadala, and Hannah Carter. The dissertation author was the primary investigator and author of this paper.

The content of Chapter 2, in full, is a reformatted reprint of the material currently being prepared for submission for publication as "VADeR's On That SNP: A New Hope For Transforming Polygenic Risk Scores And Revealing Genetic Heterogeneity" by James V. Talwar, Adam Klie, Meghana S. Pagadala, Gil Pasternak, Brent Rose, Tyler M. Seibert, and Hannah Carter. The dissertation author was the primary investigator and author of this paper.

The content of Chapter 3, in full, is a reformatted reprint of the material as it appears in

“Autoimmune alleles at the major histocompatibility locus modify melanoma susceptibility” in *The American Journal of Human Genetics*, 2023 by James V. Talwar, David Laub, Meghana S. Pagadala, Andrea Castro, McKenna Lewis, Georg E. Luebeck, Bryan R. Gorman, Cuiping Pan, Frederick N. Dong, Kyriacos Markianos, Craig C. Teerlink, Julie Lynch, Richard Hauger, Saiju Pyarajan, Philip S. Tsao, Gerald P. Morris, Rany M. Salem, Wesley K. Thompson, Kit Curtius, Maurizio Zanetti, and Hannah Carter. The dissertation author was the primary investigator and author of this paper.

VITA

- 2014 University of Pennsylvania
 Bachelor of Science in Engineering, Bioengineering
- 2023 University of California San Diego
 Master of Science, Computer Science
- 2024 University of California San Diego
 Doctor of Philosophy, Bioinformatics and Systems Biology

PUBLICATIONS

James V. Talwar, Adam Klie, Meghana S. Pagadala, Gil Pasternak, Brent Rose, Tyler M. Seibert, Hannah Carter. “VADER’s On That SNP: A New Hope For Transforming Polygenic Risk Scores And Revealing Genetic Heterogeneity.”, , 2024.

James V. Talwar, Adam Klie, Meghana S Pagadala, Hannah Carter. “GRIEVOUS: your command-line general for resolving cross-dataset genotype inconsistencies.”, *Bioinformatics*, 2024.

Adam Klie, David Laub, **James V. Talwar**, Hayden Stites, Tobias Jores, Joe J. Solvason, Emma K. Farley, Hannah Carter. “Predictive analyses of regulatory sequences with EUGENE.”, *Nature Computational Science*, 2023.

Andrea Castro, Aaron M. Goodman, Zachary Rane, **James V. Talwar**, Garrett M. Frampton, Gerald P. Morris, Scott M. Lippman, Xinlian Zhang, Razelle Kurzrock, Hannah Carter. “Autoimmune HLA Alleles and Neoepitope Presentation Predict Post-Allogenic Transplant Relapse.”, *Journal of Immunotherapy and Precision Oncology*, 2023.

James V. Talwar, David Laub, Meghana S. Pagadala, Andrea Castro, McKenna Lewis, Georg E. Luebeck, Bryan R. Gorman, Cuiping Pan, Frederick N. Dong, Kyriacos Markianos, Craig C. Teerlink, Julie Lynch, Richard Hauger, Saiju Pyarajan, Philip S. Tsao, Gerald P. Morris, Rany M. Salem, Wesley K. Thompson, Kit Curtius, Maurizio Zanetti, Hannah Carter. “Autoimmune alleles at the major histocompatibility locus modify melanoma susceptibility.”, *The American Journal of Human Genetics*, 2023.

Meghana Pagadala, Timothy J. Sears, Victoria H. Wu, Eva Pérez-Guijarro, Hyo Kim, Andrea Castro, **James V. Talwar**, Cristian Gonzalez-Colin, Steven Cao, Benjamin J. Schmiedel, Shervin Goudarzi, Divya Kirani, Jessica Au, Tongwu Zhang, Teresa Landi, Rany M. Salem, Gerald P. Morris, Olivier Harismendy, Sandip Pravin Patel, Ludmil B. Alexandrov, Jill P. Mesirov, Maurizio Zanetti, Chi-Ping Day, Chun Chieh Fan, Wesley K. Thompson, Glenn Merlino, J. Silvio Gutkind, Pandurangan Vijayanand, Hannah Carter. “Germline modifiers of the tumor immune microenvironment implicate drivers of cancer risk and immunotherapy response.”, *Nature Communications*, 2023.

Meghana S. Pagadala, Joshua A. Linscott, **James V. Talwar**, Tyler M. Seibert, Brent Rose, Julie Lynch, Matthew Panizzon, Richard Hauger, Moritz H. Hansen, Jesse D. Sammon, Matthew H. Hayn, Karim Kader, Hannah Carter, Stephen T. Ryan. “PRState: Incorporating genetic ancestry in prostate cancer risk scores for men of African ancestry.”, *BMC cancer*, 2022.

ABSTRACT OF THE DISSERTATION

Filling in the Gaps: Computational Tools, Transformer-Based Frameworks, and Integrative Analyses to Advance Complex Disease Risk Prediction

by

James Vishnu Talwar

Doctor of Philosophy in Bioinformatics and Systems Biology

University of California San Diego, 2024

Hannah Carter, Chair
Melissa Gymrek, Co-Chair

We currently live in a state of perpetual genomic epistemic uncertainty, unaware globally of the distinct predispositions and liabilities driven by our underlying encoding, particularly for complex diseases. The immense dimensionality intrinsic to genetic data, coupled with limited availability of large, comprehensive datasets, introduces practical barriers hindering a complete understanding of genotype-to-phenotype relationships. Moreover, complex phenotypes present compounding challenges, given their polygenic nature and the contributions of environmental factors.

The advent of increasingly affordable genomic profiling technologies has led to a proliferation of genomic datasets, offering new avenues to explore the genetic basis of complex

diseases. Genome-wide association studies (GWASs) and polygenic risk scores (PRSs) have been instrumental in identifying and quantifying variant contributions to disease susceptibility. PRSs, though, despite their utility, are typically formulated in an additive manner that assumes conditional independence among variants, neglecting potential epistatic interactions that may impact phenotypic outcomes. Additionally, since PRSs largely rely on single nucleotide polymorphism (SNP) level data, they may fail to capture the influence of complex variants and highly polymorphic regions not adequately represented at this resolution.

My dissertation encompasses a multifaceted approach to address these limitations. First, to both enable the formation of large high-quality composite genomic datasets from individual studies and improve cross-dataset analysis portability, I developed GRIEVOUS, a tool for genomic data harmonization that resolves artifactual noise caused by differences in variant indexing and allele assignments across datasets. Second, I introduce VADeR, a Transformer-based architecture that combines techniques from both natural language processing and computer vision to generate genomic risk scores and reveal genetic heterogeneity, demonstrated in a proof-of-concept study in prostate cancer. Finally, I investigate the influence of highly polymorphic regions on complex disease risk by assessing the role of specific HLA genes in melanoma susceptibility. Overall, this work contributes new tools, methods, and insights that advance the detection of genetic contributions to complex diseases.

Introduction

“Mind the gap.” A mantra of train authorities globally surprisingly seems suited to describe a perennial problem in precision genomics. Numerous traits exhibit marked heritability, or the fraction of phenotypic variance attributable to genetic factors, yet the full genetic basis for this heritability has remained elusive. Significant variants as identified through genome-wide association studies (GWASs) explain only a small portion of complex disease heritability.^{1;2} While a number of methods have been developed to allow for an expansion from significant to all genotyped variants yielding a boost in genetic heritability attribution,^{3–5} the heritability gap still remains sizable.

Heritability focuses on how much phenotypic variation in a population can be explained by genetics, but for complex diseases we are particularly concerned with risk, or the likelihood of an individual to develop a condition given their genetics. The eponymous nature of these diseases is not only attributable to their polygenic nature, with multiple risk loci scattered across the genome, but also to environmental contributions. Polygenic risk scores (PRSs) attempt to disentangle genetic contributions and provide a metric of disease risk through a confluence of individual risk loci. In the absence of comprehensive causative environmental exposures, diagnostic prognostication may be off the table, but as a metric for screening, genetic risk plays a critical role, especially for cancer.

Survival rates drop precipitously as cancer stage increases,⁶ making early detection paramount. PRSs have been shown to capture risk independent from clinical features (e.g., family history, age, sex) across cancer types,^{7–9} yet, similar to missing heritability, these scores often fail to capture global genetic risk. Moreover they have clinically underwhelmed, exhibiting poor population generalization.^{10–12} To advance the frontier of genetic risk and improve its clinical utility we must ask the following question: Where are novel risk factors hiding?

This dissertation advances the quest of improving genomic-based risk prediction in complex diseases by addressing critical gaps and limitations across three distinct dimensions. Chapter 1 tackles a foundational issue capable of compromising genomic data integrity: the introduction of artifactual noise arising from discrepancies in variant indexing and allele assignments across datasets. These inconsistencies undermine both the quality of composite genomic datasets assembled from individual studies and the reliability of multi-dataset analyses. To resolve this issue, I developed a command-line tool, **GRIEVOUS**, to harmonize genomic data and eliminate these inconsistencies. **GRIEVOUS** enables the creation of high-quality datasets and enhances cross-dataset portability, laying a foundation for robust and reliable genomic analyses.

Chapter 2 reimagines PRS methodologies by addressing both their inherent assumptions and limitations in capturing the sequential and spatial complexities exhibited by the genome. To this end I propose a new deep learning architecture designed to capture these properties, the **V**ision **A**dapted **D**isease **E**lucidating **T**ransformer (**VADeR**), which leverages ideas from natural language processing embeddings and Vision Transformer patch partitioning to improve genotype-to-phenotype prediction. Additionally, I introduce a paired attention-based attribution metric **D**irected **A**ttention **R**elevance from **T**ransformer **H**euristics (**DARTH**) scores to provide an interpretable lens into **VADeR**'s predictions, which not only enhances model transparency, but also facilitates the identification of specific genomic regions driving predicted risk.

Chapter 3 turns the focus towards highly polymorphic genomic regions, specifically the Human Leukocyte Antigen (HLA) locus, a region inadequately captured by traditional PRS approaches. By investigating the role of autoimmune-predisposing HLA class I alleles in melanoma risk, this chapter reveals novel risk-relevant insights that enhance our understanding of complex disease susceptibility and immune dynamics.

Together, these contributions illuminate distinct yet interconnected avenues for advancing complex disease genomic-based risk prediction and emphasize the importance of bridging knowledge gaps to deepen our understanding of genetic contributions to disease.

Chapter 1

GRIEVOUS: your command-line general for resolving cross-dataset genotype inconsistencies

1.1 Abstract

1.1.1 Summary

Harmonizing variant indexing and allele assignments across datasets is crucial for data integrity in cross-dataset studies such as multi-cohort genome-wide association studies, meta-analyses, and the development, validation, and application of polygenic risk scores. Ensuring this indexing and allele consistency is a laborious, time-consuming, and error-prone process requiring a certain degree of computational proficiency. Here, we introduce GRIEVOUS, a command-line tool for cross-dataset variant homogenization. By means of an internal database and a custom indexing methodology, GRIEVOUS identifies, formats, and aligns all biallelic single nucleotide polymorphisms (SNPs) across all summary statistic and genotype files of interest. Upon completion of dataset harmonization, GRIEVOUS can also be used to extract the maximal set of biallelic SNPs common to all datasets.

1.1.2 Availability and Implementation

GRIEVOUS and all supporting documentation and tutorials can be found at <https://github.com/jvtalwar/GRIEVOUS>. It is freely and publicly available under the MIT license and can be installed via pip.

1.2 Introduction

The explosion of genetic variant datasets, driven by increasingly affordable genomic profiling technologies, presents exciting opportunities across a number of fields, ranging from precision medicine to ecology to agriculture. However, taking full advantage of genetic data from multiple sources can be challenging when datasets use different conventions for encoding variants. Analyses requiring the integration of various datasets range from the straightforward, such as validating genetic findings or polygenic scores in new datasets, to the more complex, such as cross-dataset joint statistical analyses (e.g., genome wide association meta-analyses^{13;14}, two-sample Mendelian randomizations^{15;16}). These approaches can frequently be undermined by discrepancies in the definition of identical variants across datasets.^{13;15}

Common variants, such as single nucleotide polymorphisms (SNPs), are abundantly shared across different datasets, and can be imputed against reference panels to increase coverage.^{17;18} Quality control of SNP data is critical, and best practices include setting uncertain genotypes to missing, orienting genotypes to the forward reference DNA strand, and filtering on missingness, minor allele frequency, or deviation from Hardy Weinberg equilibrium.^{18–21} However, even after performing the steps to generate a high-quality dataset, there can still be discrepancies between datasets due to differences in conventions used to define genotypes. If left uncorrected, these discrepancies can create illusory genotypic differences where none exist. Given that common variant sets can include millions of SNPs, it can be challenging and time consuming to manually identify and resolve discrepancies.

To reduce this burden to the researcher, we developed a Generalized Realignment of Innocuous and Essential Variants Otherwise Utilized as Skewed or GRIEVOUS, a command-line tool designed to ensure cross-cohort consistency and maximal feature recovery of biallelic SNPs, the most commonly used class of variants for genetic studies²¹. Whether creating a composite cohort from smaller studies for joint analyses, or ensuring feature fidelity for validation studies or polygenic score portability, GRIEVOUS reduces the problem of variant consistency and recovery to a simple streamlined set of commands.

1.3 Material and Methods

1.3.1 Datasets

Summary statistics and genotype data were obtained for two diseases: breast cancer (BC) and prostate cancer (PC). Specifically, data consisted of one summary statistic file and two different genotype datasets, each of which diverged in genotyping and imputation methodology from the other. Each file (i.e. both summary statistics and genotypes), for each condition studied here, indexed variants in a unique manner, capturing the discordance that can exist between datasets.

For BC, summary statistics were obtained from a large-scale BC GWAS study by Michailidou et al.²² (downloaded from: <https://bcac.ccge.medschl.cam.ac.uk/bcacdata/oncoarray/oncoarray-and-combined-summary-result/gwas-summary-results-breast-cancer-risk-2017>), and the genotype datasets utilized were from the Discovery, Biology, and Risk of Inherited Variants in Breast Cancer (DRIVE; dbGaP study accession: phs001265.v1.p1) project and the UK Biobank (UKBB). For PC, summary statistics from a large-scale PC multi-ancestry GWAS meta-analysis by Conti et al.²³ were downloaded from dbGaP (study accession: phs001120.v2.p1). Genomic data from the ELucidating Loci Involved in Prostate cancer Susceptibility (ELLIPSE; dbGaP study accession: phs001120.v1.p1) consortium and UKBB were utilized as the PC genotype datasets. PC PRSs were calculated as an effect size weighted genotype summation, with effect sizes and P-values for SNP subselection derived from the Conti et al.²³ summary statistics. Finally, we note that both DRIVE and ELLIPSE were subsets of each disease’s summary statistic GWAS.

Both DRIVE and ELLIPSE were genotyped using the OncoArray microarray,²⁴ while the UKBB was genotyped using the UK Biobank Axiom Array.²⁵ To recover untyped markers in both DRIVE and ELLIPSE, we imputed genotypes with Minimac4 using the 1000 Genomes Phase 3 Version 5 reference panel, via the Michigan Imputation Server.²⁶ Details of UKBB imputation can be found in the original UKBB report by Bycroft et al.²⁵

1.3.2 Identifying Valid Variants

To address the problem of variant index validity (i.e., ensuring variant indices are unique by genomic position), GRIEVOUS reformulates a given genomic dataset as a graph. Each unique variant genomic position is defined as a node with its corresponding edges assigned to all dataset-defined indices for that variant position. As the set of all dataset-defined indices must be unique by genomic position for a variant to be valid, all valid variants will only contain self-edges. However, all nodes existing in a connected component with a size of 2 or more, correspond to invalid variants and are removed by GRIEVOUS, as there exists at least one non-unique dataset-defined index spanning multiple genomic positions for each variant in the component (and thus a non-self edge connecting these nodes in the graph). This procedure is implemented during GRIEVOUS *realign*-ment and is illustrated in **Figure A.1** as *Extract Valid SNPs*.

1.3.3 GRIEVOUS: design and framework

The current state of variant indexing can be summarized concisely as unsystematic. Datasets can use any means or mechanisms for indexing variants, ranging from custom concatenations of dataset variant information (e.g. CHR_POS_REF_ALT), to rsIDs, to any mixture in between (**Figure 1.1A**). Arbitrary assignments of reference (REF) and alternate (ALT) definitions across datasets poses a similar problem (**Figure 1.1A**). Differing genotyping technologies, and in the case of summary statistics, the characterization of effect alleles, often lead to divergences in REF/ALT definitions across datasets. Harmonization of dataset indices and allele assignments, thus first requires variant organization, a task specifically addressed by the design of GRIEVOUS (**Figure 1.1B**).

GRIEVOUS deploys a flexible internal database, backed by chromosome-level parquet files, to build a unified variant index and align allele assignments. Initialized in an empty manner, GRIEVOUS databases are continuously updated with each run, adding all hitherto unobserved biallelic SNPs identified in a given dataset (**Figure 1.1B**). The iterative nature of database variant additions ensures that each subsequent dataset is consistently oriented with the set of common biallelic SNPs found in all previously GRIEVOUS realigned datasets.

The key command behind GRIEVOUS is *realign*, which consists of three steps: *clean*,

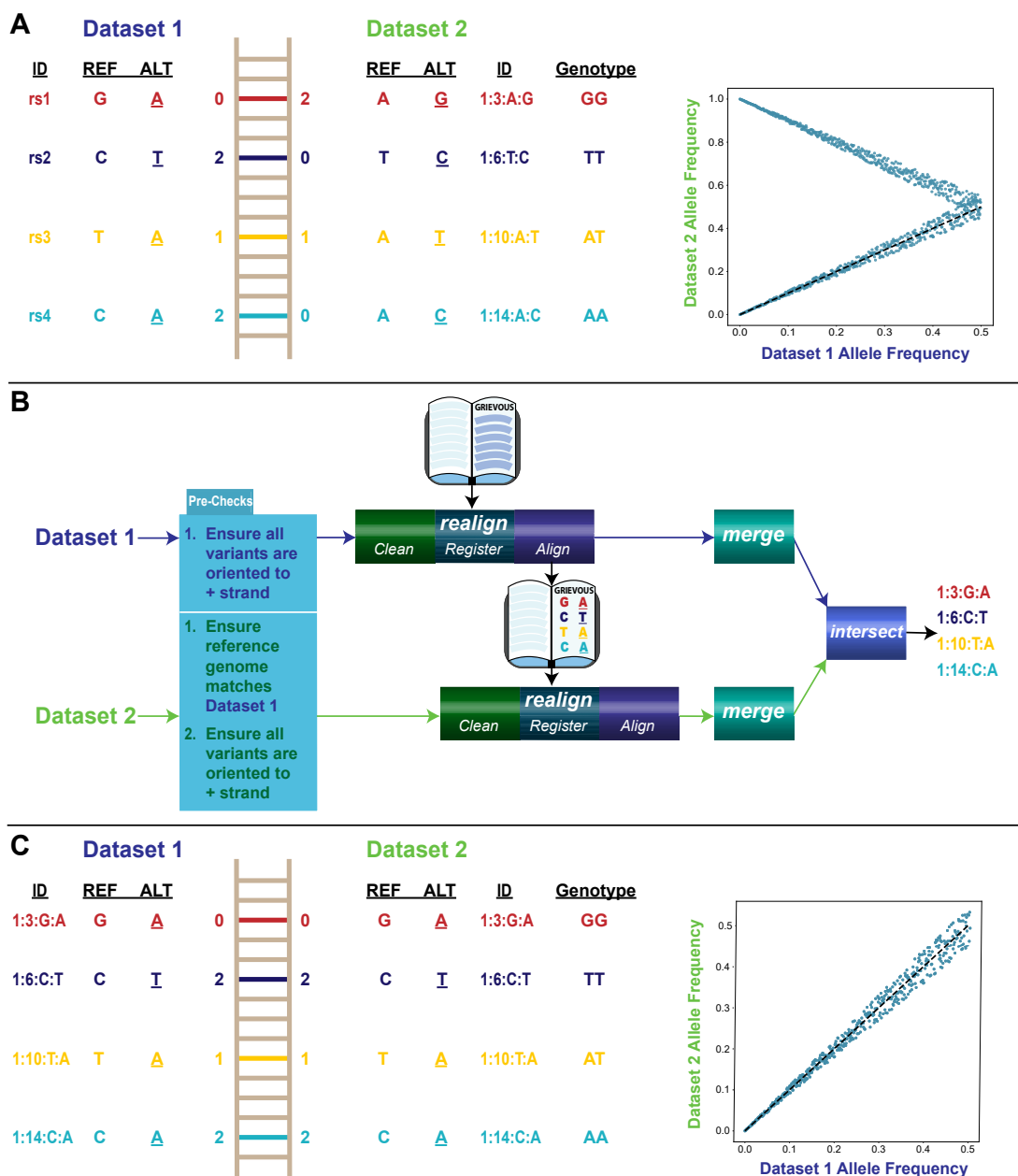


Figure 1.1. A. Example of the same genotypes described by two different genomic datasets that use different schemas for indexing and assigning variants. This complicates variant recovery and leads to spurious artifactual allele frequency divergences across datasets. B. Each genomic dataset is harmonized with GRIEVOUS. Upon ensuring all datasets use the same reference assembly and strand orientation, each dataset is passed through realign, sequentially, and then passed to merge to generate composite dataset level reports of all identified biallelic and inverted variants resulting from the GRIEVOUS realignment process. Finally, intersect is called once across all datasets, to identify the maximal set of biallelic SNPs common across all datasets. C. After harmonization with GRIEVOUS, all genomic datasets use the same schema of variant indexing, and all biallelic SNPs common across all datasets are consistently assigned, eliminating cross-dataset artifactual allele frequency divergences.

register, and *align* (**Figure A.1**). *Clean* identifies valid biallelic SNPs, by means of a graph formulation (Material and Methods: *Identifying Valid Variants*), and resolves duplications and multiple indexing issues. Specifically, variant positions are defined as nodes and dataset-specific IDs as edges. Any node for which there exists an edge to another node in the graph is considered an invalid variant, with a dataset-specific ID pointing to multiple genomic positions. *Register* performs variant indexing reformulation (i.e. a colon-separated unification of chromosome, position, REF, and ALT defined by the dataset; CHR:POS:REF:ALT) and database comparison. Dataset variants found in the reverse index assignment from a GRIEVOUS database are marked for realignment, while all variants not found in the database by genomic position are registered for addition. *Align* enacts the adjustment of SNP assignments, updates the database, and writes output files for the *realign* run. These files include reports of identified, reassigned, and duplicated variants, and for PLINK2 binary file inputs,^{27;28} reassignment files to convert the binary genomic data to the GRIEVOUS format.

GRIEVOUS can harmonize an arbitrary number of user-defined genomic datasets, whether they be PLINK2 binary files, summary statistics, or a mixture of the two (**Figure 1.1C**). Variants are processed at the chromosome-level, allowing parallelization and integration into different workflows. Additional commands, merge and intersect, combine chromosome-level reports and generate the maximal set of biallelic SNPs common to all datasets of interest, respectively (**Figure 1.1B**). GRIEVOUS also supports the creation of multiple databases for project-specific needs, enhancing parallel processing and organization.

1.4 Results

Polygenic risk scores (PRS) provide an example of an application where performance and portability depend on the fidelity of SNP assignments across datasets. They are traditionally calculated as an effect size weighted genotype summation, wherein each SNP’s effect size is multiplied by the number of corresponding alleles carried, and these products are summed across risk alleles. Thus, if differences exist in how genotypes are defined across datasets, the contribution of discrepant SNPs to the score could be calculated incorrectly, introducing noise to the PRS calculation and reducing score performance on the outcome of interest.

This problem can be avoided through careful harmonization of SNP consistency across datasets prior to PRS calculation. GRIEVOUS accomplishes this through indexing unification, assignment synchronization, and variant recovery across datasets. To demonstrate GRIEVOUS’ effectiveness, we applied it to harmonize summary statistics and genotype data for two different diseases: breast cancer (BC) and prostate cancer (PC; Material and methods: *Datasets*).

For each condition, GRIEVOUS *realign* was applied sequentially in a dataset consistent order: (i) summary statistics, (ii) DRIVE/ELLIPSE genotypes, (iii) UKBB genotypes. This process aligned the ALT allele in the genotyped datasets to the effect allele in the summary statistics for all biallelic variants, facilitating unbiased cross-disease comparison of SNP reassignments. The results are presented in **Table 1.1**.

Both ELLIPSE and DRIVE used different nomenclature for common variant indexing relative to the UKBB, a disparity that GRIEVOUS effectively managed. Post-reindexing, GRIEVOUS successfully recovered and reassigned all database-consistent biallelic SNPs across datasets. For BC, there was high concordance in variant assignment, likely because DRIVE was part of the BC GWAS. However, in PC, despite ELLIPSE being part of the PC GWAS, 43.4% of cross-dataset common variants differed in allele assignment from the summary statistics and required reassignment. These discrepancies drastically impacted PRS generalization in both PC genotype datasets. Specifically, PRSs applied to the unharmonized PC datasets exhibited an area under the receiver operating characteristic curve (AUC) below random chance. However, after harmonizing all datasets with GRIEVOUS, expected PRS performance was restored (**Table A.1**).

In summary, GRIEVOUS harmonized datasets and identified the shared set of biallelic SNPs for both diseases, demonstrating its utility in genomic dataset harmonization.

Table 1.1. Summary of datasets and GRIEVOUS realignment results. ^{SS} denotes summary statistics. * denotes the first dataset GRIEVOUS realigned for a condition, which by definition will reassign 0 SNPs. All GRIEVOUS realignments were deployed in parallel and employed multiple cores (summary statistics: 2; genotype files: 4).

Condition	Dataset	Total Number of Variants	Number of Biallelic SNPs Identified	Number of Biallelic SNPs Re-assigned	Average Run Time (Minutes)	Max Memory Usage (GB)	Cross-Dataset SNP Intersection Size
Breast Cancer	Michailidou et al. ^{SS}	11,792,542	10,413,027	0*	1.71 $\pm$ 1.08	3.00	10,295,993
	DRIVE	48,838,144	44,849,194	2	9.13 $\pm$ 4.03	9.97	
	UKBB	97,013,422	92,429,914	659	8.03 $\pm$ 3.69	14.77	
Prostate Cancer	Conti et al. ^{SS}	29,235,255	26,529,293	0*	4.66 $\pm$ 3.48	6.80	26,098,297
	ELLIPSE	48,899,508	44,905,050	11,329,170	7.22 $\pm$ 3.49	8.55	
	UKBB	97,013,422	92,424,521	11,374,523	10.42 $\pm$ 7.85	14.96	

1.5 Conclusion

The harmonization of variant indexing and assignments is crucial for the accuracy of multi-dataset studies of genetic variants such as the development, validation, and application of PRSs. Here we introduce GRIEVOUS, a command-line tool designed to simplify and expedite the process of harmonizing and recovering variants across different cohorts. GRIEVOUS efficiently unifies variant indices, and primarily focuses on the recovery of biallelic SNPs. Variants not fitting this category currently still require manual verification by the user to ensure allele assignment consistency. However, these could be supported in future releases. Before using GRIEVOUS, it is important for users to confirm that all datasets use the same reference assembly and strand orientation to ensure accurate results.

In summary, GRIEVOUS streamlines the time-consuming and meticulous process of SNP homogenization and reduces opportunities for human error.

1.6 Ancillary Details

1.6.1 Author Contributions

Original concept, J.V.T.; project supervision, H.C.; GRIEVOUS development, J.V.T.; data acquisition and processing, J.V.T., A.K., M.S.P.; GRIEVOUS analyses, J.V.T.; preparation of manuscript, J.V.T., H.C.

1.6.2 Data and Code Availability

In order to promote the widespread use of GRIEVOUS, we have made it freely available under the MIT license and easily installable via pip. The code for GRIEVOUS, which was implemented in Python, along with supporting documentation and tutorials, can be found on GitHub (<https://github.com/jvtalwar/GRIEVOUS>). The code for GRIEVOUS is also archived on Zenodo.²⁹

1.6.3 Conflicts of Interest

None declared.

1.6.4 Acknowledgements

This work was supported by National Institutes of Health (NIH) grant [R01CA269919 to H.C.] and NIH infrastructure grant [2P41GM103504-11].

The results shown here are in large part based upon data generated by the UKBB (Project #37671) and the following dbGaP studies: phs001265.v1.p1, phs001120.v2.p1, and phs001120.v1.p1. This research also used publicly available summary statistics from the Breast Cancer Association Consortium (available at: <https://bcac.ccge.medschl.cam.ac.uk/>). For phs001265.v1.p1 we acknowledge: OncoArray genotyping and phenotype data harmonization for the Discovery, Biology, and Risk of Inherited Variants in Breast Cancer (DRIVE) breast-cancer case control samples was supported by X01 HG007491 and U19 CA148065 and by Cancer Research UK (C1287/A16563). Genotyping was conducted by the Center for Inherited Disease Research (CIDR), Centre for Cancer Genetic Epidemiology, University of Cambridge, and the National Cancer Institute. The following studies contributed germline DNA from breast cancer cases and controls: the Two Sister Study (2SISTER), Breast Oncology Galicia Network (BREGAN), Copenhagen General Population Study (CGPS), Cancer Prevention Study 2 (CPSII), The European Prospective Investigation into Cancer and Nutrition (EPIC), Melbourne Collaborative Cohort Study (MCCS), Multiethnic Cohort (MEC), Nashville Breast Health Study (NBHS), Nurses Health Study (NHS), Nurses Health Study 2 (NHS2), Polish Breast Cancer Study (PBCS), Prostate Lung Colorectal and Ovarian Cancer Screening Trial (PLCO), Studies of Epidemiology and Risk Factors in Cancer Heredity (SEARCH), The Sister Study (SISTER), Swedish Mammographic Cohort (SMC), Women of African Ancestry Breast Cancer Study (WAABCS), Women’s Health Initiative (WHI). For phs001120 we acknowledge: Funding for the meta-analysis was provided by NIH grant U19CA148537. For de novo genotyping we would like to acknowledge the NCRN nurses and consultants for their work in the UKGPCS study. We thank all the patients who took part in this study. This work was supported by Cancer Research UK (grant numbers C5047/A7357, C1287/A10118, C1287/A5260, C5047/A3354, C5047/A10692, C16913/A6135 and C16913/A6835). We would also like to thank the following for funding support: Prostate Research Campaign UK (now Prostate Cancer UK), The Institute of Cancer Research and The Everyman Campaign, The National Cancer Research Network

UK, The National Cancer Research Institute (NCRI) UK. We are grateful for support of NIHR funding to the NIHR Biomedical Research Centre at The Institute of Cancer Research and The Royal Marsden NHS Foundation Trust. The MEC was supported by NIH grants CA63464, CA54281, and CA098758.

The content of Chapter 1, in full, is a reformatted reprint of the material as it appears in “GRIEVOUS: your command-line general for resolving cross-dataset genotype inconsistencies” in *Bioinformatics*, 2024 by James V. Talwar, Adam Klie, Meghana S. Pagadala, and Hannah Carter. The dissertation author was the primary investigator and author of this paper.

Chapter 2

VADeR’s On That SNP: A New Hope For Transforming Polygenic Risk Scores And Revealing Genetic Heterogeneity

2.1 Abstract

Polygenic risk scores (PRSs) serve as quantitative metrics of genetic liability for various conditions. Traditionally calculated as an effect size weighted genotype summation, this formulation assumes conditional feature independence and overlooks the potential for complex interactions among genetic variants. Transformers, a class of deep learning architectures known for capturing dependencies and interactions between features, have demonstrated remarkable predictive power across domains. In this work, we introduce VADeR, a Vision Transformer (ViT)-inspired architecture that combines techniques from both natural language processing and computer vision to capture properties exhibited by genetic data and effectively learn both local and global interactions for genotype-to-phenotype prediction. Evaluating VADeR’s performance in predicting prostate cancer risk, we found that across a range of metrics, including accuracy, average precision, and Matthews correlation coefficient, VADeR outperformed all benchmark methods demonstrating its effectiveness in the context of complex disease risk prediction. To illuminate identified drivers of disease risk by VADeR we leveraged attention-based attribution techniques to construct DARTH scores. These scores revealed distinct genetic heterogeneity captured by VADeR, with drivers of predicted risk identified in key prostate cancer risk regions including HOXB13, TMPRSS2, and MSMB loci. Overall, by effectively capturing dependencies among genetic variants and providing interpretable insights, VADeR and DARTH scores offer

a promising direction for advancing genotype-to-phenotype prediction, particularly in complex disease.

2.2 Introduction

Decrypting genomic contributions to complex phenotypes continues to endure as a formidable challenge, a solution to which currently exists beyond the scope of our predictive capabilities. In the absence of a comprehensive key to the genomic cipher, polygenic risk scores (PRSs), traditionally calculated as a genome-wide association study (GWAS) effect size weighted genotype summation, serve as approximation approaches, distilling genetic contributions from a confluence of individual loci to a metric of phenotype liability.¹⁰

Intrinsic to the genome are a unique set of complexities and modeling challenges, ones that closely parallel problems confronted in both natural language processing (NLP) and computer vision (CV). Much like in language, where sequence dependencies and contextual relationships between tokens define semantic meaning,³⁰ the genome exhibits sequential structures with long-range dependencies between variants.^{31–33} Simultaneously, similar to the properties of spatial locality and compositionality found in CV problems, where local patterns of pixels affect the global image structure,³⁴ genetic data presents non-uniform statistical dependencies across loci,^{21;35;36} requiring methods that account for both local and global interactions. These hybrid characteristics, spanning both sequential and spatial dependencies, require innovative approaches that can integrate the strengths of both NLP and CV techniques to model the intricate patterns and variations within the genome.

Transformers are a class of deep-learning architectures that have led to significant advancements in both of these domains.^{37–41} Modularly these models can be decomposed to an alternating series of residual connection joined (multi-headed-self) attention layers and feedforward networks.^{39;41} Attention provides a mechanism for updating representations from a capturing of pairwise similarity between any two elements in a sequence independent of distance, affording a powerful means of effectively capturing both short and long range dependencies. Awareness of the performance gains Transformer models can yield, coupled with recognition of the fundamental structural similarities the genome shares with language and images, has led

to the widespread experimentation and application of Transformers to problems in genomics modeling.^{31;42–44}

In the pursuit of directly disentangling genetic contributions to phenotypes and transcending the limitations of conditional independence assumed in both GWAS effect size estimation and classical additive PRS models, deep learning has been widely explored across a diverse range of datasets and phenotypic traits.^{45–49} The outcomes of such studies, many of which employ simple and shallow architectures,^{45;46;49} have been mixed; successes are frequently reported in the context of non-complex phenotypes, while results for complex traits remain less definitive.^{45;46;48;50} Moreover, reported performance improvements can be validated within the same dataset (partitioned through train/validation/test split) rather than through independent external validation, which may inadvertently overstate model generalizability.^{45;50–52} Feature selection, a mandatory prerequisite for reducing dimensionality to enable such modeling approaches, similarly has the potential to artifactually inflate performance if the properties by which features are selected are specific to the dataset.⁴⁵

Recent investigations utilizing Transformer architectures for genotype-to-phenotype prediction have reported promising results across a range of non-complex phenotypes, such as body mass index, systolic blood pressure, and lipid profiles.^{50;51} These successes, while encouraging, remain at the proof-of-concept stage, with validation performed on partitions of the same dataset rather than on independent external cohorts, which as mentioned may limit the generalizability of the findings. Moreover, these studies have not yet extended their methodologies to complex phenotypes or diseases, leaving a gap in understanding their potential broader applicability. Additionally, simplifying assumptions employed in formulating feature embeddings, such as arbitrarily grouping contiguous variants in the feature set to reach a constant dimension and enable linear projection,⁵⁰ can fail to capture the complexity of the genomic landscape. Without accounting for the variable spacing of variants within each group, employing a single linear projection effectively blurs the distinct representations of individual variants, failing to capture their specific contributions to phenotypic outcomes.

Independent of any reported performance gains deep learning approaches can afford in this space, barriers exist preventing their widespread adoption. Clinically, interpretability reigns

supreme, and classical additive summary-statistics-based PRS approaches (and their Bayesian shrinkage counterparts)^{36;53–56} exhibit the convenient property of inherent explainability.²¹ Deep learning methods, despite the theoretical performance advantages they might provide, exist in a space of opacity, as their parameter space (which can order on the order of hundreds of millions to billions) and recursive non-linearities preclude identification of feature contribution to prediction. This proves prohibitive to identifying underlying biases that may compromise model generalizability, risking steep declines in performance when encountering out-of-distribution data.⁵⁷ Moreover, extraction of novel biological insights identified by the model also remains a challenge. Ultimately, the absence of clear interpretability impedes trustworthiness, a critical factor in clinical contexts where transparency is a necessity.^{57;58}

To address these limitations, inherently explainable deep learning architectures which hierarchically structure their network topology according to a defined biological prior have been proposed.^{48;59} Though this affords an avenue to improved interpretability, it imposes a relational inductive bias on the network, similar to graph neural networks.⁶⁰ Inducing such biases can enhance performance, provided the imposed structure aligns well with the underlying data distribution. However, in the context of biological data, explicitly defining node connections inherently depends on existing knowledge, which can constrain a model’s capacity to capture relationships involving features with uncertain roles in the hierarchy. This reliance on predefined structure can also impose limitations, both by embedding the biases of those who design the architecture based on current biological understanding and by potentially overlooking aspects of the biology that remain undiscovered. Furthermore, establishing these hierarchies demands specialized domain expertise and necessitates custom architecture designs tailored to each modeling domain, limiting flexibility and scalability across diverse biological applications.

Altogether, these challenges highlight the nascent stage of deep learning, particularly Transformer-based architectures, as applied to problems in end-to-end genotype-to-phenotype prediction, especially for complex diseases. To advance the field forward we introduce the **V**ision **A**dapted **D**isease **E**lucidating **T**ransformer (VADeR), an architecture inspired by Vision Transformers (ViTs) that combines techniques from NLP embeddings and ViT patch partitioning to generate chromosomally consistent, region-specific representations. As proof-of-concept for

the viability of this architecture, we assessed its performance in prostate cancer (PC) prediction, ensuring independent test set evaluation. Across a range of metrics VADeR outperformed all benchmarks, including statistical genomics approaches with access to a much larger effective pool of information, demonstrating its effectiveness towards the task of complex disease risk prediction. Towards the goal of interpretability, we formulate **D**irected **A**ttention **R**elevance from **T**ransformer **H**euristics (DARTH) scores, an attention-based attribution metric capable of capturing personalized regional contributions driving predicted disease risk by VADeR. Analysis of these scores highlighted VADeR’s ability to capture genetic heterogeneity and identify key patterns of attributed PC risk, including the well-established HOXB13 and TMPRSS2 risk regions.^{61–64} The concordance of these interpretability findings, coupled with improved performance relative to baseline PRS methods despite limited training data relative to the number of parameters, underscores the potential of our joint DARTH-VADeR framework for both prediction and interpretability in genotype-to-phenotype modeling. These results suggest that this approach merits broader exploration across a range of genotype-to-phenotype prediction challenges.

2.3 Material and Methods

2.3.1 Datasets

The ELucidating Loci Involved in Prostate cancer Susceptibility (ELLIPSE) Consortium is a globally collaborative effort directed towards its namesake (i.e., the discovery of PC risk loci), that has made genotypes available for a large number of cases and controls available to researchers. From dbGaP (study accession: phs001120.v1.p1), we downloaded all available genotypes and corresponding phenotypes for ELLIPSE ($n = 91,644$; $n_{cases} = 54,283$; $n_{controls} = 37,361$), which were employed for formulating training and validation sets. Specifically, these training and validation sets corresponded to a random 80/20 split of ELLIPSE, ensuring consistency in the frequencies of case-control status, ancestry, and family history (for those for which it was available).

To infer ancestry, we computed the top 10 principal components with FlashPCA2⁶⁵ on 1000 Genomes and ELLIPSE individuals. These top principal components were subsequently employed as features for an ancestral k-means clustering, where fitting was performed on 1000

Genomes individuals (with each cluster corresponding to a specific ancestry). ELLIPSE ancestry was then assigned according to its nearest cluster. This procedure yielded a 99.2% agreement with self-reported ancestries in the ELLIPSE cohort. We observe that ELLIPSE exhibits a large overrepresentation of European ancestry individuals in ($n_{European} = 82,594$) corresponding to 90.1% of the dataset, despite ancestry being a documented risk factor for PC.^{7;66}

To evaluate model generalizability, we constructed a balanced independent test set ($n = 15,235$; $n_{cases} = 7,591$; $n_{controls} = 7,644$) from a subsampling of the UK Biobank (UKBB). Specifically, test set cases were defined as all male individuals of European ancestry in the UKBB with an ICD code specific to PC. Test set controls were randomly selected from all remaining male UKBB individuals of European ancestry without a cancer ICD code in an age-matched manner. Details of UKBB ancestry assignment can be found in the works of Richard et al.⁶⁷ and Pagadala et al.⁶⁸

We note that ELLIPSE and the UKBB differ in their genotyping and imputation methodologies. ELLIPSE was genotyped using the OncoArray microarray,²⁴ while the UKBB was genotyped using the UK Biobank Axiom Array.²⁵ For ELLIPSE, we imputed genotypes with Minimac 4 using the 1000 Genomes Phase 3 Version 5 reference panel, by means of the Michigan Imputation Server.²⁶ Quality control prior to ELLIPSE imputation is documented by Pagadala et al.⁶⁶ Full details of UKBB imputation can be found in the original UKBB report by Bycroft et al.²⁵ Following imputation, ELLIPSE and the UKBB contained genotypes for 48,899,508 and 97,013,422 variants respectively. All datasets and variants used in this study employed GRCh37 as the reference genome, and all coordinates reported in this paper follow this reference.

2.3.2 Formulating Feature Sets

Given the high-dimensionality of the feature space, strong correlation structure that exists between features (i.e., variants), and noise introduced due to genotyping and imputation error, using all shared variants across datasets poses several problems for any PRS methodology. Classical additive PRSs, commonly known as clumping + thresholding (C+T), narrow down the feature space by identifying and selecting the most significant genome-wide association study (GWAS) variants at a locus and omitting all less significant correlates.²¹ From this clumped

feature set, thresholding is applied across a range of GWAS p-values, further reducing the feature set and creating a candidate set of scores, with the final PRS selected based on its best performance in an independent dataset.

Rather than directly omitting features, a number of PRS methods instead leverage linkage disequilibrium (LD) to their advantage by reweighting SNP effects according to Bayesian shrinkage techniques^{36;53–56} yielding improved performances relative to C+T approaches. Though theoretically plausible that all variants could be used with these approaches, it is often recommended to constrain the feature space to HapMap3 variants. It is also worth noting that both classical and Bayesian PRSs are fundamentally dependent on base data or GWAS summary statistics.

While classical additive PRSs often underperform compared to more refined statistical genomics methods, for machine learning (ML) based approaches, feature selection through an analogous C+T technique offers a reasonable tradeoff between reducing dimensionality and maintaining predictive power, making it a practical strategy for handling complex genomic data. To identify significant variants and their effect sizes, not only for p-value-based feature selection in our VADeR and baseline ML models, but also to ensure the generation of both statistical genomics and classical PRSs, we downloaded summary statistics from dbGaP (study accession: phs001120.v2.p1) for a large-scale PC multi-ancestry GWAS meta-analysis by Conti et al.²³ It should be emphasized that ELLIPSE was a subset of this GWAS, but, perhaps more importantly, the UKBB was not, preventing any test set leakage during the feature selection phase.

Upon summary statistic acquisition, clumping was conducted in an independent dataset, the 1000 Genomes, with a radius of 125kb and an r^2 of 0.8. P-value thresholding across 5e-8, 5e-7, 5e-6, 5e-5, 5e-4 was subsequently applied to formulate candidate feature sets of 4,062, 5,231, 7,024, 10,389, and 12,017 variants, respectively. All variants employed for training, validating, and testing VADeR and baseline models passed classical quality control metrics including a minor allele frequency (MAF) exceeding 1% and an imputation quality score exceeding 0.8.²¹

Features for all models were encoded as alternate allele counts. Moreover, for VADeR and ML baseline models, these counts were standardized according to the mean and standard deviation of the training set distribution to improve training convergence. Finally, to ensure

maximal feature recovery and orientation consistency across all genomic datasets and summary statistics, all datasets were harmonized with GRIEVOUS.⁶⁹

2.3.3 VADer Architecture

Similar to Vision Transformers (ViTs),³⁹ we partition the genome into non-overlapping patches of length W , ensuring each patch is chromosomally consistent (i.e., all positions within a patch are located on the same chromosome, and do not overflow into another). All p-value significant variants in the feature set are then mapped onto these patches, with all strongly correlated features, as identified through the aforementioned clumping procedure (Material and Methods: *Formulating Feature Sets*), removed. Across p-value thresholds 5e-8, 5e-7, 5e-6, 5e-5, 5e-4, this procedure yielded 331, 436, 603, 1,008, and 1,204 genomic patches, respectively.

Given the imbalance in patch density, or more simply, the inconsistency in the number of variants across patches, mapping each genomic patch to a constant latent vector size D with a single trainable linear projection, as is classically done, is infeasible. To ameliorate this problem, VADer employs patch specific projections, with a trainable projection for each patch. While this increases the parameters relative to a single projection, it affords the theoretical advantage of a richer patch embedding at a region (essentially a patch tokenization), especially given the datatype (SNPs) and feature encoding (counts). Moreover, to account for non-linear interactions that may occur between variants within a patch, we found it advantageous to GELU activate each patch representation of dimension D after projection.

With all genomic patches having been mapped to a D dimensional representation, we prepend a learnable classification token (CLS) and add learnable positional embeddings. We also append r learnable registers to the sequence of embedded patches. Registers were included to both improve performance and interpretability, as their inclusion has been shown to prevent artifacts attributable to token recycling from appearing in attention maps in ViT networks.⁷⁰ This sequence of embedded vectors are then passed as inputs to a Transformer encoder.

The Transformer encoder^{39;41} consists of L layers, alternating between multi-headed self-attention (MSA) and a multi-layer perceptron (MLP). Within each layer, layer normalization (LN) is applied prior to MSA and after the residual connection (prior to the MLP) within each

block.^{39;71;72} Classically, GELU is the nonlinearity applied in the MLP of ViTs. However it has been shown that substituting GELU with Gated Linear Units (GLU) variants can yield improved performance.⁷³ Given this, VADeR replaces GELU with SwiGLU as the MLP activation function. VADeR also employs a learnable temperature within each MSA head, to allow for a sharpening of the attention score distribution, which has been shown to be helpful when training ViTs from scratch on small-size datasets.⁷⁴

Following the transformation of all patches through the entirety of the Transformer encoder, VADeR distills the underlying output to a single, informative individual representation using CLS pooling. This representation is then fed into a linear classification head for disease prediction. A detailed overview of the VADeR network architecture can be found in **Figure 2.1**.

Several key hyperparameters were selected through empirical evaluation of validation set performance. A detailed summary of model hyperparameters can be found in **Table 2.1**.

Table 2.1. Summary of VADeR model hyperparameters.

Hyperparameter	Value
Patch Window Length (W)	250kb
Embedding Dimension (D)	768
Number of Transformer Layers (L)	12
Number of Attention Heads	12
Number of Registers (r)	8
Patch Projection Dropout	0.2
Transformer Dropout	0.2
Patch Projection Activation	GELU
MLP Activation	SwiGLU
MLP Dimension	2,048
Query, Key, Value Dimension (d_h)	64
Learnable Temperature Initialization (τ)	8

2.3.4 Training VADeR

VADeR models across all feature sets were trained using the same procedure, employing binary cross-entropy (BCE) as the loss function and AdamW⁷⁵ as the optimizer, with $\beta_1 = 0.9$, $\beta_2 = 0.99$, $\epsilon = 10^{-8}$, and a weight decay of 0.1. The learning rate was linearly warmed up to 8×10^{-5} over the first 95 steps (approximately 5 epochs), before being decayed according to a

cosine schedule for the remainder of the 115 total epochs. We also found it beneficial to employ gradient clipping at a global maximum norm of 10 to improve training stability.

Model checkpointing occurred after every epoch, with metrics computed and stored for evaluation. For each feature set, the model that achieved the highest area under the receiver operating characteristic curve (AUC) on the validation set was selected as the best model.

All VADeR models were implemented in PyTorch and trained with an effective batch size of 4,096. To achieve this, we utilized a combination of distributed training and gradient accumulation. Distributed training was conducted across 4 A30 GPUs using PyTorch’s Distributed Data Parallel (DDP), with a local batch size set to the maximum even number of samples that could fit into GPU memory.

2.3.5 Baselines

To contextualize the prognostic potential of VADeR for complex disease risk prediction, here in PC, we optimized a number of baseline models spanning classical (i.e., C+T), Bayesian, and multiple ML approaches. We also recreated the 269 variant PC PRS reported by Conti et al.,²³ the large-scale PC GWAS meta-analysis from which we obtained our summary statistics.

Classical PRS were constructed with PRSice-2.⁷⁶ As with VADeR, p-value thresholding was conducted across 5e-8, 5e-7, 5e-6, 5e-5, 5e-4, and the 1000 Genomes was used as the LD reference. The optimal classical PRS was identified as the threshold which maximized the validation set performance (AUC).

LDpred2⁵⁶ was used as our Bayesian benchmark. As recommended by the authors, we constrained the variants under consideration by the method to HapMap3+ variants.⁷⁷ Employing the 1000 Genomes as the reference panel, we observed a marked divergence between validation (AUC = 0.83916) and test set performance (AUC = 0.66622), and a notable test set underperformance relative to C+T (**Table 2.2**). While the former is likely attributable to the fact that the validation set was used in the generation of the summary statistics by Conti et al.,²³ the latter suggests poor effect size reestimation, which is dependent upon the LD reference. Given the dependence of reestimation from LD information, we decided to use the entirety of ELLIPSE as the reference panel for LDpred2. Given that ELLIPSE made up 39.1% of the

original study, we hypothesized that it would serve as a more suitable and representative LD reference, though this came at the cost of independently assessing hyperparameter performance on the validation set. Fortunately, LDpred2-auto is able to directly infer hyperparameter values, and thus this was the approach used and reported here. We note that though LDpred2 does provide LD matrices for use in cases such as this, these matrices are computed from European individuals of the UKBB, which if used here, would result in test set leakage.

It is important to emphasize that PRSs generated with either C+T or LDpred2 make use of the effect sizes provided by the original summary statistics, while VADeR and the ML baselines (described subsequently), must naturally learn a mapping between features and labels from the training set alone. Given that a marked discrepancy exists between the training set size ($n = 73,315$) and the number of individuals used in the Conti et al. study ($n = 234,253$),²³ it could be conceived as unfair to compare approaches, as C+T and LDpred2 have access to a larger pool of effective information. To alleviate these concerns, we ran both PRSice-2 and LDpred2 again on a second set of summary statistics generated from our training set alone. These followed the same procedures as above, with the exception that for LDpred2 we used only the training set genotypes (as opposed to all of ELLIPSE) as our LD reference panel. This approach ensures a fair cross-method comparison of different approaches. However, for comprehensiveness, we report results for C+T and LDpred2 for both the Conti et al. summary statistics and the training set summary statistics, with the former being denoted by an asterisk (*).

ML baselines consisted of logistic regressions, support vector machines, random forests, XGBoosts, and fully-connected feedforward neural networks (FC-FFNs). Hyperparameters for each approach (e.g., regularization, model depth) were optimized to maximize the validation set performance (AUC) using a Tree-structured Parzen Estimator (TPE) sampler with Optuna.⁷⁸ Specifically, 100 trials were conducted across each p-value threshold feature set, resulting in 500 total trained models per baseline. The final model for each baseline ML approach was the one that resulted in the maximum validation AUC. Logistic regression, support vector machine, and random forest models were implemented with scikit-learn. XGBoost models were implemented with CUDA acceleration (on an A30 GPU) via the xgboost Python package. FC-FFN models were implemented in PyTorch, with each model utilizing a RTX 5000 GPU for training and

inference.

Finally, all PRSs without an explicit boundary for classification (i.e., C+T, LDpred2, and the recreated Conti et al. PRS), were transformed to a $[0, 1]$ prediction probability via a logistic regression fit on top of the original method test set scores.

2.3.6 DARTH Patchtribution

Self-attention, or the assignment of pairwise attention between all token pairs, is the defining building block of Transformer networks. Directly utilizing attention values as a metric of relevance (often from the last layer in the model) for Transformer explainability is a common practice as it allows for an easy and intuitive identification of attention patterns. Though convenient, this approach fails to globally capture token contributions, omitting not only intermediate attention scores, but also the contributions of other model components.⁷⁹

Methods such as attention rollout and attention flow⁸⁰ attempt to improve attention-based Transformer explainability by accounting for the intrinsic mixing of information across layers. However, each of these approaches suffers from limitations. Rollout, which linearly combines attentions along paths in the pairwise attention graph, fails to distinguish between positive and negative contributions to the decision, which leads to artifacts in the attention attribution.⁷⁹ Attention flow, which conceptualizes the attention graph as a flow network and assigns attributions according to a max-flow problem formulation, has been shown to yield improved attributions relative to rollout in specific cases, but is too slow to support large-scale evaluations.

In addition to the aforementioned limitations of the previously identified attention attributions (i.e., attention maps, rollout, and flow), it is also worth emphasizing that each of these methods (along with many other non-attention based attribution approaches) are class-agnostic. To prevent artifact introduction that might otherwise occur through using these approaches for model interpretability, we instead implemented an extension of Beyond Attention, a Transformer interpretability method proposed by Chefer et al.^{79;81} This approach provides a class-specific comprehensive capturing of information propagation through the network, and has been shown to outperform other interpretability approaches across multiple modalities.⁸¹

Patch relevance can be conceptualized as a self-contained entity prior to passing through VADER’s Transformer encoder, as at this point each patch remains a local representation, independent of the information contained in any other patch. Given this we can initialize a relevance map R_0 according to this self-containment principle as an identity matrix:

$$R_0 = I^{(p+r+1) \times (p+r+1)} \quad (2.1)$$

where p is the number of genomic patches in a given feature set. As patches pass through the Transformer, attention heads in each layer contextualize patch representations. Thus in any given layer, the attention maps can be used to update the relevance map, but must necessarily account for the varying importance and relevance of each attention head.⁸² To generate a layer-specific attention map $\bar{A} \in \mathbb{R}^{(p+r+1) \times (p+r+1)}$, one can perform gradient-weighted averaging across heads:

$$\bar{A} = \mathbb{E}_h ((\nabla A \odot A)^+) \quad (2.2)$$

where

$$A = \text{softmax} \left(\frac{Q \cdot K^T}{\tau} \right) \quad (2.3)$$

is the attention map of a given attention head, $\odot$ is the Hadamard product, $\nabla A = \frac{\delta y}{\delta A}$ with y representing the model output, and $\mathbb{E}_h$ is the mean across attention heads. Negative contributions are removed prior to averaging to capture positive relevance in accordance with the original method.^{79;81}

The circuitry of relevance within a Transformer layer is also branching, with residual connections subsequent to each MSA operation. Thus appropriately updating R_i at each layer requires accounting for this facet of the architecture and leads to the formulation of the following relevancy update rule:

$$R_{i+1} = R_i + \bar{A} \cdot R_i \quad (2.4)$$

With this rule in place, we can compute personalized per patch **D**irected **A**ttention **R**elevance

from **Transformer Heuristics (DARTH)** scores through the extraction of the CLS token row in R_L .

2.3.7 Statistical Analyses

Two-sided Mann-Whitney U tests were employed to assess whether prediction difference distributions differed between groups. Fisher’s exact tests were used to assess the significance of increases in predicted cases by VADeR between clusters. These statistical tests were implemented with the `scipy.stats` Python package. Wherever multiple hypotheses were tested, p-values were corrected by the Benjamini-Hochberg procedure.

2.4 Results

2.4.1 Evaluating VADeR’s Prognostic Potential

Transformers and their derivatives have emerged as the state-of-the-art across NLP benchmarks⁴¹ and have seen widespread adoption in both NLP and CV tasks. Their success is largely attributable to their self-attention mechanism, which allows for capturing complex, contextual relationships within data to effectively model both local and global patterns. Given that the genome shares many intrinsic properties with data of these aforementioned domains (e.g., sequential structures, long-range dependencies, non-uniform statistical relationships), Transformer-based approaches have shown significant potential across a number of genomic applications^{31;42–44} and afford an exciting avenue of investigation for tasks such as complex disease risk prediction.

Recognizing these challenges and the need for specialized solutions, we developed a new ViT-inspired³⁹ model architecture named VADeR for complex disease risk prediction (**Figure 2.1**). VADeR conceptualizes the genome as an image which can be partitioned into regions, or patches, of static length, but of differing SNP density. Unlike classic ViTs which project all image patches (of equal pixel density) through a single trainable linear projection, VADeR projects each genomic patch through its own single layer non-linear projection to generate a regional representation of consistent dimensionality (Material and Methods: *VADeR Architecture*). These patch representations are then passed through a Transformer encoder, CLS pooled, and passed to a final classification head to generate a disease prediction.

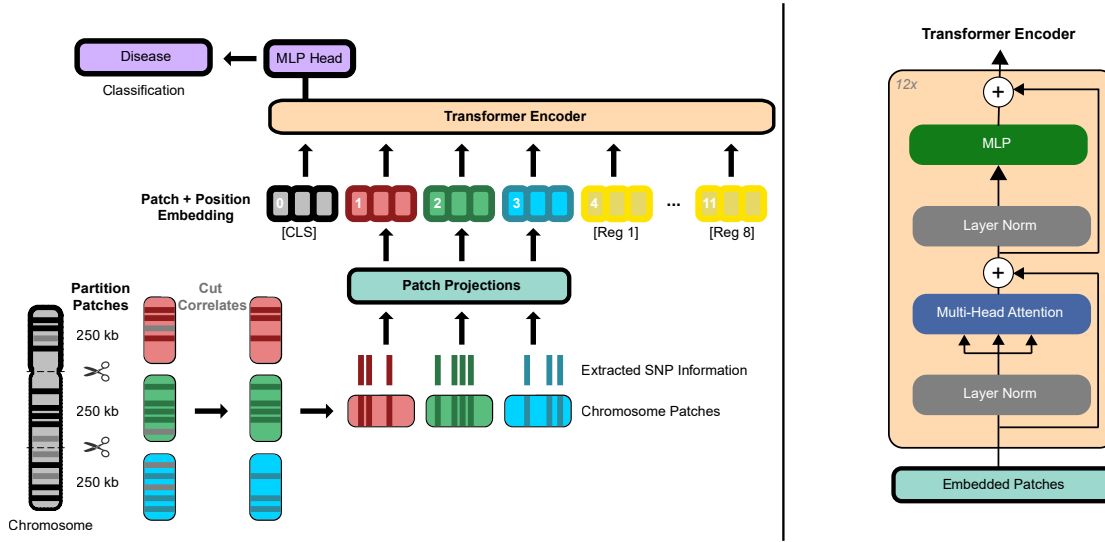


Figure 2.1. VADeR model overview. The genome is partitioned into non-overlapping patches of 250kb, ensuring chromosomal consistency. From these patches, all variants in the candidate feature set (after p-value thresholding and clumping) are extracted and passed to a patch specific projection. Positional embeddings are subsequently added to the output of these projections. A learnable classification token (CLS) and registers are then prepended and appended to the representation, respectively. The resulting genomic sequence representation is then passed through a Transformer encoder, CLS pooled, and passed to a classification head for disease prediction.

To evaluate VADeR’s capacity for complex disease prediction, we sought a highly heritable target condition for which a large non-private case-control genotype dataset existed. Prostate cancer (PC) emerged as an ideal candidate as it satisfied both of these criteria. From ELLIPSE, we constructed training and validation sets according to a random 80/20 split (Material and Methods: *Datasets*). To reduce the feature space to enable model training, we employed a clumping and thresholding based approach (Material and Methods: *Formulating Feature Sets*). Specifically, clumping employed the 1000 Genomes, and thus was independent of the training and validation sets, while p-value thresholding relied on a separate, though not completely independent, large-scale PC multi-ancestry GWAS meta-analysis by Conti et al.²³ To fairly assess model generalization, we constructed an independent test set from a subset of the UKBB (Material and Methods: *Datasets*).

To contextualize VADeR’s performance we optimized a number of baseline models including classical C+T, LDpred2,⁵⁶ and five different ML approaches (Material and Methods: *Baselines*). We also reconstructed the 269 variant PC PRS reported by Conti et al.,²³ the study from which we acquired our summary statistics. As detailed in Material and Methods: *Baselines*, it is important to highlight that VADeR and the ML baselines must learn to map genomic features to labels exclusively from the training set ($n = 73,315$), while the remaining PRS methods (i.e., C+T, LDpred2, and the 269 Conti et al. PRS) can leverage a much larger effective sample size ($n = 234,253$) by directly utilizing effect sizes from the original Conti et al. summary statistics. To ensure a fair comparison across models, we also generated training set specific C+T and LDpred2 baseline PRSs, using effect sizes derived exclusively from the training set. By evaluating these models in both contexts, we provide a contextualized and fair assessment of VADeR’s performance relative to the baselines. Those baselines that use effect sizes from the original Conti et al. summary statistics²³ are marked with an asterisk (*).

Regardless of the effect sizes used for PRS formulation, VADeR consistently outperformed all statistical genomics baselines (including the Conti et al. PRS). Specifically, compared to LDpred2*, the top-performing statistical genomics baseline across metrics, VADeR achieved absolute increases of 0.007 in AUC, 0.243% in accuracy, 0.008 in average precision (AP), 0.016 in F1 score, and 0.006 in Matthew’s correlation coefficient (MCC). These differences effectively

double across each metric when limiting statistical genomics models to training set specific effect sizes (**Table 2.2**). Against the ML baselines, VADeR emerged as the top-performing model across 3 of the 5 metrics, including accuracy, AP, and MCC, with particularly strong gains in accuracy (1.03%) and MCC (0.008). AP gains were more modest (0.0003), with XGB and logistic regression showing comparable performance. VADeR also ranked second in AUC, with performance closely matching that of logistic regression, the leading model by this metric. We note the difference in AUC between VADeR and logistic regression was statistically insignificant with a DeLong test⁸³ p-value of 0.17. Overall, VADeR demonstrated the best performance in accuracy, AP, and MCC, and ranked a close second in AUC. A detailed breakdown of model performances across metrics can be found in **Table 2.2**.

2.4.2 Linking VADeR Attributable Risk to Loci with DARTH Scores

VADeR’s improved performance across metrics, especially compared to classical baselines, gives rise to an important question: from where are these performance gains originating? Given the nature of the architecture, a natural assumption might be that VADeR is capturing complex genomic interactions or novel risk regions not fully accounted for by classical PRS methods. It is also plausible to consider that the gains are attributable to personalized contextualized risk, with VADeR identifying and emphasizing contributions of particular genomic regions which would otherwise be underweighted.

A reasonable starting place to begin to answer these questions is to assess the concordance and deviation between VADeR’s identified influential genomic region attributions and those identified by GWAS and classical PRSs. However to do so requires the establishment of a metric of model interpretability. Interpretability and explainability for deep-learning models, particularly for Transformer-based approaches, is currently an open and active area of research. Broadly speaking, attribution approaches fall into one of four buckets^{79;84}: gradient-based (e.g., Grad-CAM,⁸⁵ Integrated Gradients⁸⁶), attribution propagation (e.g., LRP,⁸⁷ DeepLIFT⁸⁸), Shapley-value based (e.g., ViT Shapley⁸⁹), and attention-based (e.g., attention maps,³⁸ attention rollout and flow⁸⁰). For Transformers, attention-based approaches are, understandably so given the architecture, the most common, but like any attribution method require wariness of the

Table 2.2. Test set performances of VADeR and baseline models across multiple metrics. $P \leq$ corresponds to the p-value threshold for which validation performance was maximized and thus used for final baseline evaluation. The best performing model for each metric is **bolded**, while the second-best model is underlined. * denotes method utilized reported summary statistics’ effect sizes, which were inferred with 3.2x the amount of data as our training set. [†] LDpred2 hyperparameter which corresponds to the proportion of causal variants.

Method	$P \leq$	# Variants	AUC	Accuracy	AP	Precision	Recall	F1 Score	MCC
C + T	5e-5	730	0.68427	0.63853	0.67102	0.63875	0.63193	0.63532	0.27704
LDpred2	0.0056 [†]	965,472	0.71266	0.65711	0.69870	<u>0.65880</u>	0.64682	0.65275	0.31421
Logistic Regression	5e-4	12,017	0.72843	0.64687	0.71322	0.60590	0.83322	0.70161	0.31761
SVM	5e-4	12,017	0.72176	0.65369	0.70652	0.61878	0.79436	0.69566	0.32114
Random Forest	5e-5	10,389	0.67829	0.50594	0.67771	0.50212	0.99657	0.66778	0.07296
XGB	5e-4	12,017	0.72635	0.64102	<u>0.71323</u>	0.59944	<u>0.84258</u>	<u>0.70051</u>	0.30937
FC-FFN	5e-5	10,389	0.72525	0.65172	0.70903	0.61521	0.80371	0.69694	0.31940
VADeR	5e-4	12,017	<u>0.72665</u>	0.66400	0.71352	0.64931	0.70808	0.67742	0.32954
Conti et al.*	N/A	269	0.70611	0.64693	0.69208	0.64747	0.63970	0.64356	0.29385
C + T*	5e-4	2,894	0.68677	0.63328	0.67246	0.63060	0.63733	0.63395	0.26659
LDpred2*	0.032 [†]	965,472	0.71949	<u>0.66157</u>	0.70528	0.65976	0.66236	0.66106	<u>0.32314</u>

intrinsic limitations of the methodology.

Direct utilization of attention maps, particularly from the final layer of the model, though frequently used, omits the contributions of the rest of the model architecture and can overemphasize redundant regions through token recycling.⁷⁰ The inclusion of registers, as we have with VADeR, has been shown to resolve the latter issue, but the former is still a concern. Approaches such as attention rollout and attention flow⁸⁰ were proposed to account for the intrinsic mixing of information across layers, but each suffers from their own limitations due to the underlying method assumptions and formulation (Material and Methods: *DARTH Patchtribution*).

To prevent the introduction of spurious artifacts that might otherwise occur through usage of these or other attribution methods, we formulated **D**irected **A**ttention **R**elevance from **T**ransformer **H**euristics (DARTH) scores based on an extension of Beyond Attention introduced by Chefer et al.^{79;81} (Material and Methods: *DARTH Patchtribution*). DARTH scores equate to VADeR’s personalized patch-specific attentional attribution fingerprint, and serve as a metric of a given region’s influence towards driving disease risk for an individual.

With the establishment of an interpretability metric for VADeR, we returned to our original objective: investigating the alignment of GWAS-identified regions of disease risk with those identified by VADeR. As DARTH scores are personalized encapsulations of regional attributed risk, a global view of regional risk was still required for such a comparison. To enable this, we averaged DARTH patch contributions across the entire test set, thus obtaining a metric of global model attributed risk. Overlaying these averaged attributions with variant significances reported in the GWAS summary statistics generated by Conti et al., we found that VADeR’s highest attributable risk regions often aligned with those containing the most significant GWAS variants, though specific differences were observed (**Figure 2.2A**). For example, VADeR’s top three global attributable regions of risk fell across consecutive patches in the chromosome 8q24 region, a well-documented locus for prostate cancer risk^{90;91} (**Figure 2.2B**). Outside of chromosome 8, high risk patches were also identified on chromosome 10 and 17. Notably, the high risk patch on chromosome 10, located at 10q11.2, encompasses both the MSMB gene and the documented prostate cancer risk variant rs10993994.^{92;93} This variant, positioned within the promoter of MSMB, modulates the gene’s expression, which is one of the most abundant proteins

secreted by the prostate.⁹⁴ Chromosome 17 exhibited multiple high-attribution patches that align in trend but deviate in distribution from the Manhattan tower significances, highlighting a different pattern of risk localization.

Assessing the concordance between VADeR’s global average attributions and GWAS-identified risk regions, though conceptually insightful, captures a perspective independent of classical risk attribution. A direct comparison against risk attributed by classical PRSs is needed to gain such a perspective. Given DARTH scores represent VADeR’s attribution at the patch level, we extracted the contributions of attributed risk at a patch resolution for a classical weighted genotype summation PRS employing the same subset of features and effect sizes from the Conti et al. study²³ (**Figure 2.2C**). As with DARTH scores, to distill personalized patch risk attributions to a global representation, we averaged these classical PRS contributions across the entire test set.

Comparing the risk attributed by each approach, we observe notable areas of concordance and divergence. Specifically, patches on chromosome 8, 17, 19, and 20 received high attributed risk across approaches, while the patch starting at chromosome 2:173.25 Mb exhibits a marked discrepancy between approaches (**Figure 2.2C**). This chromosome 2 patch, containing 73 SNPs, is the classical PRS highest risk attribution region globally, yet from VADeR’s perspective is of minimal significance. A natural assumption might be that the contributions of risk at this region effectively serve as a bias for the classical PRS, simply additively shifting the score distribution. However, we find this not to be the case, as this patch exhibits the ninth highest classical PRS patch variability ($\sigma^2 = 3.66$) across the entire test set, just below the top three chromosome 8 patches. It might next be assumed that the magnitude of classical PRS risk for this patch is attributable to the patch density (i.e., the number of SNPs in the patch). However, we find that this region falls outside of the top 20 patches in terms of patch density. Given discrepancies between top attributed patches between approaches, we questioned whether DARTH scores could be used to regularize classical PRS (i.e., narrow the feature space) and thus improve performance. Despite differences in methodology, with VADeR’s architecture geared towards capturing complex non-linear interactions, we found that reducing the number of patches used by a classical additive PRS to the top 25% and 50% of patches with the highest average DARTH

score yielded improvements in stratification potential by AUC, with gains of 0.27% and 0.84% respectively.

Given the divergences in density across patches, it is fair to question the correlation between the number of variants in a patch and VADer’s averaged DARTH scores. Though we did observe a correlation between the number of variants in a patch and its average DARTH score (pearson $r = 0.51$), we observed pronounced exceptions to this phenomenon, most notably, with two single SNP patches falling among the top 15 average DARTH scores. To our knowledge, neither of these SNPs, rs13013607 (Conti et al. GWAS $p = 7.59 \times 10^{-6}$) and rs1624540 (Conti et al. GWAS $p = 3.13 \times 10^{-5}$), an intron variant in INTS4, located on chromosomes 2 and 11 respectively, are currently documented risk variants for prostate cancer. However, their high average attribution by VADer suggests that they warrant further investigation either independently or interactively, particularly among individuals of European ancestry. We note that rs13013607, for which the nearest genes are CALM2, STPG4, and TTC7A, is over 125 Mb away from the top chromosome 2 classical PRS patch investigated above (**Figure 2.2C**), suggesting that this high average VADer attribution is independent of variant correlation between these patches.

Collapsing attributions to the average, while providing a means for evaluating global concordance and conceptual alignment between methodologies, fundamentally blurs personalized patch contributions and thus fails to elucidate unique contextualized contributions to disease risk. To initiate an investigation into VADer’s ability to capture genetic heterogeneity contributions to disease risk, we employed UMAP reduction⁹⁵ to project the 1204-dimensional DARTH manifold to two dimensions, transforming each test set individual into a distinct two-dimensional point. This projection revealed a clear gradient of VADer prediction values across the dimensions, with a predominant, dense central cluster displaying less defined substructure, alongside three distinct, separate regions, indicative of unique genetic heterogeneity driving predicted disease risk (**Figure 2.3A**).

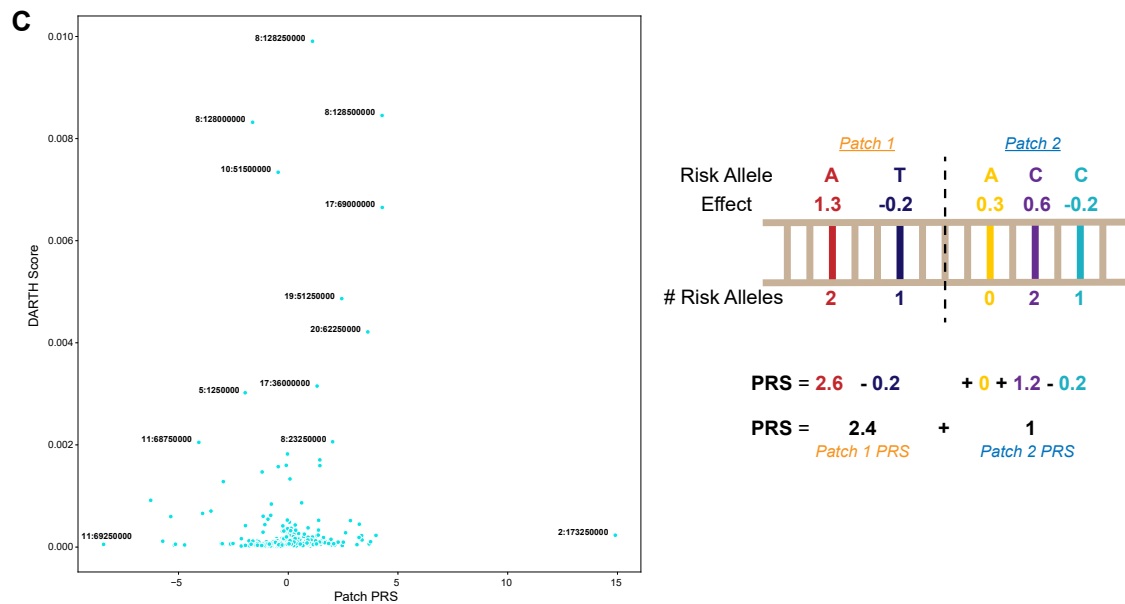
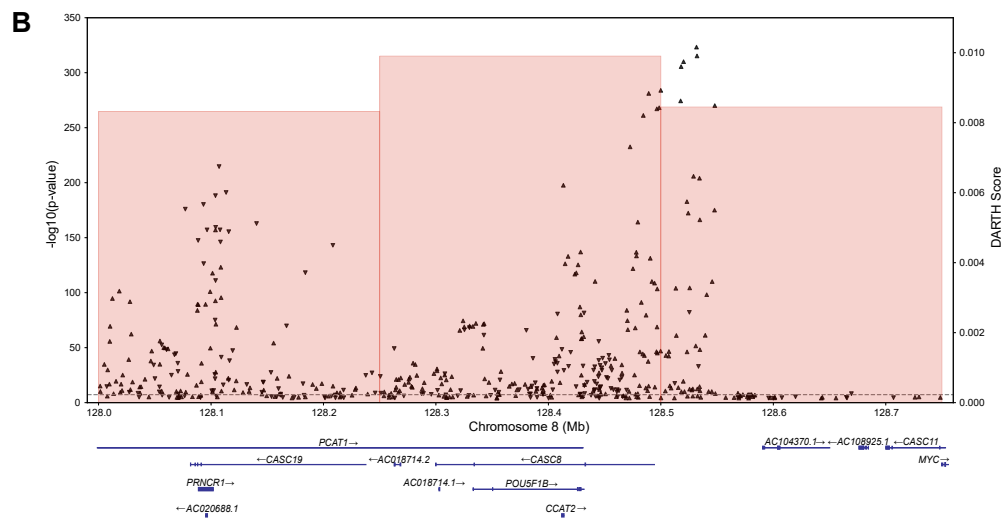
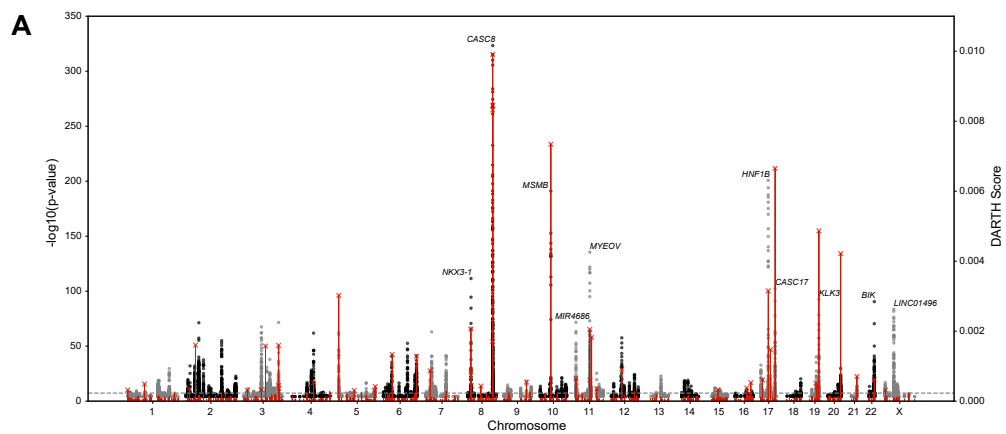
An investigation into the patch attributions driving these three regions of observed substructure led to an interesting finding: each cluster was driven by a specific patch DARTH score outlier. Individuals in cluster 1 ($n = 317$) exhibited a pronounced elevation in DARTH score

for the patch starting at chromosome 17:46.75 Mb, the region which encodes for HOXB13 (**Figure 2.3B**). This patch, which ranked 17th among all patches by global average DARTH score and 3rd among all patches on chromosome 17, exhibited the highest DARTH score bimodality coefficient (BC) of any patch in the test set (BC = 0.90), and was distinctly uniquely elevated amongst individuals in this cluster (**Figure B.1A**). Notably, the variant rs138213197, corresponding to a HOXB13 G84E mutation, is known to significantly increase PC risk in individuals of European descent,^{61;64} particularly in Nordic countries.⁹⁶ Though this variant was not in the feature set, VADeR was able to distinctly capture from the variants in this patch an increased risk in individuals in this cluster (OR = 3.19; $p = 2.19 \times 10^{-19}$) relative to individuals in the predominant, dense central cluster ($n = 12,026$). Perhaps unsurprisingly, cluster 1 individuals showed a significant increase in identity by state (IBS) at this patch relative to a matched size random sampling of non-cluster 1 individuals from the test set (cluster 1 mean patch IBS = 0.83; random sample patch IBS = 0.76). Thus though IBS across the entire feature set was equivalent across both subsets (mean IBS = 0.74), it is possible that VADeR was capable of identifying increased risk from HOXB13 from an ancestry association in this region.

Among individuals in cluster 2 ($n = 303$), the DARTH score outlier was the patch starting at chromosome 21:42.75 Mb, the region which encodes for TMPRSS2 (**Figure 2.3C**). Similar to the dominant patch in cluster 1, this patch, which was ranked 23rd among all patches by global average DARTH score, exhibited the second highest bimodality of any patch in the test set (BC = 0.88), and was predominantly elevated amongst individuals in this cluster (**Figure C.1B**). A defining molecular subtype in PC is the presence of a TMPRSS2-ETS gene fusion, predominantly involving ERG (though ETV1 and ETV4 are found to a lesser extent).^{62;63} This fusion, which is found in approximately 50% of PC cases, is a significant driver in the etiology of PC, as it results in the constitutive expression of an oncogenic transcription factor.^{63;97} Given the location of this patch, it is likely that the dominance of attributed risk here relative to other patches relates to VADeR’s identification of an increased likelihood for such a fusion.

Unlike cluster 1 where the HOXB13 patch had a five-fold increase in mean cluster DARTH score (**Figure 2.3B**), cluster 2 displayed a more distinct pattern of DARTH scores among patches. The dominant TMPRSS2 patch in cluster 2 showed only a two-fold increase in mean cluster

Figure 2.2. Comparing average attributions alignments between classical PRS and VADeR with DARTH scores. A) DARTH Manhattan plot - Averaging DARTH scores across the entire test set for each patch in the feature set (represented by red lines) exhibits a similar structure to the one formed by Conti et al. summary statistic p-values (represented by circles). The dotted line corresponds to the genome wide significance p-value of 5×10^{-8} . B) Visualizing the dominant region of the test set average DARTH attribution. Three contiguous patches on chromosome 8 spanning from 128-128.75 Mb overlay with the largest p-value peak containing regions encoding the oncogene MYC and PC associated lncRNAs PCAT1, PRNCR1, CASC8, CASC11, and CASC19. Triangle orientation for variants indicates direction of risk effect size. C) Assessing average VADeR attributions through DARTH scores against the average patch contributions to a classical additive PRS. Labeled regions correspond to the chromosome and position of a patch start point.



DARTH score relative to the next highest attributed patch (**Figure 2.3C**). Other significant patches in cluster 2 included those identified on chromosomes 8, 10, and 17. In relation to the TMPRSS2-ERG fusion-positive PC subtype, SNPs in each of these patches, in particular rs16901979 (8q24), rs1859962 (17q24), and the aforementioned rs10993994 (10q11) have all been shown to be associated with differences in the presence of this subtype.⁹⁸ The SNP rs1859962 is located in the coding region of CASC17 and is part of the feature set, within the top identified global average chromosome 17 patch (**Figure 2.2A**). Although rs16901979 was not in the feature set, it is possible that the risk attributed to this SNP was captured by the patch representations of the previously mentioned contiguous chromosome 8 patches with the highest global average VADeR attribution (**Figure 2.2B**). However, the converse is also possible. Modulation of PC, particularly for this subtype, may instead be modified broadly by this entire chromosome 8 region. Taken together these results suggest specific genetic heterogeneity identified by VADeR in PC, potentially towards TMPRSS2-ERG fusion-positive PC predisposition, though in the absence of clinical information we are unable to conclusively make this claim. We cannot also definitively eliminate that the DARTH score signature identified by VADeR for cluster 2 is completely unrelated to shared ancestry among individuals in this cluster. IBS among all pairs of individuals in cluster 2 at the TMPRSS2 patch relative to a matched size random sampling of non-cluster individuals did exhibit a significant difference (cluster 2 mean patch IBS = 0.84; random sample patch IBS = 0.79), though IBS across the entire feature set was consistent across both groups (mean IBS = 0.74). Finally, we note that like individuals in cluster 1, we observed an increased predicted risk in individuals in cluster 2 (OR = 1.34; $p = 0.016$) relative to individuals in the predominant, dense central cluster.

Unlike clusters 1 and 2, for which the dominating DARTH score patches were regions underemphasized or otherwise overlooked by a global assessment, the key identified driver of risk for individuals in cluster 3 ($n = 2,589$) was the patch beginning at chromosome 10:51.5 Mb (**Figure 2.3D**). This patch, which was the top chromosome 10 risk region and 4th highest risk region overall in our global analysis (**Figure 2.2A**), includes, as previously mentioned, both the MSMB gene and the well-documented PC risk variant rs10993994. This is particularly notable given that the protein for which MSMB encodes exhibits broad tumor suppressive properties,⁹⁹

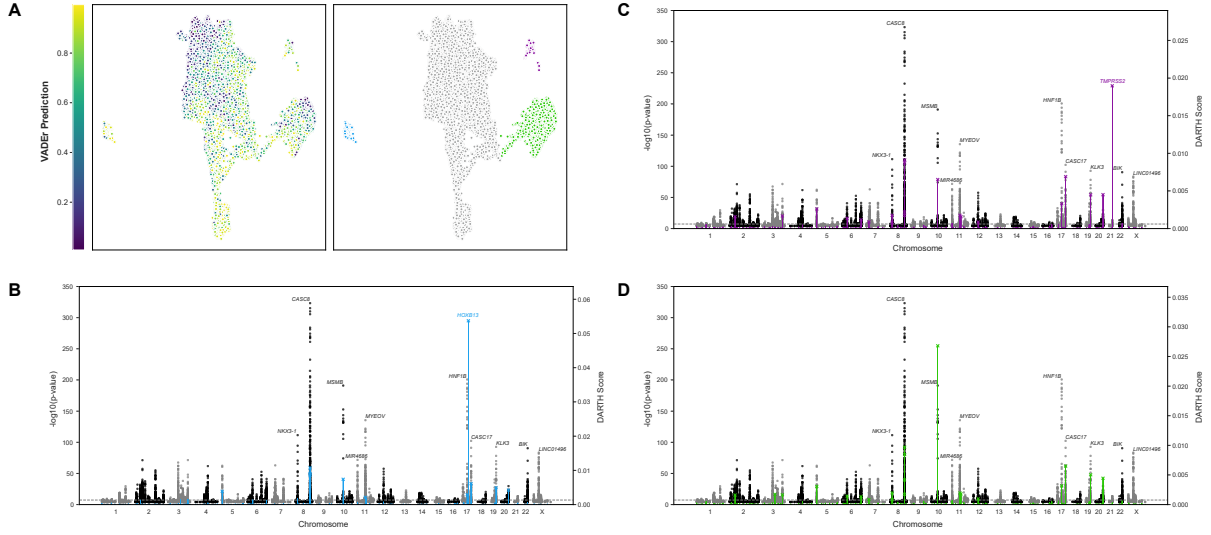


Figure 2.3. Identifying genetic substructure driving predicted risk from DARTH scores. A) UMAP of patch DARTH scores across the entire test set. Prediction scores show a clear gradient across the dimensions (left) with three distinct regions of substructure (right). B) Across individuals in cluster 1, the DARTH score distribution exhibits a notable deviation from the global average test set distribution, with the patch spanning 46.75-47 Mb on chromosome 17 containing 11 variants identified as the dominant region of relevance. C) Cluster 2 exhibits a similar phenomenon to cluster 1, with the patch spanning 42.75-43 Mb on chromosome 21 containing 30 variants identified as the dominant region of relevance. D) Unlike individuals in clusters 1 and 2, for which VADER uniquely emphasizes muted regions, the dominant driver of predicted risk in cluster 3 individuals aligns with the chromosome 10 tower spanning 51.5-51.75 Mb containing 31 variants and encompassing the gene MSMB.

and rs10993994 is a documented modulator of this gene's expression.¹⁰⁰ Though rs10993994 is in this patch, it is unlikely that the emphasis of effect attributed to this region is driven solely by contributions of this SNP. There are 31 variants encoded in this patch, 8 of which fall within the coding region of MSMB, and 3 (including rs10993994) positioned less than 1500 bp upstream of this gene, providing a number of possible cis-modulators of MSMB. Notably, several of these SNPs have been investigated for their association with MSMB expression,¹⁰¹ suggesting that multiple variants in this region may contribute to modulating gene expression and, consequently, risk. Moreover, as with clusters 1 and 2, a significantly higher mean IBS for this patch was observed relative to a matched size random sampling of non-cluster individuals (cluster 3 mean patch IBS = 0.87; random sample patch IBS = 0.78). This, in conjunction with consistent genome-wide similarity between groups (mean IBS = 0.74), underscores a distinctive contribution of a multitude of variants to attributed predicted risk by VADER.

As expected, given the MSMB patch’s high global average DARTH score and relatively low bimodality ($BC = 0.47$) compared to the top patches identified in clusters 1 and 2, its DARTH score cluster signature is less distinct, with elevated MSMB patch DARTH scores capable of being found across the test set (**Figure B.1C**). Nonetheless, within cluster 3, the MSMB patch showed a three-fold increase in mean cluster DARTH score relative to the next highest attributed patch (**Figure 2.3D**). This pronounced difference emphasizes the unique contribution of the MSMB region to risk in cluster 3 individuals, particularly in contrast to the dense central cluster, where attributed risk is primarily concentrated by the contiguous chromosome 8 patches spanning 128-128.75 Mb (**Figure B.1D-F**).

2.4.3 Empirical Validation of DARTH Scores and Regional Risk Attribution

While DARTH scores represent a metric of personalized patch-level attribution by VADER from which we can identify distinctly heterogeneous genetic contributions to disease, it remains to be empirically shown their alignment with contribution to predicted disease risk. Moreover, given the cluster signatures identified above, particularly the identified outlier patches driving the observed substructure, it remains unclear as to the mechanisms driving the observed distributions. Are the contributions of outlier patches in each cluster conditionally dependent on the rest of the feature set (i.e., the representation of one’s genome), or are the effects regional, captured by patch-level embeddings? The inflation in patch-level IBS scores for the outlier patches in clusters 1-3 is suggestive though not necessarily indicative of the latter.

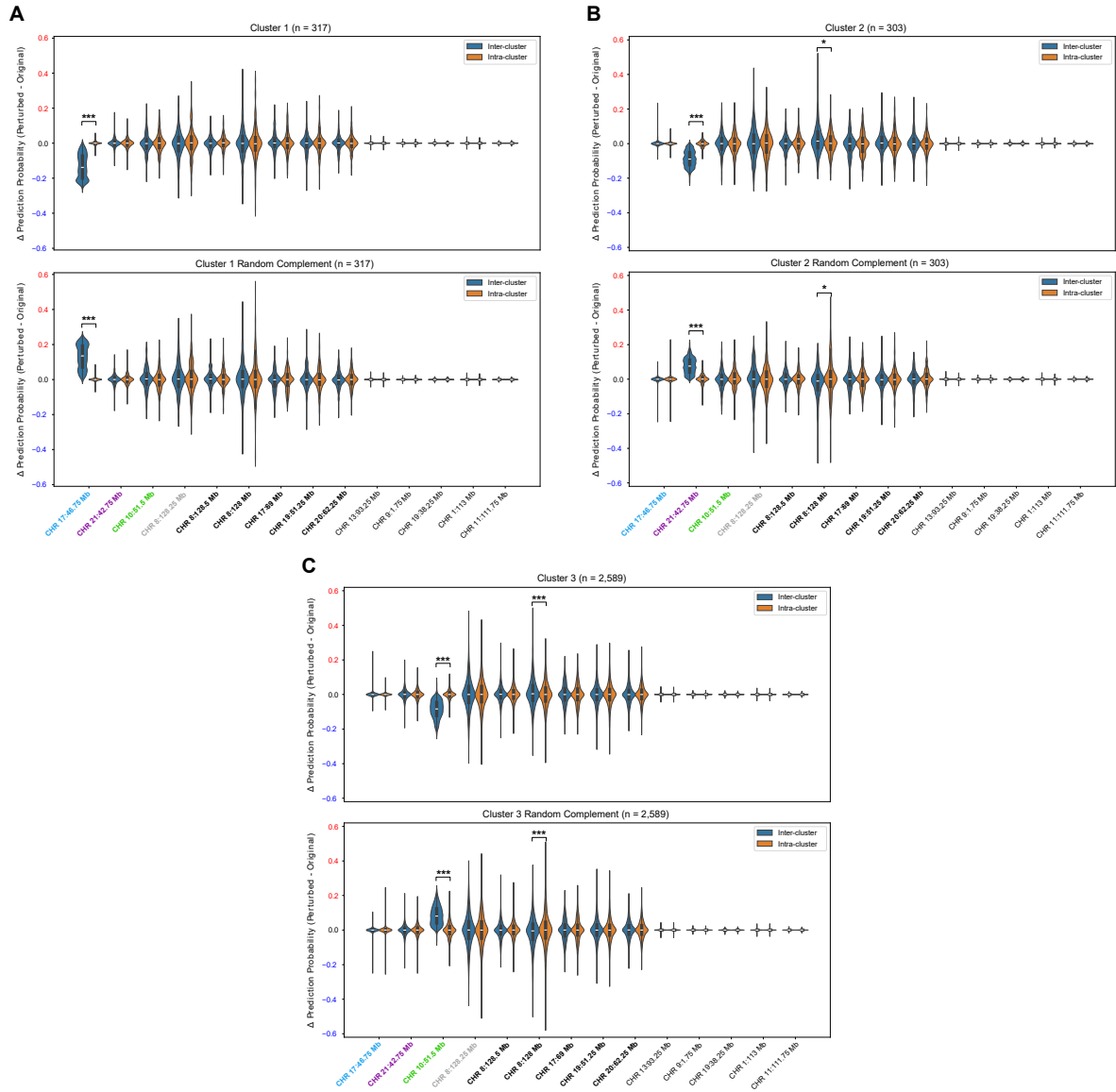
To address these outstanding uncertainties, we conducted a patch perturbation analysis. For each cluster we constructed a matched size complement from a random sampling of non-cluster individuals. With both a cluster and its assigned complement established, patch-level genotypes were then shuffled independently both within a set and between sets. Using the dominating cluster 1 HOXB13 patch and cluster 1 and its complement as an example, we would first shuffle only the genotypes captured by this patch among individuals in each set (intra-cluster). Next we would randomly shuffle this HOXB13 patch genotype between individuals in each cluster (inter-cluster), ensuring all individuals in cluster 1 and its complement received a unique HOXB13 patch genotype from outside their defined set. With both intra- and inter-cluster shuffling

completed, we could evaluate and contextualize a patch’s contribution to model prediction by computing the difference in VADeR prediction probability for each patch-shuffled genotype relative to their original genotype.

Comparison of the prediction difference distributions between inter- and intra-cluster patch shuffling further supports the regional risk hypothesis, suggesting that VADeR’s emphasis on outlier patches is attributable to its patch-level embeddings. Among cluster 1 and its matched size random complement, we observed minimal variation in PC prediction probability when performing intra-cluster shuffling for the dominant DARTH score cluster 1 HOXB13 patch. However, inter-cluster shuffling of this patch led to significant shifts in predicted risk for both groups, with cluster 1 individuals having a reduced median predicted risk of PC by 13.79% ($p = 4.73 \times 10^{-102}$) and individuals in the random complement having an increased median predicted risk of PC by 13.52% ($p = 2.29 \times 10^{-101}$; **Figure 2.4A**). Moreover, no other patches exhibited a significant distribution deviation in VADeR prediction probability difference between the two shuffling methodologies in either group.

A similar phenomenon was observed for both clusters 2 and 3 and their matched size random complements. Inter-cluster shuffling relative to intra-cluster shuffling of the dominant cluster 2 risk patch (TMPRSS2) significantly reduced predicted median risk of PC in cluster 2 individuals by 8.97% ($p = 4.77 \times 10^{-89}$) and significantly increased predicted median risk of PC in random complement individuals by 7.51% ($p = 2.52 \times 10^{-78}$; **Figure 2.4B**). Likewise, inter-cluster shuffling of the dominant cluster 3 risk patch (MSMB) led to a significant 8.35% ($p < 5 \times 10^{-4}$) decrease in predicted median risk for cluster 3 individuals and a significant 8.10% ($p < 5 \times 10^{-4}$) increase for individuals in the corresponding random complement (**Figure 2.4C**). Modest though significant distribution deviations were also observed for the chromosome 8:128 Mb patch in each of these clusters (**Figure 2.4B,C**). Encouragingly, across all analyses, we observed that variability in VADeR prediction probability difference, regardless of shuffling methodology, tracked with global average DARTH attribution scores (**Figure 2.4**), further supporting the ability of DARTH scores to capture personalized contributions to disease risk.

Figure 2.4. Quantifying patch contributions to VADeR predicted disease risk through patch genotype shuffling among individuals within a subset (Intra) and between subsets (Inter). Patches depicted here include the top DARTH score patch identified for each cluster (colored accordingly), the top 5 global average DARTH score patches (bolded), and the bottom 5 global average DARTH score patches. A) Only genotype shuffling of the identified top cluster 1 patch between populations (compared to within a population) resulted in significant deviations in risk prediction, with a 13.79% decrease (top) and 13.52% increase (bottom) observed across groups. B) Genotype shuffling of the identified top cluster 2 patch between populations resulted in significant decreases of 8.97% (top) and increases of 7.51% (bottom) in VADeR prediction probability. The patch beginning at chromosome 8:128 Mb also showed modest but significant differences between shuffling methodologies in each group ($p_{cluster2} = 0.014$; $p_{complement} = 0.018$). C) Genotype shuffling of the identified top cluster 3 patch between populations resulted in significant decreases of 8.35% (top) and increases of 8.10% (bottom) in VADeR prediction probability. Similar to cluster 2, the chromosome 8:128 Mb patch also showed modest but significant differences between shuffling methodologies in each population ($p_{cluster3} = 1.7 \times 10^{-4}$; $p_{complement} = 1.5 \times 10^{-4}$).



2.5 Discussion

In 1941, Beadle and Tatum proposed the concept of one-gene, one-enzyme, one-function which posited that simple and linear connections are expected between genotype and phenotype.¹⁰² Summary-statistic-based PRSs, which afford advantages in broad applicability and intrinsic explainability, quantify attributed genetic liability in a similar manner. However, all too frequently the assumptions of this model breakdown, failing to account for the complexity of biological systems, such as the existence and contributions of epistatic interactions among genetic variants.

In this work, we introduce VADeR, a novel ViT-inspired architecture that integrates techniques from both NLP and CV to align with the unique characteristics of genetic data, capture interactive effects, and enhance predictive power in genotype-to-phenotype prediction. As proof-of-concept, we assessed its performance in PC prediction, and observed that across a range of metrics, including accuracy, AP, and MCC, VADeR outperformed all benchmark methods. Moreover, VADeR consistently outperformed each summary-statistic-based PRS baseline, including those which leveraged a much larger effective sample size through usage of external summary statistics (**Table 2.2**).

An investigation of baselines by PRS methodology (i.e., summary-statistic-based vs. ML-based) reveals an interesting, if perhaps not unexpected, finding. Specifically, ML PRS approaches, random forest excepting, have the potential to yield notable predictive improvements over their summary-statistic-based counterparts, particularly for rank-based metrics such as AUC and AP. However, unlike VADeR, this improvement was not consistent across metrics, as LDpred2* outperformed ML baselines globally in accuracy and MCC. Despite this, it is important to highlight the improved performance potential that appropriately regularized and hyperparameter optimized ML methods, even if simplistic, might afford the problem of genotype-to-phenotype prediction, especially if trained on datasets of size comparable to those from which large-scale summary statistics are computed.

While VADeR demonstrated superior performance across a range of metrics relative to baseline methods, one may question the practical significance of these gains given their modest

scale. It is important to contextualize these improvements within the inherent limitations of complex disease heritability. In this work we investigated PC, a condition for which heritability is estimated to be approximately 57%.^{103;104} Though this estimate indicates a substantial genetic contribution, it also highlights an upper bound on the predictive power of what is achievable through genetic data alone. Thus, given the limitations imposed by disease heritability, even incremental performance gains represent meaningful advancements, and further emphasize that we are working at the margins of what is genetically predictable.

AUC is a commonly used, rank-based metric for assessing the performance of PRS models.^{21;105} Interestingly, we observed that logistic regression outperformed VADeR by this metric, though not significantly ($p = 0.17$). Regardless of the significance, it is important to contextualize the paradoxical relationship that can exist between desirable model attributes, such as identifying distinct genetic heterogeneity, and their potential to deflate performance on a metric such as AUC. For example, VADeR identified the high-risk DARTH score dominant HOXB13 and TMPRSS2 clusters, both of which exhibited an enrichment in predicted cases, which by definition corresponds to elevated prediction probabilities. Though this stratification enhances the identification of high-risk individuals, it also introduces a cluster-shifting effect, which effectively concentrates high prediction scores in a way that can reduce overall rank differentiation. This nuance highlights the limitations of relying solely on rank-based metrics when evaluating models that capture distinct high-risk genetic substructure and interactions, and emphasizes the need for a multifaceted assessment of model performance.

Independent of rank-based metric intricacies, it is noteworthy that VADeR achieved strong performance despite the limited training set size relative to the model’s complexity. With a 12-layer, 12-headed Transformer architecture and a 768-dimensional embedding space, the number of trainable model parameters (96M) far exceeds the number of training samples ($n = 73,315$) and features ($n = 12,017$). Drawing parallels to the original ViT study, which found that Transformers can underperform convolutional neural networks (CNNs) on mid-sized datasets without strong regularization or large-scale pre-training,³⁹ our findings suggest that VADeR is surprisingly well-suited to genotype-to-phenotype modeling even without extensive data, augmentations, or pre-training. Moreover, it is reasonable to anticipate that as more

genomic data becomes available, allowing for increased training set sizes, the magnitude of VADeR’s performance gains could increase, though underlying disease heritability naturally imposes an upper bound.

While VADeR and other ML-based methods afford avenues for improving PRS performance, these approaches are of limited practical and clinical utility without a means to identify, interpret, and validate factors driving prediction. To that end, we complement the VADeR architecture with a local (i.e., individual-specific), region-level interpretability metric, DARTH scores. These scores, which leverage attention-based attribution techniques introduced by Chefer et al.,^{79;81} when collapsed to the global average resolution exhibit a pattern of attributed risk that closely, though importantly, not identically aligns with GWAS-identified high-risk regions (**Figure 2.2A**). Specifically, well-documented PC risk regions at 8q24, 10q11, and 17q24 were all identified as significant drivers of predicted risk by VADeR across the test set. Additionally, VADeR identified two globally prominent single-SNP patches on chromosomes 2 (rs13013607) and 11 (rs1624540), which to our knowledge have not yet been documented as PC risk variants. This suggests these SNPs, or the regions in which they fall, may represent novel drivers of PC risk, either independently or interactively, in individuals of European descent.

Leveraging the inherent personalized properties DARTH scores provide affords an opportunity to investigate and identify unique patterns of model attributed risk. Across the test set, we observed three distinct clusters driven by risk patterns that notably deviated from both the global average and GWAS-level perspectives (**Figure 2.3**). Specifically, each of these clusters exhibited an outlier dominated signature, with a specific patch dominating the attribution distribution. For clusters 1 ($n = 317$) and 2 ($n = 303$) the key drivers of identified risk were those regions containing the well-known PC risk genes HOXB13 and TMPRSS2, respectively.^{61–64} Notably, neither of these regions were among those with the highest global average DARTH score or p-value significance, marking these regions as rare but important drivers of predicted disease risk.

In contrast, cluster 3 ($n = 2,589$) was primarily driven by the most significant GWAS and global average DARTH score region on chromosome 10, a patch containing both MSMB and the PC risk variant rs10993994.^{92;93} Although this region is well-established for its PC risk association, individuals in cluster 3, like those in clusters 1 and 2, exhibited a markedly

elevated IBS score at the DARTH score dominant region, suggesting that distinctive genotype configurations amplify attributed risk. This idea is further supported by patch perturbation analyses, which indicate that risk associated with cluster-defining outlier patches is effectively captured by patch-level embeddings.

In conclusion, our study demonstrates the efficacy of the VADeR architecture as a powerful approach for genotype-to-phenotype modeling, particularly in the context of complex diseases like PC. Coupled with our introduction of DARTH scores, a robust, personalized, and region-specific metric, we afford a means to divine VADeR identified genetic substructure. Together, VADeR and DARTH scores enabled the identification of meaningful genetic heterogeneity and distinct regions of risk that align with documented PC risk regions (e.g., HOXB13, TMPRSS2, MSMB), indicating our model is capable of effectively capturing significant genetic patterns driving disease, even when rare, which holds potential clinical and therapeutic implications. Though this study primarily focused on PC prediction in European-ancestry individuals, it serves as a proof-of-concept for the viability of our joint architecture-interpretability approach. Expanding this framework to diverse populations and other genotype-to-phenotype prediction challenges offers an important avenue for future exploration. By advancing our understanding of genetic architectures underlying complex diseases, DARTH-VADeR offers a promising path towards improving both genomic risk scores and personalized medicine strategies.

2.6 Ancillary Details

2.6.1 Author Contributions

Original concept, J.V.T.; project supervision, H.C.; VADeR development and training, J.V.T.; baseline modeling, J.V.T., A.K., G.P.; DARTH-VADeR interpretability, J.V.T.; data acquisition and processing, J.V.T., A.K., M.S.P.; prostate cancer interpretation advising, B.R., T.S.; preparation of manuscript, J.V.T., H.C.

2.6.2 Conflicts of Interest

The authors declare no competing interests.

2.6.3 Acknowledgements

The results presented here are in large part made possible by data generated by both the ELLIPSE Consortium (dbGaP study: phs001120) and the UKBB (Project #37671). This work was supported by Emerging Leader Award from The Mark Foundation for Cancer Research, grant #18-022-ELA, NIH Grant R01CA269919 to H.C., and infrastructure grant 2P41GM103504-11.

For the dbGaP study phs001120 we acknowledge: Funding for the meta-analysis was provided by NIH grant U19CA148537. For de novo genotyping we would like to acknowledge the NCRN nurses and consultants for their work in the UKGPCS study. We thank all the patients who took part in this study. This work was supported by Cancer Research UK (grant numbers C5047/A7357, C1287/A10118, C1287/A5260, C5047/A3354, C5047/A10692, C16913/A6135 and C16913/A6835). We would also like to thank the following for funding support: Prostate Research Campaign UK (now Prostate Cancer UK), The Institute of Cancer Research and The Everyman Campaign, The National Cancer Research Network UK, The National Cancer Research Institute (NCRI) UK. We are grateful for support of NIHR funding to the NIHR Biomedical Research Centre at The Institute of Cancer Research and The Royal Marsden NHS Foundation Trust. The MEC was supported by NIH grants CA63464, CA54281 and CA098758.

The content of Chapter 2, in full, is a reformatted reprint of the material currently being prepared for submission for publication as “VADeR’s On That SNP: A New Hope For Transforming Polygenic Risk Scores And Revealing Genetic Heterogeneity” by James V. Talwar, Adam Klie, Meghana S. Pagadala, Gil Pasternak, Brent Rose, Tyler M. Seibert, and Hannah Carter. The dissertation author was the primary investigator and author of this paper.

Chapter 3

Autoimmune alleles at the major histocompatibility locus modify melanoma susceptibility

3.1 Abstract

Autoimmunity and cancer represent two different aspects of immune dysfunction. Autoimmunity is characterized by breakdowns in immune self-tolerance, while impaired immune surveillance can allow for tumorigenesis. The class I major histocompatibility complex (MHC-I), which displays derivatives of the cellular peptidome for immune surveillance by CD8+ T cells, serves as a common genetic link between these conditions. As melanoma-specific CD8+ T cells have been shown to target melanocyte-specific peptide antigens more often than melanoma-specific antigens, we investigated whether vitiligo- and psoriasis-predisposing MHC-I alleles conferred a melanoma-protective effect. In individuals with cutaneous melanoma from both The Cancer Genome Atlas ($n = 451$) and an independent validation set ($n = 586$), MHC-I autoimmune-allele carrier status was significantly associated with a later age of melanoma diagnosis. Furthermore, MHC-I autoimmune-allele carriers were significantly associated with decreased risk of developing melanoma in the Million Veteran Program ($OR = 0.962$, $p = 0.024$). Existing melanoma polygenic risk scores (PRSs) did not predict autoimmune-allele carrier status, suggesting these alleles provide orthogonal risk-relevant information. Mechanisms of autoimmune protection were neither associated with improved melanoma-driver mutation association nor improved gene-level conserved antigen presentation relative to common alleles. However, autoimmune alleles showed higher affinity relative to common alleles for particular windows of melanocyte-conserved antigens

and loss of heterozygosity of autoimmune alleles caused the greatest reduction in presentation for several conserved antigens across individuals with loss of HLA alleles. Overall, this study presents evidence that MHC-I autoimmune-risk alleles modulate melanoma risk unaccounted for by current PRSs.

3.2 Introduction

The incidence of cutaneous melanoma, the most common form, has seen an increase globally, particularly in Western countries.^{106;107} Early detection is a major determinant of overall disease prognosis with the 5-year survival rate dropping precipitously from 99% to 27.3% for local versus distant disease, respectively.⁶ Models developed for early disease detection are often built around well-known environmental and host risk factors including ultraviolet radiation exposure,^{108–110} pigmentary phenotypes,^{111–113} melanocytic nevi count,^{114;115} sex,^{116;117} age,^{116;117} telomere length,^{118;119} immunosuppression,^{120;121} and family history.^{122;123} However, though cutaneous melanoma ranks among the most heritable forms of cancer with an estimated heritability of 58%,¹⁰⁴ the majority of genetic susceptibility remains unaccounted for.

Cutaneous melanoma is also considered among the most immunogenic forms of cancer. Melanoma exhibits one of the highest mutation burdens across cancers, which is driven primarily by the mutagenic influence of ultraviolet radiation exposure.^{124;125} This increases the number of neoepitopes presented to the immune system and plays an essential role in immune surveillance. Immunosuppression, though, impairs the immune system’s cytotoxic potential and is a documented risk factor for increased melanoma incidence.^{120;121} Lymphocyte infiltration and melanoma-specific antibodies have been shown to be powerful prognostic factors as well.^{126;127} Immune traits themselves show considerable heritability,^{128–130} and early investigations suggest that heritable immune alleles also contribute to melanoma risk.^{131–133} In contrast to cancer, where poor immune function is a risk factor,^{134;135} increased sensitivity of the immune system can lead to autoimmune disorders.¹³⁶ This dichotomy has led to speculation that induction of autoimmunity in cancer patients could lead to tumor regression and better immunotherapy efficacy,^{137–142} though studies investigating the relationship between autoimmunity and cancer risk have returned mixed findings.^{143–145} If autoimmune alleles can enhance host anti-tumor

immune responses when immunotherapy is administered, it is possible that they could also enhance anti-tumor immunity more generally. Thus, we speculated that autoimmune alleles, particularly those related to T-cell responses directed against melanocyte or melanoma-specific antigens, might modify polygenic risk for melanoma in ways that might not be captured by current polygenic risk scores (PRS).

The class I major histocompatibility complex (MHC-I) represents a fundamental component of the antigen-directed immune response common to cancer and autoimmunity. MHC-I binds and displays peptide antigens derived primarily from intracellular proteins on the cell surface for immune surveillance by CD8+ T cells.¹⁴⁶ In cancer, neopeptides containing somatic mutations unique to the tumor genome, when displayed by MHC-I, can be recognized as foreign by CD8+ T cells, triggering the release of cytotoxic granules.^{147;148} The peptide-binding specificity of MHC-I is determined by three highly polymorphic genes, HLA-A, HLA-B, and HLA-C, encoded at the Human Leukocyte Antigen (HLA) locus on chromosome 6. The specific set of HLA alleles carried by an individual has been found to impose selective constraints on the developing tumor genome^{149–151} and modify response to immunotherapy.^{152–154} MHC-I loss, often due to deletion or mutation of HLA genes, is one mechanism of immune evasion during tumor development.^{155;156}

MHC-I also plays a role in several skin-specific autoimmune disorders. In vitiligo, destruction of melanocytes and consequent loss of skin pigmentation is mediated by CD8+ T-cell responses to self-antigens displayed by MHC-I.^{157;158} Another skin autoimmune disorder involving CD8+ T-cell responses is psoriasis,^{159;160} which is characterized by dermal leukocyte infiltration and hyperproliferation of keratinocytes.¹⁶⁰ CD8+ T cells in psoriasis have been shown to target melanocytes in patients carrying particular MHC-I alleles.^{161;162} Both of these conditions also share a risk variant rs9468925 in HLA-B/HLA-C,¹⁶³ supporting a shared MHC-I driven etiology, which is consistent with disease co-occurrence findings in clinical investigations.^{164;165}

Intriguingly, multiple vitiligo risk alleles implicated by genome-wide association studies have been shown to exhibit protection from cutaneous melanoma,^{131;132} and emergence of vitiligo during immunotherapy treatment of melanoma patients has been associated with better responses.^{166–169} Though psoriasis has also been documented as an immune-related adverse effect of immunotherapy, the association with melanoma prognosis is far less characterized.^{170;171}

Altogether these findings support the potential for class I autoimmune HLA alleles to modify melanoma risk and inform risk prediction. To further investigate this possibility, we evaluated autoimmune HLA carrier status in cutaneous melanoma samples from The Cancer Genome Atlas (TCGA). Skin autoimmune alleles were associated with a significantly later age at diagnosis among melanoma cases, which was recapitulated in an independent validation set of 586 individuals assembled from the UK Biobank and 4 other published melanoma genome sequencing studies. Moreover, these findings generalized to melanoma incidence with autoimmune HLA carriers significantly associating with a decreased risk of developing melanoma in the Million Veteran Program cohort. Finally, we investigated the peptide-specificity of autoimmune alleles for a set of 215 melanoma-specific driver mutations spanning 172 genes, and for conserved and cancer antigens previously implicated in melanoma-directed immunity. These analyses highlight a protective role for MHC-I in the context of melanoma development.

3.3 Material and Methods

3.3.1 Datasets

TCGA skin cutaneous melanoma tumors (SKCM) were used as the discovery set ($N = 470$). Cases were retained if they had appropriate clinical information for downstream analysis (i.e., age of diagnosis; 11 did not have this information), were microsatellite-stable (3 MSI tumors were removed), and were at least 20 years of age at time of diagnosis (5 were <20). Individuals below 20 years of age were excluded due to their increased likelihood of harboring rare predisposing risk variants. After filtering there were 451 tumor samples, the majority of which were from individuals of European ancestry ($N = 433/451$). Other ancestries represented were Asian ($N = 12$) and African American ($N = 1$) with the remaining few unknown. MHC-I genotypes were called using the exome based methods POLYSOLVER and HLA-HD.^{172;173}

An independent validation set of melanoma cases with germline WXS/WGS data was built from 5 separate melanoma studies. Two of these studies (Hugo et al.¹⁷⁴: SRP067938, SRP090294, Van Allen et al.¹⁷⁵: phs000452.v2.p1.c2) focused on melanoma response to immune-checkpoint inhibitors (ICPI). Two of these studies (Melanoma Exome Sequencing: phs000933.v2.p1.c1^{176;177}, The genetic and transcriptomic evolution of melanoma: phs001550.v2.p1.c1¹⁷⁸), were broader

melanoma studies focusing on the genetic basis of sun-exposed melanoma and melanoma evolution. The final study consisted of individuals from the UKBB with WXS/WGS with ICD10 codes: C433, C434, C436, and C437. We did not have ancestry information for the first 4 studies, and for the UKBB individuals, the majority were of European ancestry ($N = 236/239$). Validation set individuals were also filtered by age to exclude individuals under 20 years old. MHC-I genotypes for the validation set were called using HLA-HD.¹⁷³ In total there were 586 individuals in our validation set for which we were able to infer MHC-I autoimmune (AI) allele carrier status. Out of these 586, we were able to obtain fully-resolved HLA types (i.e., HLA-A, HLA-B, and HLA-C genotypes) for 559 individuals. Discovery and validation sets are described in **Table S1**.

Cases from the Melanostrum Consortium ($N = 3,001$) were used to evaluate PRS generalizability from absolute risk (i.e., cases vs. controls) to relative risk (i.e., age-specific effects). These cases were collected from countries of European ancestry including Italy ($N = 1,239$), Spain ($N = 1,024$), Greece ($N = 716$), and Cyprus ($N = 22$). Genotypes were available at 204 risk SNPs used in the development of the PRS by Gu et al.⁹ Cases from this dataset were filtered to only include individuals >20 years in age.

The Million Veteran Program (MVP; $N = 187,292$) was used to assess the generalizability of our findings to melanoma incidence. Both controls ($N = 171,878$) and cases ($N = 15,414$) were constructed from individuals both >20 years of age at the time of last follow-up and of European descent as determined by HARE (Harmonized Ancestry and Race/Ethnicity).¹⁷⁹ This was done to match the predominantly European composition of the discovery and validation datasets and to maximize the comparability of effect sizes relative to PRS SNPs measured in the Melanostrum study (Material and Methods: PRS Implementation). Controls excluded individuals with non-melanoma cancer diagnoses as determined by PheCode.¹⁸⁰ MHC-I alleles for the MVP were called using HIBAG¹⁸¹ with the multi-ethnic IKMB and Axiom UK Biobank array models.¹⁸² MVP 1.0 Axiom array design and genotype QC details are described elsewhere.¹⁸³

To identify putative conserved antigens we leveraged RNA-seq from a set of healthy melanocytes (dbGaP: phs001500.v1.p1) derived from newborn foreskins ($N = 106$) and version 7 of the Genotype-Tissue Expression (GTEx) project (dbGaP: phs000424.v7.p2).

3.3.2 Quantifying immune infiltration

To quantify immune infiltrates we ran CIBERSORTx¹⁸⁴ with the LM22 signature matrix¹⁸⁵ on the bulk RNA sequencing available for TCGA melanomas, which we downloaded from the TCGA GDC portal. RNA-seq reads were realigned and quantified by means of Sailfish¹⁸⁶ (with default parameters). Infiltration analyses were confined to primary tumors (N = 103) where the tumor immune microenvironment would be most relevant to the risk of developing melanoma. The effect of AI alleles on CD8+ T-cell infiltrates was modeled by multivariable regression, first including raw tumor mutation burden (TMB) as a covariate, and later including neoantigen burden at weak and strong cutoffs and BRAF p.Val600Glu mutation status.

3.3.3 Identifying driver mutations

We downloaded WES-based mutation calls from the TCGA GDC portal from four different mutation callers: VarScan,¹⁸⁷ MuSE,¹⁸⁸ MuTect,¹⁸⁹ and SomaticSniper.¹⁹⁰ RNA Variant Allelic Fraction (VAF) was obtained with bam-readcount. From these mutation calls we focused on single nucleotide variants, which dominate the landscape of melanoma and account for the majority of driver events.^{191–193} A mutation was considered to be a potential driver if it: 1) altered protein sequence, 2) was found in both the DNA and RNA in at least one individual, and 3) had a median DNA and RNA variant allele fraction (VAF) percentile less than or equal to 40%. DNA mutations were only considered at a patient-specific level if they were called by at least two of the mutation callers mentioned above. In total 51,062 mutations satisfied these criteria.

We further filtered the list of putative drivers based on recurrence. Specifically, if a specific mutation was detected in 4 or more different tumors we categorized it as a likely driver. In total 109 mutations satisfied this criterion (0.215% of the 51,062 candidate mutations). For those mutations that failed to reach this recurrence, we calculated mutation-specific contributions to melanoma pathogenicity using scores from a melanoma-specific CHASM classifier.^{194;195} Mutations with a CHASM score greater than or equal to 0.9 were deemed to be likely melanoma drivers; 106 mutations satisfied this criterion (0.206% of the 51,062 candidate mutations). Combining these recurrent and predicted driver singleton mutations yielded a final set of 215 melanoma drivers.

3.3.4 Identification and differential expression of conserved antigens

GTEX V7 contains 11,688 RNA-seq samples from 714 donors across 53 tissue types and was aligned with STAR v2.4.2a to GENCODE v19 and quantified with RNA-SeQC v1.1.8. RNA-seq reads from healthy melanocytes were aligned with STAR v2.5.0b to GENCODE v19 and quantified with RSEM v1.2.31. After quantification, both melanocyte and GTEX datasets were filtered such that genes with ≤ 0.5 RSEM or with counts < 6 in $> 93\%$ of samples were removed. Additionally, ribosomal RNA, Y chromosomal, and histone genes were removed. Ribosomal and histone mRNA are not polyadenylated. Notably, the melanocyte dataset is exclusively male, while GTEX is not, which would potentially lead to Y chromosomal genes being falsely identified as stably expressed downstream.

Across healthy melanocytes and each tissue type in GTEX, genes were scored as stably expressed genes (SEGs) using the output from the scoring method described in scMerge.¹⁹⁶ Briefly, the method first fits the expression of each gene from each tissue sample to a Gamma-Gaussian mixture. For the expression of a gene x_i , the Gamma component corresponds to samples with low expression and the Gaussian component corresponds to samples with high expression. This mixture has the joint density function

$$f(x_i; \alpha_i, \beta_i, \mu_i, \sigma_i, \lambda_i) = \lambda_i \times \text{Gamma}(\alpha_i, \beta_i) + (1 - \lambda_i) \times N(\mu_i, \sigma_i) \quad (3.1)$$

where α_i and β_i are the shape and rate parameters of the gamma component, μ_i and σ_i are the mean and standard deviation of the Gaussian component, and the mixing proportion λ_i is bounded by $[0, 1]$. The SEG scoring method also takes into account the proportion of zeros in each gene ω_i for each tissue. Each gene is then scored by the percentile ranks of its mixing proportion λ_i , coefficient of variation (CV) σ_i/μ_i , and proportion of zeros ω_i such that the average percentile rank across all three metrics is minimal. The highest scoring genes have lower mixing proportions, CVs, and proportion of zeros.

Using the aforementioned method, each gene from every tissue was fit to a Gamma-Gaussian distribution using scMerge v1.6.0 and given a score from 0 to 1 for stable expression. Then, the set of genes across all GTEX tissues that had scores > 0.69 was removed from the

set of genes in melanocytes that had scores >0.69 . The threshold of 0.69 was chosen based on the observation that the scores of canonical melanocyte genes were, with the exception of DCT/TYRP2, all ~ 0.7 and by visual inspection of the distributions of scores across tissues (**Figure C.1**). This process yielded 4 canonical melanocyte genes (PMEL, MLANA, TYRP1, TYR) and 48 additional protein coding genes for downstream analysis. The 48 non-canonical melanocyte genes were also passed to PANTHER¹⁹⁷ for Reactome pathway¹⁹⁸ overrepresentation analysis using Fisher’s Exact test, which identified several genes as members of the folate metabolism pathway (FDR 0.0288).

To determine differential expression of conserved antigens in melanoma relative to melanocytes, we applied the same expression quantification pipeline and gene filtering steps to both healthy melanocytes and TCGA SKCM samples. Specifically, we quantified HLA-allele specific expression using HLApers v1.0¹⁹⁹ and the Kallisto v0.44.0²⁰⁰ pipeline for HLApers. Reads were aligned to GENCODE v30²⁰¹ and IMGT HLA v3.41.0.²⁰² After quantification, both datasets were filtered in the same way as the conserved antigen identification pipeline described above. Additionally, 11 samples with missing age of diagnosis were removed from the TCGA SKCM set. We then performed a differential expression analysis with DESeq2 v1.30.1²⁰³ conditioned on ancestry. However, several potentially relevant covariates were also incompatible across these two datasets. Namely, age, sex, tumor type (primary or metastatic), tumor purity,²⁰⁴ melanocytic plasticity score,²⁰⁵ and TIDE score.²⁰⁶ Due to these discrepant covariates, we also checked that no covariates were associated with significant differential expression for any conserved antigens. The only gene subject to differential expression was MAGEA10 with a -1.3 ± 0.4 log fold change (LFC) in primary vs. metastatic melanoma. This is substantially less than the +8 LFC of MAGEA10 observed in melanocytes vs. melanoma (Results: *Investigating Mechanisms of MHC-I Autoimmune Allele Protection*).

3.3.5 Predicting binding affinities

MHC-I allele binding affinities were computed across the available 2,915 unique MHC-I alleles for both driver mutations and conserved antigens. Since driver mutations altered protein sequence, we evaluated MHC-I alleles’ ability to present neoepitopes by generating all unique

8-11mers found in a mutation relative to the wild-type (corresponding to the set of novel peptides a MHC-I allele can present to the immune system). To circumvent cross-allele and cross-peptide variabilities that are inherent in predicted IC50 comparisons, we used percentile ranks relative to a random set of peptides provided by NetMHCpan-4.1²⁰⁷ to approximate binding affinity for every MHC-I allele peptide pair. These percentile rank scores correspond to how strongly an allele binds a particular peptide relative to a set of random natural peptides. From peptide level rank scores, MHC-I mutation specific binding affinities were assigned according to the best rank score, the minimum allele-specific rank score across all unique 8-11mers for a mutation.

For conserved antigens, we partitioned proteins into their entire set of 8-11mers across the full length of the protein. MHC-I allele peptide pair percentile rank scores were again generated using NetMHCpan-4.1. Several metrics were derived from the percentile rank scores for downstream analyses. For broad and gene-level comparisons, we defined an allele-specific conserved antigen repertoire as the set of 8-11mers presented at or below a given percentile rank. For position-wise presentability (**Figure C.2**) we used the best percentile rank of all overlapping peptides at each position along a protein. We also defined the fraction of a gene presentable (FGP) by a given HLA allele as the fraction of 8-11mers from the encoded protein predicted to bind at a percentile rank less than 0.5.

3.3.6 HLA population allele representations

To compare HLA-B and HLA-C autoimmune (AI) alleles to common non-AI alleles, defined as those alleles with a population frequency $\geq 1\%$ as given by the National Marrow Donor Program (NMDP; 19 common HLA-B alleles, 13 common HLA-C alleles), we established maximum and minimum population allele representations. For driver neoantigens, a maximum population allele representation was assigned to have coverage at each rank equating to the coverage of the best presenting common allele at that rank. Similarly a minimum population allele representation was assigned to have coverage at each rank equating to the coverage of the worst presenting common allele at the rank. For maximum and minimum population allele representations in conserved antigens (CA), the metric of assignment was the fraction of CA peptides bound as opposed to coverage, as was consistent with CA analyses.

3.3.7 PRS implementation

We implemented the melanoma PRS developed by Gu et al.⁹ comprising 204 SNPs detected in subjects from Northern Europe, Australia, and the United States, and validated in subjects from Southern Europe. For the discovery set we were able to extract 190 of the 204 risk SNP genotypes using PLINK.²⁷ For the validation set, datasets lacking sufficient SNP data for PRS construction were excluded leaving 239 individuals for which 201 of the 204 risk variants were extracted using PLINK2.²⁸ For the Melanostrum Consortium ($N = 3,001$), all 204 risk SNPs were extracted. Risk SNP minor allele frequencies (MAFs) were compared across datasets to ensure no significant differences (**Figure C.3**). For the MVP we were able to extract 202 of the 204 risk variants (rs2025016 and rs6833655 were missing) using PLINK2.²⁸ The final PRS for each dataset was generated as a weighted sum across extracted risk SNPs, ensuring SNPs were oriented to the correct allele, in each dataset in accordance with the optimal melanoma risk model by Gu et al.⁹

3.3.8 Multistage carcinogenesis model for melanoma

Dating back to the 1950's Armitage-Doll model²⁰⁸ of cancer incidence and those created soon after by Knudsen and Moolgavkar²⁰⁹ and others, multistage models of cancer are among the most developed mathematical methods for defining carcinogenesis and determining timescales of tumor formation in human populations.^{210–214} These models assume evolutionary stages from normal cells to development of clinically detected symptomatic cancers. These stages typically include intermediate premalignant and preclinical malignant stages that represent field cancerization dynamics of stochastically growing and shrinking clonal populations in a tissue. These models can be described mathematically as stochastic multi-type branching processes with probabilities of events occurring with certain rates (**Figure C.4A**). By calculating an age-dependent hazard function for cancer incidence using solutions to equations from the probability generating functions starting from birth, we can calibrate these models to fit hazard rates derived from cancer incidence registry data such as Surveillance, Epidemiology, and End Results (SEER) in the US.²¹⁵ Importantly, this modeling framework provides a link between cell-level dynamics and population-level incidence data so that we can estimate parameters governing clonal growth,

dwell times, and mutational “hits” in at-risk individuals.

In previous work we found that the “two-stage” model (2 “hits” for development of a first malignant cell) shown in **Figure C.4** is closely approximated by a model that includes an effective malignant transformation rate and a characteristic lag-time or “sojourn” time between malignant transformation and clinical detection (see Luebeck et al. 2013 for mathematical details²¹⁶). Here we created a two-stage model for melanoma incidence that adjusts for birth trends, similar to methods used previously in esophageal squamous cell carcinoma (ESCC).²¹⁷ In this way, our models capture trends for both age and birth (and thus calendar period) to enable robust estimation via Markov Chain Monte Carlo (MCMC) simulation of cell-level parameters for tumor evolution by sex and race/ethnicity (in the absence of genetic ancestry; see **Figure C.4B** for examples of model fits). We obtained estimates and 95% confidence intervals for tumor sojourn times in males and females via MCMC posterior estimates for the lag-time parameter. Chains were run for 100,000 cycles with a 4,000 cycle burn-in and checked for convergence. All code for hazard function calculation and parameter estimation was written in Fortran. The ICD-O-3 codes used for extraction of SEER data melanoma, all races combined, from SEER*Stat include: 8720/3, 8721/3, 8722/3, 8723/3, 8726/3, 8727/3, 8728/3, 8730/3, 8740/3, 8741/3, 8742/3, 8743/3, 8744/3, 8745/3, 8746/3, 8761/3, 8770/3, 8771/3, 8772/3, 8773/3, 8774/3, 8780/3, 8790/3.

3.3.9 Personalizing sojourn time estimates from clock-like mutational signatures

To personalize sojourn time estimates, we combined the sex-specific estimates returned by our two-stage model for melanoma incidence (Material and Methods: *Multistage Carcinogenesis Model for Melanoma*) with tumor specific genetic marks in TCGA. Specifically, we quantified the per patient UV-corrected SBS1 signature by counting CpG>TpG mutations after excluding CpCpG and TpCpG mutations. This adjusted SBS1 has previously been used as a molecular clock for melanoma.^{218;219} After extracting UV-corrected SBS1 CpG>TpG mutations in TCGA (N = 397/451), we fit a linear model predicting age of diagnosis from these clock-like mutational signatures. Personalized predicted age of onset was then estimated by:

$$\begin{aligned} \text{Age of Onset} = & \text{Age of Diagnosis} - (\text{Sex Specific Tumor Sojourn Time Estimate} \\ & + (\text{C} > \text{T Burden} - \text{median}(\text{C} > \text{T Burden})) \cdot \beta_{\text{C} > \text{T mutation}}) \end{aligned} \quad (3.2)$$

where sex-specific tumor sojourn time estimate equates to the outputs of our two-stage melanoma incidence model, C>T burden equates to the number of UV-corrected SBS1 CpG>TpG mutations one carries, and $\beta_{\text{C} > \text{T mutation}}$ equates to the coefficient of the linear model predicting age of diagnosis from UV-corrected SBS1 CpG>TpG mutations. For those individuals for which UV-corrected SBS1 CpG>TpG mutations could not be inferred, we substituted the median UV-mutation spectrum corrected CpG>TpG mutation burden as their C>T burden, leading to a shift equivalent to the aforementioned sex-specific sojourn time estimates. Note here that personalized sojourn time estimates equate to the shift (i.e., subtraction) from age of diagnosis.

To assess the validity of these personalized sojourn time estimates, we ensured these estimates fell within expected ranges. Specifically, we compared our personalized sojourn time estimates to the expected time of whole genome duplication (WGD) to melanoma diagnosis as reported by Gerstung et al.²¹⁸ We calculated WGD status for TCGA individuals from ploidy, which was computed as the length weighted sum of genomic segments with constant total copy number. For more details, we direct the reader to the work by Steele et al.²²⁰ from which these values were obtained.

3.3.10 Calculating HLA loss of heterozygosity

Inference of HLA loss of heterozygosity (LOH) was determined by means of the computational tool LOHHLA.¹⁵⁵ As input we provided normal and tumor DNA BAM files, consensus HLA calls (obtained from POLYSOLVER and HLA-HD),^{172;173} and purity and ploidy estimates as returned by Sequenza.²²¹ Upon completion of LOHHLA, HLA alleles were classified as lost if they exhibited a copy number <0.5, in accordance with LOHHLA documentation. We also employed an additional quality control step ensuring all alleles classified as lost had a reported p-value <0.05. Across skin cutaneous melanoma tumors (SKCM) in the TCGA, we observed 90 incidences of HLA LOH across 56 individuals. For the 7 autoimmune alleles investigated in this

study, we observed 16 incidences of autoimmune HLA LOH across 12 autoimmune allele carriers.

3.3.11 Statistical analyses

All box plot statistical tests comparing age of diagnosis effects between groups were assessed using the default Mann-Whitney U statistical test. Leave-one-out analysis was conducted by narrowing the AI allele set into all 7 unique sets of 6 AI alleles and stratifying individuals accordingly. Performing leave-one-out analysis by dropping all carriers of each allele yielded similar results, with AI status significantly associated with a later age of diagnosis in each holdout set. T-tests were used to compare PRS distributions across AI-allele status in both discovery and validation sets. Fisher’s exact tests were used to evaluate associations between HLA-proximal PRS SNPs and MHC-I AI allele carrier status. These statistical tests were all implemented via the default `scipy.stats` Python package. Regression analyses were modeled using ordinary least squares linear models through the `statsmodels.formula.api` Python package.²²² Effect sizes for AI alleles and PRS SNPs in the TCGA were calculated using Cliff’s D. Effect sizes for AI alleles and PRS SNPs in the case-control MVP were reported as odds ratios calculated from logistic regression through the `statsmodels.api` Python package. Wherever multiple hypotheses were tested, p-values were corrected by the Benjamini-Hochberg procedure, implemented through the `statsmodels.stats.multitest` package in Python.

3.4 Results

3.4.1 MHC-I autoimmune alleles associate with a later age of diagnosis in melanoma

To investigate the relationship between autoimmunity and melanoma development, we sought evidence that MHC-I autoimmune alleles could provide protection from melanoma. We defined a set of 8 MHC-I alleles based on documented links to skin autoimmune conditions. This set included 3 alleles linked to psoriasis (HLA-B*27:05, HLA-B*57:01, HLA-C*12:03),^{223–227} 2 alleles linked to vitiligo (HLA-A*02:01, HLA-B*13:02),^{228–233} and 1 allele linked to both conditions (HLA-C*06:02).^{232–235} We also included 2 alleles (HLA-B*39:06, HLA-B*51:01) with postulated psoriasis associations,^{223;236} but strong associations with other autoimmune conditions, specifically type 1 diabetes^{237–239} and Behcet’s disease,²⁴⁰ respectively (**Table S2**). Among

these, all but HLA-A*02:01 are associated with a narrow specificity for antigen^{241;242} and have a population allele frequency <10% (**Table S3**).

Using individuals with skin cutaneous melanoma tumors (SKCM) from the TCGA as our discovery set (**Table S1**), we called MHC-I genotypes using two exome based methods: POLYSOLVER and HLA-HD (Material and Methods: *Datasets*).^{172;173} Autoimmune allele frequencies in the discovery set were present at the population distributions as reported by the NMDP²⁴³ (**Table S3**). Individuals under 20 years of age were also excluded from further analysis given their increased likelihood of harboring rare germline predisposing risk variants.

To evaluate the effect of carrying an autoimmune (AI) allele, we partitioned the discovery set into two groups: those with at least one AI allele and those lacking any of these alleles. We first assessed the potential for AI allele carrier status to be confounded by sex or UV exposure, two factors that influence melanoma incidence. Males have a well-documented higher risk of developing melanoma,^{244;245} however, we did not observe any sex specific age differences (**Figure C.5A**, $p = 0.247$). UV-associated mutational signatures correlated strongly with overall tumor mutation burden in the discovery set (Pearson $R = 0.984$; **Figure C.5B**). There were also no significant differences in mutation burden or UV-associated mutational signatures between AI allele carriers and non-carriers (**Figure C.5B-D**; A*02:01 included: $p_{\text{mutation}} = 0.365$, $p_{UV} = 0.186$; A*02:01 excluded: $p_{\text{mutation}} = 0.352$, $p_{UV} = 0.415$).

We hypothesized that among individuals diagnosed with melanoma, a protective effect would manifest as delayed disease onset relative to individuals without these alleles. As HLA-A*02:01 differs from the other 7 AI alleles in antigen presentation properties^{241;242} and population frequency, and may promote higher expression of the immune suppressive checkpoint TIM-3 by T cells relative to the other AI alleles,²⁴⁶ we considered AI carrier status both excluding and including HLA-A*02:01. In the discovery set, AI allele carrier status excluding HLA-A*02:01 was significantly associated with a median later age of melanoma diagnosis of 5 years (**Figure 3.1A**, $p = 0.002$). Upon including HLA-A*02:01, this association was lost (**Figure C.5E**, $p = 0.316$). Evaluation of T-cell infiltration levels estimated by CIBERSORTx¹⁸⁴ (Material and Methods: *Quantifying Immune Infiltration*) indicated that primary tumors carrying AI alleles had significantly lower levels of regulatory T cells (**Figure 3.1B**, $p = 0.048$). However, tumors

in HLA-A*02:01 carriers showed no reduction in regulatory T cells (**Figure C.5F-G**). For subsequent analyses we therefore defined AI allele carrier status based on the 7 non-HLA-A*02:01 AI alleles; individuals carrying only HLA-A*02:01 were considered non-carriers.

To further probe allele-specific contributions we performed a leave-one-out analysis, evaluating AI age effects across all 7 single allele exceptions. Regardless of which allele was held out, we consistently observed a significant relationship between AI status and age of diagnosis (**Figure C.6**), with a 5 year later median age of melanoma diagnosis (median age without AI = 57; median age with AI = 62). We did not observe significant differences at the level of HLA supertype, suggesting that the effects are specific to individual autoimmune alleles (**Figure C.7**); the largest observed effect was for the B44 supertype, to which none of the AI alleles belong, however it was not significant after multiple testing correction ($p_{adj} = 0.199$, median earlier age of diagnosis difference = 4 years).

For the 7-allele definition of carrier status, later age of diagnosis was conserved across sex with a median age difference of 6 years in males and 4 years in females (**Figure 3.1C**, $p_{male} = 0.011$, $p_{female} = 0.033$). We used an ordinary least squares (OLS) regression to model the effect of carrying an AI allele on age at diagnosis with tumor thickness, sex, and mutation burden as covariates and these findings remained significant with a predicted 4.76 delayed years to melanoma diagnosis (**Figure 3.1D**, $p = 0.003$). As tumor thickness was only available for a subset of tumors ($N = 347$), we also evaluated the effect of AI allele carrier status with primary vs. metastatic disease, sex, and mutation burden as covariates ($N = 416$), and sex and mutation burden only as covariates ($N = 451$). Both analyses yielded similar results, with a predicted 3.62 (**Figure C.8A**, $p = 0.015$) and 4.07 (**Figure C.8B**, $p = 0.005$) delayed years to melanoma diagnosis respectively. We further evaluated whether carrying multiple AI alleles had an additive effect. Using a linear model to predict age of diagnosis as a function of an individual's total number of autoimmune alleles, the discovery set showed a significant 2.6 delayed years to melanoma diagnosis per autoimmune allele (**Figure 3.1E**, $p = 0.002$; CI = [0.935, 4.262]).

We also observed that the number of autoimmune alleles significantly associated with increased CD8+ T-cell infiltration in primary tumors (**Figure 1F**). In a linear model predicting CD8+ T-cell infiltration as a function of an individual's total number of autoimmune alleles, the

per MHC-I AI allele CD8+ T-cell infiltrate increase was 2.7% ($p = 0.040$; CI = [0.127%, 5.282%]). When we included tumor mutation burden as a covariate, the effect of AI alleles on increased CD8+ T-cell infiltration remained significant, yielding a 2.65% increase per MHC-I AI allele carried ($p = 0.045$; CI = [0.056%, 5.253%]). Together, these analyses suggest that the delayed diagnosis associated with AI alleles could be driven by a more robust CD8+ T-cell response less impeded by regulatory T-cell activity.

To confirm these findings, we investigated the age of melanoma diagnosis and autoimmune allele presence in an independent validation set of 586 individuals diagnosed with cutaneous melanoma, compiled from 4 published dbGaP studies^{174–178} and the UKBioBank. Although AI allele frequencies once again matched closely with the expected distribution (**Table S3**) we observed notable differences in the distribution of sex and age (**Table S1**), especially for the UKBioBank. Whereas in TCGA the sex and age distributions of melanoma diagnosis closely matched the rates reported by SEER (a M/F ratio of 1.637 in TCGA, versus an M/F ratio of 1.624 in SEER), females were overrepresented in the UKBioBank (an M/F ratio of 0.498) and female sex across the entire validation set was significantly associated with an earlier age of melanoma diagnosis independent of AI status (**Figure C.9A**; $p = 0.003$, median earlier age of diagnosis difference = 3 years). This suggests an intrinsic selection bias within the studies from which our validation set was compiled. If these biases relate to unmeasured environmental risk factors, they could mask the contribution of genetic risk factors. Despite this, we again observed that having at least one autoimmune-linked MHC-I allele was significantly associated with a later age of diagnosis (**Figure 3.1G**, $p = 0.044$; median age separation = 1 year). While the direction of this effect was conserved across males with a median age difference of 1 year, we did not observe any significant differences in females (**Figure C.9B**; $p_{male} = 0.055$, $p_{female} = 0.236$).

As most validation samples lacked tumor sequencing data, it was not possible to estimate UV exposure. Thus, our regression analysis was limited to the presence of an AI allele and sex. Validation individuals with at least one AI allele had a predicted 2.10 delayed years to melanoma diagnosis relative to those without any of these alleles ($p = 0.082$; CI = [-0.265, 4.457]). By comparison, the discovery set showed a predicted 4.0 delayed years to melanoma diagnosis when only AI status and sex are considered ($p = 0.006$; CI = [1.154, 6.850]). In the validation group

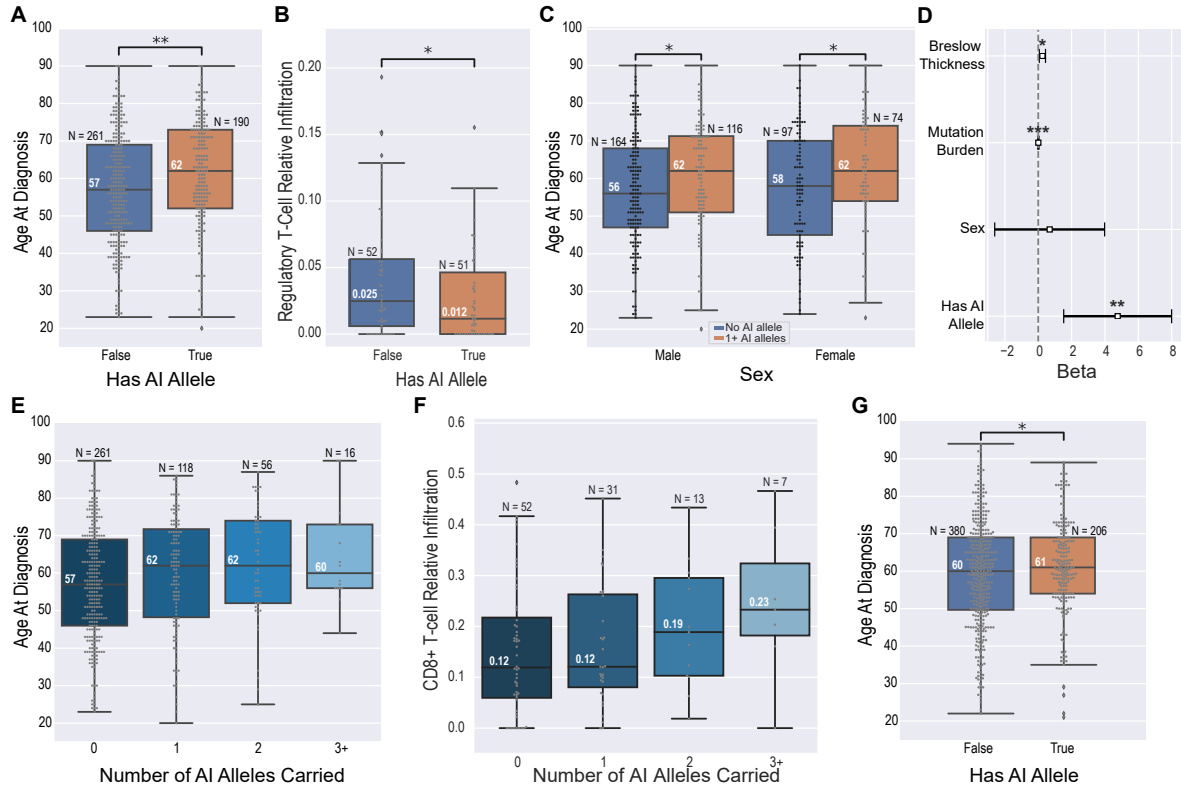


Figure 3.1. Effect of MHC-I AI alleles on melanoma age at diagnosis and immune infiltration. A) TCGA: having at least one MHC-I linked AI allele is associated with a significant median later age of melanoma diagnosis of 5 years ($p = 0.002$). B) TCGA: carriers of at least one MHC-I AI allele exhibit a significant reduction in regulatory T cells in primary melanoma tumors ($p = 0.048$). C) TCGA: MHC-I-linked AI allele age of melanoma diagnosis effect is conserved across sex ($p_{male} = 0.011$, $p_{female} = 0.033$). D) TCGA: effect sizes and significance of coefficients in a multivariable model fitted to age at diagnosis. Having at least one MHC-I-linked AI allele associated significantly with a 4.76-year delay to melanoma diagnosis after controlling for tumor thickness, sex, and mutation burden ($p_{autoimmune} = 0.003$, $CI = [1.581, 7.943]$). E) TCGA: age of diagnosis significantly increases with an individual's total number of MHC-I autoimmune alleles ($p = 0.002$). F) TCGA: infiltrating CD8+ T cell levels increase with the number of AI alleles carried ($p = 0.040$). G) Validation: having at least one MHC-I linked AI allele is associated with a significant median later age of melanoma diagnosis of 1.0 year ($p = 0.044$).

individuals with fully-resolved HLA types ($N = 559/586$), we observed a predicted 0.727 year delay to melanoma diagnosis per autoimmune allele, though these results did not reach statistical significance ($p = 0.328$; $CI = [-0.731, 2.184]$).

Finally, we attempted to identify other HLA alleles associated with age at diagnosis by comparing individuals with a particular allele to all individuals without that allele. While none of the associations were significant after multiple hypothesis testing correction, we did note that HLA-B*27:02 showed an earlier median age of diagnosis in both datasets (**Figure C.10**; SKCM-TCGA: 10 years earlier, $N_{Has\ HLA-B*27:02} = 15$, $p = 0.050$, $p_{adj} = 0.483$; Validation Set: 11 years earlier, $N_{Has\ HLA-B*27:02} = 3$, $p = 0.033$, $p_{adj} = 0.388$). However, given the limited dataset sizes with sequencing data, coupled with the highly polymorphic nature of the HLA, this analysis is underpowered and may warrant further investigation in larger datasets.

3.4.2 MHC-I autoimmune alleles modify melanoma risk

Polygenic risk scores (PRS) use information about genetic risk factors to predict individual disease risk. We evaluated the utility of incorporating MHC-I AI allele carrier status into a risk scoring framework in the context of the PRS developed by Gu et al.⁹ (Material and Methods: *PRS Implementation*). This PRS comprises 204 SNPs, of which 16 are found on chromosome 6, though none fall within the HLA class-I region (**Figure 3.2A**). Although the HLA class-I genes are not among the genes associated with each SNP as reported by Gu et al.,⁹ we assessed whether HLA-proximal PRS SNPs (i.e., those SNPs within 3MB of the HLA-coding region) associated with MHC-I AI allele carrier status across our discovery set, but observed no significant relationship ($p_{adj} = 0.235$).

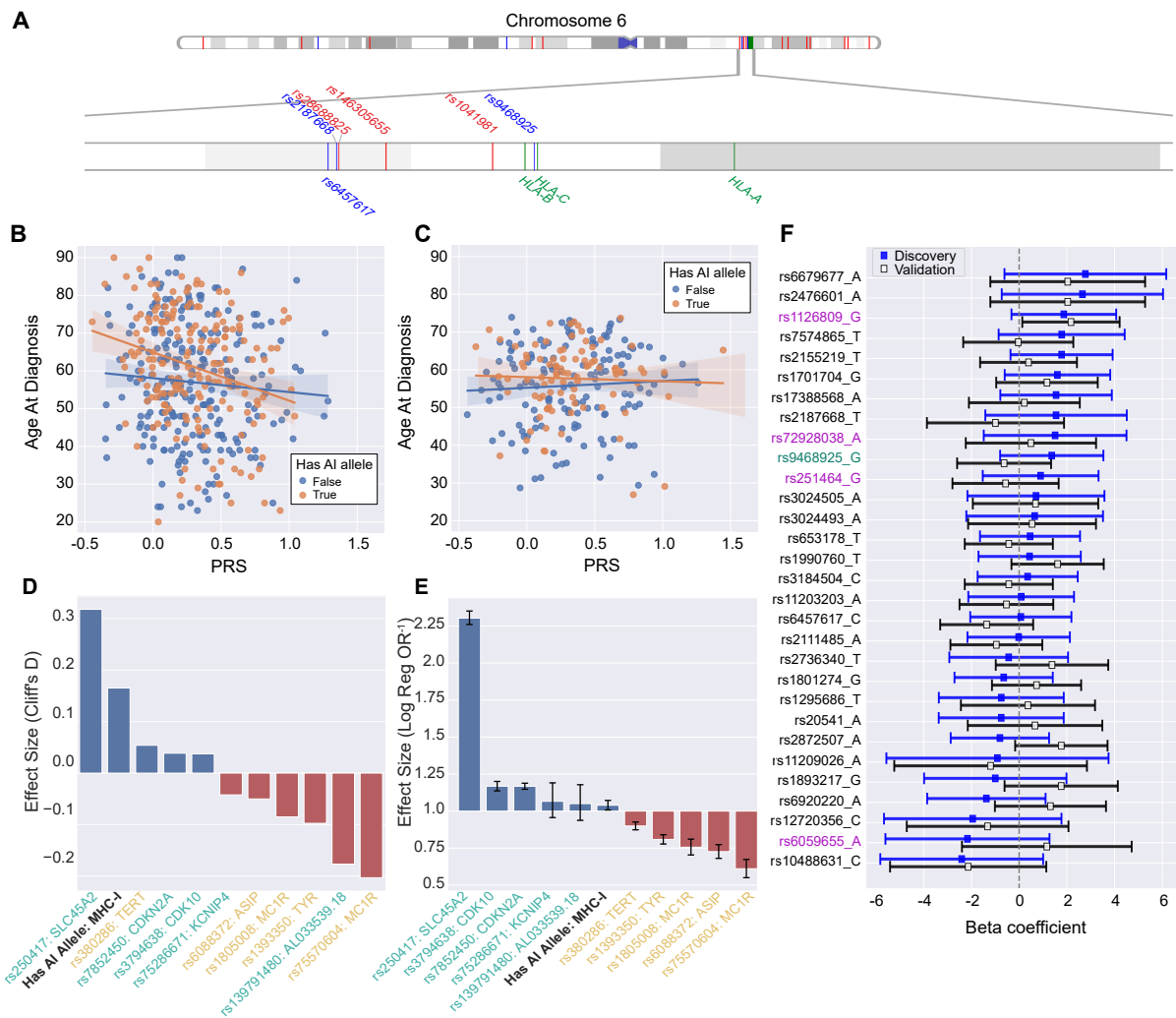
As the PRS was developed to stratify cases from controls, we ensured that it could be generalized to capture age-specific effects in a case-only setting through regression analysis. We observed that higher PRS associated significantly with earlier age of diagnosis in the TCGA (discovery: $p = 0.002$, $Beta = -7.499$, $CI = [-12.141, -2.857]$), but not in the validation set (validation: $p = 0.710$, $Beta = 0.7551$, $CI = [-3.237, 4.747]$) consistent with unmeasured confounders masking genetic risk factors. Given the disparity in PRS generalization, we evaluated the PRS age stratification in melanoma cases from the Melanostrum Consortium ($N = 3,001$), the

original PRS validation set.⁹ Here we observed that higher risk scores were again biased toward an earlier age of diagnosis ($p = 0.055$, Beta = -1.780, CI = [-3.598, 0.038]). While the MAF of the PRS SNPs generally correlated well across TCGA, Melanostrum and the validation set, we did observe certain SNPs with a MAF difference ≥ 0.1 across datasets. Four SNPs exhibited this difference between our discovery set and Melanostrum (rs1464510, rs187989493, rs7041168, rs7164220). Between our validation set and Melanostrum, seven SNPs exhibited this MAF gap (rs187989493, rs1393350, rs7164220, rs13338146, rs12919293, rs75570604, rs2092180) including the maximum PRS effect SNP (rs75570604). Finally, between our discovery and validation sets one SNP exhibited a MAF difference ≥ 0.1 (rs1464510) (**Figure C.3**). These discrepancies may explain the varying performance of PRS in age stratification across datasets.

We next evaluated the relationship between PRS and age at diagnosis in AI allele carriers versus non-carriers. As expected, PRS score distributions did not differ between those with and without MHC-I AI alleles in either dataset (**Figure C.11**; $p_{disc} = 0.999$, $p_{val} = 0.901$). Increases in PRS showed a greater negative effect on age at diagnosis in those with an autoimmune allele relative to those without one in the discovery set (**Figure 3.2B**). However, interaction effects from a linear model between PRS and AI MHC-I allele genotype in the discovery set, while large in magnitude, did not reach significance ($p = 0.055$, Beta = -9.124, CI = [-18.432, 0.183]) and did not reproduce in the validation set (**Figure 3.2C**).

We also sought to assess whether AI carrier status was associated with melanoma incidence. In a large case-control dataset composed of individuals from the Million Veteran Program²⁴⁷ (MVP; N = 187,292; Material and Methods: *Datasets*), AI MHC-I carriers were less likely to have a melanoma diagnosis (OR = 0.962, $p = 0.024$), as would be expected if MHC-I AI alleles conferred a protective effect against melanomagenesis. Each additional autoimmune allele carried further decreased the likelihood of melanoma (Beta = -0.031, $p = 0.003$). To quantify the magnitude of AI MHC-I allele carrier status relative to individual PRS SNPs, we compared effect sizes in both the discovery set and the MVP. PRS stratification in the MVP yielded an AUC of 0.61, in line with, but less than the reported 0.644 AUC in the Melanostrum Consortium⁹ (Material and Methods: *PRS Implementation*). Compared to the five largest PRS weight SNPs in both directions (i.e., melanoma predisposing and protective), MHC-I AI allele status exhibited

Figure 3.2. Contextualization and quantification of MHC-I AI alleles' effect on melanoma risk A) A PhenoGram plot of chromosome 6 shows melanoma PRS SNPs (red) and AI-associated SNPs (blue) fall outside of the HLA-coding region (green). The distance from the closest PRS SNP (rs1041981) to the class I-coding region (HLA-B) is 215,819 bp. The closest AI SNP (rs9468925) falls in between HLA-C/HLA-B. B) TCGA: age of diagnosis as a function of PRS and AI allele presence. Individuals with AI MHC-I alleles show a steeper decrease in age of diagnosis as PRS increases ($p = 0.055$, Beta = 9.124, CI = [18.432, 0.183]). C) Validation: age of diagnosis as a function of PRS and AI allele presence. Individuals with AI MHC-I alleles again show a steeper decrease in age of diagnosis as PRS increases, but interaction effects were insignificant ($p = 0.499$, Beta = 2.878, CI = [11.254, 5.498]). D) TCGA: Cliff's D for having at least one MHC-I-linked AI allele, the five strongest PRS melanoma-predisposing SNPs mapped to their nearest gene (yellow), and the five strongest PRS melanoma-protective SNPs mapped to their nearest gene (turquoise). E) MVP: inverse odds ratios (1/OR) for MHC-I AI allele presence, the five strongest PRS melanoma-predisposing SNPs mapped to their nearest gene (yellow), and the five strongest PRS melanoma-protective SNPs mapped to their nearest gene (turquoise). F) AI SNP effect on age of diagnosis: AI SNPs showed varying associations with age of diagnosis ranging from strongly protective (high beta) to strongly predisposing (low beta). Overall these effects were highly variable and did not reach statistical significance after multiple-hypothesis correction. Joint vitiligo-melanoma-associated SNPs are marked in red, the lone HLA-C/HLA-B psoriasis and vitiligo SNP is marked in green, and the remaining 25 broad AI SNPs are marked in black. Error bars correspond to ± 2 standard deviations.



the fourth largest effect size by magnitude and second largest positive effect size in the discovery set (**Figure 3.2D**, Cliff’s $D = 0.165$). In the MVP, the effect of having an MHC-I linked AI allele was more modest, with the smallest odds ratio by magnitude relative to the 10 largest weight PRS SNPs (**Figure 3.2E**, $OR = 0.962$, $CI = [0.930, 0.995]$).

We next evaluated whether these trends extended to non-MHC AI risk SNPs. In total we examined 30 AI SNPs, including four with established vitiligo-melanoma associations either as the joint lead risk SNP for both conditions (rs1126809, rs6059655) or in strong linkage disequilibrium with known cutaneous melanoma risk SNPs (rs72928038, rs251464),¹³³ and one (rs9468925) that is associated with both psoriasis and vitiligo and falls in between HLA-C/HLA-B.¹⁶³ The remaining 25 AI SNPs are broadly associated with autoimmunity according to the criteria outlined by Chat et al.¹⁴¹ (i.e., associated with at least three autoimmune conditions, and at least one of which surpassed a significance threshold of $p = 10^{-7}$) and were previously investigated in the context of improved immune-checkpoint inhibitor efficacy. While coefficients for the relationship between AI SNP genotype and age of melanoma diagnosis ranged from strongly protective (e.g., rs6679677; $Beta_{Disc} = 2.775$; $Beta_{Val} = 2.036$) to strongly predisposing (e.g., rs10488631; $Beta_{Disc} = -2.407$; $Beta_{Val} = -2.137$) across datasets, overall these effects exhibited large variability and were not significant after multiple-hypothesis testing (**Figure 3.2F**). In contrast to MHC-I AI alleles, including non-MHC-I AI SNPs as covariates with PRS did not improve prediction of age at diagnosis.

3.4.3 MHC-I autoimmune alleles associate with delayed onset of melanoma

To further evaluate whether delayed diagnosis observed for AI alleles corresponds to a delay in onset of melanomagenesis, we sought to impute individual-specific sojourn time, the time period from tumor initiation to diagnosis. Age at onset can then be estimated by subtracting sojourn time from age at diagnosis. To obtain an individualized estimate of sojourn time, we combined an average sojourn time estimated from population statistics with a tumor specific estimate based on the clock-like mutational signature, SBS1.²¹⁹

To estimate the expected time between the initial transformed malignant cell (in a surviving malignant clone that escapes extinction) and clinical detection, we developed a cell-

based stochastic branching process model for the development of independent premalignant clones (such as nevi) that can arise and clonally expand in normal skin epithelium (Material and Methods: *Multistage Carcinogenesis Model for Melanoma*). Each cell in these clones has the propensity to transform to a malignant cell with a certain probability, the malignant clone population can expand in size or go extinct through a stochastic birth-death process, and clinical detection may occur with a size-based detection probability. Mathematically, the expectation of the lag-time variable, or the time between the founder cell of a persistent malignant clone and clinical detection, can be interpreted as the average “age” or sojourn time of the detected tumor.²¹⁶

To obtain the estimated average, we analyzed SEER9 melanoma age- and group-specific incidence data from 1975-2018.²¹⁵ The hazard function from a “two-stage” model corresponded to the best fit to SEER incidence for both males and females (**Figure C.4**). Adding additional stages to the model (i.e., more than 2 rate-limiting events or “hits” such as driver mutations required before malignant transformation) did not improve the fits. With estimated model parameters, we found that the expected tumor sojourn time in males was 8.35 years (Markov chain Monte Carlo [MCMC] 95% CI = [6.61, 9.73]) and similarly in females was 9.64 years (95% CI = [8.48, 10.66]). Previous studies have estimated melanoma doubling times that can be used to then calculate the corresponding tumor sojourn time. With a mean doubling time of 144 days,²⁴⁸ the mean growth rate in an exponentially growing tumor is approximately 1.76 per year. Assuming malignant tumors are detected on average at 10^8 or 10^9 cells in size, this implies a melanoma sojourn time of 10.5 years and 11.8 years, respectively. This is in line with our above estimates, along with those found in a previous modeling study of melanoma doubling times (mean = 3.78 months).²⁴⁹ Although estimates may vary based on patient-specific factors, our findings suggest that it takes approximately a decade on average for a melanoma to be detected after it is first initiated in an individual.

We next approximated an individual sojourn time based on SBS1-associated mutation burden, such that the median sojourn time would correspond to the estimated baseline. Mutational signature SBS1 has been used previously as a molecular clock in melanoma, after first excluding the subset of mutations that overlap with the SBS7 signature of UV-driven mutagenesis,

specifically mutations at CpCpG and TpCpG DNA contexts.²¹⁸ We compared the UV-corrected SBS1 CpG>TpG mutations to the uncorrected SBS1 in TCGA, confirming that it showed increased correlation with age of diagnosis (**Figure C.12A-B**; $r_{C>T} = 0.230$, $r_{SBS1} = 0.112$; $\rho_{C>T} = 0.238$, $\rho_{SBS1} = 0.211$) and greatly reduced correlation with UV mutagenesis as quantified by SBS7 (**Figure C.12C-D**; $r_{C>T} = 0.284$, $r_{SBS1} = 0.816$; $\rho_{C>T} = 0.353$, $\rho_{SBS1} = 0.434$). Fitting a linear model to predict age at diagnosis from UV-corrected SBS1 mutation burden, each mutation equated to approximately 0.85 years. We then estimated individualized sojourn times by matching the median UV-corrected SBS1 burden to the estimated average sex-specific time and adding or subtracting time according to how much larger or smaller the observed UV-corrected SBS1 burden was relative to the median (Material and Methods: *Personalizing Sojourn Time Estimates from Clock-Like Mutational Signatures*).

To assess whether the estimates obtained are reasonable, we evaluated the sojourn time for individuals with ploidy profiles indicating whole genome duplication (WGD) events.²¹⁸ Gerstung et al.²¹⁸ estimated that WGD occurs approximately 5.7 years prior to melanoma diagnosis with a range of 3.2 to 16 years. We reasoned that if our estimate of sojourn time based on UV-corrected SBS1 provided a good approximation, then the individualized sojourn time should exceed the timing of WGD in those tumors where it occurred. Estimating WGD events from ploidy information in TCGA, we found that sojourn estimates were significantly longer, and generally exceeded 5.7 years, for the subset of tumors with WGD (**Figure C.12E**).

Subtracting the individual sojourn time estimates from age of diagnosis, we further partitioned our discovery set into PRS quintiles and stratified by AI-carrier status. AI carrier status exhibited significant later predicted ages of onset for the lowest ($p = 0.011$) and second-highest risk quintiles ($p = 0.019$) (**Figure C.12F**). Across quintiles we observed the median predicted onset age ranged from 42-56, within the current recommended melanoma screening range. We did not observe longer sojourn times in AI carriers in TCGA, suggesting that AI alleles do not change the time from onset to diagnosis (**Figure S.12G**). In the future, further customizing melanoma onset estimates with epigenetic marks has the potential to improve estimates of age at onset and inform optimal screening based on germline genetic risk factors.

3.4.4 Investigating mechanisms of MHC-I autoimmune allele protection

One possible explanation for a protective effect of selected HLA alleles would be a stronger affinity for neoantigens or tumor associated antigens that cannot easily be suppressed by melanomas. In psoriasis, melanocyte antigens such as ADAMTS-like protein 5 presented by HLA-C*06:02 (one of the seven AI alleles) can induce a targeted CD8+ T-cell response against melanocytes.¹⁶² Similarly, in vitiligo, CD8+ T cells target antigens from melanosomal proteins such as PMEL, MLANA/MART1, TYR, TYRP1, and DCT.^{230;250;251} Interestingly melanoma-specific CD8+ T cells appear to recognize peptides derived from these conserved melanocytic antigens more often than melanoma-specific antigens.^{230;252–254} Given this, a protective effect in cancer could indicate that AI alleles mediate more effective immune surveillance against conserved cancer antigens (i.e., self-antigens overexpressed in tumors) or even against somatic mutations that promote tumor development.

We reasoned that a protective effect manifesting as delayed age of diagnosis would require more effective immune surveillance against early driver mutations or conserved melanocyte-specific antigens expressed by melanomas. For neoantigens, we identified a set of 215 mutations that exhibited joint high DNA and RNA variant allelic fraction coupled with either recurrence or driver likelihood as predicted by the CHASM algorithm^{194;195} (Material and Methods: *Identifying Driver Mutations*). This included well-known driver mutations in BRAF, NRAS and CDKN2A (**Figure 3.3A-B**), with BRAF p.Val600Glu (V600E) being the most frequent mutation across the discovery set at 164 unique occurrences (**Figure 3.3A**). For conserved antigens, we included peptides derived from antigens associated with melanocytes^{230;250} and melanoma, including the melanoma antigen gene (MAGE) family^{255–261} and genes constitutively expressed in melanocytes (Material and Methods: *Identification and Differential Expression of Conserved Antigens*) resulting in 87 genes (**Table S4**). We evaluated whether MHC-I AI alleles could better expose driver neoantigens and conserved antigens for immune surveillance than 2,908 other common MHC-I alleles based on NetMHCPan-4.1 binding affinity scores for 8-11mer peptides, using 0.5 and 2 percentile ranks as strong and weak binding cutoffs respectively (Material and Methods: *Predicting Binding Affinities*).²⁶²

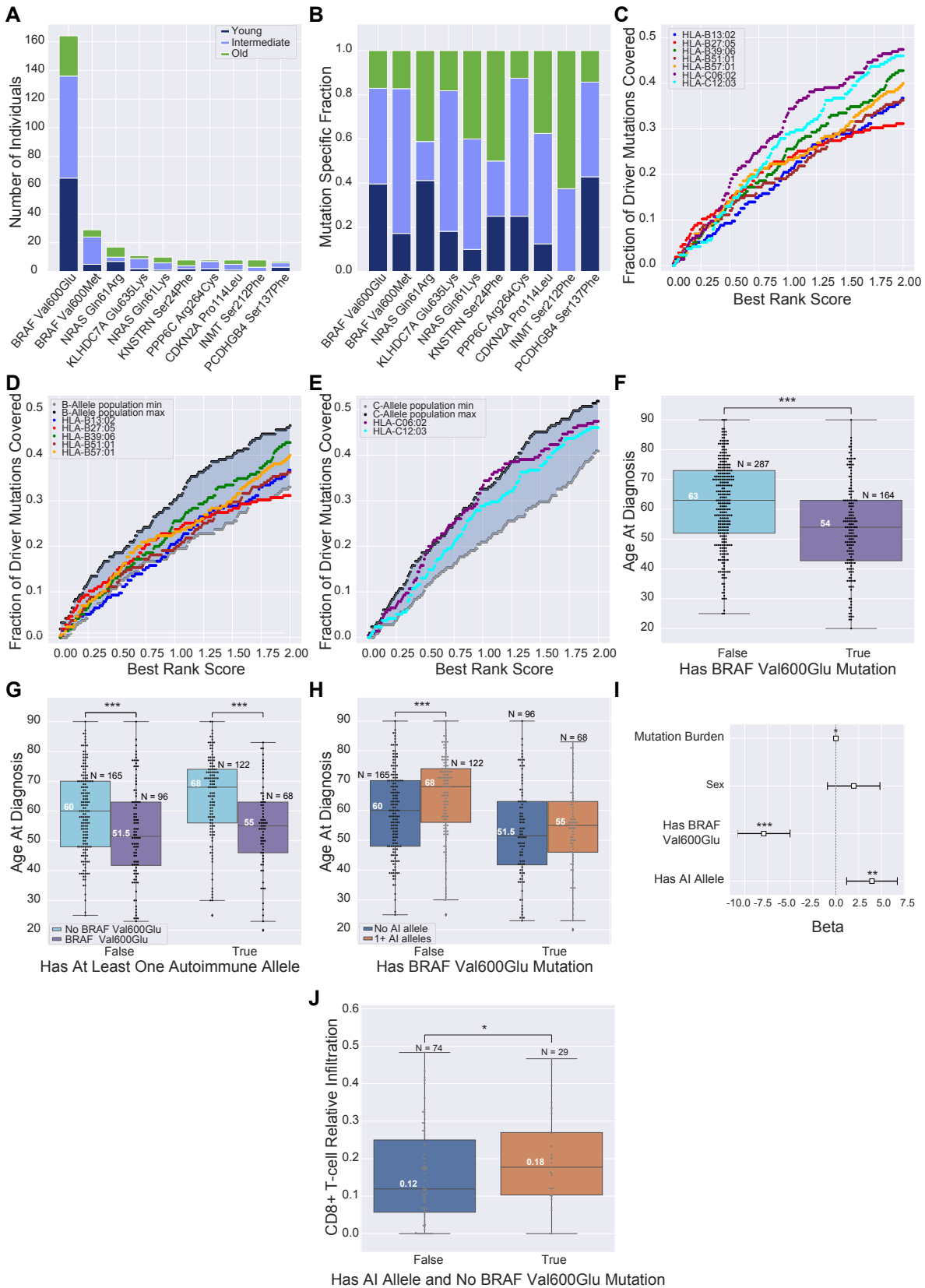
There was variability in the potential of MHC-I AI alleles to present peptides containing

driver neoantigens in general, with HLA-C*06:02 and HLA-C*12:03 presenting the largest proportion of mutations across all affinities, and HLA-B*27:05 presenting the largest proportion at stronger affinities (**Figure 3.3C**). As compared to common alleles (Material and Methods: *HLA Population Allele Representations*), MHC-I AI alleles did not show particularly better binding affinity for neoantigens for either HLA-B or HLA-C alleles (**Figure 3.3D-E**), though HLA-C*06:02 had coverage close to and briefly exceeding the common allele maximum (**Figure 3.3E**). In general, maximum and minimum population B-alleles exhibited greater coverage than their corresponding C-alleles at lower rank scores, but C-alleles outperformed B-alleles as the rank score exceeded 0.5 (**Figure C.13A**).

There were also differences in which mutations generated neopeptides with the best specificity. While common and AI alleles on average exhibited specific mutation binding preferences (**Figure C.14**) among the 215 drivers, there was no single mutation that was more effectively presented by all MHC-I AI alleles versus common alleles. One mutation, BRAF p.Val600Glu, showed significant age differences with mutation carriers being diagnosed with melanoma on average 9 years earlier than those without (**Figure 3.3F**, $p_{adj} = 7.58 \times 10^{-6}$). However, rather than observing a correlation between lack of AI allele presence and having a BRAF p.Val600Glu mutation, BRAF p.Val600Glu status significantly shifted age of diagnosis earlier regardless of AI allele status (**Figure 3.3G**). Moreover, it appeared to counter the AI allele protective effect with a reduced age gap between those with and without AI alleles in BRAF p.Val600Glu tumors (**Figure 3.3H**). Regression analysis showed similar results with p.Val600Glu mutation presence having a larger effect size than AI carrier status by almost 4 years (**Figure 3.3I**). Finally, we assessed whether either factor in isolation associated with increased CD8+ T-cell infiltration. We observed that while AI carrier status and BRAF p.Val600Glu mutation status shifted CD8+ T-cell infiltration levels insignificantly toward higher and lower means respectively (**Figure C.15**, $p_{autoimmune} = 0.090$, $p_{BRAF\ p.Val600Glu} = 0.147$), the conjunction of AI carrier status and the absence of a BRAF p.Val600Glu mutation significantly associated with increased CD8+ T-cell immune infiltration levels (**Figure 3.3J**, $p = 0.041$). Replacing TMB with neoantigen burden and BRAF p.Val600Glu status gave an estimated increase in CD8+ T-cell infiltration of 2.75% per MHC-I AI allele carried ($p = 0.039$).

Figure 3.3. Identification and coverage of melanoma driver mutations including BRAF p.Val600Glu and its impact on age at diagnosis and immune infiltration

A) Frequency of the ten most recurrent mutations across the TCGA by age group. Young (<50) and old (≥ 69) age groups correspond to the bottom and top 30% of individuals by age, respectively. The intermediate ($50 \leq x < 69$) age group corresponds to the remaining 40% of individuals. B) Relative age group distribution of the ten most recurrent mutations across the TCGA. C) Fraction of driver mutations presented by AI alleles as a function of best rank score. D) Fraction of driver mutations presented by HLA-B AI alleles relative to common ($\geq 1\%$ population frequency; 19 alleles) HLA-B alleles. Maximum and minimum population allele coverage corresponds to the maximum and minimum fraction of driver mutations capable of being presented across common alleles at each best rank score, respectively. E) Fraction of driver mutations presented by HLA-C AI alleles relative to common ($\geq 1\%$ population frequency; 13 alleles) HLA-C alleles. F) TCGA: individuals with a BRAF p.Val600Glu mutation show a significant 9-years-earlier age of melanoma diagnosis relative to those without this mutation ($p_{adj} = 7.58 \times 10^{-6}$). G) TCGA: BRAF p.Val600Glu significantly reduces melanoma age of diagnosis across individuals independent of AI allele presence with median earlier ages of diagnosis of 8.5 years in those without an AI allele ($p = 7.39 \times 10^{-4}$) and 13 years in those with an AI allele ($p = 8.12 \times 10^{-7}$). H) TCGA: individuals with an AI allele show a significant median later age of diagnosis of 8 years in the absence of a BRAF p.Val600Glu mutation ($p = 2.98 \times 10^{-4}$). However BRAF p.Val600Glu mutation presence appears to counter the AI-allele-protective effect with a loss of significance between those with and without AI alleles ($p = 0.232$). I) TCGA: BRAF p.Val600Glu mutation presence is significantly associated with 7.91 earlier years to melanoma diagnosis ($p_{BRAF\ p.Val600Glu} = 7.12 \times 10^{-8}$, CI = [5.075, 10.750]), while having at least one MHC-I-linked AI allele is significantly associated with 3.97 delayed years to melanoma diagnosis ($p_{autoimmune} = 4.59 \times 10^{-3}$, CI = [1.230, 6.700]) after controlling for sex and mutation burden. Mutation burden also remained significant but had a minimal contribution to age of diagnosis with 0.003 delayed years until melanoma diagnosis ($p_{mutation} = 0.014$, CI = [0.001, 0.005]). J) TCGA: having at least one MHC-I-linked AI allele coupled with the absence of a BRAF p.Val600Glu mutation significantly associated with a 6% increase in CD8+ T cell infiltration in primary melanoma tumors ($p = 0.041$).



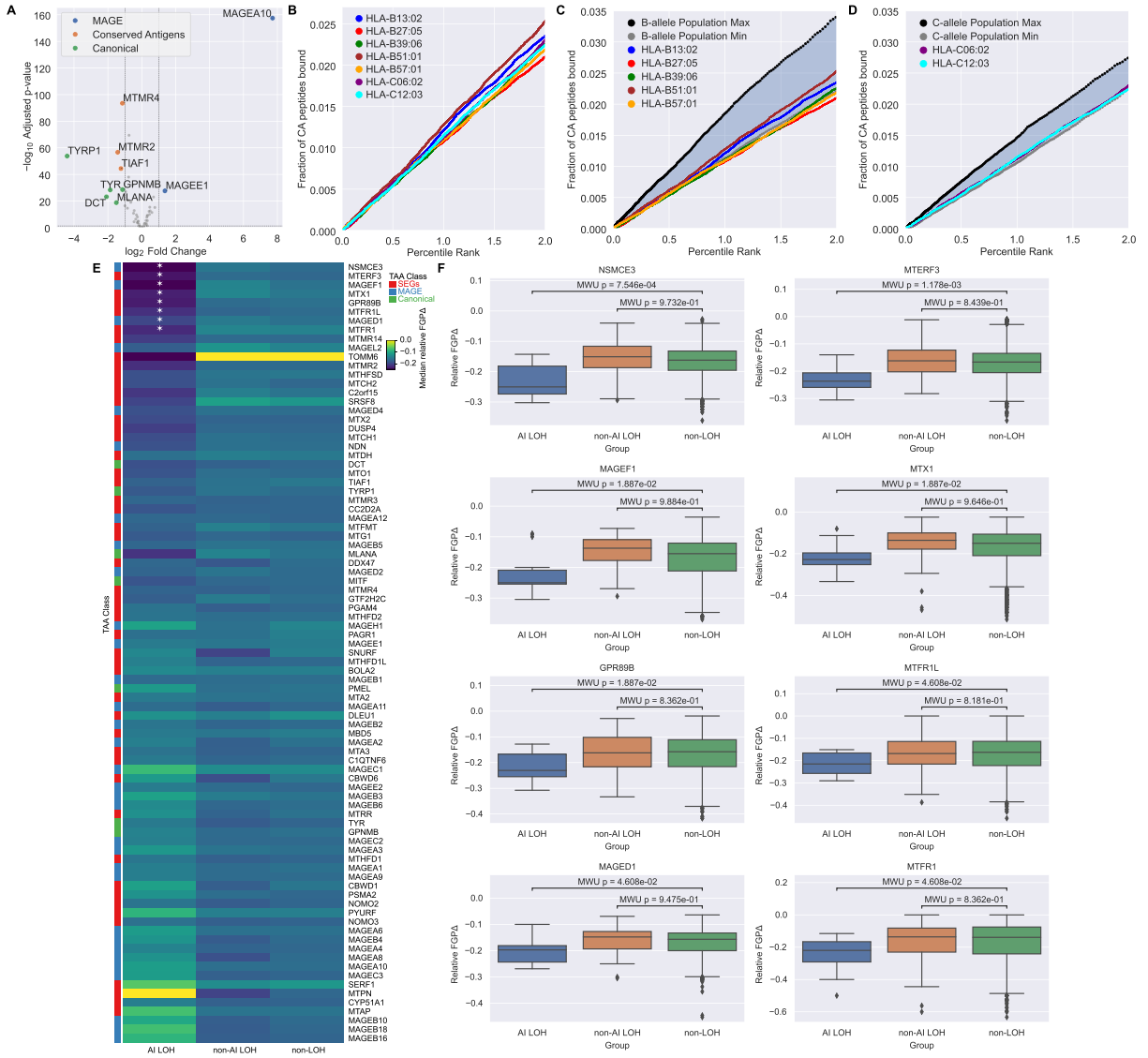
Across AI alleles, only three were predicted to present BRAF p.Val600Glu. HLA-B*27:05 was the only allele with a predicted affinity below the strong binding cutoff (best rank score = 0.22). HLA-B*27:05-restricted cytotoxic T-cell responses have been observed against p.Val600Glu.²⁶³ HLA-B*39:06 and HLA-B*57:01 had scores of 1.78 and 0.61 respectively, showing potential for weaker p.Val600Glu binding. We found no association with occurrence (OR = 1.04, $p = 0.843$) or expression of BRAF p.Val600Glu ($p = 0.336$) in individuals carrying AI alleles in general, or those carrying one or more of these three AI alleles specifically. An association might have been expected if the mutant allele was subject to strong counter selection by immune surveillance. Notably, BRAF p.Val600Glu has been suggested to avoid immune surveillance by accelerating internalization of cell surface MHC-I.²⁶⁴

We next compared the affinity of MHC-I AI alleles for conserved antigens versus common alleles. Here, we considered differences in NetMHCpan-4.1 affinities both regionally and on average for 87 proteins, including known melanoma cancer antigens^{230;250;255-261} and 52 genes that were both stably and specifically expressed in melanocytes and expressed in melanomas (**Table S4**). These 52 included four well-known melanocyte genes, PMEL, MLANA, TYRP1, and TYR. The rest were enriched for the folate metabolism pathway which is important for DNA repair in melanocytes.²⁶⁵ Evaluating expression in melanomas relative to melanocytes, we observed that two MAGE genes, MAGEA10 and MAGEE1, were significantly upregulated and that several canonical melanocyte and stably expressed genes were downregulated (**Figure 3.4A**). This is consistent with reports that MAGE genes are specific to reproductive tissues²⁶⁶ and tumors²⁵⁵⁻²⁶¹ and canonical melanocyte genes such as TYRP1 are minimally expressed, if not undetectable, in melanoma.^{253;267;268}

To evaluate broad allele-specific differences in conserved antigen-derived peptide repertoires, we compared the fraction of 8-11mers from each conserved antigen predicted to bind at a given percentile rank (Material and Methods: *Predicting Binding Affinities*). Across AI alleles we observed minimal variation with HLA-B*51:01 being the most promiscuous AI allele and HLA-B*13:02 being the second most promiscuous across the binding range (**Figure 3.4B**). However, in general, AI alleles exhibited a narrower conserved antigen repertoire relative to common alleles (**Figure 3.4C-D**). HLA-B generally presented a broader range of peptides across

Figure 3.4. Conserved antigen coverage characterization including HLA LOH effects

A) Differential expression of conserved antigens between normal melanocytes and melanoma. Labeled genes have an adjusted p value < 0.05 and an absolute fold change > 2 . B) Fraction of peptides from conserved antigens (CAs) presented by each AI allele. C) Fraction of peptides from conserved antigens (CAs) presented by HLA-B AI alleles relative to common HLA-B alleles ($\geq 1\%$ population frequency; 19 alleles). Maximum and minimum population allele representations correspond to the maximum and minimum fraction of conserved antigen peptides bound across common alleles at each rank score, respectively. D) Fraction of peptides from conserved antigens (CAs) presented by HLA-C AI alleles relative to common HLA-C alleles ($\geq 1\%$ population frequency; 13 alleles). E) Median relative decrease in the fraction of a gene presented (FGP) by respective removal of AI alleles that were affected by loss of heterozygosity (AI LOH), non-AI alleles with LOH, and other alleles that were not affected by LOH from the same individuals. The * symbols denote decreases in FGP that are significantly lower than the effect of losing other alleles on the same genotype not affected by LOH by Mann-Whitney U test with a false discovery rate < 0.05 . F) Distributions of the relative decrease in FGP for the eight genes with the greatest loss in presentation attributable to observed LOH of AI alleles.



conserved antigens than HLA-C (**Figure C.13B**). Considering individual conserved genes, AI alleles were generally predicted to bind a smaller fraction of derived peptides than common alleles. Comparison of the distributions of the top 10 differences by mean at a gene-specific level showed that almost all MHC-I AI alleles present conserved antigens no better than common alleles (**Figure C.16**), with the notable exception of HLA-B*27:05’s affinity for SRSF8 and TOMM6 (**Figure C.16A-B**).

Since MHC-I presented peptides are not uniformly sampled across the entire parent protein²⁶⁹ we revisited our analysis with a focus on regional differences by calculating position-wise best percentile ranks across all 8-11mers overlapping a position in each gene (**Figure C.2A**). We did not observe any region of a conserved antigen that was consistently more effectively presented by MHC-I AI alleles relative to common alleles, however there were short regions where specific MHC-I AI alleles had stronger affinity than common alleles (**Figure C.2B**). In addition to binding SRSF8 more generally, HLA-B*27:05 also exhibited the greatest position-wise advantage to SRSF8 over common alleles (**Figure C.2B**). This is in spite of having the smallest conserved antigen repertoire overall (**Figure 3.4B-C**). Several proteins had regions with affinity differences between HLA-B AI and common alleles, including the canonical melanocyte genes PMEL, GPNMB, and TYR and the MAGE genes MAGEA10, MAGEE1, MAGED4, and MAGEA8 (**Figure C.2B, Figure C.17**). However, regional affinity advantages of AI alleles over common alleles tended to be allele, gene, and position specific with little overlap across AI alleles. Overall, these analyses suggest that the protective effect of MHC-I AI alleles does not derive from a shared specificity for a conserved antigen or neoantigen, but cannot rule out that differences in specificity contribute to the effect.

Finally, we looked for trends suggesting selective pressure to evade immune responses on the basis of AI-allele presentation of conserved antigens. While we did not detect consistent downregulation of conserved antigens in tumors with AI alleles, loss of heterozygosity (LOH) of AI alleles accounted for the greatest loss of potential to present peptides derived from conserved antigens in multiple cases (Material and Methods: *Calculating HLA Loss of Heterozygosity*). In particular, for tumor antigens NSMCE3, MTERF3, MAGEF1, MTX1, GPR89B, MTFR1L, MAGED1, and MTFR1, loss of AI alleles resulted in a significant reduction in potential to

present peptides from these proteins relative to other HLA alleles carried by the same individuals (**Figure 3.4E-F**, $p_{adj} < 0.05$). This suggests that LOH affecting AI alleles could allow tumors to evade T-cell responses directed at certain conserved antigens.

3.5 Discussion

Immunosurveillance has been implicated in melanomagenesis prevention,^{120;121;134} yet HLA contributions to melanoma risk have largely remained uncharacterized. Here we investigated whether predisposing MHC-I alleles for CD8+ T-cell driven skin-associated autoimmune disorders (vitiligo and psoriasis) could protect against melanoma. Our findings support this hypothesis, with AI-allele carriers exhibiting a significant later age of melanoma diagnosis in the TCGA and a decreased risk of developing melanoma among individuals in the MVP. Moreover, AI-allele specific protection appears not only to be uncaptured by current melanoma PRS,⁹ but also can augment PRS performance in relative risk stratification. While at least 6 vitiligo risk SNPs are protective from melanoma,^{131;132} our results show that autoimmune risk effects can be extended to MHC-I as well, and suggest a broader space of joint autoimmune predisposition and melanoma protection.

We further investigated potential mechanisms linking MHC-I AI alleles to delayed age at diagnosis. Immune activity against melanoma has been linked to high mutation burden due to UV exposure,^{124;125} suggesting neoantigens could provide a potential substrate. Focusing on mutations that drive melanomagenesis, we did not see obvious differences in the affinity for MHC-I AI alleles for neopeptides relative to other alleles. While AI alleles did not seem to interact with specific mutations, we did note that BRAF p.Val600Glu was associated with an earlier age at diagnosis independent of AI allele carrier status. Potential of autoimmune alleles to present BRAF p.Val600Glu did not appear to impact the incidence of the mutation, which is consistent with reports that this mutation associates with impaired immune surveillance.²⁶⁴

As CD8+ T cells target healthy melanocytes through conserved antigens in both psoriasis and vitiligo, responses against conserved antigens could also provide an explanation. We observed positions better presented by AI alleles in genes spanning all subcategories of conserved antigens including canonical melanocyte genes, melanoma antigen genes, and genes stably expressed in

melanocytes. In particular, certain amino acids in PMEL and TYR (**Figure C.2B**), which are melanocyte-specific antigens recognized by CD8+ T cells in melanoma-associated vitiligo,^{230;250} were better presented by AI alleles. However, we did not observe position-wise advantages for AI alleles in other CD8+ T-cell targets, such as MLANA or DCT. MHC-I AI alleles also uniquely target positions in the melanoma antigen genes MAGEA10 and MAGEE1 which are upregulated in melanoma relative to melanocytes (**Figure 3.4A**, **Figure C.2B**). Finally, we observed positions uniquely targeted by AI alleles within melanocyte stably expressed genes. Notably, SRSF8 exhibited the position most favored by AI alleles and is overall substantially better targeted by the AI allele, HLA-B*27:05 (**Figure C.16**). This suggests that melanocyte-specific conserved antigens, other than those already identified in melanoma-associated vitiligo, may also contribute to the protective effect of AI alleles. Additionally, while canonical melanocyte genes have been known to be recognized by CD8+ T cells, they are typically downregulated in melanoma (**Figure 3.4A**) or otherwise inconsistently expressed.^{270;271} In contrast, the tumor-specific antigens considered in this study exhibit stable expression across melanocytes and melanoma and might serve as more consistent immune targets.

Altogether, these findings support that AI alleles' unique immunopeptidomes could contribute to their protective effect against melanoma. However, further investigation is needed. In particular, some studies have suggested that stability of the MHC-I-peptide complex distinguishes AI alleles from other MHC-I alleles,^{272–274} so it is possible that the mechanism is not fully dependent on antigen specificity. There is also evidence that different alleles result in qualitatively different T-cell responses. While the 7 AI alleles studied here associated with lower levels of regulatory T-cell infiltration (**Figure 3.1B**), HLA-A*02:01 did not (**Figure C.5F-G**), despite documented associations with vitiligo. This may relate to increased expression of TIM-3 by CD8+ T cells induced by HLA-A*02:01 restriction.²⁴⁶ TIM-3 interacts with galectin-9 expressed by regulatory T cells (and certain tumor cells) to induce CD8+ T-cell apoptosis.^{275–277} In contrast, CD8+ T cells restricted by HLA-B*27 or HLA-B*57, upon epitope recognition, do not upregulate TIM-3,²⁴⁶ but instead upregulate granzyme B²⁷⁸ killing the regulatory T cells they encounter. More generally, it has been observed that healthy individuals carrying HLA-B*27/B*57 alleles have lower levels of regulatory T cells.²⁴⁶ Overall, protective effects were not shared by broader

HLA supertype groupings supporting that allele-specific characteristics, whether related to unique antigen specificities, stability or some other characteristic, are likely to account for any protective effect.

In general, AI alleles presented a narrower peptide repertoire compared to common alleles (**Figure 3.4C-D**). Notably, HLA-B*27:05 and HLA-B*57:01, both of which covered a smaller fraction of potential conserved peptides than the minimum HLA-B population allele representation, are known for their fastidiousness, or narrow peptide-binding repertoire, as studied in the context of progression from HIV infection to AIDS.²⁷⁹ Košmrlj et al. also observed that fastidious alleles present peptides not found in common alleles' repertoires. We similarly observed that the most fastidious AI alleles, HLA-B*27:05 and HLA-B*57:01, exhibited the greatest position-wise advantages over common alleles. Additionally, fastidious class I alleles are expressed on the cell-surface at much higher levels than their promiscuous class I counterparts,²⁸⁰ and therefore likely offer more opportunities for both neoepitope and self-antigen presentation from their binding repertoires. Moreover, CD8+ T cells restricted by fastidious alleles (e.g., HLA-B*57:01) may also be more cross-reactive as they are subject to less negative selection in the thymus given the narrower set of self-peptides to which they are exposed.²⁷⁹ Interestingly, we note that outside of autoimmune associations four of our seven AI-alleles (HLA-B*27:05, HLA-B*51:01, HLA-C*06:02 and HLA-B*57:01) are among the strongest HIV-protective alleles.^{224;281;282}

Peptide-MHC (pMHC) affinity is driven largely by MHC-I sequence variation where specific polymorphisms shape the binding pockets in the peptide-binding groove. Unsurprisingly, given the skin-specific autoimmune associations for which these alleles were selected, certain AI-alleles have similar binding pocket characteristics. For example, HLA-C*06:02 and HLA-C*12:03 both share a strongly negative E-pocket.²⁸³ HLA-C*06:02 shares an electronegative B-pocket with HLA-B*27:05 as well.^{283;284} An allele's affinity-based presentable peptide repertoire though is an idealistic representation that fails to account for antigen processing pathway contributions. Ultimately the space of bound peptides presented on the cell-surface is far narrower. Endoplasmic reticulum aminopeptidases (ERAP) 1 and 2 play an essential role in MHC-I antigen processing. They function to clip peptides to the appropriate size for MHC-I binding,^{285;286} but can also destroy potential ligands through overtrimming.²⁸⁷⁻²⁹⁰ ERAP1 and ERAP2 are prominent risk

factors in MHC-I linked autoimmune conditions and both are associated with psoriasis risk.^{291–293} Epistatic effects between ERAP1 and AI-risk alleles have been observed, particularly with HLA-C*06:02 in psoriasis,²⁹¹ and suggests this risk interaction is tied to an increased likelihood of specific autoantigens making it to the cell surface. Taken together, this suggests skin-specific autoimmune predisposing ERAP genes may also confer melanoma protection and is an interesting area for further pursuit.

Work by Chowell et al.¹⁵² and Cummings et al.²⁹⁴ showed that in melanoma the B44 supertype associates with extended survival after treatment with immune-checkpoint inhibitors (ICPI). While none of our AI alleles fall within this supertype, we surprisingly observed that the B44 supertype trended towards an earlier age of diagnosis in our discovery set (**Figure C.7**). The B44 supertype has an electropositive B-pocket, with an affinity for negatively charged residues at P2 such as glutamic acid (Glu).²⁹⁵ Given this strong glutamic acid affinity, it may be that in the context of ICPI, the B44 supertype is capable of inducing an immune response through the binding and presentation of the highly recurrent BRAF p.Val600Glu mutation. In contrast, the MHC-I AI alleles associated with later age at diagnosis do not bind negatively charged residues at P2. In fact HLA-B*27:05 and HLA-C*06:02 have electronegative B-pockets with strong affinities for the positively charged arginine at position 2,^{283;284} and effectively serve as B44 antonyms. This suggests peptide repertoire differences between the B44 supertype and the AI allele set, and further suggests the potential for a dichotomy between MHC-I associated melanoma protection and ICPI response, which is associated with somatic mutation presentation and high tumor mutation burden.

In conclusion, our study supports that skin-specific autoimmune MHC alleles have a protective effect in melanoma. Additional assessments of these alleles' effect on age of onset in further independent cohorts, as they become available, would be of future interest. We also note the potential for MHC-I mediated autoimmunity to interact with cancer development more broadly. Selecting alleles associated with autoimmune disease affecting the tissue type under investigation may yield similar findings across cancer types. For example, some MHC-I alleles are associated with multiple autoimmune conditions, such as HLA-B*27:05 in both psoriasis and ankylosing spondylitis.^{223;224;226;227;296;297} Ankylosing spondylitis is a targeted

form of spinal arthritis primarily affecting the entheses and leads to bone erosion and broad vertebral fusion,²⁹⁸ and makes for a conceivable autoimmune counterpart to osteosarcomas. Our study did not address MHC-II alleles which are generally expressed more specifically by antigen presenting cells, but have nonetheless been implicated in both skin autoimmune disorders and immunosurveillance. Taken together, tissue-specific MHC AI carrier status may broaden the scope of the autoimmune-cancer risk interplay and remains an interesting area for further exploration.

3.6 Ancillary Details

3.6.1 Author Contributions

Original concept, J.V.T. and H.C.; project supervision, H.C.; project planning and design, J.V.T. and H.C.; data acquisition, processing, and analysis, J.V.T., D.L., M.S.P., and A.C.; melanoma sojourn time estimation, M.L., G.E.L., and K.C.; statistical advising, W.K.T.; UKBB data acquisition, R.M.S.; MVP data acquisition and processing, B.R.G., C.P., F.N.D., K.M., R.H., C.C.T., J.L., S.P., and P.S.T.; immunological interpretation advising, G.P.M. and M.Z.; preparation of the manuscript, J.V.T., D.L., and H.C.

3.6.2 Data and Code availability

All data for this project were obtained from public sources. Discovery Set: Data were obtained from The Cancer Genome Atlas (TCGA) Research Network (<http://cancergenome.nih.gov/>). Normal exome sequences and clinical data were downloaded from the GDC on June 23-26th, 2018 and April 25th, respectively, using the gdc-client v1.3.0. Somatic mutations were accessed from the NCI Genomic Data Commons (<https://portal.gdc.cancer.gov/>) on May 14th, 2017. Genotype calls for TCGA were accessed from GDC on April 26, 2019. Validation and Stably Expressed Gene Datasets: dbGaP data from datasets phs000452.v2.p1.c2, phs001550.v2.p1.c1, phs000933.v2.p1.c1 and phs001500.v1.p1.c1 were obtained using AsperaConnect v3.9.5.172984. SRA toolkit v2.9.2 was used to obtain WXS/WGS data from the Sequence Read Archive (SRA) (including from the following studies: SRP067938, SRP090294). UKBB data was retrieved under project ID 37671. Million Veteran Program: Genotype and phenotype information were

obtained through application. Melanostrum Dataset: Genotype data from Gu et al.⁹ were obtained by direct communication with the authors. Code for all analyses can be found at <https://github.com/cartercompbio/MelMHC/>

3.6.3 Conflicts of Interest

G.P.M. receives research support from Thermo Fisher, CareDx, and PIRCHE. M.Z. is a board member of Invectys Inc.

3.6.4 Acknowledgements

This work was supported by a grant from the Harry J. Lloyd Charitable Trust (20191857) to H.C., an Emerging Leader Award from The Mark Foundation for Cancer Research (18-022-ELA) to H.C., NIH (National Institutes of Health) grant RO1 CA220009 grant to M.Z. and H.C., NIH grant R01 MH122688-02 grant to W.K.T., an NIH National Library of Medicine training grant (T15LM011271) to A.C., and NCI (National Cancer Institute) grant (U24CA248138). Computational infrastructure was funded by NIH grant 2P41GM103504-11. The results shown here are in large part based upon data generated by the TCGA Research Network (<https://www.cancer.gov/tcga>), the UKBB (project #37671), and the following studies: dbGaP: phs000452.v2.p1.c2, phs000933.v2.p1.c1, phs001550.v2.p1.c1, phs001500.v1.p1.c1, phs000424.v7.p2, and GEO: GSE78220. This research also used data from the Million Veteran Program (MVP), Office of Research and Development, Veterans Health Administration. This research was supported by the Million Veteran Program MVP022 award #I01 CX001727 (PI: R.H.). MVP data access was provided under the Genisis Core Project. This publication does not represent the views of the Department of Veterans Affairs or the United States Government. The graphical abstract was created with BioRender.com.

The content of Chapter 3, in full, is a reformatted reprint of the material as it appears in “Autoimmune alleles at the major histocompatibility locus modify melanoma susceptibility” in *The American Journal of Human Genetics*, 2023 by James V. Talwar, David Laub, Meghana S. Pagadala, Andrea Castro, McKenna Lewis, Georg E. Luebeck, Bryan R. Gorman, Cuiping Pan, Frederick N. Dong, Kyriacos Markianos, Craig C. Teerlink, Julie Lynch, Richard Hauger, Saiju Pyarajan, Philip S. Tsao, Gerald P. Morris, Rany M. Salem, Wesley K. Thompson, Kit Curtius,

Maurizio Zanetti, and Hannah Carter. The dissertation author was the primary investigator and author of this paper.

Epilogue

We exist as the embodiment of an intricate genomic tapestry woven by the master and cryptic artisan that is nature. Genomic threads are inextricably intertwined, crafted into a composite of phenotypes shaped by complex and enigmatic mechanisms. This dissertation serves as a trifocal lens through which we can sharpen our focus on the entanglement of genomic patterns driving complex disease susceptibility. Each viewing zone affords a different resolution that uniquely aims to advance our understanding of the subtle interdependencies that link genetic variation to the diverse phenotypic manifestations of complex conditions. In the subsequent sections, I provide a chapter-by-chapter perspective, detailing conclusions, limitations, and future directions inspired by this work.

4.1 **GRIEVOUS: Conclusions, Limitations, and Future Directions**

Injection of artifactual noise due to discrepant variant indexing and allele assignments across datasets compromises both the integrity of large composite genomic datasets formulated from individual studies and the fidelity of any multi-dataset analysis. Identifying and faithfully resolving these discrepancies between or across datasets classically required a rigorously laborious, time-consuming, and error-prone investment on behalf of a computationally proficient investigator. The creation and availability of GRIEVOUS, my command-line tool for cross-dataset variant homogenization elaborated on in Chapter 1, not only markedly enhances the efficiency of cross-dataset variant consistency, but also democratizes access to high-quality data harmonization, empowering researchers with varying computational expertise to engage in complex genomic studies. Consequently, this allows for the assembly of large, high-quality datasets and enables faithful multi-dataset genomic analyses, both of which improve our capacity to identify genomic

contributors to complex disease.

GRIEVOUS represents a significant step forward in addressing the challenges of cross-dataset variant homogenization. However, like any tool, it is not without limitations. In its current build, GRIEVOUS only resolves artifactual noise introduced by discrepant variant indexing and allele assignments across genomic datasets for the most commonly used class of variants for genetic studies, biallelic SNPs. Variants of other categories (e.g., indels, multiallelic) fall outside of the scope of the tool and still require manual verification. While this is a limitation, it also affords an opportunity for future improvement, by broadening the scope of variants handled to ideally completely abstract this problem from researchers.

GRIEVOUS also highlights a broader challenge within the genomic research landscape: the lack of integration across tools and file formats. While powerful computational tools such as PLINK2 help facilitate genomic file manipulation and analyses, the pre-analysis stage still demands the use of multiple disparate tools, each with unique file requirements and formats (e.g., imputation, reference assembly consistency, GRIEVOUS harmonization). This fragmentation affords an opportunity for the development of a unified, fully integrated framework which if implemented could streamline data quality control, reduce processing time, and further enhance genomic analysis accessibility. Furthermore, GRIEVOUS itself, currently written in Python and employing chromosome-level Parquet files for its internal databases, could benefit from performance optimizations such as re-implementation in a more computationally efficient language (e.g., Rust) and improvements to its database structure. Such enhancements would position GRIEVOUS as an even more effective tool within this envisioned unified framework, empowering researchers to overcome existing barriers and further advancing the fidelity and efficiency of complex genomic analyses.

4.2 VADeR: Conclusions, Limitations, and Future Directions

The unique blend of sequential and spatial complexities exhibited by the genome closely parallels properties found in both language and vision, properties which are not fully captured by additive summary-statistics-based PRS approaches (and their Bayesian shrinkage counterparts). This highlights the need for a paradigm shift in PRS methodology. Specifically, the development

of innovative modeling frameworks capable of capturing local and global patterns underlying the genome’s sequential and spatial dependencies offers a promising path forward to advancing genotype-to-phenotype prediction. Transformers, a class of deep-learning architectures which have revolutionized the fields of NLP and CV, provide a compelling foundation for this effort.

To that end, in Chapter 2 I introduce the **V**ision **A**dapted **D**isease **E**lucidating **T**ransformer (VADeR), a novel architecture inspired by Vision Transformers³⁹ (ViTs) that blends techniques from NLP embeddings and ViT patch partitioning. This design enables the model to capture sequential and spatial dependencies within the context of genotype-to-phenotype prediction. Moreover, recognizing the paramount importance of interpretability in clinical applications, I also developed **D**irected **A**ttention **R**elevance from **T**ransformer **H**euristics (DARTH) scores, an attention-based interpretability metric. DARTH scores enable the attribution of model predictions to specific genomic regions, providing personalized insights into the genetic contributions driving predicted disease risk. By combining advanced modeling capabilities with enhanced interpretability, this DARTH-VADeR framework addresses key limitations of existing PRS methods, and paves the way for more accurate and clinically actionable genotype-to-phenotype predictions.

VADeR represents an inductive bias-free approach to genotype-to-phenotype modeling, relying solely on the definition of a patch window length, W . As discussed in Chapter 2, architectures that embed relational inductive biases through the incorporation of biological priors can offer advantages in both inherent explainability and performance, particularly when training data is limited. However, such designs come with tradeoffs. Specifically, inducing these biases can enhance performance if the imposed structure aligns well with the underlying data distribution. However, in the context of biological data, explicitly defining node connections based on existing knowledge risks constraining the model’s ability to uncover relationships involving features with uncertain or poorly understood roles. This reliance on predefined structure can also embed biases reflective of current biological understanding, potentially overlooking aspects of the biology that remain undiscovered. Nevertheless, it would be valuable to compare VADeR’s performance against architectures such as hierarchical Transformers and graph neural networks that incorporate established biological priors. Such comparisons could inform future integrations,

where approaches are merged to leverage the strengths of each. For example, DARTH scores inferred from VADeR could be used to refine or expand the structure of prior-based models.

In Chapter 2, the investigation focused on prostate cancer as a proof-of-concept evaluation for our DARTH-VADeR framework. However, as previously mentioned, the number of trainable model parameters ($n = 96M$) far exceeded the number of training samples ($n = 73,315$) and features ($n = 12,017$), which likely constrained VADeR’s performance potential. As larger genomic datasets become available, it would be valuable to assess the performance gains achievable with more extensive training, especially in relation to the upper bound imposed by disease heritability. Furthermore, expanding investigations beyond individuals of European ancestry to include diverse populations would provide a more comprehensive evaluation of our approach. Such studies could assess model robustness across ancestries and yield further novel insights into genetic heterogeneity.

Improvements in feature encoding and representation present a promising avenue for boosting VADeR’s performance. Currently, VADeR collapses feature representations to allele counts, a simplification that overlooks haplotype-specific contributions to disease and the potential for detecting genomic imprinting. Incorporating phased data and leveraging a Siamese architecture could address these limitations by capturing inheritance patterns more effectively. However, for such an approach to succeed, it is essential that phasing is consistent across all genomic variants on all chromosomes, ensuring that input representations accurately reflect true inheritance patterns. Additionally, broadening the feature encoding from SNPs and allele counts to contiguous genomic sequences (e.g., including flanking variants around a locus) may prove advantageous by enhancing the richness of the genomic context provided to the model. However, this also would introduce additional challenges in data loading and dimensionality, requiring careful consideration of computational feasibility.

Defining genomic position presents both a significant challenge and opportunity. Linearizing the genome by chromosome and location is overly simplistic and fails to capture the complex three-dimensional structure, where spatial proximity in three dimensions can bring seemingly distant positions into close functional contact. Moreover, chromosome numbers are effectively arbitrary in the context of genome conformation and provide little meaningful information

for positional encoding. To navigate these challenges, VADeR employed learnable positional embeddings for its patch representations, offering a flexible and data-driven approach. However, further advancements may be possible through methods that more faithfully reflect the true genomic position, in the context of the underlying three-dimensional organization. While deducing true genomic positions is inherently difficult, customized techniques such as rotary positional embeddings tailored to genomic topology could offer potential advantages. Additionally, VADeR’s learned positional embeddings themselves may hold insights into genomic positioning, though this would require a detailed investigation beyond the scope of our investigations. Regardless, exploring opportunities in global genomic positional encodings and their ability to impact model performance represents a promising avenue for future research.

We currently live in an exciting era, where advances in NLP, CV, and generative modeling are beginning to transform clinical applications. This presents an unprecedented opportunity to develop multimodal or cross-modal architectures that seamlessly integrate diverse sources of clinical information including clinical notes, imaging data (where available), and genomics. For instance, multimodal Transformers could be fine-tuned using parameter-efficient techniques in conjunction with VADeR to create a unified prediction framework. Such an approach would enable the comprehensive integration of all available data modalities, leverage state-of-the-art models, and likely yield substantial performance gains in the task of complex disease prediction.

Finally, it is worth mentioning that one promising technique not explored here is employing distillation during training (similar to Data-efficient image Transformers, or DeiT’s). In this work, VADeR outperformed all statistical genomics baselines with a strictly supervised training procedure. However, such a finding may not generalize across phenotypes, especially for those with limited training data. In such cases, statistical genomics methods such as LDpred2 may outperform VADeR and the other machine learning baselines investigated here. However, employing a (hard-label) distillation training procedure, where VADeR (or more generally one’s machine learning method of choice) is trained to match the prediction of the top-performing statistical genomics method, presents a promising avenue for improving performance in data-limited scenarios. This strategy merits serious consideration for future applications where training data is sparse.

In summary, DARTH-VADeR lays the groundwork for devising more accurate and interpretable genotype-to-phenotype predictions. Future research should focus on broadening the scope of investigations (i.e., across ancestries and conditions), expanding datasets, evaluating and improving feature representations, refining positional embeddings, exploring integrative multi-modal models, and leveraging distillation techniques. These efforts provide practical pathways to further unravel the genomic tapestry, deepen our understanding of the mechanisms driving genotype contributions to complex phenotypes, and enhance predictive performance.

4.3 MHC-I Autoimmune Utility from Liability (MAUL): Conclusions, Limitations, and Future Directions

Building upon the advancements in genomic data quality and polygenic risk modeling discussed in Chapters 1 and 2, Chapter 3 shifts the focus towards highly polymorphic genomic regions inadequately captured by traditional PRSs. The Human Leukocyte Antigen (HLA) locus on chromosome 6 is one such notable polymorphic region. Encoding the highly polymorphic HLA-A, HLA-B, and HLA-C genes, these genes determine the peptide-binding specificity of the class I major histocompatibility complex (MHC-I), a protein integral to immune surveillance. While the extreme polymorphicity of the HLA locus encapsulates the diversity of immune responses, it also introduces challenges in identifying its contributions to disease risk. Variations in this region can predispose individuals to complex diseases in distinct ways, often at opposite ends of the immune dysfunction spectrum, such as autoimmunity and cancer. As MHC-I presents intracellular peptide antigens to CD8+ T cells it has been conceptualized as a fulcrum of immune dysfunction, tipping the balance toward heightened immune activity in autoimmune conditions or impaired immune surveillance in cancer. Chapter 3 leverages this idea to identify and investigate specific autoimmune-predisposing HLA class I alleles that confer a protective effect against melanoma, providing evidence that these alleles offer orthogonal, risk-relevant information beyond what is captured by existing melanoma PRSs.

These findings open two key corridors for future exploration on the autoimmune-cancer frontier. The first pertains to the broader implications of autoimmune-predisposing class I HLA alleles to modify cancer susceptibility across other cancer types. Here, we focused on

the relationship between cutaneous melanoma (a highly immunogenic cancer) and HLA alleles associated with skin-specific autoimmune conditions. Expanding this investigation to other cancer and tissue-specific autoimmune HLA allele pairs is a logical next step. One that has the potential to elucidate commonalities in methods of cancer protection, reveal conserved patterns across protective alleles, and expand our understanding of how HLA diversity shapes immune responses in the context of both cancer and autoimmunity.

The second corridor involves extending the scope beyond class I HLA alleles to other components of the antigen presentation pathway. Class II HLA alleles, which play a key role in presenting extracellular-derived peptides (as well as peptides derived from self-protein degradation in the endosomal pathway) to CD4+ T cells,¹⁴⁶ may similarly exhibit protective effects in melanoma or other cancers. Additionally, non-HLA genes involved in antigen processing, such as ERAP1 and ERAP2, are known autoimmune risk factors and may also modulate cancer susceptibility. Investigating these genes, alongside autoimmune-associated pathways outside the antigen presentation system, could further elucidate the broader interplay between autoimmunity and cancer. In conclusion, expanding upon the autoimmune-cancer interplay foundations highlighted in this Chapter represents a promising direction of future research with the potential to deepen our understanding of the mechanisms driving immune dysfunction and its dual role in complex disease risk predisposition and protection.

Appendix A

Supplemental Material for Chapter 1

A.1 Supplementary Figures

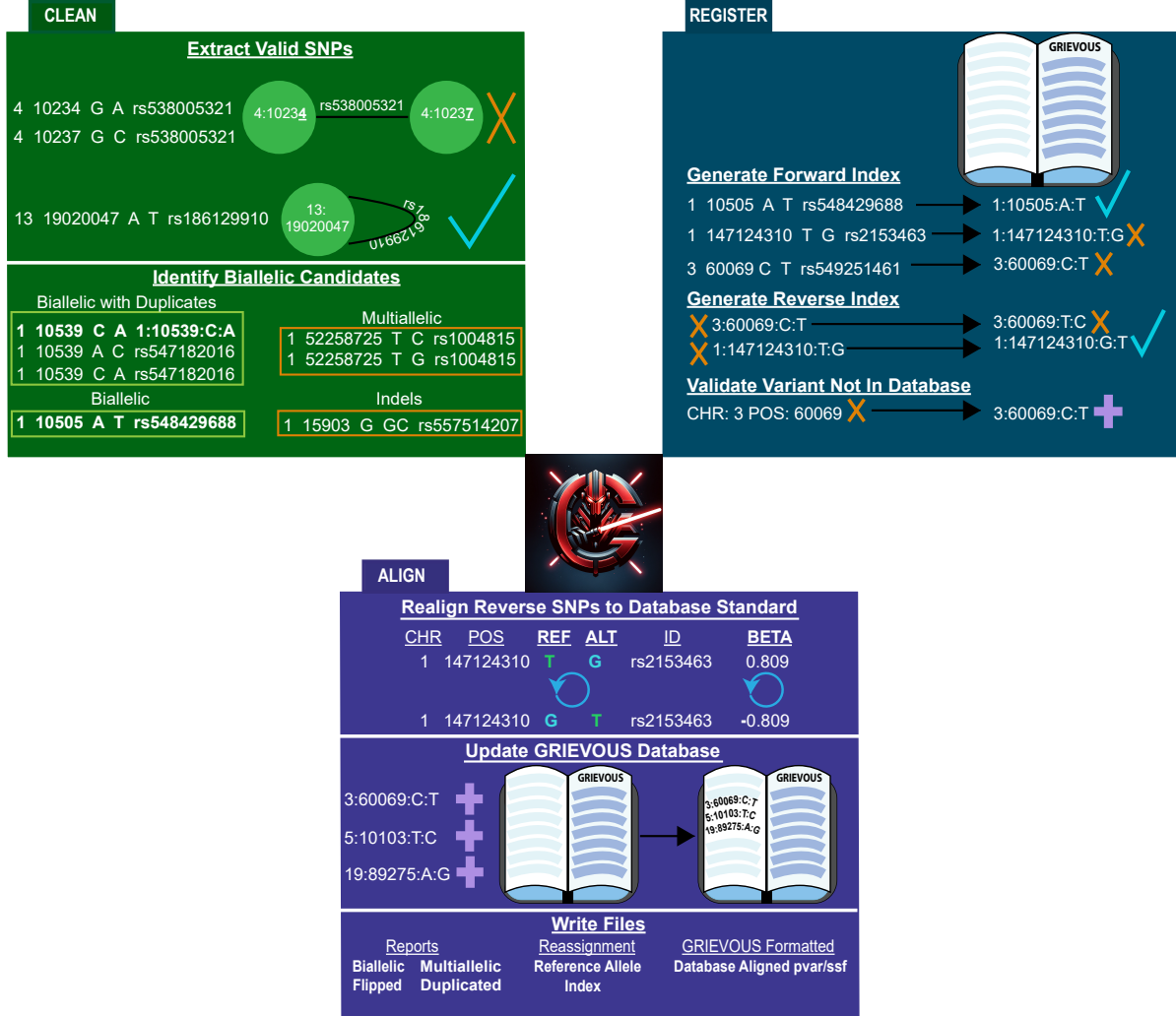


Figure A.1. Dataset realignment with GRIEVOUS can be conceptualized as a three-step workflow. The first step, *clean*, identifies and extracts valid biallelic variant candidates. The second, *register*, generates GRIEVOUS variant indices and performs variant database comparisons, identifying all biallelic variants requiring realignment and all biallelic SNP database additions. Finally, *align*, realigns REF and ALT alleles (and corresponding effect sizes for summary statistics) for those SNPs *register* identified as database reversed. Subsequently, *align* updates the GRIEVOUS database with all biallelic SNPs identified in the current dataset and writes dataset-level reports, reassignment files, and GRIEVOUS formatted files.

A.2 Supplementary Tables

Table A.1. Polygenic risk score (PRS) area under the receiver operating characteristic curve (AUC) performances for prostate cancer (PC) genotype datasets before and after harmonization with GRIEVOUS. Current PC PRS AUCs fall in the range of 0.59-0.64, on par with PRS performances after harmonization with GRIEVOUS.

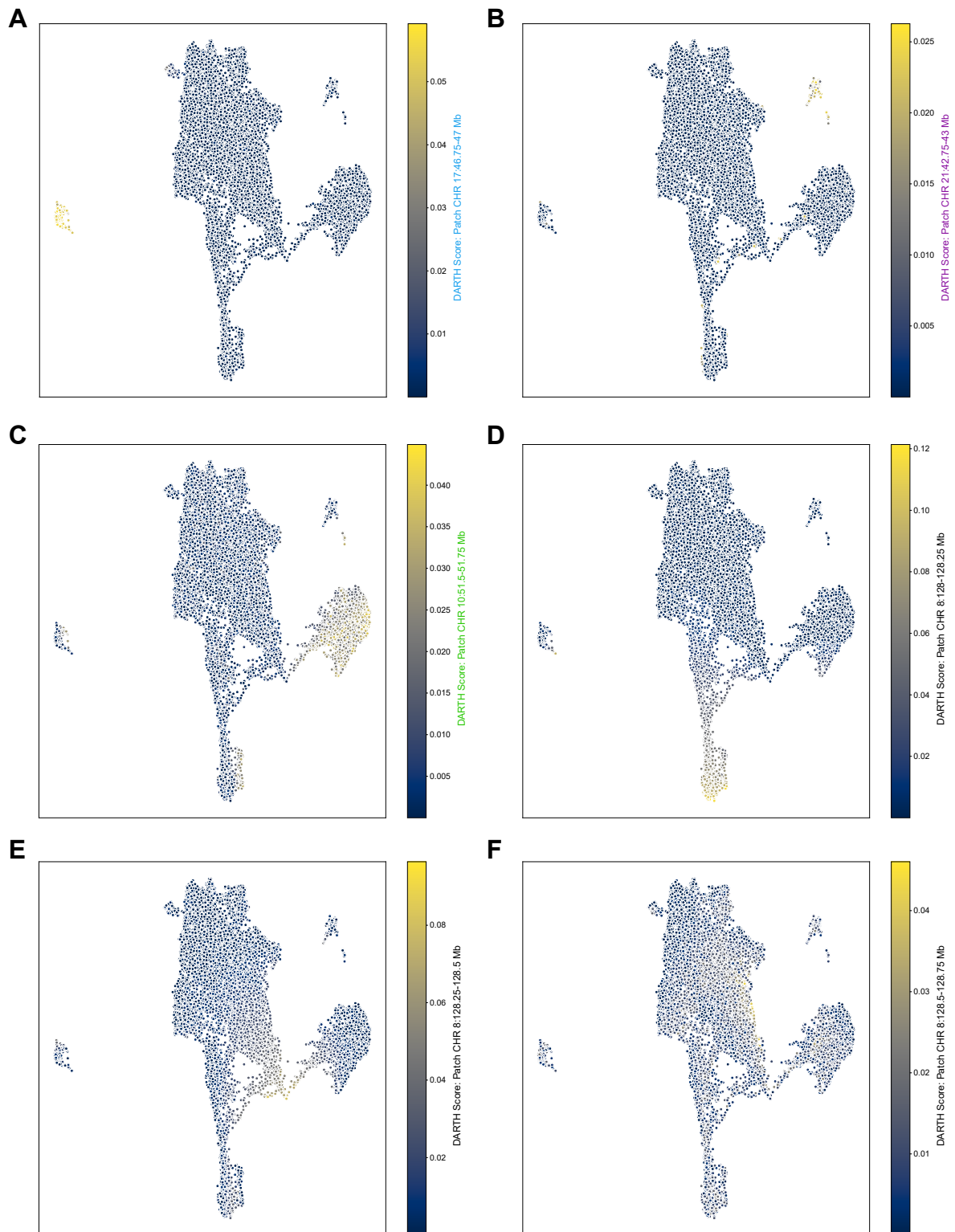
P-value Threshold	Number of PC PRS SNPs	Before: ELLIPSE PRS AUC	After: ELLIPSE PRS AUC	Before: UKBB PRS AUC	After: UKBB PRS AUC
5×10^{-8}	21,915	0.43673	0.63296	0.45324	0.62266
5×10^{-7}	28,420	0.43212	0.63744	0.44743	0.62364
5×10^{-6}	39,290	0.43230	0.63947	0.44056	0.62427
5×10^{-5}	58,326	0.43055	0.64101	0.43210	0.62585
5×10^{-4}	110,287	0.42048	0.64312	0.42799	0.62436

Appendix B

Supplemental Material for Chapter 2

B.1 Supplementary Figures

Figure B.1. Projecting personalized patch-specific DARTH scores across the UMAP dimensions for A) the top cluster 1 HOXB13 patch spanning 46.75-47 Mb (GRCh37) on chromosome 17 B) the top cluster 2 TMPRSS2 patch spanning 42.75-43 Mb on chromosome 21 C) the top cluster 3 MSMB patch spanning 51.5-51.75 Mb on chromosome 10 and D-F) the top 3 patches with the highest global average DARTH score which span consecutively from 128-128.75 Mb on chromosome 8.



Appendix C

Supplemental Material for Chapter 3

All Supplemental Tables can be found in the published version of this material at the following location: [https://www.cell.com/ajhg/fulltext/S0002-9297\(23\)00170-2](https://www.cell.com/ajhg/fulltext/S0002-9297(23)00170-2)

C.1 Supplementary Figures

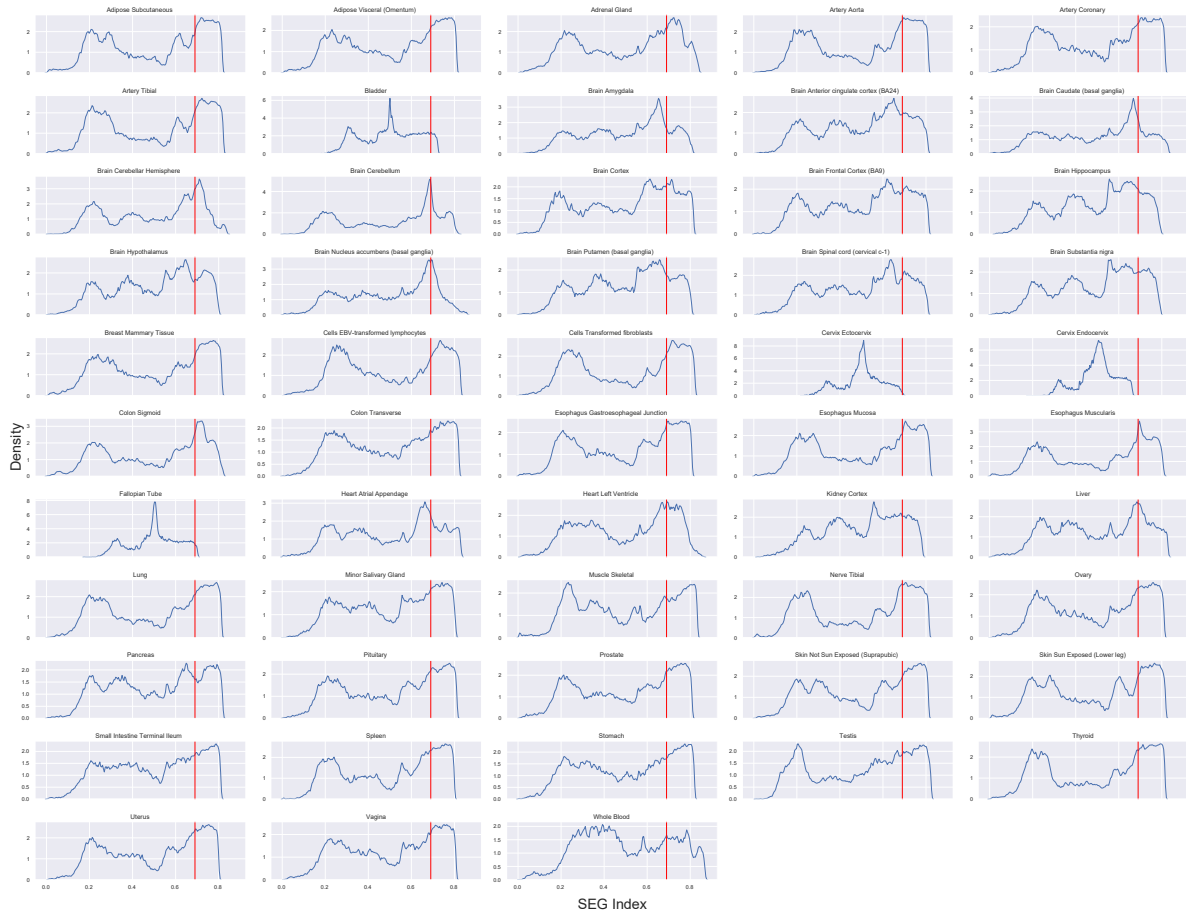


Figure C.1. Distributions of stably expressed gene (SEG) scores shown for all 53 tissues in GTEx. The threshold of 0.69 was selected and is shown in red.

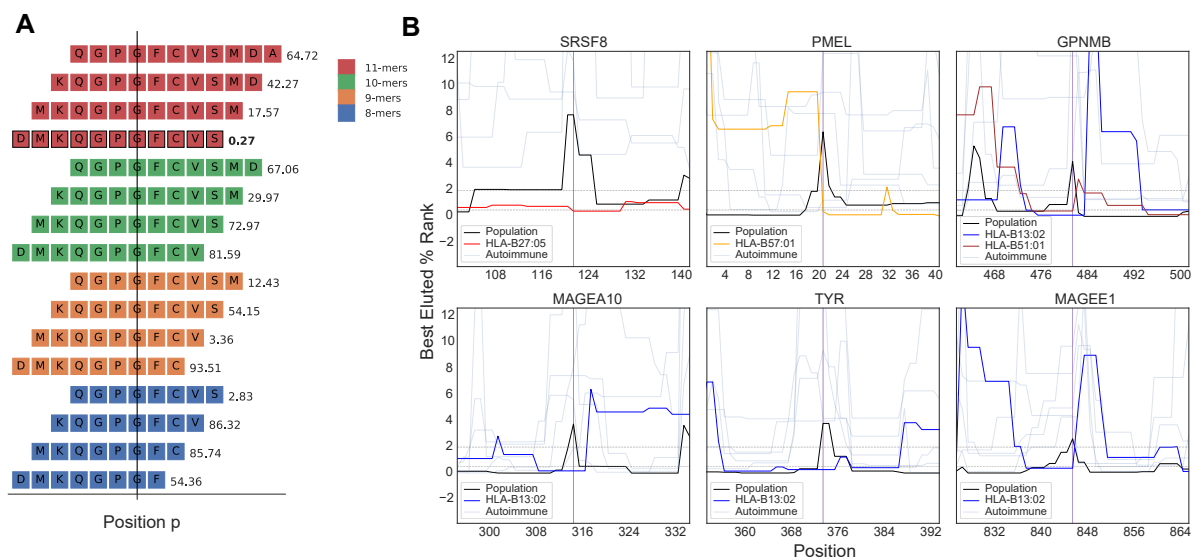


Figure C.2. A) Position-wise best percentile rank scores were computed from predictions for 8-11mers by taking the best percentile rank of all overlapping peptides at any given position. B) Greatest position-wise differences in best percentile rank between HLA-B autoimmune and common alleles amongst the top 2 differences (SRSF8 and the canonical melanocyte gene PMEL) and several DE conserved antigens (GPNMB, MAGEA10, TYR, MAGEE1). Common alleles are shown in black and autoimmune alleles are shown in transparent blue unless they are predicted to elute at better percentile ranks than common alleles. Vertical purple lines demarcate positions where autoimmune alleles are predicted to elute at better percentile ranks than common alleles. Plots show up to ± 20 amino acids of the demarcated positions.

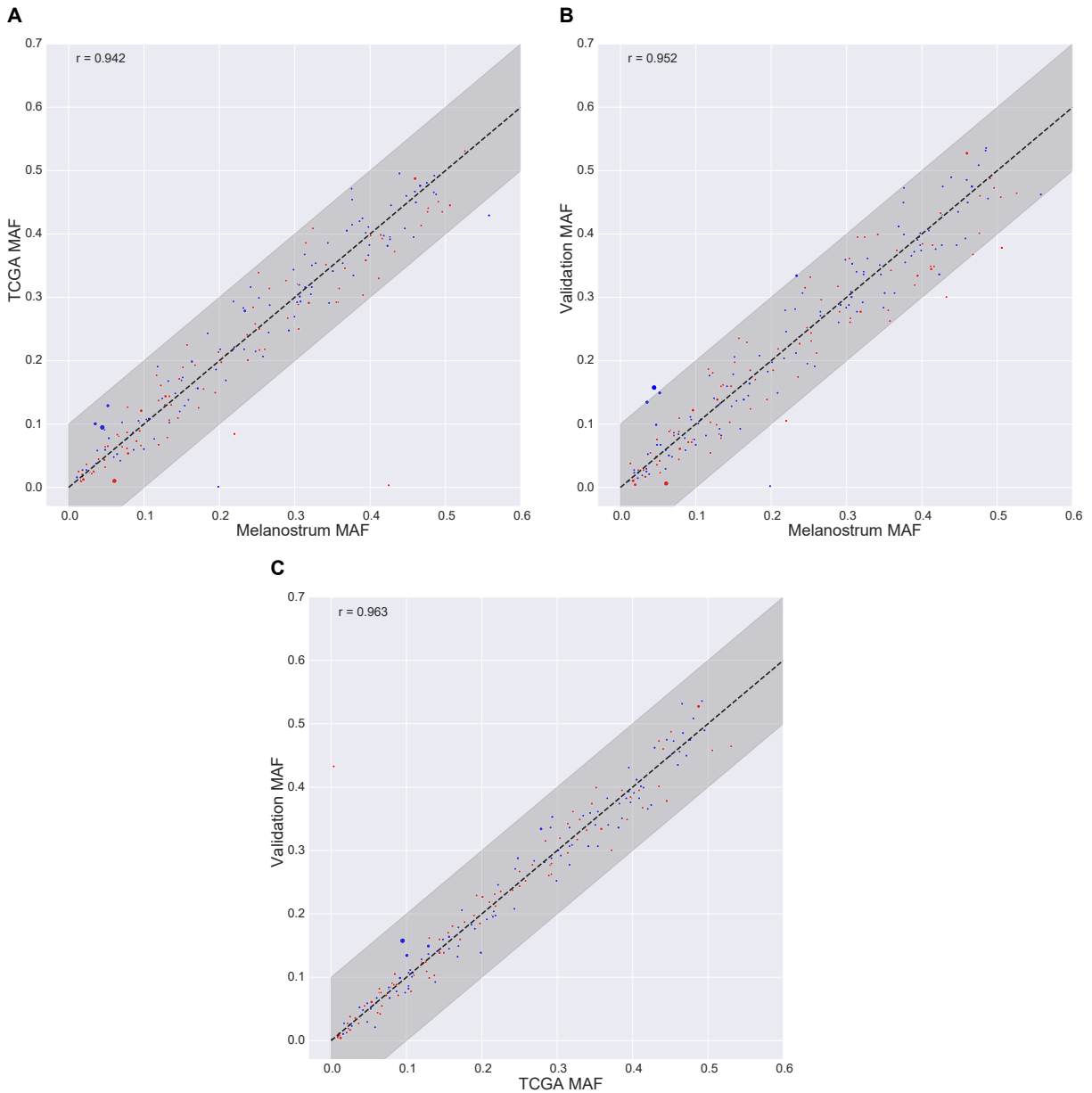


Figure C.3. PRS SNP minor allele frequencies (MAFs) relationship between A) Discovery and Melanostrum (Pearson $R = 0.942$), B) Validation and Melanostrum (Pearson $R = 0.952$), and C) Validation and Discovery (Pearson $R = 0.963$). Red points are PRS SNPs with a protective effect in melanoma (negative PRS weight), while blue points are PRS SNPs with a predisposing melanoma effect (positive PRS weight). Points are sized according to the magnitude of their weight in the PRS. SNP effects ranged in PRS weight from -0.4 to 0.49.

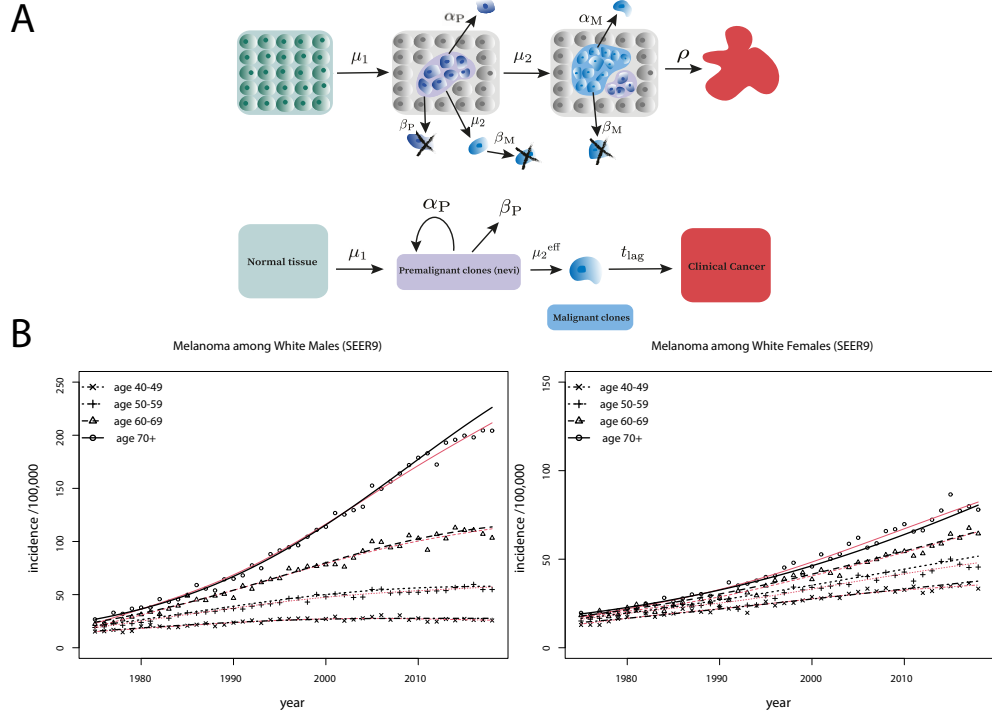


Figure C.4. A) *Upper row*: Multistage clonal expansion (MSCE) model - illustration of a stochastic realization of the multi-type branching process. In the two-stage MSCE model, the first ‘hit’ occurs when normal cells (green) undergo asymmetric division (due to mutation, for example) and create one premalignant daughter cell with rate μ_1 and one normal cell. Premalignant cells can then clonally expand with rates α_P for division and β_P for cell death/differentiation (purple clones, e.g., dysplastic nevi). The second ‘hit’ occurs at rate μ_2 per cell per year wherein a premalignant cell creates a malignant daughter cell (due to secondary mutation, for example) and 1 premalignant cell. α_M and β_M represent cell division and death/differentiation, respectively, for malignant cells (blue). Clinical detection occurs for malignant clones with rate ρ . *Bottom row*: Mathematical approximation of the two-stage model includes an effective mutation rate μ_{eff}^2 for transformation of a malignant cell that survives and a tumor sojourn time/ ‘lag-time’ representing the time between the first persistent malignant cell and the clinically detected cancer. B) Hazard curves from the MSCE models for melanoma incidence/100,000 (black lines) correspond to US SEER data from 1975-2018 (black shapes) in white females (right) and white males (left) separately by age groups, as represented by the different symbols, over calendar years. Calibrated parameters from the twostage model yield similar average lag-time (tumor sojourn time) for males and females of ~ 10 years. For comparison, non-parametric fits using 4th-order smoothing splines shown in red.

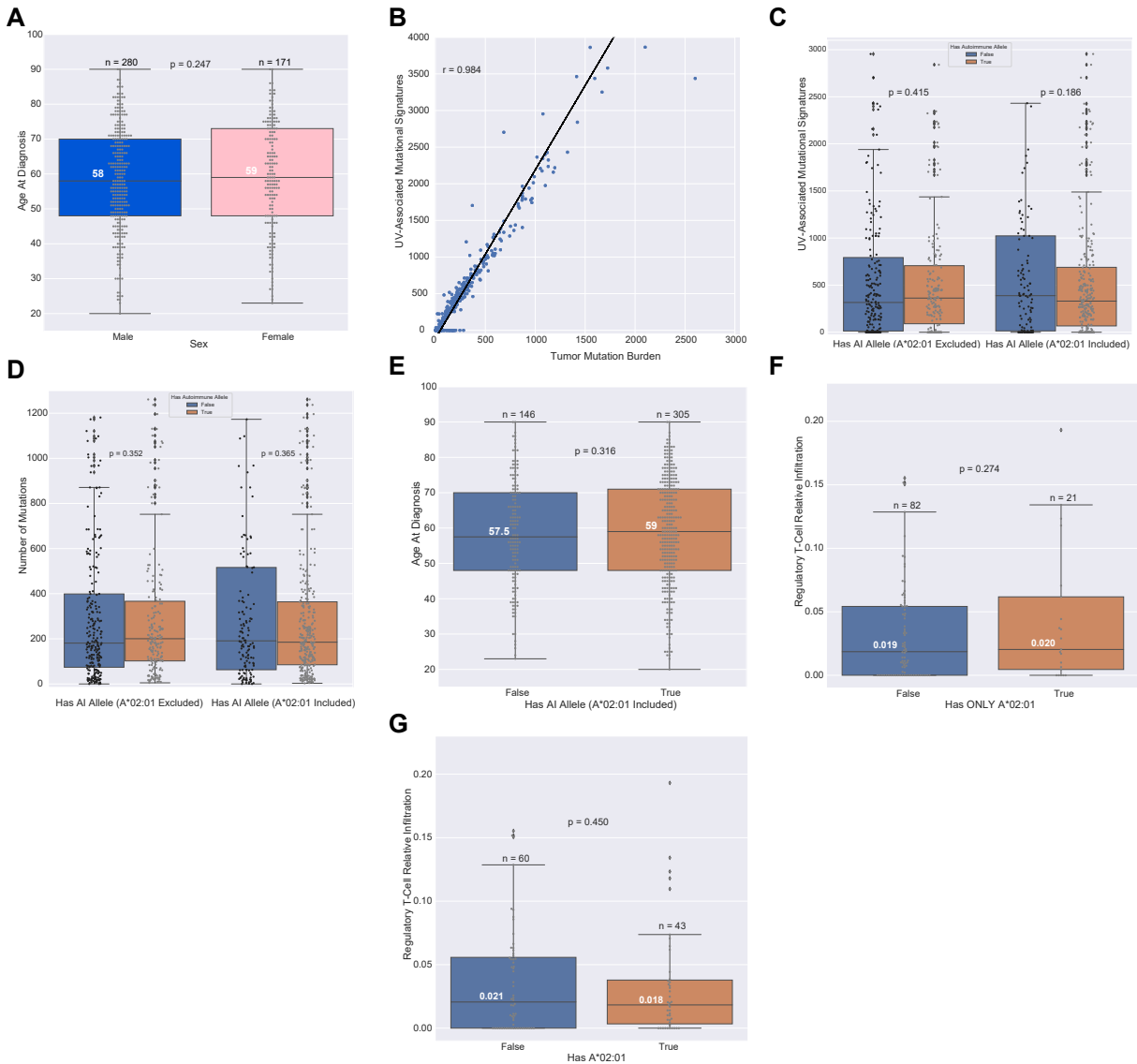
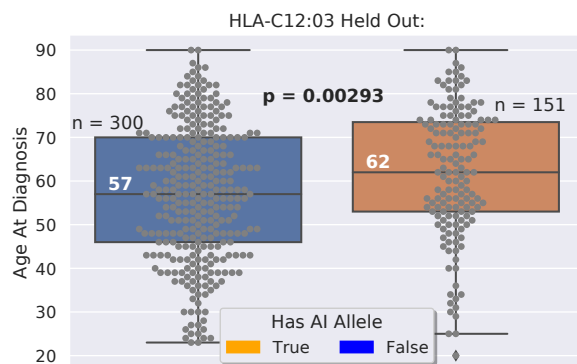
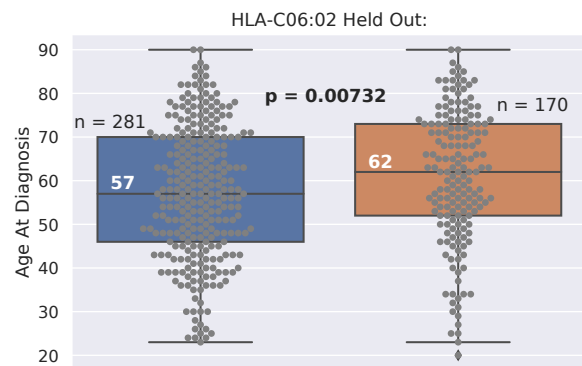
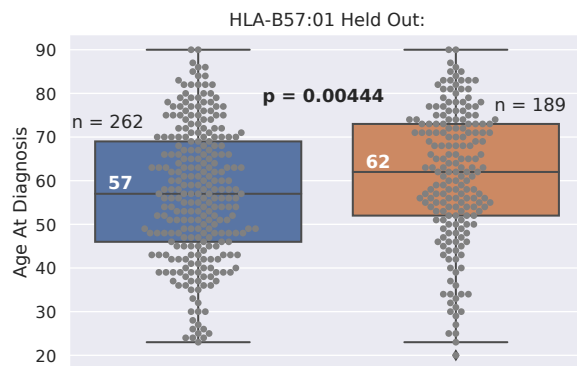
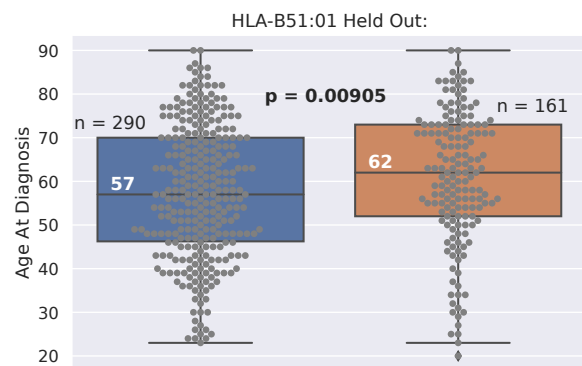
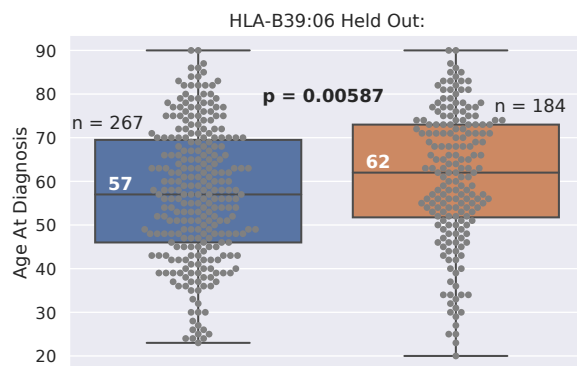
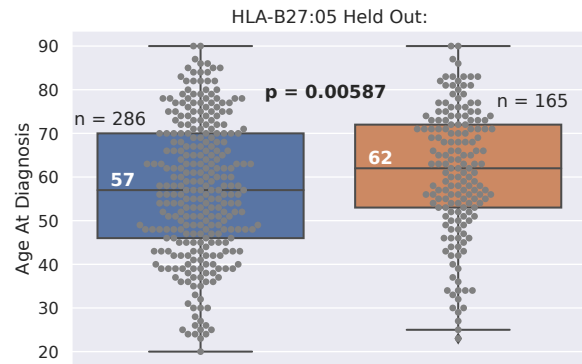
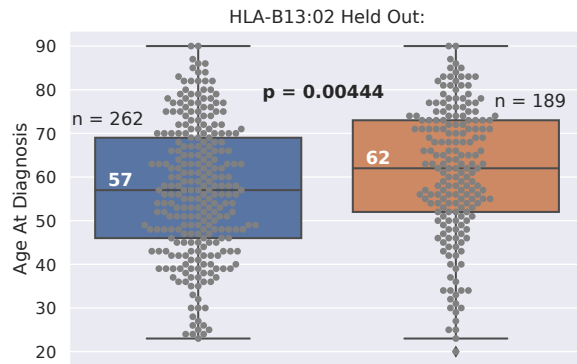


Figure C.5. Covariates and associations with melanoma in the TCGA. A) Boxplots comparing sex and melanoma age of diagnosis. B) Correlation between total mutation burden and UV-signature associated mutations. C) Boxplots comparing the number of UV-signature associated mutations across those with and without a MHC-I linked autoimmune allele. D) Boxplots comparing the number of mutations across those with and without a MHC-I linked autoimmune allele. E) Effect of MHC-I autoimmune alleles on age at diagnosis in melanoma including HLA-A*02:01. F) Carriers of HLA-A*02:01 and no other autoimmune alleles do not exhibit a significant difference in regulatory T-cells in primary melanoma tumors. G) Carriers of HLA-A*02:01, independent of other alleles carried, do not exhibit a significant difference in regulatory T-cells in primary melanoma tumors.

Figure C.6. Effect of omitting each of the 7 individual autoimmune alleles from carrier status in TCGA. For each autoimmune allele, individuals carrying only the excluded allele were assigned to non-AI allele carrier status. All p-values are multiple hypotheses corrected by means of the Benjamini-Hochberg method.



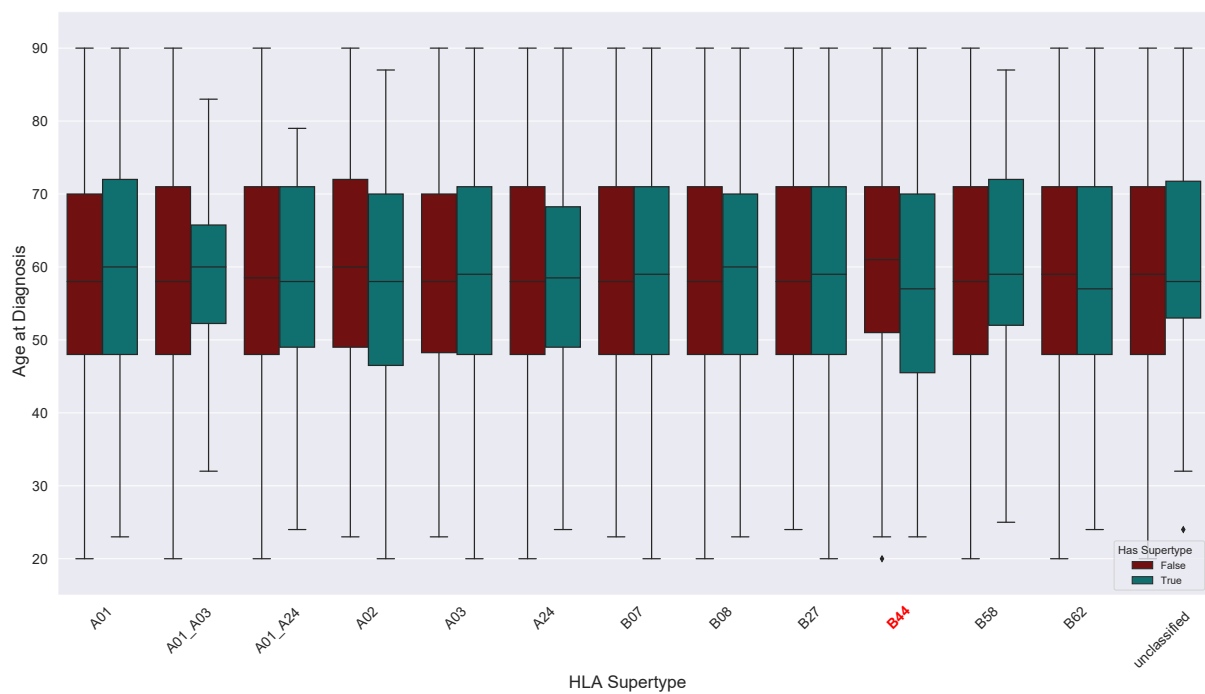


Figure C.7. Effect of HLA supertype on age at diagnosis in TCGA. No significant age differences were observed across any supertype, though those with the B44 supertype (red) trended towards an earlier age of diagnosis ($p = 0.199$, median earlier age of diagnosis difference = 4 years).

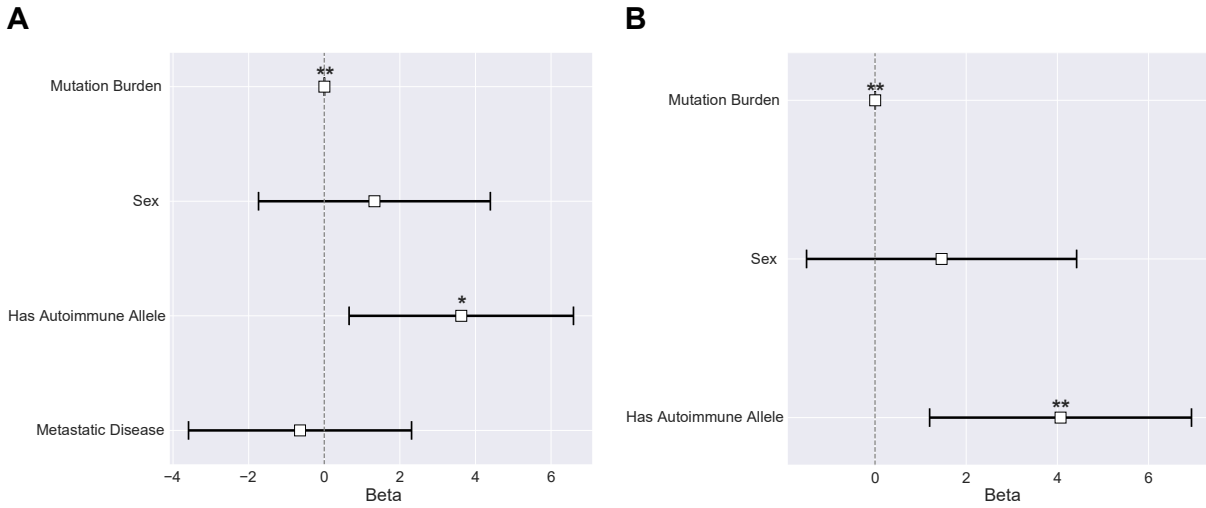


Figure C.8. Regression modeling of autoimmune allele effects on melanoma age of diagnosis in TCGA. A) Having at least one MHC-I linked autoimmune allele is significantly associated with 3.62 delayed years to melanoma diagnosis after controlling for primary vs. metastatic disease, sex, and mutation burden ($n = 416$, $p_{\text{autoimmune}} = 0.015$). Those with an AJCC pathologic tumor stage of II or below were assigned a primary disease label, while those with a tumor stage of III or higher were assigned a metastatic disease label. B) Having at least one MHC-I linked autoimmune allele is significantly associated with 4.07 delayed years to melanoma diagnosis after controlling for sex and mutation burden ($n = 451$, $p_{\text{autoimmune}} = 0.005$).

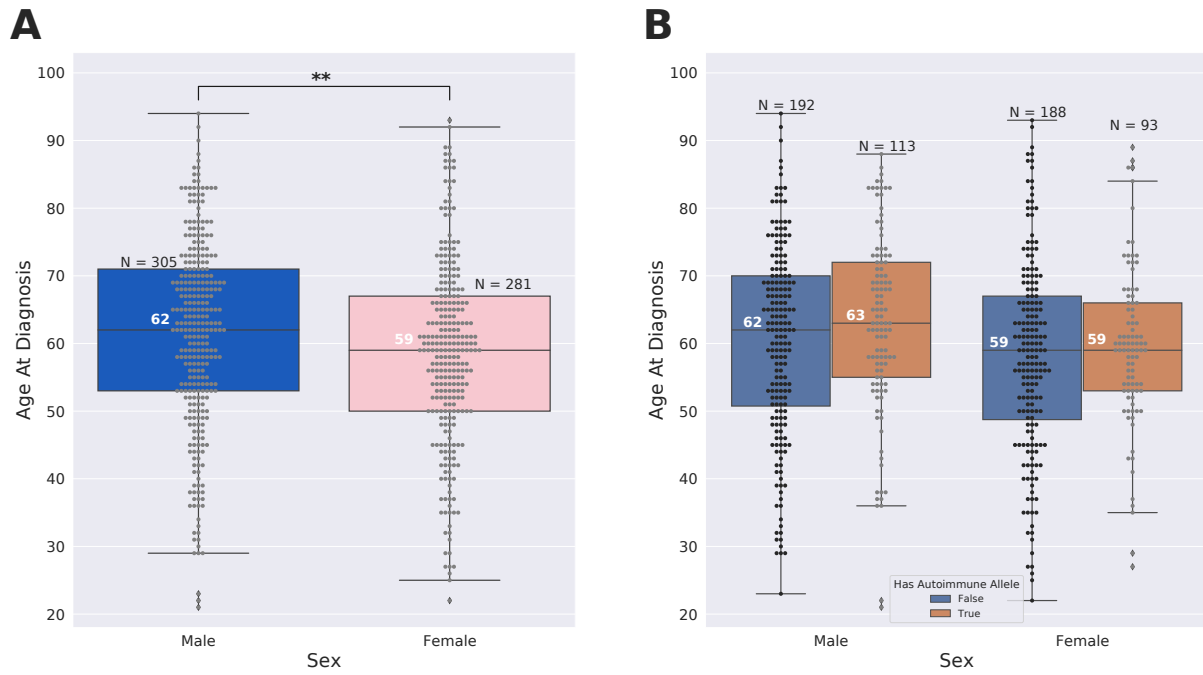


Figure C.9. A) Females in the validation set showed a significant earlier age of diagnosis relative to males ($p = 0.003$, median difference = 3 years) B) Direction of AI carrier status effect in males is consistent with later age of diagnosis findings ($p_{males} = 0.055$, $p_{females} = 0.236$).

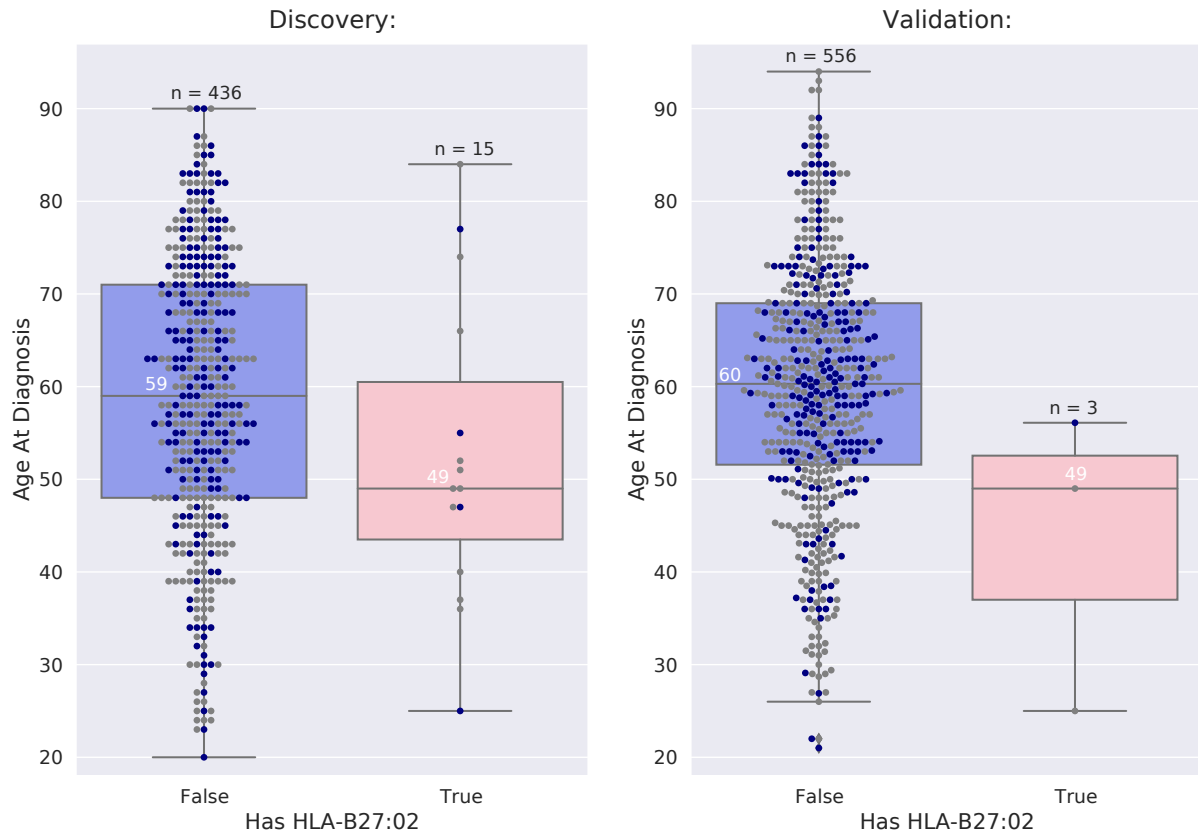


Figure C.10. HLA-B*27:02 was associated with an earlier age of diagnosis across both discovery and validation sets. Blue points correspond to MHC-I autoimmune allele carriers, while gray points correspond to those without any MHC-I AI alleles.

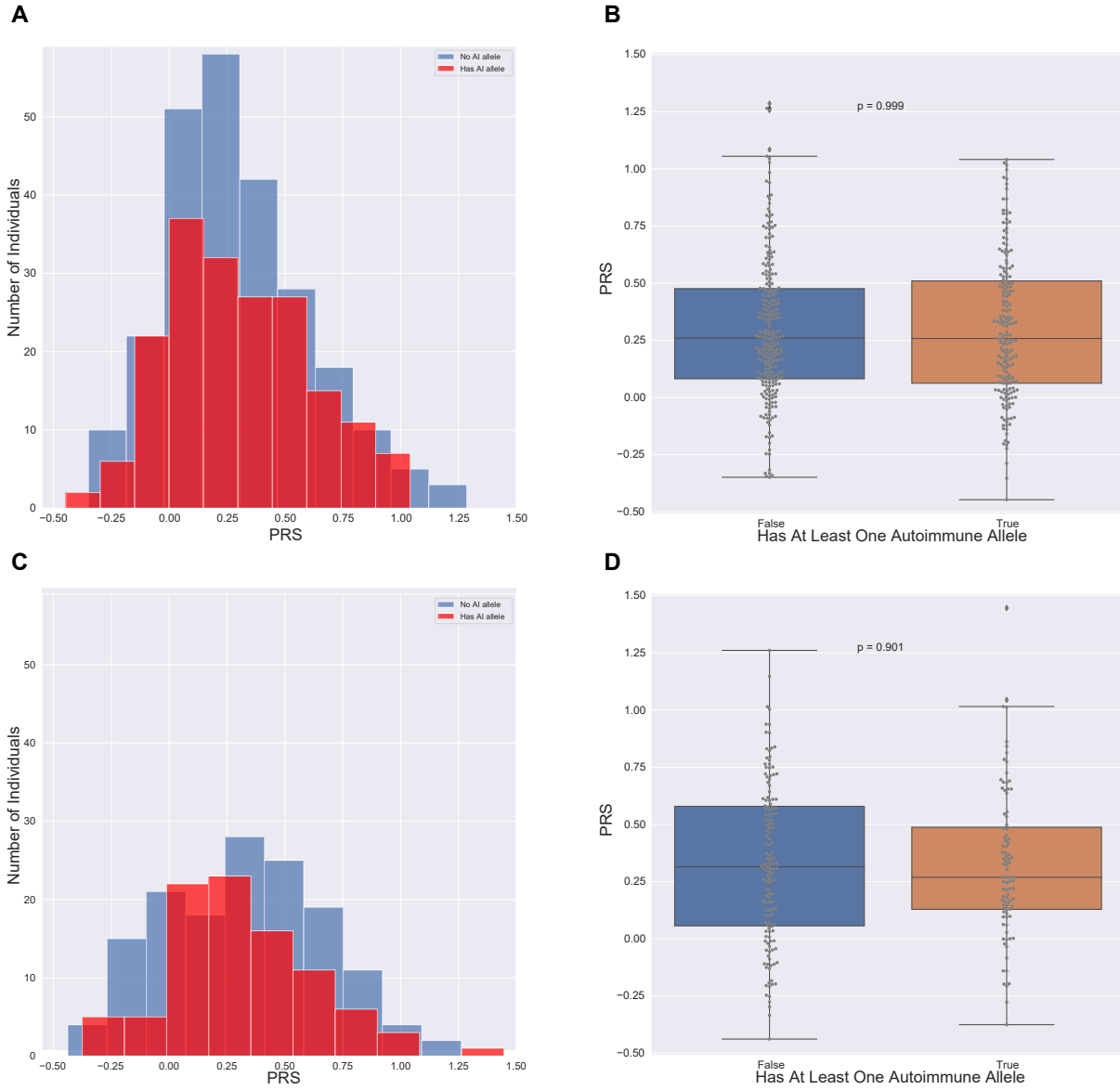
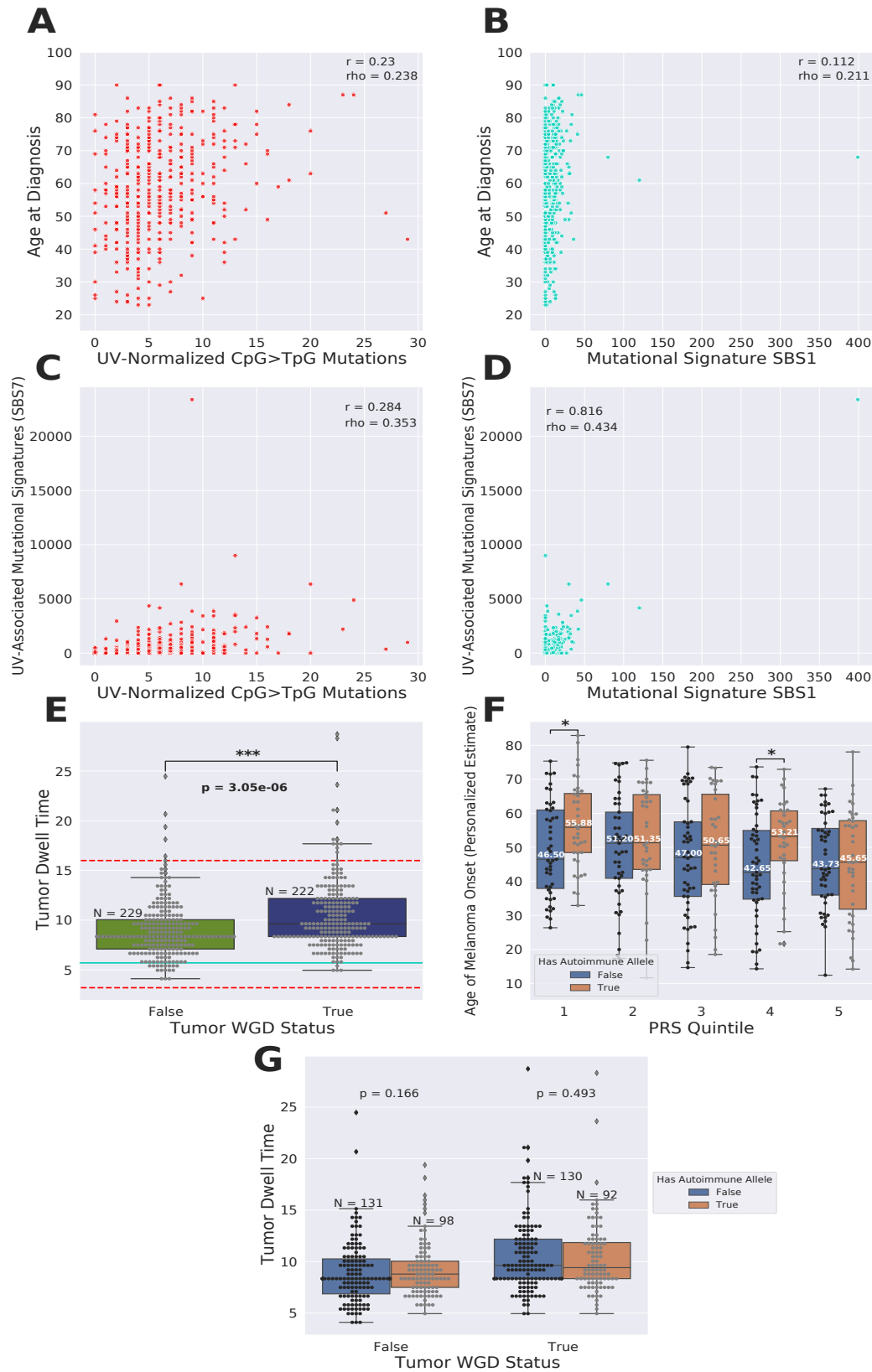


Figure C.11. PRS and MHC-I autoimmune alleles exhibit independent effects with age. A) *TCGA*: Histogram of PRS stratified across individuals with and without MHC-I autoimmune alleles. B) *TCGA*: Boxplots stratifying PRS across individuals with and without MHC-I autoimmune alleles ($p = 0.999$, T-test). C) *Validation*: Histogram of PRS stratified across individuals with and without MHC-I autoimmune alleles. D) *Validation*: Boxplots stratifying PRS across individuals with and without MHC-I autoimmune alleles ($p = 0.901$, T-test).

Figure C.12. *TCGA*: A) Correlation between UV-mutation spectrum corrected CpG>TpG mutations and age of diagnosis. B) Correlation between mutational signature SBS1 and age of diagnosis. C) Correlation between UV-mutation spectrum corrected CpG>TpG mutations and UV-associated mutational signature SBS7. D) Correlation between mutational signature SBS1 and UV-associated mutational signature SBS7. E) Tumor dwell time in WGD and non-WGD melanomas. The solid blue line at 5.7 years equates to literature-reported estimated timing from WGD to melanoma diagnosis. The dotted red lines at 3.2 and 16 years represent the literature-reported WGD to melanoma diagnosis range. F) Boxplots of personalized estimated age of melanoma onset across PRS quintiles. Autoimmune MHC-I carrier status exhibited significant later predicted ages of onset for the lowest (PRS Quintile 1; $p = 0.011$) and second-highest risk quintiles (PRS Quintile 4; $p = 0.019$). G) Autoimmune MHC-I carriers do not show longer tumor dwell times, regardless of tumor WGD status.



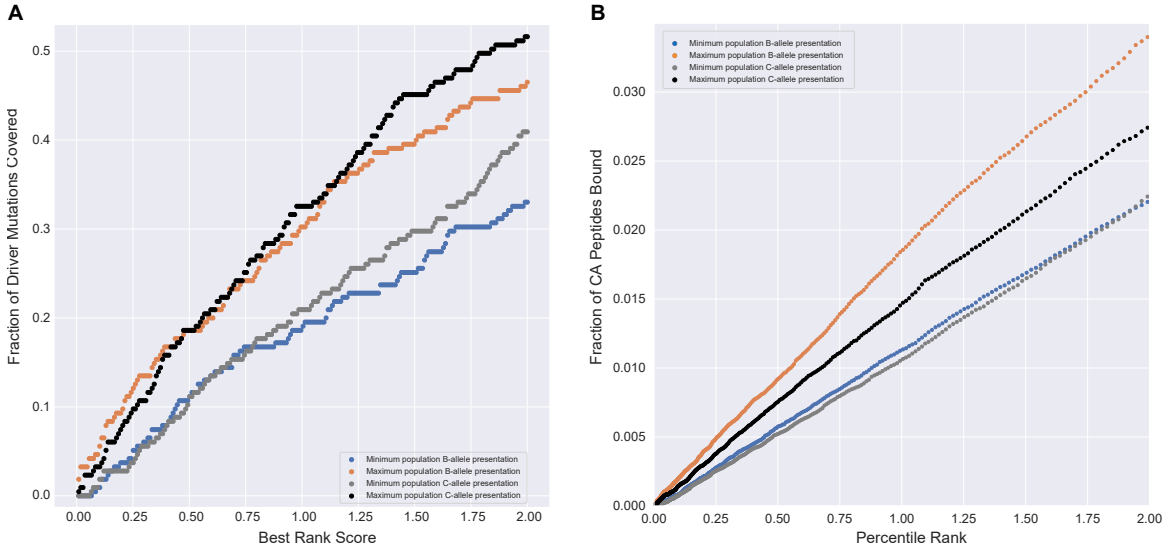


Figure C.13. A) Maximum and minimum HLA-B population allele representations (Material and Methods: *HLA Population Allele Representations*) exhibit greater neopeptide coverage than HLA-C maximum and minimum population allele representations at lower rank scores. As rank scores exceed the classical strong binding threshold (0.5), HLA-C population allele representations generally exhibit greater neopeptide coverage. B) For conserved antigens (CA), the maximum and minimum population allele representations for HLA-B presents a greater fraction of peptides than the maximum and minimum population allele representations for HLA-C. This is most pronounced in the maximum representations, where HLA-B presents a greater fraction of CA peptides compared to HLA-C as percentile rank increases.

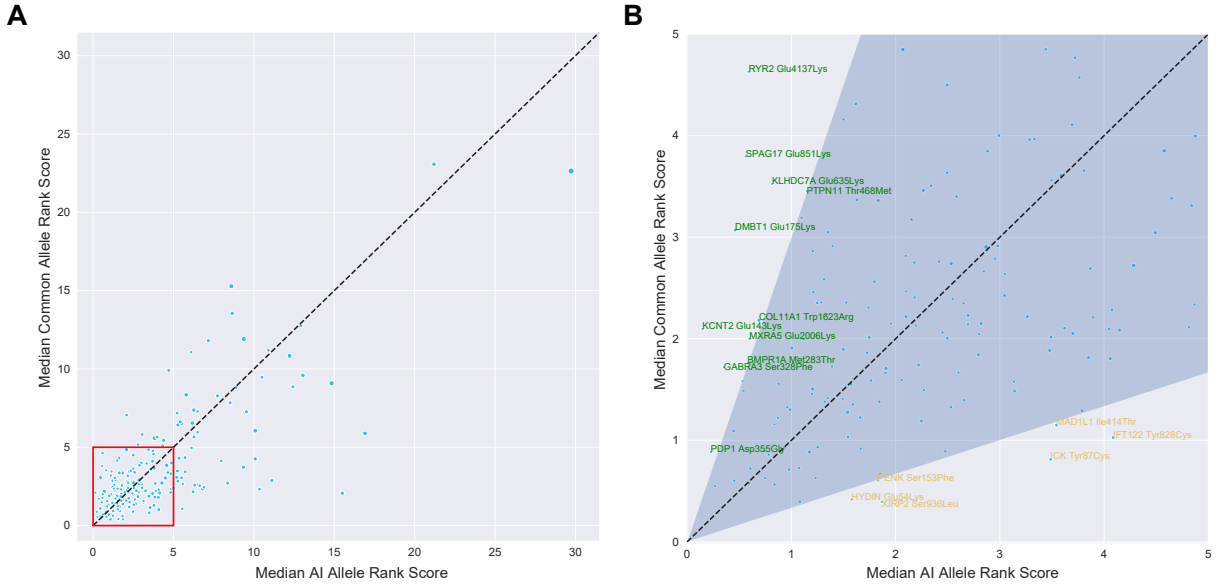


Figure C.14. Predicted median driver mutation rank scores for both common and autoimmune MHC-I alleles. The varying point size reflects a mutation's standard deviation across all autoimmune alleles, with larger points having larger standard deviations. A) Predicted median rank scores for all driver mutations ($n = 215$) for both common and AI alleles. B) Predicted median rank scores for driver mutations in a presentable range (outlined in red in panel A). Mutations with a 3-fold median rank score discrepancy between AI and common alleles in either direction are labeled, and fall outside the shaded region. Green mutations correspond to mutations that are better presented by the median autoimmune allele, while yellow mutations correspond to mutations that are better presented by the median common allele.

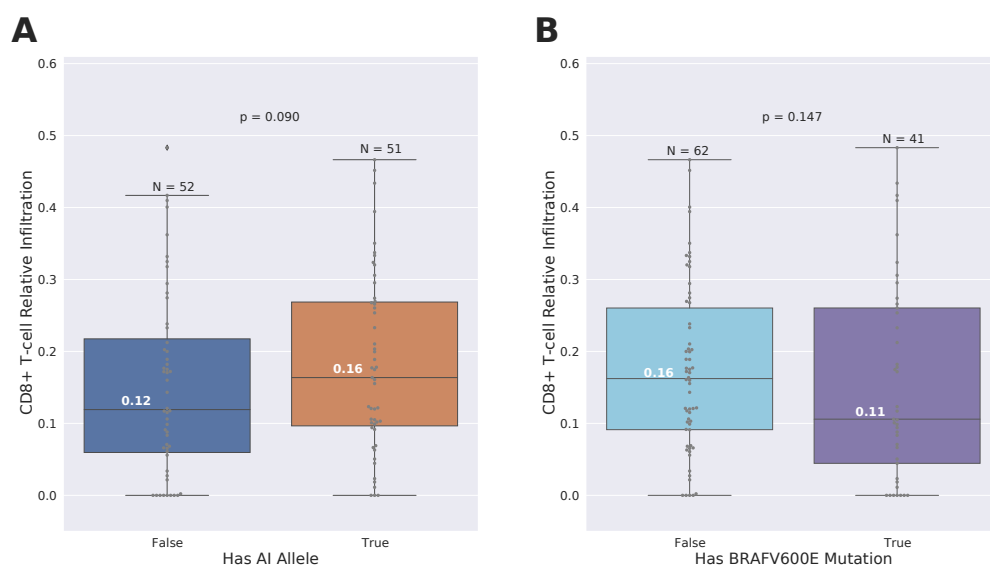


Figure C.15. *TCGA*: A) CD8+ T-Cell infiltration in carriers of at least one MHC-I autoimmune allele against non-MHC-I autoimmune allele carriers in primary melanoma tumors. B) CD8+ T-Cell infiltration in individuals with a BRAF p.Val600Glu mutation against individuals without a BRAF p.Val600Glu mutation in primary melanoma tumors.

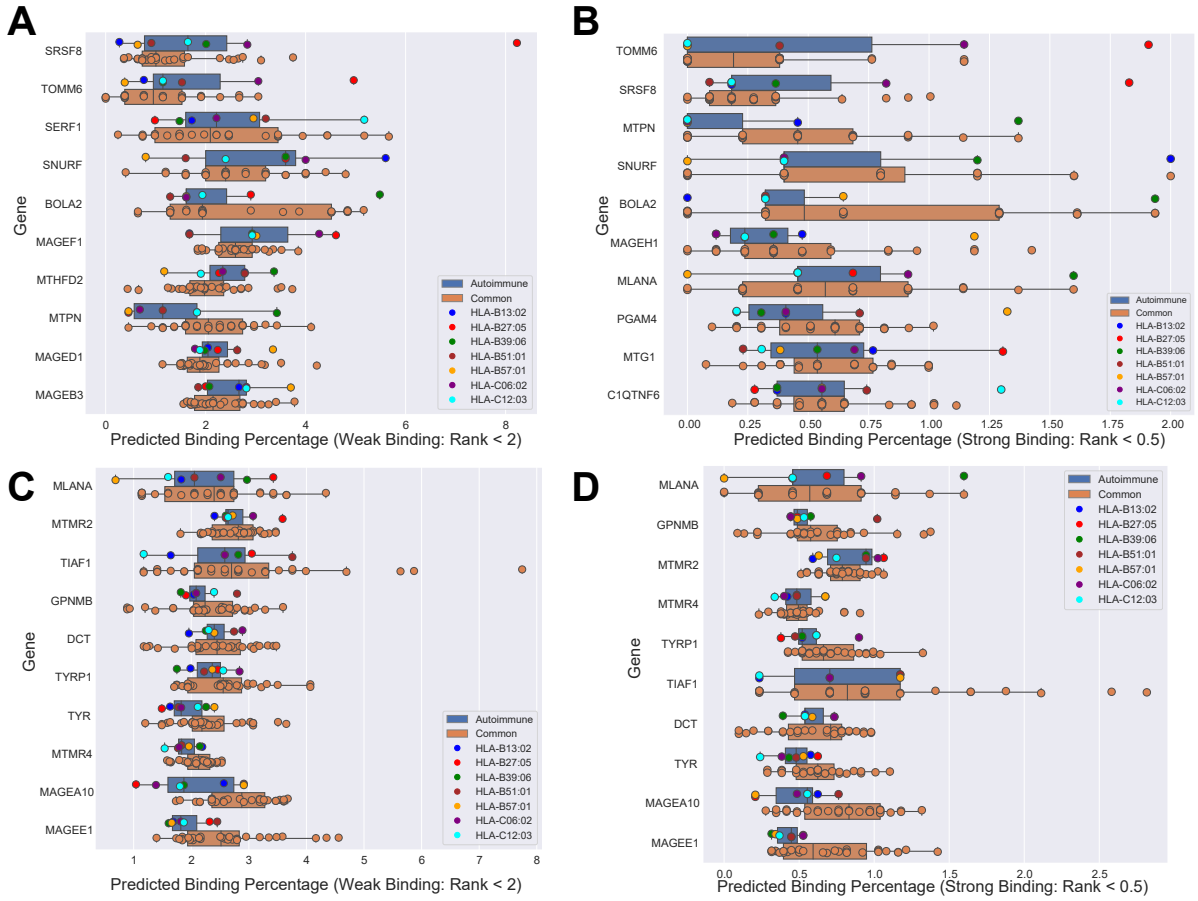


Figure C.16. Distributions of the predicted binding repertoire of conserved antigens between MHC-I autoimmune and common alleles. A-B) Top 10 conserved antigens by difference in mean percentage binding (autoimmune - common). Percent binding at the A) 2% rank threshold and B) 0.5% rank threshold. C-D) Differentially expressed conserved antigens sorted as in panels A-B showing percentage binding at the C) 2% rank threshold and D) 0.5% rank threshold.

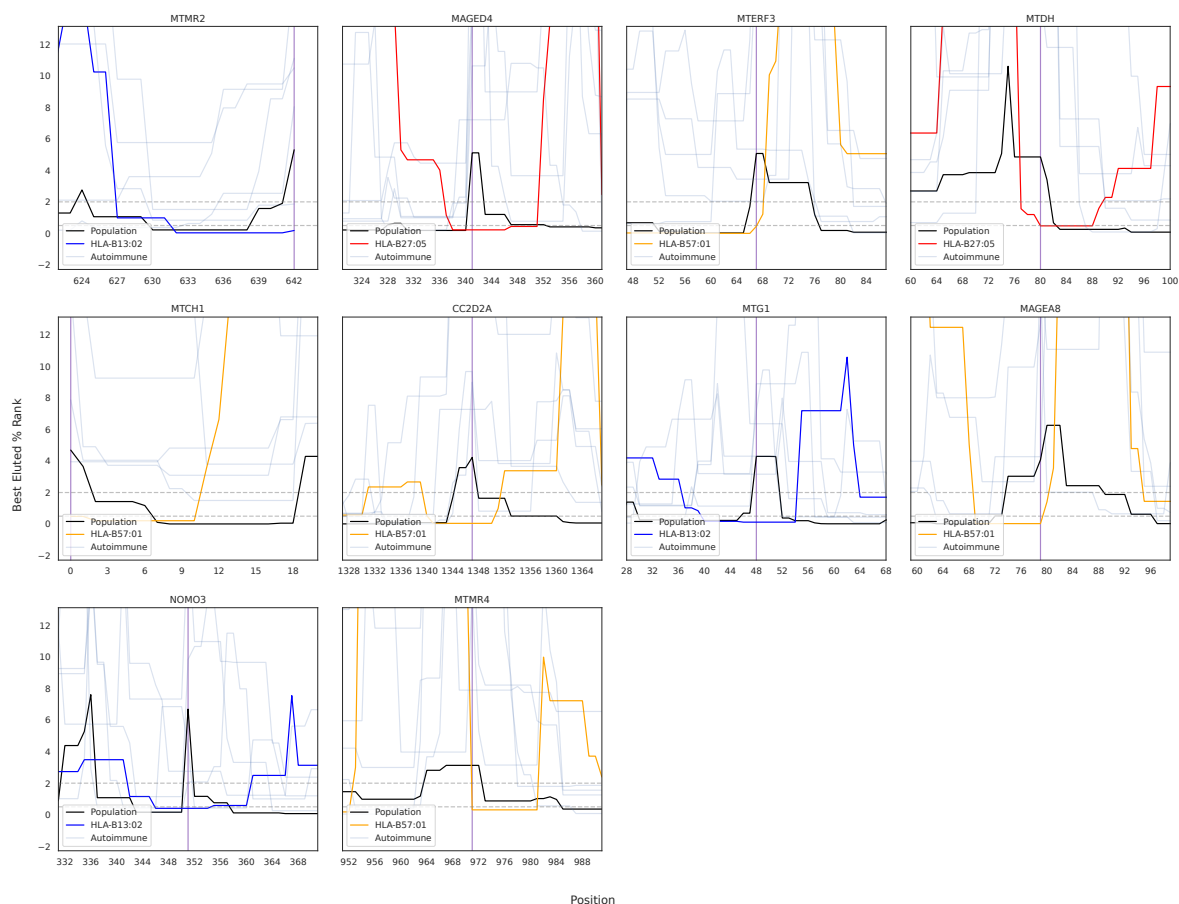


Figure C.17. Greatest position-wise differences in predicted binding to conserved antigens between HLA-B autoimmune and common alleles amongst the top 8 differences (MAGED4, MTERF3, MTDH, MTCH1, CC2D2A, MTG1, MAGEA8, NOMO3) and DE conserved antigens (MTMR2, MTMR4) not shown in Figure S2B. Common alleles are shown in black and autoimmune alleles are shown in transparent blue unless they are predicted to elute at better percentile ranks than common alleles. Vertical purple lines demarcate positions where autoimmune alleles are predicted to elute at better percentile ranks than common alleles. Plots show up to ± 20 amino acids of the demarcated positions

Bibliography

- [1] Brendan Maher. Personal genomes: The case of the missing heritability. *Nature*, 456(7218): 18–21, November 2008.
- [2] Evan E Eichler, Jonathan Flint, Greg Gibson, Augustine Kong, Suzanne M Leal, Jason H Moore, and Joseph H Nadeau. Missing heritability and strategies for finding the underlying causes of complex disease. *Nat. Rev. Genet.*, 11(6):446–450, June 2010.
- [3] Sang Hong Lee, Naomi R Wray, Michael E Goddard, and Peter M Visscher. Estimating missing heritability for disease from genome-wide association studies. *Am. J. Hum. Genet.*, 88(3):294–305, March 2011.
- [4] David Golan, Eric S Lander, and Saharon Rosset. Measuring missing heritability: inferring the contribution of common variants. *Proc. Natl. Acad. Sci. U. S. A.*, 111(49):E5272–81, December 2014.
- [5] Brendan K Bulik-Sullivan, Po-Ru Loh, Hilary K Finucane, Stephan Ripke, Jian Yang, Nick Patterson, Mark J Daly, Alkes L Price, and Benjamin M Neale. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.*, 47(3):291–295, February 2015.
- [6] Howlader N, Noone AM, Krapcho M, Miller D, Brest A, Yu M, Ruhl J, Tatalovich Z, Mariotto A, Lewis DR, Chen HS, Feuer EJ, Cronin KA (eds). SEER cancer statistics review, 1975-2017.
- [7] Brian T Helfand, James Kearns, Carly Conran, and Jianfeng Xu. Clinical validity and utility of genetic risk scores in prostate cancer. *Asian J. Androl.*, 18(4):509–514, July 2016.
- [8] Paige Maas, Myrto Barrdahl, Amit D Joshi, Paul L Auer, Mia M Gaudet, Roger L Milne, Fredrick R Schumacher, William F Anderson, David Check, Subham Chattopadhyay, Laura Baglietto, Christine D Berg, Stephen J Chanock, David G Cox, Jonine D Figueroa, Mitchell H Gail, Barry I Graubard, Christopher A Haiman, Susan E Hankinson, Robert N Hoover, Claudine Isaacs, Laurence N Kolonel, Loic Le Marchand, I-Min Lee, Sara Lindström, Kim Overvad, Isabelle Romieu, Maria-Jose Sanchez, Melissa C Southey, Daniel O Stram, Rosario Tumino, Tyler J VanderWeele, Walter C Willett, Shumin Zhang, Julie E Buring, Federico Canzian, Susan M Gapstur, Brian E Henderson, David J Hunter, Graham G Giles, Ross L Prentice, Regina G Ziegler, Peter Kraft, Montse Garcia-Closas, and Nilanjan Chatterjee. Breast cancer risk from modifiable and nonmodifiable risk factors among white

- women in the united states. *JAMA Oncol*, 2(10):1295–1302, October 2016.
- [9] Fangyi Gu, Ting-Huei Chen, Ruth M Pfeiffer, Maria Concetta Fargnoli, Donato Calista, Paola Ghiorzo, Ketty Peris, Susana Puig, Chiara Menin, Arcangela De Nicolo, Monica Rodolfo, Cristina Pellegrini, Lorenza Pastorino, Evangelos Evangelou, Tongwu Zhang, Xing Hua, Curt T DellaValle, D Timothy Bishop, Stuart MacGregor, Mark I Iles, Matthew H Law, Anne Cust, Kevin M Brown, Alexander J Stratigos, Eduardo Nagore, Stephen Chanock, Jianxin Shi, Melanoma Meta-Analysis Consortium, MelaNostrum Consortium, and Maria Teresa Landi. Combining common genetic variants and non-genetic risk factors to predict risk of cutaneous melanoma, 2018.
 - [10] Ali Torkamani, Nathan E Wineinger, and Eric J Topol. The personal and clinical utility of polygenic risk scores. *Nat. Rev. Genet.*, 19(9):581–590, May 2018.
 - [11] Amit Sud, Clare Turnbull, and Richard Houlston. Will polygenic risk scores for cancer ever be clinically useful? *npj Precision Oncology*, 5(1):1–5, May 2021.
 - [12] Francisco M De La Vega and Carlos D Bustamante. Polygenic risk scores: a biased prediction? *Genome Med.*, 10(1):100, December 2018.
 - [13] Eleftheria Zeggini and John P A Ioannidis. Meta-analysis in genome-wide association studies. *Pharmacogenomics*, 10(2):191–201, February 2009.
 - [14] Evangelos Evangelou and John P A Ioannidis. Meta-analysis methods for genome-wide association studies and beyond. *Nat. Rev. Genet.*, 14(6):379–389, May 2013.
 - [15] Fernando Pires Hartwig, Neil Martin Davies, Gibran Hemani, and George Davey Smith. Two-sample mendelian randomization: avoiding the downsides of a powerful, widely applicable but potentially fallible technique. *Int. J. Epidemiol.*, 45(6):1717–1726, December 2016.
 - [16] Eleanor Sanderson, M Maria Glymour, Michael V Holmes, Hyunseung Kang, Jean Morrison, Marcus R Munafò, Tom Palmer, C Mary Schooling, Chris Wallace, Qingyuan Zhao, and George Davey Smith. Mendelian randomization. *Nature Reviews Methods Primers*, 2(1):1–21, February 2022.
 - [17] Elisabeth M van Leeuwen, Alexandros Kanterakis, Patrick Deelen, Mathijs V Kattenberg, P Eline Slagboom, Paul I W de Bakker, Cisca Wijmenga, Morris A Swertz, Dorret I Boomsma, Cornelia M van Duijn, Lennart C Karssen, and Jouke Jan Hottenga. Population-specific genotype imputations using minimac or IMPUTE2. *Nat. Protoc.*, 10(9):1285–1296, July 2015.
 - [18] Jonathan R I Coleman, Jack Euesden, Hamel Patel, Amos A Folarin, Stephen Newhouse, and Gerome Breen. Quality control, imputation and analysis of genome-wide genotyping data from the illumina HumanCoreExome microarray. *Brief. Funct. Genomics*, 15(4):298–304, July 2016.

- [19] Carl A Anderson, Fredrik H Pettersson, Geraldine M Clarke, Lon R Cardon, Andrew P Morris, and Krina T Zondervan. Data quality control in genetic case-control association studies. *Nat. Protoc.*, 5(9):1564–1573, August 2010.
- [20] Andries T Marees, Hilde de Kluiver, Sven Stringer, Florence Vorspan, Emmanuel Curis, Cynthia Marie-Claire, and Eske M Derks. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *Int. J. Methods Psychiatr. Res.*, 27(2): e1608, June 2018.
- [21] Shing Wan Choi, Timothy Shin-Heng Mak, and Paul F O’Reilly. Tutorial: a guide to performing polygenic risk score analyses. *Nat. Protoc.*, 15(9):2759–2772, September 2020.
- [22] Kyriaki Michailidou, Sara Lindström, Joe Dennis, Jonathan Beesley, Shirley Hui, Siddhartha Kar, Audrey Lemaçon, Penny Soucy, Dylan Glubb, Asha Rostamianfar, Manjeet K Bolla, Qin Wang, Jonathan Tyrer, Ed Dicks, Andrew Lee, Zhaoming Wang, Jamie Allen, Renske Keeman, Ursula Eilber, Juliet D French, Xiao Qing Chen, Laura Fachal, Karen McCue, Amy E McCart Reed, Maya Ghoussaini, Jason S Carroll, Xia Jiang, Hilary Finucane, Marcia Adams, Muriel A Adank, Habibul Ahsan, Kristiina Aittomäki, Hoda Anton-Culver, Natalia N Antonenkova, Volker Arndt, Kristan J Aronson, Banu Arun, Paul L Auer, François Bacot, Myrto Barrdahl, Caroline Baynes, Matthias W Beckmann, Sabine Behrens, Javier Benitez, Marina Bermisheva, Leslie Bernstein, Carl Blomqvist, Natalia V Bogdanova, Stig E Bojesen, Bernardo Bonanni, Anne-Lise Børresen-Dale, Judith S Brand, Hiltrud Brauch, Paul Brennan, Hermann Brenner, Louise Brinton, Per Broberg, Ian W Brock, Annegien Brooks, Angela Brooks-Wilson, Sara Y Brucker, Thomas Brüning, Barbara Burwinkel, Katja Butterbach, Qiuyin Cai, Hui Cai, Trinidad Caldés, Federico Canzian, Angel Carracedo, Brian D Carter, Jose E Castelao, Tsun L Chan, Ting-Yuan David Cheng, Kee Seng Chia, Ji-Yeob Choi, Hans Christiansen, Christine L Clarke, NBCS Collaborators, Margriet Collée, Don M Conroy, Emilie Cordina-Duverger, Sten Cornelissen, David G Cox, Angela Cox, Simon S Cross, Julie M Cunningham, Kamila Czene, Mary B Daly, Peter Devilee, Kimberly F Doheny, Thilo Dörk, Isabel Dos-Santos-Silva, Martine Dumont, Lorraine Durcan, Miriam Dwek, Diana M Eccles, Arif B Ekici, A Heather Eliassen, Carolina Ellberg, Mingajeva Elvira, Christoph Engel, Mikael Eriksson, Peter A Fasching, Jonine Figueroa, Dieter Flesch-Janys, Olivia Fletcher, Henrik Flyger, Lin Fritschi, Valerie Gaborieau, Marike Gabrielson, Manuela Gago-Dominguez, Yu-Tang Gao, Susan M Gapstur, José A García-Sáenz, Mia M Gaudet, Vassilios Georgoulas, Graham G Giles, Gord Glendon, Mark S Goldberg, David E Goldgar, Anna González-Neira, Grethe I Grenaker Alnæs, Mervi Grip, Jacek Gronwald, Anne Grundy, Pascal Guénel, Lothar Haeberle, Eric Hahnen, Christopher A Haiman, Niclas Håkansson, Ute Hamann, Nathalie Hamel, Susan Hankinson, Patricia Harrington, Steven N Hart, Jaana M Hartikainen, Mikael Hartman, Alexander Hein, Jane Heyworth, Belynda Hicks, Peter Hillemanns, Dona N Ho, Antoinette Hollestelle, Maartje J Hooning, Robert N Hoover, John L Hopper, Ming-Feng Hou, Chia-Ni Hsiung, Guanmengqian Huang, Keith Humphreys, Junko Ishiguro, Hidemi Ito, Motoki Iwasaki, Hiroji Iwata, Anna Jakubowska, Wolfgang Janni, Esther M John, Nichola Johnson, Kristine Jones, Michael Jones, Arja Jukkola-Vuorinen, Rudolf Kaaks, Maria Kabisch, Katarzyna Kaczmarek, Daehee Kang, Yoshio Kasuga, Michael J Kerin, Sofia Khan, Elza Khusnutdinova, Johanna I Kiiski, Sung-Won Kim, Julia A Knight, Veli-Matti Kosma, Vessela N

Kristensen, Ute Krüger, Ava Kwong, Diether Lambrechts, Loic Le Marchand, Eunjung Lee, Min Hyuk Lee, Jong Won Lee, Chuen Neng Lee, Flavio Lejbkiewicz, Jingmei Li, Jenna Lilyquist, Annika Lindblom, Jolanta Lissowska, Wing-Yee Lo, Sibylle Loibl, Jirong Long, Artitaya Lophatananon, Jan Lubinski, Craig Luccarini, Michael P Lux, Edmond S K Ma, Robert J MacInnis, Tom Maishman, Enes Makalic, Kathleen E Malone, Ivana Maleva Kostovska, Arto Mannermaa, Siranoush Manoukian, Joann E Manson, Sara Margolin, Shivaani Mariapun, Maria Elena Martinez, Keitaro Matsuo, Dimitrios Mavroudis, James McKay, Catriona McLean, Hanne Meijers-Heijboer, Alfons Meindl, Primitiva Menéndez, Usha Menon, Jeffery Meyer, Hui Miao, Nicola Miller, Nur Aishah Mohd Taib, Kenneth Muir, Anna Marie Mulligan, Claire Mulot, Susan L Neuhausen, Heli Nevanlinna, Patrick Neven, Sune F Nielsen, Dong-Young Noh, Børge G Nordestgaard, Aaron Norman, Olufunmilayo I Olopade, Janet E Olson, Håkan Olsson, Curtis Olswold, Nick Orr, V Shane Pankratz, Sue K Park, Tjoung-Won Park-Simon, Rachel Lloyd, Jose I A Perez, Paolo Peterlongo, Julian Peto, Kelly-Anne Phillips, Mila Pinchev, Dijana Plaseska-Karanfilska, Ross Prentice, Nadege Presneau, Darya Prokofyeva, Elizabeth Pugh, Katri Pylkäs, Brigitte Rack, Paolo Radice, Nazneen Rahman, Gadi Rennert, Hedy S Rennert, Valerie Rhenius, Atocha Romero, Jane Romm, Kathryn J Ruddy, Thomas Rüdiger, Anja Rudolph, Matthias Ruebner, Emiel J T Rutgers, Emmanouil Saloustros, Dale P Sandler, Suleeporn Sangrajang, Elinor J Sawyer, Daniel F Schmidt, Rita K Schmutzler, Andreas Schneeweiss, Minouk J Schoemaker, Fredrick Schumacher, Peter Schürmann, Rodney J Scott, Christopher Scott, Sheila Seal, Caroline Seynaeve, Mitul Shah, Priyanka Sharma, Chen-Yang Shen, Grace Sheng, Mark E Sherman, Martha J Shrubsole, Xiao-Ou Shu, Ann Smeets, Christof Sohn, Melissa C Southey, John J Spinelli, Christa Stegmaier, Sarah Stewart-Brown, Jennifer Stone, Daniel O Stram, Harald Surowy, Anthony Swerdlow, Rulla Tamimi, Jack A Taylor, Maria Tengström, Soo H Teo, Mary Beth Terry, Daniel C Tessier, Somchai Thanasitthichai, Kathrin Thöne, Rob A E M Tollenaar, Ian Tomlinson, Ling Tong, Diana Torres, Thérèse Truong, Chiu-Chen Tseng, Shoichiro Tsugane, Hans-Ulrich Ulmer, Giske Ursin, Michael Untch, Celine Vachon, Christi J van Asperen, David Van Den Berg, Ans M W van den Ouweland, Lizet van der Kolk, Rob B van der Luit, Daniel Vincent, Jason Vollenweider, Quinten Waisfisz, Shan Wang-Gohrke, Clarice R Weinberg, Camilla Wendt, Alice S Whittemore, Hans Wildiers, Walter Willett, Robert Winqvist, Alicja Wolk, Anna H Wu, Lucy Xia, Taiki Yamaji, Xiaohong R Yang, Cheng Har Yip, Keun-Young Yoo, Jyh-Cherng Yu, Wei Zheng, Ying Zheng, Bin Zhu, Argyrios Ziogas, Elad Ziv, ABCTB Investigators, ConFab/AOCS Investigators, Sunil R Lakhani, Antonis C Antoniou, Arnaud Droit, Irene L Andrulis, Christopher I Amos, Fergus J Couch, Paul D P Pharoah, Jenny Chang-Claude, Per Hall, David J Hunter, Roger L Milne, Montserrat García-Closas, Marjanka K Schmidt, Stephen J Chanock, Alison M Dunning, Stacey L Edwards, Gary D Bader, Georgia Chenevix-Trench, Jacques Simard, Peter Kraft, and Douglas F Easton. Association analysis identifies 65 new breast cancer risk loci. *Nature*, 551(7678):92–94, November 2017.

- [23] David V Conti, Burcu F Darst, Lilit C Moss, Edward J Saunders, Xin Sheng, Alisha Chou, Fredrick R Schumacher, Ali Amin Al Olama, Sara Benlloch, Tokhir Dadaev, Mark N Brook, Ali Sahimi, Thomas J Hoffmann, Atushi Takahashi, Koichi Matsuda, Yukihide Momozawa, Masashi Fujita, Kenneth Muir, Artitaya Lophatananon, Peggy Wan, Loic Le Marchand, Lynne R Wilkens, Victoria L Stevens, Susan M Gapstur, Brian D Carter, Johanna Schleutker, Teuvo L J Tammela, Csilla Sipeky, Anssi Auvinen, Graham G

Giles, Melissa C Southey, Robert J MacInnis, Cezary Cybulski, Dominika Wokołorczyk, Jan Lubiński, David E Neal, Jenny L Donovan, Freddie C Hamdy, Richard M Martin, Børge G Nordestgaard, Sune F Nielsen, Maren Weischer, Stig E Bojesen, Martin Andreas Røder, Peter Iversen, Jyotsna Batra, Suzanne Chambers, Leire Moya, Lisa Horvath, Judith A Clements, Wayne Tilley, Gail P Risbridger, Henrik Gronberg, Markus Aly, Robert Szulkin, Martin Eklund, Tobias Nordström, Nora Pashayan, Alison M Dunning, Maya Ghoussaini, Ruth C Travis, Tim J Key, Elio Riboli, Jong Y Park, Thomas A Sellers, Hui-Yi Lin, Demetrius Albanes, Stephanie J Weinstein, Lorelei A Mucci, Edward Giovannucci, Sara Lindstrom, Peter Kraft, David J Hunter, Kathryn L Penney, Constance Turman, Catherine M Tangen, Phyllis J Goodman, Ian M Thompson, Robert J Hamilton, Neil E Fleshner, Antonio Finelli, Marie-Élise Parent, Janet L Stanford, Elaine A Ostrander, Milan S Geybels, Stella Koutros, Laura E Beane Freeman, Meir Stampfer, Alicja Wolk, Niclas Håkansson, Gerald L Andriole, Robert N Hoover, Mitchell J Machiela, Karina Dalsgaard Sørensen, Michael Borre, William J Blot, Wei Zheng, Edward D Yeboah, James E Mensah, Yong-Jie Lu, Hong-Wei Zhang, Ninghan Feng, Xueying Mao, Yudong Wu, Shan-Chao Zhao, Zan Sun, Stephen N Thibodeau, Shannon K McDonnell, Daniel J Schaid, Catharine M L West, Neil Burnet, Gill Barnett, Christiane Maier, Thomas Schnoeller, Manuel Luedeke, Adam S Kibel, Bettina F Drake, Olivier Cussenot, Géraldine Cancel-Tassin, Florence Menegaux, Thérèse Truong, Yves Akoli Koudou, Esther M John, Eli Marie Grindedal, Lovise Maehle, Kay-Tee Khaw, Sue A Ingles, Mariana C Stern, Ana Vega, Antonio Gómez-Caamaño, Laura Fachal, Barry S Rosenstein, Sarah L Kerns, Harry Ostrer, Manuel R Teixeira, Paula Paulo, Andreia Brandão, Stephen Watya, Alexander Lubwama, Jeannette T Bensen, Elizabeth T H Fontham, James Mohler, Jack A Taylor, Manolis Kogevinas, Javier Llorca, Gemma Castaño-Vinyals, Lisa Cannon-Albright, Craig C Teerlink, Chad D Huff, Sara S Strom, Luc Multigner, Pascal Blanchet, Laurent Brureau, Radka Kaneva, Chavdar Slavov, Vanio Mitev, Robin J Leach, Brandi Weaver, Hermann Brenner, Katarina Cuk, Bernd Holleccek, Kai-Uwe Saum, Eric A Klein, Ann W Hsing, Rick A Kittles, Adam B Murphy, Christopher J Logothetis, Jeri Kim, Susan L Neuhausen, Linda Steele, Yuan Chun Ding, William B Isaacs, Barbara Nemesure, Anselm J M Hennis, John Carpten, Hardev Pandha, Agnieszka Michael, Kim De Ruyck, Gert De Meerleer, Piet Ost, Jianfeng Xu, Azad Razack, Jasmine Lim, Soo-Hwang Teo, Lisa F Newcomb, Daniel W Lin, Jay H Fowke, Christine Neslund-Dudas, Benjamin A Rybicki, Marija Gamulin, Davor Lessel, Tomislav Kulis, Nawaid Usmani, Sandeep Singhal, Parliament, Matthew, Frank Claessens, Steven Joniau, Thomas Van den Broeck, Manuela Gago-Dominguez, Jose Esteban Castela, Maria Elena Martinez, Samantha Larkin, Paul A Townsend, Claire Aukim-Hastie, William S Bush, Melinda C Aldrich, Dana C Crawford, Shiv Srivastava, Jennifer C Cullen, Gyorgy Petrovics, Graham Casey, Monique J Roobol, Guido Jenster, Ron H N van Schaik, Jennifer J Hu, Maureen Sanderson, Rohit Varma, Roberta McKean-Cowdin, Mina Torres, Nicholas Mancuso, Sonja I Berndt, Stephen K Van Den Eeden, Douglas F Easton, Stephen J Chanock, Michael B Cook, Fredrik Wiklund, Hidewaki Nakagawa, John S Witte, Rosalind A Eeles, Zsafia Kote-Jarai, and Christopher A Haiman. Trans-ancestry genome-wide association meta-analysis of prostate cancer identifies new susceptibility loci and informs genetic risk prediction. *Nat. Genet.*, 53(1):65–75, January 2021.

- [24] Christopher I Amos, Joe Dennis, Zhaoming Wang, Jinyoung Byun, Fredrick R Schumacher, Simon A Gayther, Graham Casey, David J Hunter, Thomas A Sellers, Stephen B Gruber,

- Alison M Dunning, Kyriaki Michailidou, Laura Fachal, Kimberly Doheny, Amanda B Spurdle, Yafang Li, Xiangjun Xiao, Jane Romm, Elizabeth Pugh, Gerhard A Coetzee, Dennis J Hazelett, Stig E Bojesen, Charlisce Caga-Anan, Christopher A Haiman, Ahsan Kamal, Craig Luccarini, Daniel Tessier, Daniel Vincent, François Bacot, David J Van Den Berg, Stefanie Nelson, Stephen Demetriades, David E Goldgar, Fergus J Couch, Judith L Forman, Graham G Giles, David V Conti, Heike Bickeböllner, Angela Risch, Melanie Waldenberger, Irene Brüske-Hohlfeld, Belynda D Hicks, Hua Ling, Lesley McGuffog, Andrew Lee, Karoline Kuchenbaecker, Penny Soucy, Judith Manz, Julie M Cunningham, Katja Butterbach, Zsolia Kote-Jarai, Peter Kraft, Liesel FitzGerald, Sara Lindström, Marcia Adams, James D McKay, Catherine M Phelan, Sara Benlloch, Linda E Kelemen, Paul Brennan, Marjorie Riggan, Tracy A O'Mara, Hongbing Shen, Yongyong Shi, Deborah J Thompson, Marc T Goodman, Sune F Nielsen, Andrew Berchuck, Sylvie Laboissiere, Stephanie L Schmit, Tameka Shelford, Christopher K Edlund, Jack A Taylor, John K Field, Sue K Park, Kenneth Offit, Mads Thomassen, Rita Schmutzler, Laura Ottini, Rayjean J Hung, Jonathan Marchini, Ali Amin Al Olama, Ulrike Peters, Rosalind A Eeles, Michael F Seldin, Elizabeth Gillanders, Daniela Seminara, Antonis C Antoniou, Paul D P Pharoah, Georgia Chenevix-Trench, Stephen J Chanock, Jacques Simard, and Douglas F Easton. The OncoArray consortium: A network for understanding the genetic architecture of common cancers. *Cancer Epidemiol. Biomarkers Prev.*, 26(1):126–135, January 2017.
- [25] Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, Adrian Cortes, Samantha Welsh, Alan Young, Mark Effingham, Gil McVean, Stephen Leslie, Naomi Allen, Peter Donnelly, and Jonathan Marchini. The UK biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, October 2018.
- [26] Sayantan Das, Lukas Forer, Sebastian Schönherr, Carlo Sidore, Adam E Locke, Alan Kwong, Scott I Vrieze, Emily Y Chew, Shawn Levy, Matt McGue, David Schlessinger, Dwight Stambolian, Po-Ru Loh, William G Iacono, Anand Swaroop, Laura J Scott, Francesco Cucca, Florian Kronenberg, Michael Boehnke, Gonçalo R Abecasis, and Christian Fuchsberger. Next-generation genotype imputation service and methods. *Nat. Genet.*, 48(10):1284–1287, October 2016.
- [27] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A R Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I W de Bakker, Mark J Daly, and Pak C Sham. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, 81(3):559–575, September 2007.
- [28] Christopher C Chang, Carson C Chow, Laurent Cam Tellier, Shashaank Vattikuti, Shaun M Purcell, and James J Lee. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*, 4(1), February 2015.
- [29] James V Talwar. jvtalwar/GRIEVOUS: GRIEVOUS version 0.1.5.
- [30] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*,

June 2019.

- [31] Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R Ledsam, Agnieszka Grabska-Barwinska, Kyle R Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods*, 18(10):1196–1203, October 2021.
- [32] Molly Gasperini, Jacob M Tome, and Jay Shendure. Towards a comprehensive catalogue of validated and target-linked human enhancers. *Nat. Rev. Genet.*, 21(5):292–310, May 2020.
- [33] Ana Pombo and Niall Dillon. Three-dimensional genome architecture: players and mechanisms. *Nat. Rev. Mol. Cell Biol.*, 16(4):245–257, April 2015.
- [34] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. CvT: Introducing convolutions to vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, October 2021.
- [35] Montgomery Slatkin. Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. *Nat. Rev. Genet.*, 9(6):477–485, June 2008.
- [36] Bjarni J Vilhjálmsson, Jian Yang, Hilary K Finucane, Alexander Gusev, Sara Lindström, Stephan Ripke, Giulio Genovese, Po-Ru Loh, Gaurav Bhatia, Ron Do, Tristan Hayeck, Hong-Hee Won, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Discovery, Biology, and Risk of Inherited Variants in Breast Cancer (DRIVE) study, Sekar Kathiresan, Michele Pato, Carlos Pato, Rulla Tamimi, Eli Stahl, Noah Zaitlen, Bogdan Pasaniuc, Gillian Belbin, Eimear E Kenny, Mikkel H Schierup, Philip De Jager, Nikolaos A Patsopoulos, Steve McCarroll, Mark Daly, Shaun Purcell, Daniel Chasman, Benjamin Neale, Michael Goddard, Peter M Visscher, Peter Kraft, Nick Patterson, and Alkes L Price. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *Am. J. Hum. Genet.*, 97(4):576–592, October 2015.
- [37] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *arXiv [cs.CL]*, May 2020.
- [38] Mathilde Caron, Hugo Touvron, Ishan Misra, Herve Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9630–9640. IEEE, October 2021.
- [39] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly,

Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

- [40] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Lukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O’Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea

- Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, C J Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. GPT-4 technical report. *arXiv [cs.CL]*, March 2023.
- [41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł Ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I Guyon, U V Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, and R Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.
 - [42] Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Oteri, Christian Dallago, Evan Trop, Hassan Sirelkhatim, Guillaume Richard, Marcin Skwark, Karim Beguir, Marie Lopez, and Thomas Pierrot. The nucleotide transformer: Building and evaluating robust foundation models for human genomics. *bioRxiv*, page 2023.01.11.523679, January 2023.
 - [43] Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. *Bioinformatics*, 37(15):2112–2120, August 2021.
 - [44] Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana Davuluri, and Han Liu. DNABERT-2: Efficient foundation model and benchmark for multi-species genome. *arXiv [q-bio.GN]*, June 2023.
 - [45] Adrien Badré, Li Zhang, Wellington Muchero, Justin C Reynolds, and Chongle Pan. Deep neural network improves the estimation of polygenic risk scores for breast cancer. *J. Hum. Genet.*, 66(4):359–369, April 2021.
 - [46] Pau Bellot, Gustavo de Los Campos, and Miguel Pérez-Enciso. Can deep learning improve genomic prediction of complex human traits? *Genetics*, 210(3):809–819, November 2018.
 - [47] Han Li, Jianyang Zeng, Michael P Snyder, and Sai Zhang. PRS-net: Interpretable polygenic risk scores via geometric learning. In *Lecture Notes in Computer Science*, Lecture notes in computer science, pages 377–380. Springer Nature Switzerland, Cham, 2024.
 - [48] Arno van Hilten, Steven A Kushner, Manfred Kayser, M Arfan Ikram, Hicab H H Adams, Caroline C W Klaver, Wiro J Niessen, and Gennady V Roshchupkin. GenNet framework: interpretable deep learning for predicting phenotypes from genetic data. *Communications Biology*, 4(1):1–9, September 2021.
 - [49] Xiaopu Zhou, Yu Chen, Fanny C F Ip, Yuanbing Jiang, Han Cao, Ge Lv, Huan Zhong, Jiahang Chen, Tao Ye, Yuwen Chen, Yulin Zhang, Shuangshuang Ma, Ronnie M N Lo,

- Estella P S Tong, Alzheimer’s Disease Neuroimaging Initiative, Vincent C T Mok, Timothy C Y Kwok, Qihao Guo, Kin Y Mok, Maryam Shoai, John Hardy, Lei Chen, Amy K Y Fu, and Nancy Y Ip. Deep learning-based polygenic risk analysis for alzheimer’s disease prediction. *Commun. Med. (Lond.)*, 3(1):49, April 2023.
- [50] Costa Georgantas, Zoltán Kutalik, and Jonas Richiardi. Deep learning for polygenic risk prediction. *medRxiv*, page 2024.04.19.24306079, April 2024.
- [51] Ingoo Lee, Zach Wallace, Sungjoon Park, Hojung Nam, Amit R Majithia, and Trey Ideker. Mechanistic genotype-phenotype translation using hierarchical transformers. *bioRxiv*, page 2024.10.23.619940, October 2024.
- [52] Jiajie Peng, Zhijie Bao, Jingyi Li, Ruijiang Han, Yuxian Wang, Lu Han, Jinghao Peng, Tao Wang, Jianye Hao, Zhongyu Wei, and Xuequn Shang. DeepRisk: A deep learning approach for genome-wide assessment of common disease risk. *Fundam Res*, 4(4):752–760, July 2024.
- [53] Tian Ge, Chia-Yen Chen, Yang Ni, Yen-Chen Anne Feng, and Jordan W Smoller. Polygenic prediction via bayesian regression and continuous shrinkage priors. *Nat. Commun.*, 10(1):1776, April 2019.
- [54] Luke R Lloyd-Jones, Jian Zeng, Julia Sidorenko, Loïc Yengo, Gerhard Moser, Kathryn E Kemper, Huanwei Wang, Zhili Zheng, Reedik Magi, Tõnu Esko, Andres Metspalu, Naomi R Wray, Michael E Goddard, Jian Yang, and Peter M Visscher. Improved polygenic prediction by bayesian multiple regression on summary statistics. *Nat. Commun.*, 10(1):5086, November 2019.
- [55] Paul J Newcombe, Christopher P Nelson, Niles J Samani, and Frank Dudbridge. A flexible and parallelizable approach to genome-wide polygenic risk scores. *Genet. Epidemiol.*, 43(7):730–741, October 2019.
- [56] Florian Privé, Julyan Arbel, and Bjarni J Vilhjálmsson. LDpred2: better, faster, stronger. *Bioinformatics*, 36(22-23):5424–5431, April 2021.
- [57] Christina B Azodi, Jiliang Tang, and Shin-Han Shiu. Opening the black box: Interpretable machine learning for geneticists. *Trends Genet.*, 36(6):442–455, June 2020.
- [58] David S Watson. Interpretable machine learning for genomics. *Hum. Genet.*, 141(9):1499–1513, September 2022.
- [59] Brent M Kuenzi, Jisoo Park, Samson H Fong, Kyle S Sanchez, John Lee, Jason F Kreisberg, Jianzhu Ma, and Trey Ideker. Predicting drug response and synergy using a deep learning model of human cancer cells. *Cancer Cell*, 38(5):672–684.e6, November 2020.
- [60] Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan

- Faulkner, Caglar Gulcehre, Francis Song, Andrew Ballard, Justin Gilmer, George Dahl, Ashish Vaswani, Kelsey Allen, Charles Nash, Victoria Langston, Chris Dyer, Nicolas Heess, Daan Wierstra, Pushmeet Kohli, Matt Botvinick, Oriol Vinyals, Yujia Li, and Razvan Pascanu. Relational inductive biases, deep learning, and graph networks. *arXiv [cs.LG]*, June 2018.
- [61] Charles M Ewing, Anna M Ray, Ethan M Lange, Kimberly A Zuhlke, Christiane M Robbins, Waibhav D Tembe, Kathleen E Wiley, Sarah D Isaacs, Dorhyun Johng, Yunfei Wang, Chris Bizon, Guifang Yan, Marta Gielzak, Alan W Partin, Vijayalakshmi Shanmugam, Tyler Izatt, Shripad Sinari, David W Craig, S Lilly Zheng, Patrick C Walsh, James E Montie, Jianfeng Xu, John D Carpten, William B Isaacs, and Kathleen A Cooney. Germline mutations in HOXB13 and prostate-cancer risk. *N. Engl. J. Med.*, 366(2):141–149, January 2012.
- [62] Ken J Kron, Alexander Murison, Stanley Zhou, Vincent Huang, Takafumi N Yamaguchi, Yu-Jia Shiah, Michael Fraser, Theodorus van der Kwast, Paul C Boutros, Robert G Bristow, and Mathieu Lupien. TMPRSS2-ERG fusion co-opts master transcription factors and activates NOTCH signaling in primary prostate cancer. *Nat. Genet.*, 49(9):1336–1345, September 2017.
- [63] Jianghua Wang, Yi Cai, and Michael Ittmann. TMPRSS2/ERG fusions. In *Encyclopedia of Cancer*, pages 2986–2988. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.
- [64] Jun Wei, Zhuqing Shi, Rong Na, Chi-Hsiung Wang, W Kyle Resurreccion, S Lilly Zheng, Peter J Hulick, Kathleen A Cooney, Brian T Helfand, William B Isaacs, and Jianfeng Xu. Germline HOXB13 G84E mutation carriers and risk to twenty common types of cancer: results from the UK biobank. *Br. J. Cancer*, 123(9):1356–1359, October 2020.
- [65] Gad Abraham, Yixuan Qiu, and Michael Inouye. FlashPCA2: principal component analysis of biobank-scale genotype datasets. *Bioinformatics*, 33(17):2776–2778, September 2017.
- [66] Meghana S Pagadala, Joshua A Linscott, James V Talwar, Tyler M Seibert, Brent Rose, Julie Lynch, Matthew Panizzon, Richard Hauger, Moritz H Hansen, Jesse D Sammon, Matthew H Hayn, Karim Kader, Hannah Carter, and Stephen T Ryan. PRState: Incorporating genetic ancestry in prostate cancer risk scores for men of african ancestry. *BMC Cancer*, 22(1):1289, December 2022.
- [67] Erin L Richard, Linda K McEvoy, Ian J Deary, Gail Davies, Steven Y Cao, Eyal Oren, John E Alcaraz, Andrea Z LaCroix, Jan Bressler, and Rany M Salem. Markers of kidney function, genetic variation related to cognitive function, and cognitive performance in the UK biobank. *BMC Nephrol.*, 23(1):159, April 2022.
- [68] Meghana Pagadala, Timothy J Sears, Victoria H Wu, Eva Pérez-Guijarro, Hyo Kim, Andrea Castro, James V Talwar, Cristian Gonzalez-Colin, Steven Cao, Benjamin J Schmiedel, Shervin Goudarzi, Divya Kirani, Jessica Au, Tongwu Zhang, Teresa Landi, Rany M Salem, Gerald P Morris, Olivier Harismendy, Sandip Pravin Patel, Ludmil B Alexandrov, Jill P

- Mesirov, Maurizio Zanetti, Chi-Ping Day, Chun Chieh Fan, Wesley K Thompson, Glenn Merlino, J Silvio Gutkind, Pandurangan Vijayanand, and Hannah Carter. Germline modifiers of the tumor immune microenvironment implicate drivers of cancer risk and immunotherapy response. *Nat. Commun.*, 14(1):2744, May 2023.
- [69] James V Talwar, Adam Klie, Meghana S Pagadala, and Hannah Carter. GRIEVOUS: your command-line general for resolving cross-dataset genotype inconsistencies. *Bioinformatics*, 40(8):btac489, August 2024.
- [70] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. *arXiv [cs.CV]*, September 2023.
- [71] Alexei Baevski and Michael Auli. Adaptive input representations for neural language modeling. *ICLR*, 2019.
- [72] Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F Wong, and Lidia S Chao. Learning deep transformer models for machine translation. *ACL*, 2019.
- [73] Noam Shazeer. GLU variants improve transformer. *arXiv [cs.LG]*, February 2020.
- [74] Seunghoon Lee, Seunghyun Lee, and Byung Cheol Song. Improving vision transformers to learn small-size dataset from scratch. *IEEE Access*, 10:123212–123224, 2022.
- [75] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *ICLR*, 2019.
- [76] Shing Wan Choi and Paul F O’Reilly. PRSice-2: Polygenic risk score software for biobank-scale data. *Gigascience*, 8(7), July 2019.
- [77] Florian Privé, Clara Albiñana, Julyan Arbel, Bogdan Pasaniuc, and Bjarni J Vilhjálmsson. Inferring disease architecture and predictive ability with LDpred2-auto. *Am. J. Hum. Genet.*, 110(12):2042–2055, December 2023.
- [78] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD ’19, pages 2623–2631, New York, NY, USA, July 2019. Association for Computing Machinery.
- [79] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 782–791. IEEE, June 2021.
- [80] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. *arXiv [cs.LG]*, May 2020.
- [81] Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting

- bi-modal and encoder-decoder transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, October 2021.
- [82] Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. *Association for Computational Linguistics*, 2019.
 - [83] E R DeLong, D M DeLong, and D L Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44(3):837–845, September 1988.
 - [84] Rojina Kashefi, Leili Barekatain, Mohammad Sabokrou, and Fatemeh Aghaeipoor. Explainability of vision transformers: A comprehensive review and new perspectives. *arXiv [cs.CV]*, November 2023.
 - [85] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626. IEEE, October 2017.
 - [86] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR, 2017.
 - [87] Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek. Layer-wise relevance propagation for neural networks with local renormalization layers. In *Artificial Neural Networks and Machine Learning – ICANN 2016*, Lecture notes in computer science, pages 63–71. Springer International Publishing, Cham, 2016.
 - [88] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML’17, pages 3145–3153. JMLR.org, August 2017.
 - [89] Ian Covert, Chanwoo Kim, and Su-In Lee. Learning to estimate shapley values with vision transformers. *arXiv [cs.CV]*, June 2022.
 - [90] Sarah Benaffif, Zsofia Kote-Jarai, Rosalind A Eeles, and PRACTICAL Consortium. A review of prostate cancer genome-wide association studies (GWAS). *Cancer Epidemiol. Biomarkers Prev.*, 27(8):845–857, August 2018.
 - [91] Rosalind Eeles, Chee Goh, Elena Castro, Elizabeth Bancroft, Michelle Guy, Ali Amin Al Olama, Douglas Easton, and Zsofia Kote-Jarai. The genetic epidemiology of prostate cancer and its clinical implications. *Nat. Rev. Urol.*, 11(1):18–31, January 2014.

- [92] Rosalind A Eeles, Zsofia Kote-Jarai, Graham G Giles, Ali Amin Al Olama, Michelle Guy, Sarah K Jugurnauth, Shani Mulholland, Daniel A Leongamornlert, Stephen M Edwards, Jonathan Morrison, Helen I Field, Melissa C Southey, Gianluca Severi, Jenny L Donovan, Freddie C Hamdy, David P Dearnaley, Kenneth R Muir, Charmaine Smith, Melisa Bagnato, Audrey T Ardern-Jones, Amanda L Hall, Lynne T O'Brien, Beatrice N Gehr-Swain, Rosemary A Wilkinson, Angie Cox, Sarah Lewis, Paul M Brown, Sameer G Jhavar, Malgorzata Tymrakiewicz, Artitaya Lophatananon, Sarah L Bryant, UK Genetic Prostate Cancer Study Collaborators, British Association of Urological Surgeons' Section of Oncology, UK ProtecT Study Collaborators, Alan Horwich, Robert A Huddart, Vincent S Khoo, Christopher C Parker, Christopher J Woodhouse, Alan Thompson, Tim Christmas, Chris Ogden, Cyril Fisher, Charles Jamieson, Colin S Cooper, Dallas R English, John L Hopper, David E Neal, and Douglas F Easton. Multiple newly identified loci associated with prostate cancer susceptibility. *Nat. Genet.*, 40(3):316–321, March 2008.
- [93] Gilles Thomas, Kevin B Jacobs, Meredith Yeager, Peter Kraft, Sholom Wacholder, Nick Orr, Kai Yu, Nilanjan Chatterjee, Robert Welch, Amy Hutchinson, Andrew Crenshaw, Geraldine Cancel-Tassin, Brian J Staats, Zhaoming Wang, Jesus Gonzalez-Bosquet, Jun Fang, Xiang Deng, Sonja I Berndt, Eugenia E Calle, Heather Spencer Feigelson, Michael J Thun, Carmen Rodriguez, Demetrius Albanes, Jarmo Virtamo, Stephanie Weinstein, Fredrick R Schumacher, Edward Giovannucci, Walter C Willett, Olivier Cussenot, Antoine Valeri, Gerald L Andriole, E David Crawford, Margaret Tucker, Daniela S Gerhard, Joseph F Fraumeni, Jr, Robert Hoover, Richard B Hayes, David J Hunter, and Stephen J Chanock. Multiple loci identified in a genome-wide association study of prostate cancer. *Nat. Genet.*, 40(3):310–315, March 2008.
- [94] P A Abrahamsson, H Lilja, S Falkmer, and L B Wadström. Immunohistochemical distribution of the three predominant secretory proteins in the parenchyma of hyperplastic and neoplastic prostate glands. *Prostate*, 12(1):39–46, 1988.
- [95] Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv [stat.ML]*, February 2018.
- [96] Jianfeng Xu, Ethan M Lange, Lingyi Lu, Siqun L Zheng, Zhong Wang, Stephen N Thibodeau, Lisa A Cannon-Albright, Craig C Teerlink, Nicola J Camp, Anna M Johnson, Kimberly A Zuhlke, Janet L Stanford, Elaine A Ostrander, Kathleen E Wiley, Sarah D Isaacs, Patrick C Walsh, Christiane Maier, Manuel Luedeke, Walther Vogel, Johanna Schleutker, Tiina Wahlfors, Teuvo Tammela, Daniel Schaid, Shannon K McDonnell, Melissa S DeRycke, Geraldine Cancel-Tassin, Olivier Cussenot, Fredrik Wiklund, Henrik Grönberg, Ros Eeles, Doug Easton, Zsofia Kote-Jarai, Alice S Whittemore, Chih-Lin Hsieh, Graham G Giles, John L Hopper, Gianluca Severi, William J Catalona, Diptasri Mandal, Elisa Ledet, William D Foulkes, Nancy Hamel, Lovise Mahle, Pal Moller, Isaac Powell, Joan E Bailey-Wilson, John D Carpten, Daniela Seminara, Kathleen A Cooney, William B Isaacs, and International Consortium for Prostate Cancer Genetics. HOXB13 is a susceptibility gene for prostate cancer: results from the international consortium for prostate cancer genetics (ICPCG). *Hum. Genet.*, 132(1):5–14, January 2013.

- [97] P Adamo and M R Lodomery. The oncogene ERG: a key factor in prostate cancer. *Oncogene*, 35(4):403–414, January 2016.
- [98] Manuel Luedeke, Antje E Rinckleb, Liesel M FitzGerald, Milan S Geybels, Johanna Schleutker, Rosalind A Eeles, Manuel R Teixeira, Lisa Cannon-Albright, Elaine A Ostrander, Steffen Weikert, Kathleen Herkommer, Tiina Wahlfors, Tapio Visakorpi, Katri A Leinonen, Teuvo L J Tammela, Colin S Cooper, Zsafia Kote-Jarai, Sandra Edwards, Chee L Goh, Frank McCarthy, Chris Parker, Penny Flohr, Paula Paulo, Carmen Jerónimo, Rui Henrique, Hans Krause, Sven Wach, Verena Lieb, Tilman T Rau, Walther Vogel, Rainer Kuefer, Matthias D Hofer, Sven Perner, Mark A Rubin, Archana M Agarwal, Doug F Easton, Ali Amin Al Olama, Sara Benlloch, PRACTICAL consortium, Josef Hoegel, Janet L Stanford, and Christiane Maier. Prostate cancer risk regions at 8q24 and 17q24 are differentially associated with somatic TMPRSS2:ERG fusion status. *Hum. Mol. Genet.*, 25(24):5490–5499, December 2016.
- [99] L Beke, M Nuytten, A Van Eynde, M Beullens, and M Bollen. The gene encoding the prostatic tumor suppressor PSP94 is a target for repression by the polycomb group protein EZH2. *Oncogene*, 26(31):4590–4595, July 2007.
- [100] Bao-Li Chang, Scott D Cramer, Fredrik Wiklund, Sarah D Isaacs, Victoria L Stevens, Jielin Sun, Shelly Smith, Kristen Pruett, Lina M Romero, Kathleen E Wiley, Seong-Tae Kim, Yi Zhu, Zheng Zhang, Fang-Chi Hsu, Aubrey R Turner, Jan Adolfsson, Wennuan Liu, Jin Woo Kim, David Duggan, John Carpten, S Lilly Zheng, Carmen Rodriguez, William B Isaacs, Henrik Grönberg, and Jianfeng Xu. Fine mapping association study and functional analysis implicate a SNP in MSMB at 10q11 as a causal variant for prostate cancer risk. *Hum. Mol. Genet.*, 18(7):1368–1375, April 2009.
- [101] Z Kote-Jarai, D Leongamornlert, M Tymrakiewicz, H Field, M Guy, A A Al Olama, J Morrison, L O’Brien, R Wilkinson, A Hall, E Sawyer, K Muir, F Hamdy, J Donovan, D Neal, D Easton, and R Eeles. Mutation analysis of the MSMB gene in familial prostate cancer. *Br. J. Cancer*, 102(2):414–418, January 2010.
- [102] G W Beadle and E L Tatum. Genetic control of biochemical reactions in neurospora. *Proc. Natl. Acad. Sci. U. S. A.*, 27(11):499–506, November 1941.
- [103] Jacob B Hjelmborg, Thomas Scheike, Klaus Holst, Axel Skytthe, Kathryn L Penney, Rebecca E Graff, Eero Pukkala, Kaare Christensen, Hans-Olov Adami, Niels V Holm, Elizabeth Nuttall, Steinbjorn Hansen, Mikael Hartman, Kamila Czene, Jennifer R Harris, Jaakko Kaprio, and Lorelei A Mucci. The heritability of prostate cancer in the nordic twin study of cancer. *Cancer Epidemiol. Biomarkers Prev.*, 23(11):2303–2310, November 2014.
- [104] Lorelei A Mucci, Jacob B Hjelmborg, Jennifer R Harris, Kamila Czene, David J Havelick, Thomas Scheike, Rebecca E Graff, Klaus Holst, Sören Möller, Robert H Unger, Christina McIntosh, Elizabeth Nuttall, Ingunn Brandt, Kathryn L Penney, Mikael Hartman, Peter Kraft, Giovanni Parmigiani, Kaare Christensen, Markku Koskenvuo, Niels V Holm, Kauko Heikkilä, Eero Pukkala, Axel Skytthe, Hans-Olov Adami, Jaakko Kaprio, and Nordic Twin

- Study of Cancer (NorTwinCan) Collaboration. Familial risk and heritability of cancer among twins in nordic countries. *JAMA*, 315(1):68–76, January 2016.
- [105] Cathryn M Lewis and Evangelos Vassos. Polygenic risk scores: from research tools to clinical instruments. *Genome Med.*, 12(1):44, May 2020.
 - [106] Z Ali, N Yousaf, and J Larkin. Melanoma epidemiology, biology and prognosis. *EJC Suppl*, 11(2):81–91, September 2013.
 - [107] Natalie H Matthews, Wen-Qing Li, Abrar A Qureshi, Martin A Weinstock, and Eunyoung Cho. Epidemiology of melanoma. In William H Ward and Jeffrey M Farma, editors, *Cutaneous Melanoma: Etiology and Therapy*. Codon Publications, Brisbane (AU), February 2018.
 - [108] B A Gilchrest, M S Eller, A C Geller, and M Yaar. The pathogenesis of melanoma induced by ultraviolet radiation. *N. Engl. J. Med.*, 340(17):1341–1348, April 1999.
 - [109] C Suzanne Lea, Joseph A Scotto, Patricia A Buffler, Judith Fine, Raymond L Barnhill, and Marianne Berwick. Ambient UVB and melanoma risk in the united states: a case-control analysis. *Ann. Epidemiol.*, 17(6):447–453, June 2007.
 - [110] Lilit Garibyan and David E Fisher. How sunlight causes melanoma. *Curr. Oncol. Rep.*, 12(5):319–326, September 2010.
 - [111] Jennifer Y Lin and David E Fisher. Melanocyte biology and skin pigmentation. *Nature*, 445(7130):843–850, February 2007.
 - [112] Jonathan L Rees. Genetics of hair and skin color. *Annu. Rev. Genet.*, 37:67–90, 2003.
 - [113] Catherine M Olsen, Heidi J Carroll, and David C Whiteman. Estimating the attributable fraction for melanoma: a meta-analysis of pigmentary characteristics and freckling. *Int. J. Cancer*, 127(10):2430–2445, November 2010.
 - [114] Yu-Mei Chang, Julia A Newton-Bishop, D Timothy Bishop, Bruce K Armstrong, Veronique Bataille, Wilma Bergman, Marianne Berwick, Paige M Bracci, J Mark Elwood, Marc S Ernstoff, Adèle C Green, Nelleke A Gruis, Elizabeth A Holly, Christian Ingvar, Peter A Kanetsky, Margaret R Karagas, Loïc Le Marchand, Rona M Mackie, Håkan Olsson, Anne Østerlind, Timothy R Rebbeck, Kristian Reich, Peter Sasieni, Victor Siskind, Anthony J Swerdlow, Linda Titus-Ernstoff, Michael S Zens, Andreas Ziegler, and Jennifer H Barrett. A pooled analysis of melanocytic nevus phenotype and the risk of cutaneous melanoma at different latitudes. *Int. J. Cancer*, 124(2):420–428, January 2009.
 - [115] Catherine M Olsen, Heidi J Carroll, and David C Whiteman. Estimating the attributable fraction for cancer: A meta-analysis of nevi and melanoma. *Cancer Prev. Res.*, 3(2): 233–245, February 2010.

- [116] L Schuchter, D J Schultz, M Synnestvedt, B J Trock, D Guerry, D E Elder, R Elenitsas, W H Clark, and A C Halpern. A prognostic model for predicting 10-year survival in patients with primary melanoma. the pigmented lesion group. *Ann. Intern. Med.*, 125(5): 369–375, September 1996.
- [117] Lisa H Williams, Andrew R Shors, William E Barlow, Cam Solomon, and Emily White. Identifying persons at highest risk of melanoma using self-assessed risk factors. *J. Clin. Exp. Dermatol. Res.*, 2(6), 2011.
- [118] Jiali Han, Abrar A Qureshi, Jennifer Prescott, Qun Guo, Li Ye, David J Hunter, and Immaculata De Vivo. A prospective study of telomere length and the risk of skin cancer. *J. Invest. Dermatol.*, 129(2):415–421, February 2009.
- [119] Veronique Bataille, Bernet S Kato, Mario Falchi, Jeffrey Gardner, Masayuki Kimura, Marko Lens, Ursula Perks, Ana M Valdes, Dot C Bennett, Abraham Aviv, and Tim D Spector. Nevus size and number are associated with telomere length and represent potential markers of a decreased senescence in vivo. *Cancer Epidemiol. Biomarkers Prev.*, 16(7):1499–1502, July 2007.
- [120] Adèle C Green and Catherine M Olsen. Increased risk of melanoma in organ transplant recipients: systematic review and meta-analysis of cohort studies. *Acta Derm. Venereol.*, 95(8):923–927, November 2015.
- [121] M H Greene, T I Young, and W H Clark, Jr. Malignant melanoma in renal-transplant recipients. *Lancet*, 1(8231):1196–1199, May 1981.
- [122] Catherine M Olsen, Heidi J Carroll, and David C Whiteman. Familial melanoma: a meta-analysis and estimates of attributable fraction. *Cancer Epidemiol. Biomarkers Prev.*, 19(1):65–73, January 2010.
- [123] D Ford, J M Bliss, A J Swerdlow, B K Armstrong, S Franceschi, A Green, E A Holly, T Mack, R M MacKie, and A Osterlind. Risk of cutaneous melanoma associated with a family history of the disease. the international melanoma analysis group (IMAGE). *Int. J. Cancer*, 62(4):377–381, August 1995.
- [124] Christopher Greenman, Philip Stephens, Raffaella Smith, Gillian L Dalglish, Christopher Hunter, Graham Bignell, Helen Davies, Jon Teague, Adam Butler, Claire Stevens, Sarah Edkins, Sarah O’Meara, Imre Vastrik, Esther E Schmidt, Tim Avis, Syd Barthorpe, Gurpreet Bhamra, Gemma Buck, Bhudipa Choudhury, Jody Clements, Jennifer Cole, Ed Dicks, Simon Forbes, Kris Gray, Kelly Halliday, Rachel Harrison, Katy Hills, Jon Hinton, Andy Jenkinson, David Jones, Andy Menzies, Tatiana Mironenko, Janet Perry, Keiran Raine, Dave Richardson, Rebecca Shepherd, Alexandra Small, Calli Tofts, Jennifer Varian, Tony Webb, Sofie West, Sara Widaa, Andy Yates, Daniel P Cahill, David N Louis, Peter Goldstraw, Andrew G Nicholson, Francis Brasseur, Leendert Looijenga, Barbara L Weber, Yoke-Eng Chiew, Anna deFazio, Mel F Greaves, Anthony R Green, Peter Campbell, Ewan Birney, Douglas F Easton, Georgia Chenevix-Trench, Min-Han Tan, Sok Kean Khoo,

- Bin Tean Teh, Siu Tsan Yuen, Suet Yi Leung, Richard Wooster, P Andrew Futreal, and Michael R Stratton. Patterns of somatic mutation in human cancer genomes. *Nature*, 446(7132):153–158, March 2007.
- [125] Ludmil B Alexandrov, Serena Nik-Zainal, David C Wedge, Samuel A J Aparicio, Sam Behjati, Andrew V Biankin, Graham R Bignell, Niccolò Bolli, Ake Borg, Anne-Lise Børresen-Dale, Sandrine Boyault, Birgit Burkhardt, Adam P Butler, Carlos Caldas, Helen R Davies, Christine Desmedt, Roland Eils, Jórunn Erla Eyfjörð, John A Foekens, Mel Greaves, Fumie Hosoda, Barbara Hutter, Tomislav Ilicic, Sandrine Imbeaud, Marcin Imielinski, Natalie Jäger, David T W Jones, David Jones, Stian Knappskog, Marcel Kool, Sunil R Lakhani, Carlos López-Otín, Sancha Martin, Nikhil C Munshi, Hiromi Nakamura, Paul A Northcott, Marina Pajic, Elli Papaemmanuil, Angelo Paradiso, John V Pearson, Xose S Puente, Keiran Raine, Manasa Ramakrishna, Andrea L Richardson, Julia Richter, Philip Rosenstiel, Matthias Schlesner, Ton N Schumacher, Paul N Span, Jon W Teague, Yasushi Totoki, Andrew N J Tutt, Rafael Valdés-Mas, Marit M van Buuren, Laura van 't Veer, Anne Vincent-Salomon, Nicola Waddell, Lucy R Yates, Jessica Zucman-Rossi, P Andrew Futreal, Ultan McDermott, Peter Lichter, Matthew Meyerson, Sean M Grimmond, Reiner Siebert, Elías Campo, Tatsuhiro Shibata, Stefan M Pfister, Peter J Campbell, and Michael R Stratton. Signatures of mutational processes in human cancer. *Nature*, 500(7463):415–421, August 2013.
- [126] C G Clemente, M C Mihm, Jr, R Bufalino, S Zurrida, P Collini, and N Cascinelli. Prognostic value of tumor infiltrating lymphocytes in the vertical growth phase of primary cutaneous melanoma. *Cancer*, 77(7):1303–1310, April 1996.
- [127] D Morton, F R Eilber, R A Malmgren, and W C Wood. Immunological factors which influence response to immunotherapy in malignant melanoma. *Surgery*, 68(1):158–63; discussion 163–4, July 1970.
- [128] Valeria Orrù, Maristella Steri, Gabriella Sole, Carlo Sidore, Francesca Viridis, Mariano Dei, Sandra Lai, Magdalena Zoledziewska, Fabio Busonero, Antonella Mulas, Matteo Floris, Wiesława I Mentzen, Silvana A M Urru, Stefania Olla, Michele Marongiu, Maria G Piras, Monia Lobina, Andrea Maschio, Maristella Pitzalis, Maria F Urru, Marco Marcelli, Roberto Cusano, Francesca Deidda, Valentina Serra, Manuela Oppo, Rosella Pilu, Frederic Reinier, Riccardo Berutti, Luca Pireddu, Ilenia Zara, Eleonora Porcu, Alan Kwong, Christine Brennan, Brendan Tarrier, Robert Lyons, Hyun M Kang, Sergio Uzzau, Rossano Atzeni, Maria Valentini, Davide Firinu, Lidia Leoni, Gianluca Rotta, Silvia Naitza, Andrea Angius, Mauro Congia, Michael B Whalen, Chris M Jones, David Schlessinger, Gonçalo R Abecasis, Edoardo Fiorillo, Serena Sanna, and Francesco Cucca. Genetic variants regulating immune cell levels in health and disease. *Cell*, 155(1):242–256, September 2013.
- [129] Massimo Mangino, Mario Roederer, Margaret H Beddall, Frank O Nestle, and Tim D Spector. Innate and adaptive immune traits are differentially affected by genetic and environmental factors. *Nat. Commun.*, 8(1):1–7, January 2017.
- [130] Andrew R Marderstein, Manik Uppal, Akanksha Verma, Bhavneet Bhinder, Zakieh Tayyebi,

- Jason Mezey, Andrew G Clark, and Olivier Elemento. Demographic and genetic factors influence the abundance of infiltrating immune cells in human tissues. *Nat. Commun.*, 11(1):1–14, May 2020.
- [131] Ying Jin, Stanca A Birlea, Pamela R Fain, Tracey M Ferrara, Songtao Ben, Sheri L Riccardi, Joanne B Cole, Katherine Gowan, Paulene J Holland, Dorothy C Bennett, Rosalie M Luiten, Albert Wolkerstorfer, J P Wietze van der Veen, Anke Hartmann, Saskia Eichner, Gerold Schuler, Nanja van Geel, Jo Lambert, E Helen Kemp, David J Gawkrödger, Anthony P Weetman, Alain Taïeb, Thomas Jouary, Khaled Ezzedine, Margaret R Wallace, Wayne T McCormack, Mauro Picardo, Giovanni Leone, Andreas Overbeck, Nanette B Silverberg, and Richard A Spritz. Genome-wide association analyses identify 13 new susceptibility loci for generalized vitiligo. *Nat. Genet.*, 44(6):676–680, May 2012.
- [132] Ying Jin, Genevieve Andersen, Daniel Yorgov, Tracey M Ferrara, Songtao Ben, Kelly M Brownson, Paulene J Holland, Stanca A Birlea, Janet Siebert, Anke Hartmann, Anne Lienert, Nanja van Geel, Jo Lambert, Rosalie M Luiten, Albert Wolkerstorfer, J P Wietze van der Veen, Dorothy C Bennett, Alain Taïeb, Khaled Ezzedine, E Helen Kemp, David J Gawkrödger, Anthony P Weetman, Sulev Kõks, Ele Prans, Külli Kingo, Maire Karelson, Margaret R Wallace, Wayne T McCormack, Andreas Overbeck, Silvia Moretti, Roberta Colucci, Mauro Picardo, Nanette B Silverberg, Mats Olsson, Yan Valle, Igor Korobko, Markus Böhm, Henry W Lim, Iltefat Hamzavi, Li Zhou, Qing-Sheng Mi, Pamela R Fain, Stephanie A Santorico, and Richard A Spritz. Genome-wide association studies of autoimmune vitiligo identify 23 new risk loci and highlight key pathways and regulatory variants. *Nat. Genet.*, 48(11):1418–1424, November 2016.
- [133] Maria Teresa Landi, D Timothy Bishop, Stuart MacGregor, Mitchell J Machiela, Alexander J Stratigos, Paola Ghiorzo, Myriam Brossard, Donato Calista, Jiyeon Choi, Maria Concetta Fargnoli, Tongwu Zhang, Monica Rodolfo, Adam J Trower, Chiara Menin, Jacobo Martinez, Andreas Hadjisavvas, Lei Song, Irene Stefanaki, Richard Scolyer, Rose Yang, Alisa M Goldstein, Miriam Potrony, Katerina P Kypreou, Lorenza Pastorino, Paola Queirolo, Cristina Pellegrini, Laura Cattaneo, Matthew Zawistowski, Pol Gimenez-Xavier, Arantxa Rodriguez, Lisa Elefanti, Siranoush Manoukian, Licia Rivoltini, Blair H Smith, Maria A Loizidou, Laura Del Regno, Daniela Massi, Mario Mandala, Kiarash Khosrotehrani, Lars A Akslen, Christopher I Amos, Per A Andresen, Marie-Françoise Avril, Esther Azizi, H Peter Soyer, Veronique Bataille, Bruna Dalmasso, Lisa M Bowdler, Kathryn P Burdon, Wei V Chen, Veryan Codd, Jamie E Craig, Tadeusz Debniak, Mario Falchi, Shenying Fang, Eitan Friedman, Sarah Simi, Pilar Galan, Zaida Garcia-Casado, Elizabeth M Gillanders, Scott Gordon, Adele Green, Nelleke A Gruis, Johan Hansson, Mark Harland, Jessica Harris, Per Helsing, Anjali Henders, Marko Hočevár, Veronica Höiom, David Hunter, Christian Ingvar, Rajiv Kumar, Julie Lang, G Mark Lathrop, Jeffrey E Lee, Xin Li, Jan Lubiński, Rona M Mackie, Maryrose Malt, Josep Malvehy, Kerrie McAloney, Hamida Mohamdi, Anders Molven, Eric K Moses, Rachel E Neale, Srdjan Novaković, Dale R Nyholt, Håkan Olsson, Nicholas Orr, Lars G Fritsche, Joan Anton Puig-Butille, Abrar A Qureshi, Graham L Radford-Smith, Juliette Randerson-Moor, Celia Requena, Casey Rowe, Nilesh J Samani, Marianna Sanna, Dirk Schadendorf, Hans-Joachim Schulze, Lisa A Simms, Mark Smithers, Fengju Song, Anthony J Swerdlow, Nienke van der Stoep, Nicole A Kukutsch,

- Alessia Visconti, Leanne Wallace, Sarah V Ward, Lawrie Wheeler, Richard A Sturm, Amy Hutchinson, Kristine Jones, Michael Malasky, Aurelie Vogt, Weiyin Zhou, Karen A Pooley, David E Elder, Jiali Han, Belynda Hicks, Nicholas K Hayward, Peter A Kanetsky, Chad Brummett, Grant W Montgomery, Catherine M Olsen, Caroline Hayward, Alison M Dunning, Nicholas G Martin, Evangelos Evangelou, Graham J Mann, Georgina Long, Paul D P Pharoah, Douglas F Easton, Jennifer H Barrett, Anne E Cust, Goncalo Abecasis, David L Duffy, David C Whiteman, Helen Gogas, Arcangela De Nicolo, Margaret A Tucker, Julia A Newton-Bishop, GenoMEL Consortium, Q-MEGA and QTWIN Investigators, ATHENS Melanoma Study Group, 23andMe, SDH Study Group, IBD Investigators, Essen-Heidelberg Investigators, AMFS Investigators, MelaNostrum Consortium, Ketty Peris, Stephen J Chanock, Florence Dumenais, Kevin M Brown, Susana Puig, Eduardo Nagore, Jianxin Shi, Mark M Iles, and Matthew H Law. Genome-wide association meta-analyses combining multiple risk phenotypes provide insights into the genetic architecture of cutaneous melanoma susceptibility. *Nat. Genet.*, 52(5):494–504, May 2020.
- [134] Alexandre Corthay. Does the immune system naturally protect against cancer? *Front. Immunol.*, 5:197, May 2014.
- [135] Michele W L Teng, Jeremy B Swann, Catherine M Koebel, Robert D Schreiber, and Mark J Smyth. Immune-mediated dormancy: an equilibrium with cancer. *J. Leukoc. Biol.*, 84(4): 988–993, October 2008.
- [136] Peter K Gregersen and Timothy W Behrens. Genetics of autoimmune diseases—disorders of immune homeostasis. *Nat. Rev. Genet.*, 7(12):917–928, December 2006.
- [137] N K Nanda and E E Sercarz. Induction of anti-self-immunity to cure cancer. *Cell*, 82(1): 13–17, July 1995.
- [138] D M Pardoll. Inducing autoimmune disease to treat cancer. *Proc. Natl. Acad. Sci. U. S. A.*, 96(10):5340–5342, May 1999.
- [139] Richard A Spritz. The genetics of generalized vitiligo: autoimmune pathways and an inverse relationship with malignant melanoma. *Genome Med.*, 2(10):78, October 2010.
- [140] Arabella Young, Zoe Quandt, and Jeffrey A Bluestone. The balancing act between cancer immunity and autoimmunity in response to immunotherapy. *Cancer Immunol Res*, 6(12): 1445–1452, December 2018.
- [141] Vylyny Chat, Robert Ferguson, Danny Simpson, Esther Kazlow, Rebecca Lax, Una Moran, Anna Pavlick, Dennie Frederick, Genevieve Boland, Ryan Sullivan, Antoni Ribas, Keith Flaherty, Iman Osman, Jeffrey Weber, and Tomas Kirchhoff. Autoimmune genetic risk variants as germline biomarkers of response to melanoma immune-checkpoint inhibition. *Cancer Immunol. Immunother.*, 68(6):897–905, June 2019.
- [142] Zia Khan, Flavia Di Nucci, Antonia Kwan, Christian Hammer, Sanjeev Mariathasan, Vincent Rouilly, Jonathan Carroll, Magnus Fontes, Sergio Ley Acosta, Ellie Guardino,

- Haiyin Chen-Harris, Tushar Bhangale, Ira Mellman, Jonathan Rosenberg, Thomas Powles, Julie Hunkapiller, G Scott Chandler, and Matthew L Albert. Polygenic risk for skin autoimmunity impacts immune checkpoint blockade in bladder cancer. *Proc. Natl. Acad. Sci. U. S. A.*, 117(22):12288–12294, June 2020.
- [143] J Askling, C M Fored, L Brandt, E Baecklund, L Bertilsson, N Feltelius, L Cöster, P Geborek, L T Jacobsson, S Lindblad, J Lysholm, S Rantapää-Dahlqvist, T Saxne, and L Klareskog. Risks of solid cancers in patients with rheumatoid arthritis and after treatment with tumour necrosis factor antagonists. *Ann. Rheum. Dis.*, 64(10):1421–1426, October 2005.
 - [144] Lebin Song, Yi Wang, Jiayi Zhang, Ninghong Song, Xiaoyun Xu, and Yan Lu. The risks of cancer development in systemic lupus erythematosus (SLE) patients: a systematic review and meta-analysis. *Arthritis Res. Ther.*, 20(1):270, December 2018.
 - [145] H E Teulings, M Overkamp, E Ceylan, L Nieuweboer-Krobotova, J D Bos, T Nijsten, A W Wolkerstorfer, R M Luiten, and J P W van der Veen. Decreased risk of melanoma and nonmelanoma skin cancer in patients with vitiligo: a survey among 1307 patients and their partners. *Br. J. Dermatol.*, 168(1):162–171, January 2013.
 - [146] Kenneth L Rock, Eric Reits, and Jacques Neefjes. Present yourself! by MHC class I and MHC class II molecules. *Trends Immunol.*, 37(11):724–737, November 2016.
 - [147] Aurélie Durgeau, Yasemin Virk, Stéphanie Corgnac, and Fathia Mami-Chouaib. Recent advances in targeting CD8 T-cell immunity for more effective cancer immunotherapy. *Front. Immunol.*, 9:14, January 2018.
 - [148] Daniel J Woodsworth, Valentin Dunsing, and Daniel Coombs. Design parameters for granzyme-mediated cytotoxic lymphocyte target-cell killing and specificity. *Biophys. J.*, 109(3):477–488, August 2015.
 - [149] Rachel Marty, Saghar Kaabinejadian, David Rossell, Michael J Slifker, Joris van de Haar, Hatice Billur Engin, Nicola de Prisco, Trey Ideker, William H Hildebrand, Joan Font-Burgada, and Hannah Carter. MHC-I genotype restricts the oncogenic mutational landscape. *Cell*, 171(6):1272–1283.e15, November 2017.
 - [150] Rachel Marty Pyke, Wesley Kurt Thompson, Rany M Salem, Joan Font-Burgada, Maurizio Zanetti, and Hannah Carter. Evolutionary pressure against MHC class II binding cancer mutations. *Cell*, 175(2):416–428.e13, October 2018.
 - [151] Fan Yang, Dae-Kyum Kim, Hidewaki Nakagawa, Shuto Hayashi, Seiya Imoto, Lincoln Stein, and Frederick P Roth. Quantifying immune-based counterselection of somatic mutations. *PLoS Genet.*, 15(7):e1008227, July 2019.
 - [152] Diego Chowell, Luc G T Morris, Claud M Grigg, Jeffrey K Weber, Robert M Samstein, Vladimir Makarov, Fengshen Kuo, Sviatoslav M Kendall, David Requena, Nadeem Riaz,

- Benjamin Greenbaum, James Carroll, Edward Garon, David M Hyman, Ahmet Zehir, David Solit, Michael Berger, Ruhong Zhou, Naiyer A Rizvi, and Timothy A Chan. Patient HLA class I genotype influences cancer response to checkpoint blockade immunotherapy. *Science*, 359(6375):582–587, February 2018.
- [153] Diego Chowell, Chirag Krishna, Federica Pierini, Vladimir Makarov, Naiyer A Rizvi, Fengshen Kuo, Luc G T Morris, Nadeem Riaz, Tobias L Lenz, and Timothy A Chan. Evolutionary divergence of HLA class I genotype impacts efficacy of cancer immunotherapy. *Nat. Med.*, 25(11):1715–1720, November 2019.
- [154] 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. MHC-I genotype and tumor mutational burden predict response to immunotherapy. *Genome Med.*, 12(1):45, May 2020.
- [155] Nicholas McGranahan, Rachel Rosenthal, Crispin T Hiley, Andrew J Rowan, Thomas B K Watkins, Gareth A Wilson, Nicolai J Birkbak, Selvaraju Veeriah, Peter Van Loo, Javier Herrero, Charles Swanton, and TRACERx Consortium. Allele-specific HLA loss and immune escape in lung cancer evolution. *Cell*, 171(6):1259–1271.e11, November 2017.
- [156] Andrea Castro, Kivilcim Ozturk, Rachel Marty Pyke, Su Xian, Maurizio Zanetti, and Hannah Carter. Elevated neoantigen levels in tumors with somatic mutations in the HLA-a, HLA-B, HLA-C and B2M genes. *BMC Med. Genomics*, 12(Suppl 6):107, July 2019.
- [157] Jasper G van den Boorn, Debby Konijnenberg, Trees A M DelleMijn, J P Wietze van der Veen, Jan D Bos, Cornelis J M Melief, Florry A Vyth-Dreese, and Rosalie M Luiten. Autoimmune destruction of skin melanocytes by perilesional T cells from vitiligo patients. *J. Invest. Dermatol.*, 129(9):2220–2232, September 2009.
- [158] Wolf-Henning Boehncke and Nicolo Costantino Brembilla. Autoreactive T-lymphocytes in inflammatory skin diseases. *Front. Immunol.*, 10:1198, May 2019.
- [159] Curdin Conrad, Onur Boyman, Giulia Tonel, Adrian Tun-Kyi, Ute Laggner, Antonin de Fougères, Victor Kotlianski, Humphrey Gardner, and Frank O Nestle. $\alpha 1 \beta 1$ integrin is crucial for accumulation of epidermal T cells and the development of psoriasis. *Nat. Med.*, 13(7):836–842, July 2007.
- [160] Yihua Cai, Chris Fleming, and Jun Yan. New insights of T cells in the pathogenesis of psoriasis. *Cell. Mol. Immunol.*, 9(4):302–309, July 2012.
- [161] Jörg Christoph Prinz. Human leukocyte antigen-class I alleles and the autoreactive T cell response in psoriasis pathogenesis. *Front. Immunol.*, 9:954, April 2018.
- [162] Akiko Arakawa, Katherina Siewert, Julia Stöhr, Petra Besgen, Song-Min Kim, Geraldine Rühl, Jens Nickel, Sigrid Vollmer, Peter Thomas, Stefan Krebs, Stefan Pinkert, Michael Spannagl, Kathrin Held, Claudia Kammerbauer, Robert Besch, Klaus Dornmair, and

- Jörg C Prinz. Melanocyte antigen triggers autoimmunity in human psoriasis. *J. Exp. Med.*, 212(13):2203–2212, December 2015.
- [163] Kun-Ju Zhu, Yong-Mei Lv, Xian-Yong Yin, Zai-Xing Wang, Liang-Dan Sun, Su-Min He, Hui Cheng, Da-Yan Hu, Zheng Zhang, Yang Li, Xian-Bo Zuo, You-Wen Zhou, Sen Yang, Xing Fan, Xue-Jun Zhang, and Feng-Yu Zhang. Psoriasis regression analysis of MHC loci identifies shared genetic variants with vitiligo. *PLoS One*, 6(11):e23089, November 2011.
 - [164] Khalifa E Sharquie, Husam Ali Salman, and Aseel K Yaseen. Psoriasis and vitiligo are close relatives. *Clin. Cosmet. Investig. Dermatol.*, 10:341–345, August 2017.
 - [165] Sanket Vashist, Vikram K Mahajan, Karaninder S Mehta, Pushpinder S Chauhan, Rajinder S Yadav, Satya Bhushan Sharma, Vikas Sharma, Anuj Sharma, Bhumika Chowdhary, and Prabal Kumar. Association of psoriasis with autoimmune disorders: Results of a pilot study. *Indian Dermatol. Online J.*, 11(5):753–759, September 2020.
 - [166] S A Rosenberg and D E White. Vitiligo in patients with melanoma: normal tissue antigens can be targets for cancer immunotherapy. *J. Immunother. Emphasis Tumor Immunol.*, 19(1):81–84, January 1996.
 - [167] Camille Hua, Lise Boussemart, Christine Mateus, Emilie Routier, Céline Boutros, Hugo Cazenave, Roxane Viollet, Marina Thomas, Séverine Roy, Naima Benannoune, Gorana Tomasic, Jean-Charles Soria, Stéphane Champiat, Matthieu Texier, Emilie Lanoy, and Caroline Robert. Association of vitiligo with tumor response in patients with metastatic melanoma treated with pembrolizumab. *JAMA Dermatol.*, 152(1):45–51, January 2016.
 - [168] Yasuhiro Nakamura, Ryota Tanaka, Yuri Asami, Yukiko Teramoto, Taichi Imamura, Sayuri Sato, Hiroshi Maruyama, Yasuhiro Fujisawa, Taisuke Matsuya, Manabu Fujimoto, and Akifumi Yamamoto. Correlation between vitiligo occurrence and clinical benefit in advanced melanoma patients treated with nivolumab: A multi-institutional retrospective study. *J. Dermatol.*, 44(2):117–122, February 2017.
 - [169] Alice Indini, Lorenza Di Guardo, Carolina Cimminiello, Michele Prisciandaro, Giovanni Randon, Filippo De Braud, and Michele Del Vecchio. Immune-related adverse events correlate with improved survival in patients undergoing anti-PD1 immunotherapy for metastatic melanoma. *J. Cancer Res. Clin. Oncol.*, 145(2):511–521, February 2019.
 - [170] Lars Hofmann, Andrea Forschner, Carmen Loquai, Simone M Goldinger, Lisa Zimmer, Selma Ugurel, Maria I Schmidgen, Ralf Gutzmer, Jochen S Utikal, Daniela Göppner, Jessica C Hassel, Friedegund Meier, Julia K Tietze, Ioannis Thomas, Carsten Weishaupt, Martin Leverkus, Renate Wahl, Ursula Dietrich, Claus Garbe, Michael C Kirchberger, Thomas Eigentler, Carola Berking, Anja Gesierich, Angela M Krackhardt, Dirk Schadendorf, Gerold Schuler, Reinhard Dummer, and Lucie M Heinzerling. Cutaneous, gastrointestinal, hepatic, endocrine, and renal side-effects of anti-PD-1 therapy. *Eur. J. Cancer*, 60:190–209, June 2016.

- [171] J Bonigen, C Raynaud-Donzel, J Hureauux, N Kramkimel, A Blom, G Jeudy, A-L Breton, T Hubiche, C Bedane, D Legoupil, A Pham-Ledard, J Charles, M Pérol, E Gérard, P Combemale, D Bonnet, M-L Sigal, E Mahé, and Groupe de Recherche sur le Psoriasis and the Groupe Cancérologie Cutanée of the Société Française de Dermatologie the GEM Resopso, Apsoderm and the Groupe Français de Pneumo-Cancérologie. Anti-PD1-induced psoriasis: a study of 21 patients. *J. Eur. Acad. Dermatol. Venereol.*, 31(5):e254–e257, May 2017.
- [172] Sachet A Shukla, Michael S Rooney, Mohini Rajasagi, Grace Tiao, Philip M Dixon, Michael S Lawrence, Jonathan Stevens, William J Lane, Jamie L Dellagatta, Scott Steelman, Carrie Sougnez, Kristian Cibulskis, Adam Kiezun, Nir Hacohen, Vladimir Brusic, Catherine J Wu, and Gad Getz. Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes. *Nat. Biotechnol.*, 33(11):1152–1158, November 2015.
- [173] Shuji Kawaguchi, Koichiro Higasa, Masakazu Shimizu, Ryo Yamada, and Fumihiko Matsuda. HLA-HD: An accurate HLA typing algorithm for next-generation sequencing data. *Hum. Mutat.*, 38(7):788–797, 2017.
- [174] Willy Hugo, Jesse M Zaretsky, Lu Sun, Chunying Song, Blanca Homet Moreno, Siwen Hu-Lieskovan, Beata Berent-Maoz, Jia Pang, Bartosz Chmielowski, Grace Cherry, Elizabeth Seja, Shirley Lomeli, Xiangju Kong, Mark C Kelley, Jeffrey A Sosman, Douglas B Johnson, Antoni Ribas, and Roger S Lo. Genomic and transcriptomic features of response to anti-PD-1 therapy in metastatic melanoma. *Cell*, 165(1):35–44, March 2016.
- [175] Eliezer M Van Allen, Diana Miao, Bastian Schilling, Sachet A Shukla, Christian Blank, Lisa Zimmer, Antje Sucker, Uwe Hillen, Marnix H Geukes Foppen, Simone M Goldinger, Jochen Utikal, Jessica C Hassel, Benjamin Weide, Katharina C Kaehler, Carmen Loquai, Peter Mohr, Ralf Gutzmer, Reinhard Dummer, Stacey Gabriel, Catherine J Wu, Dirk Schadendorf, and Levi A Garraway. Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. *Science*, 350(6257):207–211, October 2015.
- [176] Michael Krauthammer, Yong Kong, Antonella Bacchiocchi, Perry Evans, Natapol Pornputtapong, Cen Wu, James P McCusker, Shuangge Ma, Elaine Cheng, Robert Straub, Merdan Serin, Marcus Bosenberg, Stephan Ariyan, Deepak Narayan, Mario Sznol, Harriet M Kluger, Shrikant Mane, Joseph Schlessinger, Richard P Lifton, and Ruth Halaban. Exome sequencing identifies recurrent mutations in NF1 and RASopathy genes in sun-exposed melanomas. *Nat. Genet.*, 47(9):996–1002, September 2015.
- [177] Rossitza Lazova, Natapol Pornputtapong, Ruth Halaban, Marcus Bosenberg, Yalai Bai, Hao Chai, and Michael Krauthammer. Spitz nevi and spitzoid melanomas: exome sequencing and comparison with conventional melanocytic nevi and melanomas. *Mod. Pathol.*, 30(5): 640–649, May 2017.
- [178] A Hunter Shain, Iwei Yeh, Ivanka Kovalyshyn, Aravindhan Sriharan, Eric Talevich, Alexander Gagnon, Reinhard Dummer, Jeffrey North, Laura Pincus, Beth Ruben, William Rickaby, Corrado D’Arrigo, Alistair Robson, and Boris C Bastian. The genetic evolution

- of melanoma from precursor lesions. *N. Engl. J. Med.*, 373(20):1926–1936, November 2015.
- [179] Huaying Fang, Qin Hui, Julie Lynch, Jacqueline Honerlaw, Themistocles L Assimes, Jie Huang, Marijana Vujkovic, Scott M Damrauer, Saiju Pyarajan, J Michael Gaziano, Scott L DuVall, Christopher J O’Donnell, Kelly Cho, Kyong-Mi Chang, Peter W F Wilson, Philip S Tsao, VA Million Veteran Program, Yan V Sun, and Hua Tang. Harmonizing genetic ancestry and self-identified race/ethnicity in genome-wide association studies. *Am. J. Hum. Genet.*, 105(4):763–772, October 2019.
 - [180] Joshua C Denny, Lisa Bastarache, Marylyn D Ritchie, Robert J Carroll, Raquel Zink, Jonathan D Mosley, Julie R Field, Jill M Pulley, Andrea H Ramirez, Erica Bowton, Melissa A Basford, David S Carrell, Peggy L Peissig, Abel N Kho, Jennifer A Pacheco, Luke V Rasmussen, David R Crosslin, Paul K Crane, Jyotishman Pathak, Suzette J Bielinski, Sarah A Pendergrass, Hua Xu, Lucia A Hindorff, Rongling Li, Teri A Manolio, Christopher G Chute, Rex L Chisholm, Eric B Larson, Gail P Jarvik, Murray H Brilliant, Catherine A McCarty, Iftikhar J Kullo, Jonathan L Haines, Dana C Crawford, Daniel R Masys, and Dan M Roden. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat. Biotechnol.*, 31(12):1102–1111, December 2013.
 - [181] X Zheng, J Shen, C Cox, J C Wakefield, M G Ehm, M R Nelson, and B S Weir. HIBAG—HLA genotype imputation with attribute bagging. *Pharmacogenomics J.*, 14(2):192–200, May 2013.
 - [182] Frauke Degenhardt, Mareike Wendorff, Michael Wittig, Eva Ellinghaus, Lisa W Datta, John Schembri, Siew C Ng, Elisa Rosati, Matthias Hübenthal, David Ellinghaus, Eun Suk Jung, Wolfgang Lieb, Shifteh Abedian, Reza Malekzadeh, Jae Hee Cheon, Pierre Ellul, Ajit Sood, Vandana Midha, B K Thelma, Sunny H Wong, Stefan Schreiber, Keiko Yamazaki, Michiaki Kubo, Gabrielle Boucher, John D Rioux, Tobias L Lenz, Steven R Brant, and Andre Franke. Construction and benchmarking of a multi-ethnic reference panel for the imputation of HLA class I and II alleles. *Hum. Mol. Genet.*, 28(12):2078–2092, June 2019.
 - [183] Haley Hunter-Zinck, Yunling Shi, Man Li, Bryan R Gorman, Sun-Gou Ji, Ning Sun, Teresa Webster, Andrew Liem, Paul Hsieh, Poornima Devineni, Purushotham Karnam, Xin Gong, Lakshmi Radhakrishnan, Jeanette Schmidt, Themistocles L Assimes, Jie Huang, Cuiping Pan, Donald Humphries, Mary Brophy, Jennifer Moser, Sumitra Muralidhar, Grant D Huang, Ronald Przygodzki, John Concato, John M Gaziano, Joel Gelernter, Christopher J O’Donnell, Elizabeth R Hauser, Hongyu Zhao, Timothy J O’Leary, VA Million Veteran Program, Philip S Tsao, and Saiju Pyarajan. Genotyping array design and data quality control in the million veteran program. *Am. J. Hum. Genet.*, 106(4):535–548, April 2020.
 - [184] Aaron M Newman, Chloé B Steen, Chih Long Liu, Andrew J Gentles, Aadel A Chaudhuri, Florian Scherer, Michael S Khodadoust, Mohammad S Esfahani, Bogdan A Luca, David Steiner, Maximilian Diehn, and Ash A Alizadeh. Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat. Biotechnol.*, 37(7):773–782, May 2019.

- [185] Aaron M Newman, Chih Long Liu, Michael R Green, Andrew J Gentles, Weiguo Feng, Yue Xu, Chuong D Hoang, Maximilian Diehn, and Ash A Alizadeh. Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods*, 12(5):453–457, March 2015.
- [186] Rob Patro, Stephen M Mount, and Carl Kingsford. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nat. Biotechnol.*, 32(5):462–464, April 2014.
- [187] Daniel C Koboldt, Qunyuan Zhang, David E Larson, Dong Shen, Michael D McLellan, Ling Lin, Christopher A Miller, Elaine R Mardis, Li Ding, and Richard K Wilson. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.*, 22(3):568–576, March 2012.
- [188] Yu Fan, Liu Xi, Daniel S T Hughes, Jianjun Zhang, Jianhua Zhang, P Andrew Futreal, David A Wheeler, and Wenyi Wang. MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. *Genome Biol.*, 17(1):178, August 2016.
- [189] Kristian Cibulskis, Michael S Lawrence, Scott L Carter, Andrey Sivachenko, David Jaffe, Carrie Sougnez, Stacey Gabriel, Matthew Meyerson, Eric S Lander, and Gad Getz. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.*, 31(3):213–219, March 2013.
- [190] David E Larson, Christopher C Harris, Ken Chen, Daniel C Koboldt, Travis E Abbott, David J Dooling, Timothy J Ley, Elaine R Mardis, Richard K Wilson, and Li Ding. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics*, 28(3):311–317, February 2012.
- [191] Eran Hodis, Ian R Watson, Gregory V Kryukov, Stefan T Arold, Marcin Imielinski, Jean-Philippe Theurillat, Elizabeth Nickerson, Daniel Auclair, Liren Li, Chelsea Place, Daniel Dicara, Alex H Ramos, Michael S Lawrence, Kristian Cibulskis, Andrey Sivachenko, Douglas Voet, Gordon Saksena, Nicolas Stransky, Robert C Onofrio, Wendy Winckler, Kristin Ardlie, Nikhil Wagle, Jennifer Wargo, Kelly Chong, Donald L Morton, Katherine Stemke-Hale, Guo Chen, Michael Noble, Matthew Meyerson, John E Ladbury, Michael A Davies, Jeffrey E Gershenwald, Stephan N Wagner, Dave S B Hoon, Dirk Schadendorf, Eric S Lander, Stacey B Gabriel, Gad Getz, Levi A Garraway, and Lynda Chin. A landscape of driver mutations in melanoma. *Cell*, 150(2):251–263, July 2012.
- [192] Cancer Genome Atlas Network. Genomic classification of cutaneous melanoma. *Cell*, 161(7):1681–1696, June 2015.
- [193] Jake R Conway, Felix Dietlein, Amaro Taylor-Weiner, Saud AlDubayan, Natalie Vokes, Tanya Keenan, Brendan Reardon, Meng Xiao He, Claire A Margolis, Jason L Weirather, Rizwan Haq, Bastian Schilling, F Stephen Hodi, Dirk Schadendorf, David Liu, and Eliezer M Van Allen. Integrated molecular drivers coordinate biological and clinical states in melanoma. *Nat. Genet.*, 52(12):1373–1383, December 2020.

- [194] Hannah Carter, Sining Chen, Leyla Isik, Svitlana Tyekucheva, Victor E Velculescu, Kenneth W Kinzler, Bert Vogelstein, and Rachel Karchin. Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. *Cancer Res.*, 69(16):6660–6667, August 2009.
- [195] Collin Tokheim and Rachel Karchin. CHASMplus reveals the scope of somatic missense mutations driving human cancers. *Cell Syst*, 9(1):9–23.e8, July 2019.
- [196] Yingxin Lin, Shila Ghazanfar, Kevin Y X Wang, Johann A Gagnon-Bartsch, Kitty K Lo, Xianbin Su, Ze-Guang Han, John T Ormerod, Terence P Speed, Pengyi Yang, and Jean Yee Hwa Yang. scMerge leverages factor analysis, stable expression, and pseudoreplication to merge multiple single-cell RNA-seq datasets. *Proceedings of the National Academy of Sciences*, 116(20):9775–9784, May 2019.
- [197] Huaiyu Mi, Dustin Ebert, Anushya Muruganujan, Caitlin Mills, Laurent-Philippe Albou, Tremayne Mushayamaha, and Paul D Thomas. PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. *Nucleic Acids Res.*, 49(D1):D394–D403, January 2021.
- [198] Bijay Jassal, Lisa Matthews, Guilherme Viteri, Chuqiao Gong, Pascual Lorente, Antonio Fabregat, Konstantinos Sidiropoulos, Justin Cook, Marc Gillespie, Robin Haw, Fred Loney, Bruce May, Marija Milacic, Karen Rothfels, Cristoffer Sevilla, Veronica Shamovsky, Solomon Shorser, Thawfeek Varusai, Joel Weiser, Guanming Wu, Lincoln Stein, Henning Hermjakob, and Peter D’Eustachio. The reactome pathway knowledgebase. *Nucleic Acids Res.*, 48(D1):D498–D503, January 2020.
- [199] Vitor R C Aguiar, Cibele Masotti, Anamaria A Camargo, and Diogo Meyer. HLApers: HLA typing and quantification of expression with personalized index. In Sebastian Boegel, editor, *Bioinformatics for Cancer Immunotherapy: Methods and Protocols*, Methods in Molecular Biology, pages 101–112. Springer US, New York, NY, 2020.
- [200] Nicolas L Bray, Harold Pimentel, Páll Melsted, and Lior Pachter. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.*, 34(5):525–527, May 2016.
- [201] Adam Frankish, Mark Diekhans, Anne-Maud Ferreira, Rory Johnson, Irwin Jungreis, Jane Loveland, Jonathan M Mudge, Cristina Sisu, James Wright, Joel Armstrong, If Barnes, Andrew Berry, Alexandra Bignell, Silvia Carbonell Sala, Jacqueline Chrast, Fiona Cunningham, Tomás Di Domenico, Sarah Donaldson, Ian T Fiddes, Carlos García Girón, Jose Manuel Gonzalez, Tiago Grego, Matthew Hardy, Thibaut Hourlier, Toby Hunt, Osagie G Izuogu, Julien Lagarde, Fergal J Martin, Laura Martínez, Shamika Mohanan, Paul Muir, Fabio C P Navarro, Anne Parker, Baikang Pei, Fernando Pozo, Magali Ruffier, Bianca M Schmitt, Eloise Stapleton, Marie-Marthe Suner, Irina Sycheva, Barbara Uszczynska-Ratajczak, Jinuri Xu, Andrew Yates, Daniel Zerbino, Yan Zhang, Bronwen Aken, Jyoti S Choudhary, Mark Gerstein, Roderic Guigó, Tim J P Hubbard, Manolis Kellis, Benedict Paten, Alexandre Reymond, Michael L Tress, and Paul Flicek. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.*, 47(D1):D766–D773, January 2019.

- [202] James Robinson, Dominic J Barker, Xenia Georgiou, Michael A Cooper, Paul Flicek, and Steven G E Marsh. IPD-IMGT/HLA database. *Nucleic Acids Res.*, 48(D1):D948–D955, January 2020.
- [203] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, 15(12):550, December 2014.
- [204] Dvir Aran, Marina Sirota, and Atul J Butte. Systematic pan-cancer analysis of tumour purity. *Nat. Commun.*, 6(1):8971, December 2015.
- [205] Eva Pérez-Guijarro, Howard H Yang, Romina E Araya, Rajaa El Meskini, Helen T Michael, Suman Kumar Vodnala, Kerrie L Marie, Cari Smith, Sung Chin, Khiem C Lam, Andres Thorkelsson, Anthony J Iacovelli, Alan Kulaga, Anyen Fon, Aleksandra M Michalowski, Willy Hugo, Roger S Lo, Nicholas P Restifo, Shyam K Sharan, Terry Van Dyke, Romina S Goldszmid, Zoe Weaver Ohler, Maxwell P Lee, Chi-Ping Day, and Glenn Merlino. Multimodel preclinical platform predicts clinical response of melanoma to immunotherapy. *Nat. Med.*, 26(5):781–791, May 2020.
- [206] Peng Jiang, Shengqing Gu, Deng Pan, Jingxin Fu, Avinash Sahu, Xihao Hu, Ziyi Li, Nicole Traugh, Xia Bu, Bo Li, Jun Liu, Gordon J Freeman, Myles A Brown, Kai W Wucherpfennig, and X Shirley Liu. Signatures of T cell dysfunction and exclusion predict cancer immunotherapy response. *Nat. Med.*, 24(10):1550–1558, October 2018.
- [207] Birkir Reynisson, Bruno Alvarez, Sinu Paul, Bjoern Peters, and Morten 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.*, 48(W1):W449–W454, July 2020.
- [208] P Armitage and R Doll. The age distribution of cancer and a multi-stage theory of carcinogenesis. *Br. J. Cancer*, 8(1):1–12, March 1954.
- [209] S H Moolgavkar and A G Knudson, Jr. Mutation and cancer: a model for human carcinogenesis. *J. Natl. Cancer Inst.*, 66(6):1037–1052, June 1981.
- [210] S H Moolgavkar, A Dewanji, and D J Venzon. A stochastic two-stage model for cancer risk assessment. I. the hazard function and the probability of tumor. *Risk Anal.*, 8(3):383–392, September 1988.
- [211] S H Moolgavkar and G Luebeck. Two-event model for carcinogenesis: biological, mathematical, and statistical considerations. *Risk Anal.*, 10(2):323–341, June 1990.
- [212] Andrew F Brouwer, Rafael Meza, and Marisa C Eisenberg. Parameter estimation for multistage clonal expansion models from cancer incidence data: A practical identifiability analysis. *PLoS Comput. Biol.*, 13(3):e1005431, March 2017.
- [213] Rafael Meza, Jihyoun Jeon, Suresh H Moolgavkar, and E Georg Luebeck. Age-specific

- incidence of cancer: Phases, transitions, and biological implications. *Proc. Natl. Acad. Sci. U. S. A.*, 105(42):16284–16289, October 2008.
- [214] Wai-Yuan Tan. *Stochastic models of carcinogenesis*. CRC Press, 2020.
 - [215] SEER research data 1975-2018 when using SEER*Stat: Surveillance, epidemiology, and end results (SEER) program (www.seer.cancer.gov) SEER*Stat database: Incidence - SEER research data, 9 registries, nov 2020 sub (1975-2018) - linked to county attributes - time dependent (1990-2018) income/rurality, 1969-2019 counties, national cancer institute, DCCPS, surveillance research program, released april 2021, based on the november 2020 submission.
 - [216] E Georg Luebeck, Kit Curtius, Jihyoun Jeon, and William D Hazelton. Impact of tumor progression on cancer incidence curves. *Cancer Res.*, 73(3):1086–1096, February 2013.
 - [217] Georg E Luebeck, Thomas L Vaughan, Kit Curtius, and William D Hazelton. Modeling historic incidence trends implies early field cancerization in esophageal squamous cell carcinoma. *PLoS Comput. Biol.*, 17(5):e1008961, May 2021.
 - [218] Moritz Gerstung, Clemency Jolly, Ignaty Leshchiner, Stefan C Dentre, Santiago Gonzalez, Daniel Rosebrock, Thomas J Mitchell, Yulia Rubanova, Pavana Anur, Kaixian Yu, Maxime Tarabichi, Amit Deshwar, Jeff Wintersinger, Kortine Kleinheinz, Ignacio Vázquez-García, Kerstin Haase, Lara Jerman, Subhajit Sengupta, Geoff Macintyre, Salem Malikic, Nilgun Donmez, Dimitri G Livitz, Marek Cmero, Jonas Demeulemeester, Steven Schumacher, Yu Fan, Xiaotong Yao, Juhee Lee, Matthias Schlesner, Paul C Boutros, David D Bowtell, Hongtu Zhu, Gad Getz, Marcin Imielinski, Rameen Beroukhim, S Cenk Sahinalp, Yuan Ji, Martin Peifer, Florian Markowitz, Ville Mustonen, Ke Yuan, Wenyi Wang, Quaid D Morris, Paul T Spellman, David C Wedge, and Peter Van Loo. The evolutionary history of 2,658 cancers. *Nature*, 578(7793):122–128, February 2020.
 - [219] Ludmil B Alexandrov, Philip H Jones, David C Wedge, Julian E Sale, Peter J Campbell, Serena Nik-Zainal, and Michael R Stratton. Clock-like mutational processes in human somatic cells. *Nat. Genet.*, 47(12):1402–1407, November 2015.
 - [220] Christopher D Steele, Ammal Abbasi, S M Ashiqul Islam, Amy L Bowes, Azhar Khandekar, Kerstin Haase, Shadi Hames-Fathi, Dolapo Ajayi, Annelien Verfaillie, Pawan Dhama, Alex McLatchie, Matt Lechner, Nicholas Light, Adam Shlien, David Malkin, Andrew Feber, Paula Proszek, Tom Lesluyes, Fredrik Mertens, Adrienne M Flanagan, Maxime Tarabichi, Peter Van Loo, Ludmil B Alexandrov, and Nischalan Pillay. Signatures of copy number alterations in human cancer. *Nature*, 606(7916):984–991, June 2022.
 - [221] F Favero, T Joshi, A M Marquard, N J Birkbak, M Krzystanek, Q Li, Z Szallasi, and A C Eklund. Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. *Ann. Oncol.*, 26(1):64–70, January 2015.
 - [222] Skipper Seabold and Josef Perktold. Statsmodels: Econometric and statistical modeling

- with python. In *Proceedings of the 9th Python in Science Conference*, volume 57, page 61, 2010.
- [223] Yukinori Okada, Buhm Han, Lam C Tsoi, Philip E Stuart, Eva Ellinghaus, Trilokraj Tejasvi, Vinod Chandran, Fawnda Pellett, Remy Pollock, Anne M Bowcock, Gerald G Krueger, Michael Weichenthal, John J Voorhees, Proton Rahman, Peter K Gregersen, Andre Franke, Rajan P Nair, Gonçalo R Abecasis, Dafna D Gladman, James T Elder, Paul I W de Bakker, and Soumya Raychaudhuri. Fine mapping major histocompatibility complex associations in psoriasis and its clinical subtypes. *Am. J. Hum. Genet.*, 95(2):162–172, August 2014.
 - [224] Haoyan Chen, Genki Hayashi, Olivia Y Lai, Alexander Dilthey, Peter J Kuebler, Tami V Wong, Maureen P Martin, Marcelo A Fernandez Vina, Gil McVean, Matthias Wabl, Kieron S Leslie, Toby Maurer, Jeffrey N Martin, Steven G Deeks, Mary Carrington, Anne M Bowcock, Douglas F Nixon, and Wilson Liao. Psoriasis patients are enriched for genetic variants that protect against HIV-1 disease. *PLoS Genet.*, 8(2):e1002514, February 2012.
 - [225] Cynthia Helms, Nancy L Saccone, Li Cao, Jil A Wright Daw, Kai Cao, Tony M Hsu, Patricia Taillon-Miller, Shenghui Duan, Derek Gordon, Brandon Pierce, Jurg Ott, John Rice, Marcelo A Fernandez-Vina, Pui-Yan Kwok, Alan Menter, and Anne M Bowcock. Localization of PSORS1 to a haplotype block harboring HLA-C and distinct from corneodesmosin and HCR. *Hum. Genet.*, 118(3-4):466–476, December 2005.
 - [226] Robert Winchester, Gregory Minevich, Valeria Steshenko, Brian Kirby, David Kane, David A Greenberg, and Oliver FitzGerald. HLA associations reveal genetic heterogeneity in psoriatic arthritis and in the psoriasis phenotype. *Arthritis Rheum.*, 64(4):1134–1144, April 2012.
 - [227] Oliver FitzGerald, Muhammad Haroon, Jon T Giles, and Robert Winchester. Concepts of pathogenesis in psoriatic arthritis: genotype determines clinical phenotype. *Arthritis Res. Ther.*, 17:115, May 2015.
 - [228] G S Ogg, P Rod Dunbar, P Romero, J L Chen, and V Cerundolo. High frequency of skin-homing melanocyte-specific cytotoxic T lymphocytes in autoimmune vitiligo. *J. Exp. Med.*, 188(6):1203–1208, September 1998.
 - [229] K S Lang, C C Caroli, A Muhm, D Wernet, A Moris, B Schitteck, E Knauss-Scherwitz, S Stevanovic, H G Rammensee, and C Garbe. HLA-A2 restricted, melanocyte-specific CD8(+) T lymphocytes detected in vitiligo patients are related to disease activity and are predominantly directed against MelanA/MART1. *J. Invest. Dermatol.*, 116(6):891–897, June 2001.
 - [230] B Palermo, R Campanelli, S Garbelli, S Mantovani, E Lantelme, V Brazzelli, M Ardigò, G Borroni, M Martinetti, C Badulli, A Necker, and C Giachino. Specific cytotoxic T lymphocyte responses against melan-a/MART1, tyrosinase and gp100 in vitiligo by the use

- of major histocompatibility complex/peptide tetramers: the role of cellular immunity in the etiopathogenesis of vitiligo. *J. Invest. Dermatol.*, 117(2):326–332, August 2001.
- [231] Ying Jin, Tracey Ferrara, Katherine Gowan, Cherie Holcomb, Melinda Rastrou, Henry A Erlich, Pamela R Fain, and Richard A Spritz. Next-generation DNA re-sequencing identifies common variants of TYR and HLA-a that modulate the risk of generalized vitiligo via antigen presentation. *J. Invest. Dermatol.*, 132(6):1730–1733, June 2012.
- [232] Xue-Jun Zhang, Hong-Sheng Liu, Yan-Hua Liang, Liang-Dan Sun, Ji-Yun Wang, Sen Yang, Jiang-Bo Liu, Min Gao, Ping-Ping He, Yong Cui, and Qing Yang. Association of HLA class I alleles with vitiligo in chinese hans. *J. Dermatol. Sci.*, 35(2):165–168, August 2004.
- [233] Cheng Quan, Yun-Qing Ren, Lei-Hong Xiang, Liang-Dan Sun, Ai-E Xu, Xing-Hua Gao, Hong-Duo Chen, Xiong-Ming Pu, Ri-Na Wu, Chao-Zhao Liang, Jia-Bin Li, Tian-Wen Gao, Jian-Zhong Zhang, Xiu-Li Wang, Jun Wang, Rong-Ya Yang, Ling Liang, Jian-Bin Yu, Xian-Bo Zuo, Sheng-Quan Zhang, Shu-Mei Zhang, Gang Chen, Xiao-Dong Zheng, Pan Li, Jun Zhu, Yong-Wei Li, Xiao-Dong Wei, Wei-Song Hong, Ying Ye, Yong Zhang, Wei-Su Wu, Hui Cheng, Pu-Ling Dong, Da-Yan Hu, Yang Li, Min Li, Xin Zhang, Hua-Yang Tang, Xian-Fa Tang, Sheng-Xin Xu, Su-Min He, Yong-Mei Lv, Min Shen, Hong-Quan Jiang, Ying Wang, Kai Li, Xiao-Jing Kang, Yu-Qin Liu, Li Sun, Zhi-Fang Liu, Shao-Qiong Xie, Cheng-Yao Zhu, Qiang Xu, Jin-Ping Gao, Wen-Long Hu, Cheng Ni, Ting-Meng Pan, Yun Li, Sha Yao, Cai-Feng He, Yang-Sheng Liu, Ze-Ying Yu, Xian-Yong Yin, Feng-Yu Zhang, Sen Yang, Youwen Zhou, and Xue-Jun Zhang. Genome-wide association study for vitiligo identifies susceptibility loci at 6q27 and the MHC. *Nat. Genet.*, 42(7):614–618, July 2010.
- [234] Rajan P Nair, Philip E Stuart, Ioana Nistor, Ravi Hiremagalore, Nicholas V C Chia, Stefan Jenisch, Michael Weichenthal, Gonçalo R Abecasis, Henry W Lim, Enno Christophers, John J Voorhees, and James T Elder. Sequence and haplotype analysis supports HLA-C as the psoriasis susceptibility 1 gene. *Am. J. Hum. Genet.*, 78(5):827–851, May 2006.
- [235] J E Gudjonsson, A Karason, A Antonsdottir, E H Runarsdottir, V B Hauksson, R Upmanyu, J Gulcher, K Stefansson, and H Valdimarsson. Psoriasis patients who are homozygous for the HLA-Cw*0602 allele have a 2.5-fold increased risk of developing psoriasis compared with Cw6 heterozygotes. *Br. J. Dermatol.*, 148(2):233–235, February 2003.
- [236] Toshiyuki Yamamoto, Hiroo Yokozeki, and Kiyoshi Nishioka. Psoriasis arthropathy and HLA-B51: report of 5 cases. *J. Dermatol.*, 32(7):606–610, July 2005.
- [237] Sergey Nejentsev, Joanna M M Howson, Neil M Walker, Jeffrey Szeszko, Sarah F Field, Helen E Stevens, Pamela Reynolds, Matthew Hardy, Erna King, Jennifer Masters, John Hulme, Lisa M Maier, Deborah Smyth, Rebecca Bailey, Jason D Cooper, Gloria Ribas, R Duncan Campbell, David G Clayton, John A Todd, and Wellcome Trust Case Control Consortium. Localization of type 1 diabetes susceptibility to the MHC class I genes HLA-B and HLA-a. *Nature*, 450(7171):887–892, December 2007.
- [238] Janelle A Noble, Ana Maria Valdes, Michael D Varney, Joyce A Carlson, Priscilla Moon-

- samy, Anna Lisa Fear, Julie A Lane, Eva Lavant, Rebecca Rappner, Anthony Louey, Patrick Concannon, Josyf C Mychaleckyj, Henry A Erlich, and Type 1 Diabetes Genetics Consortium. HLA class I and genetic susceptibility to type 1 diabetes: results from the type 1 diabetes genetics consortium. *Diabetes*, 59(11):2972–2979, November 2010.
- [239] J M M Howson, N M Walker, D Clayton, J A Todd, and Type 1 Diabetes Genetics Consortium. Confirmation of HLA class II independent type 1 diabetes associations in the major histocompatibility complex including HLA-B and HLA-a. *Diabetes Obes. Metab.*, 11 Suppl 1:31–45, February 2009.
- [240] S Ohno, M Ohguchi, S Hirose, H Matsuda, A Wakisaka, and M Aizawa. Close association of HLA-Bw51 with behcet’s disease, 1982.
- [241] Jim Kaufman. Generalists and specialists: A new view of how MHC class I molecules fight infectious pathogens. *Trends Immunol.*, 39(5):367–379, May 2018.
- [242] Arman A Bashirova, Mathias Viard, Vivek Naranbhai, Alba Grifoni, Wilfredo Garcia-Beltran, Marjan Akdag, Yuko Yuki, Xiaojiang Gao, Colm O’hUigin, Malini Raghavan, Steven Wolinsky, Jay H Bream, Priya Duggal, Jeremy Martinson, Nelson L Michael, Gregory D Kirk, Susan P Buchbinder, David Haas, James J Goedert, Steven G Deeks, Jacques Fellay, Bruce Walker, Philip Goulder, Peter Cresswell, Tim Elliott, Alessandro Sette, Jonathan Carlson, and Mary Carrington. HLA tapasin independence: broader peptide repertoire and HIV control. *Proceedings of the National Academy of Sciences*, 117(45):28232–28238, 2020.
- [243] Faviel F Gonzalez-Galarza, Antony McCabe, Eduardo J Melo dos Santos, James Jones, Louise Takeshita, Nestor D Ortega-Rivera, Glenda M Del Cid-Pavon, Kerry Ramsbottom, Gurpreet Ghattaoraya, Ana Alfirevic, Derek Middleton, and Andrew R Jones. Allele frequency net database (AFND) 2020 update: gold-standard data classification, open access genotype data and new query tools. *Nucleic Acids Res.*, 48(D1):D783–D788, November 2019.
- [244] K Magnus. Prognosis in malignant melanoma of the skin. significance of stage of disease, anatomical site, sex, age and period of diagnosis. *Cancer*, 40(1):389–397, July 1977.
- [245] Christopher A Natale, Jinyang Li, Junqian Zhang, Ankit Dahal, Tzvete Dentchev, Ben Z Stanger, and Todd W Ridky. Activation of G protein-coupled estrogen receptor signaling inhibits melanoma and improves response to immune checkpoint blockade. *Elife*, 7, January 2018.
- [246] Shokrollah Elahi, Warren L Dinges, Nicholas Lejarcegui, Kerry J Laing, Ann C Collier, David M Koelle, M Juliana McElrath, and Helen Horton. Protective HIV-specific CD8+ T cells evade treg cell suppression. *Nat. Med.*, 17(8):989–995, July 2011.
- [247] John Michael Gaziano, John Concato, Mary Brophy, Louis Fiore, Saiju Pyarajan, James Breeling, Stacey Whitbourne, Jennifer Deen, Colleen Shannon, Donald Humphries, Peter

- Guarino, Mihaela Aslan, Daniel Anderson, Rene LaFleur, Timothy Hammond, Kendra Schaa, Jennifer Moser, Grant Huang, Sumitra Muralidhar, Ronald Przygodzki, and Timothy J O’Leary. Million veteran program: A mega-biobank to study genetic influences on health and disease. *J. Clin. Epidemiol.*, 70:214–223, February 2016.
- [248] J Andrew Carlson. Tumor doubling time of cutaneous melanoma and its metastasis. *Am. J. Dermatopathol.*, 25(4):291–299, August 2003.
- [249] Katherine Kay, Keith Dolcy, Robert Bies, and Dhaval K Shah. Estimation of solid tumor doubling times from progression-free survival plots using a novel statistical approach. *AAPS J.*, 21(2):27, February 2019.
- [250] Hansje-Eva Teulings, Karin J Willemsen, Iris Glykofridis, Gabrielle Krebbers, Lisa Komen, Marije W Kroon, E Helen Kemp, Albert Wolkerstorfer, J P Wietze van der Veen, Rosalie M Luiten, and Esther P M Tjin. The antibody response against MART-1 differs in patients with melanoma-associated leucoderma and vitiligo. *Pigment Cell Melanoma Res.*, 27(6):1086–1096, November 2014.
- [251] Michelle Rodrigues, Khaled Ezzedine, Iltefat Hamzavi, Amit G Pandya, John E Harris, and Vitiligo Working Group. New discoveries in the pathogenesis and classification of vitiligo. *J. Am. Acad. Dermatol.*, 77(1):1–13, July 2017.
- [252] Y Kawakami, S Eliyahu, C H Delgado, P F Robbins, K Sakaguchi, E Appella, J R Yannelli, G J Adema, T Miki, and S A Rosenberg. Identification of a human melanoma antigen recognized by tumor-infiltrating lymphocytes associated with in vivo tumor rejection. *Proc. Natl. Acad. Sci. U. S. A.*, 91(14):6458–6462, July 1994.
- [253] Y T Chen, E Stockert, S Tsang, K A Coplan, and L J Old. Immunophenotyping of melanomas for tyrosinase: implications for vaccine development. *Proc. Natl. Acad. Sci. U. S. A.*, 92(18):8125–8129, August 1995.
- [254] R F Wang, E Appella, Y Kawakami, X Kang, and S A Rosenberg. Identification of TRP-2 as a human tumor antigen recognized by cytotoxic T lymphocytes. *J. Exp. Med.*, 184(6):2207–2216, December 1996.
- [255] F Brasseur, D Rimoldi, D Liénard, B Lethé, S Carrel, F Arienti, L Suter, R Vanwijck, A Bourlond, and Y Humblet. Expression of MAGE genes in primary and metastatic cutaneous melanoma. *International Journal of Cancer*, 63(3):375–380, November 1995.
- [256] M Mori, H Inoue, K Mimori, K Shibuta, K Baba, H Nakashima, M Haraguchi, K Tsuji, H Ueo, G F Barnard, and T Akiyoshi. Expression of MAGE genes in human colorectal carcinoma. *Ann. Surg.*, 224(2):183–188, August 1996.
- [257] D L Scarcella, C W Chow, M F Gonzales, C Economou, F Brasseur, and D M Ashley. Expression of MAGE and GAGE in high-grade brain tumors: a potential target for specific immunotherapy and diagnostic markers. *Clin. Cancer Res.*, 5(2):335–341, February 1999.

- [258] Kouhei Tajima, Yuichi Obata, Hiromi Tamaki, Masahiro Yoshida, Yao-Tseng Chen, Matthew J Scanlan, Lloyd J Old, Hiroyuki Kuwano, Takashi Takahashi, Toshitada Takahashi, and Tetsuya Mitsudomi. Expression of cancer/testis (CT) antigens in lung cancer. *Lung Cancer*, 42(1):23–33, October 2003.
- [259] Ali O Gure, Ramon Chua, Barbara Williamson, Mithat Gonen, Cathy A Ferrera, Sacha Gnjjatic, Gerd Ritter, Andrew J G Simpson, Yao-T Chen, Lloyd J Old, and Nasser K Altorki. Cancer-testis genes are coordinately expressed and are markers of poor outcome in non-small cell lung cancer. *Clin. Cancer Res.*, 11(22):8055–8062, November 2005.
- [260] Catherine Barrow, Judy Browning, Duncan MacGregor, Ian D Davis, Sue Sturrock, Achim A Jungbluth, and Jonathan Cebon. Tumor antigen expression in melanoma varies according to antigen and stage. *Clin. Cancer Res.*, 12(3 Pt 1):764–771, February 2006.
- [261] Yeong-Dae Kim, Hae-Rim Park, Myung-Ha Song, Dong-Hoon Shin, Chang-Hun Lee, Min-Ki Lee, and Sang-Yull Lee. Pattern of cancer/testis antigen expression in lung cancer patients. *Int. J. Mol. Med.*, 29(4):656–662, April 2012.
- [262] Morten Nielsen and Massimo 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(1):33, March 2016.
- [263] Mads Hald Andersen, Joachim Fensterle, Selma Ugurel, Sine Reker, Roland Houben, Per Guldberg, Thomas G Berger, Dirk Schadendorf, Uwe Trefzer, Eva-B Bröcker, Per Thor Straten, Ulf R Rapp, and Jürgen C Becker. Immunogenicity of constitutively active V599EBRaf. *Cancer Res.*, 64(15):5456–5460, August 2004.
- [264] Sherille D Bradley, Zeming Chen, Brenda Melendez, Amjad Talukder, Jahan S Khalili, Tania Rodriguez-Cruz, Shujuan Liu, Mayra Whittington, Wanleng Deng, Fenge Li, Chantale Bernatchez, Laszlo G Radvanyi, Michael A Davies, Patrick Hwu, and Gregory 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(6):602–609, June 2015.
- [265] J D Williams, Elaine L Jacobson, H Kim, M Kim, and M K Jacobson. Folate in skin cancer prevention. *Subcell. Biochem.*, 56:181–197, 2012.
- [266] Andrew J G Simpson, Otavia L Caballero, Achim Jungbluth, Yao-Tseng Chen, and Lloyd J Old. Cancer/testis antigens, gametogenesis and cancer. *Nat. Rev. Cancer*, 5(8):615–625, August 2005.
- [267] Dong Fang, James Hallman, Namrata Sangha, Timothy E Kute, James A Hammarback, Wain L White, and Vijayasaraadhi Setaluri. Expression of microtubule-associated protein 2 in benign and malignant melanocytes. *Am. J. Pathol.*, 158(6):2107–2115, June 2001.
- [268] J Vachtenheim, H Novotna, and G Ghanem. Transcriptional repression of the microph-

- thalmia gene in melanoma cells correlates with the unresponsiveness of target genes to ectopic microphthalmia-associated transcription factor. *J. Invest. Dermatol.*, 117(6): 1505–1511, December 2001.
- [269] Ana Marcu, Leon Bichmann, Leon Kuchenbecker, Daniel Johannes Kowalewski, Lena Katharina Freudenmann, Linus Backert, Lena Mühlenbruch, András Szolek, Maren Lübke, Philipp Wagner, Tobias Engler, Sabine Matovina, Jian Wang, Mathias Hauri-Hohl, Roland Martin, Konstantina Kapolou, Juliane Sarah Walz, Julia Velz, Holger Moch, Luca Regli, Manuela Silginer, Michael Weller, Markus W Löffler, Florian Erhard, Andreas Schlosser, Oliver Kohlbacher, Stefan Stevanović, Hans-Georg Rammensee, and Marian Christoph Neidert. HLA ligand atlas: a benign reference of HLA-presented peptides to improve T-cell-based cancer immunotherapy. *J Immunother Cancer*, 9(4), April 2021.
- [270] Mariana Aris, Mariana Rodríguez Zubieta, Marina Colombo, Juan Martín Arriaga, Michele Bianchini, Myriam Alperovich, Alicia I Bravo, María M Barrio, and José Mordoh. MART-1- and gp100-expressing and -non-expressing melanoma cells are equally proliferative in tumors and clonogenic in vitro. *J. Invest. Dermatol.*, 132(2):365–374, February 2012.
- [271] F Journe, H Id Boufker, L Van Kempen, M-D Galibert, M Wiedig, F Salès, A Theunis, D Nonclercq, A Frau, G Laurent, A Awada, and G Ghanem. TYRP1 mRNA expression in melanoma metastases correlates with clinical outcome. *Br. J. Cancer*, 105(11):1726–1732, November 2011.
- [272] Marek Wieczorek, Esam T Abualrous, Jana Sticht, Miguel Álvaro Benito, Sebastian Stolzenberg, Frank Noé, and Christian Freund. Major histocompatibility complex (MHC) class I and MHC class II proteins: Conformational plasticity in antigen presentation. *Front. Immunol.*, 8:292, March 2017.
- [273] Gergely Bodis, Victoria Toth, and Andreas Schwarting. Role of human leukocyte antigens (HLA) in autoimmune diseases. *Rheumatol Ther*, 5(1):5–20, June 2018.
- [274] Hiroko Miyadera, Jun Ohashi, Åke Lernmark, Toshio Kitamura, and Katsushi Tokunaga. Cell-surface MHC density profiling reveals instability of autoimmunity-associated HLA. *J. Clin. Invest.*, 125(1):275–291, January 2015.
- [275] Ri-Yao Yang, Gabriel A Rabinovich, and Fu-Tong Liu. Galectins: structure, function and therapeutic potential. *Expert Rev. Mol. Med.*, 10:e17, June 2008.
- [276] Nicole B Crux and Shokrollah Elahi. Human leukocyte antigen (HLA) and immune regulation: How do classical and non-classical HLA alleles modulate immune response to human immunodeficiency virus and hepatitis C virus infections? *Front. Immunol.*, 8:832, July 2017.
- [277] Feng Wang, Wentao He, Hongmin Zhou, Jin Yuan, Ke Wu, Lijun Xu, and Zhonghua Klaus Chen. The tim-3 ligand galectin-9 negatively regulates CD8+ alloreactive T cell and prolongs survival of skin graft. *Cell. Immunol.*, 250(1):68–74, November 2007.

- [278] Stephen A Migueles, Christine M Osborne, Cassandra Royce, Alex A Compton, Rohan P Joshi, Kristin A Weeks, Julia E Rood, Amy M Berkley, Jonah B Sacha, Nancy A Coglian-Shutta, Margaret Lloyd, Gregg Roby, Richard Kwan, Mary McLaughlin, Sara Stallings, Catherine Rehm, Marie A O'Shea, Joann Mican, Beverly Z Packard, Akira Komoriya, Sarah Palmer, Ann P Wiegand, Frank Maldarelli, John M Coffin, John W Mellors, Claire W Hallahan, Dean A Follman, and Mark Connors. Lytic granule loading of CD8+ T cells is required for HIV-infected cell elimination associated with immune control. *Immunity*, 29(6):1009–1021, December 2008.
- [279] Andrej Košmrlj, Elizabeth L Read, Ying Qi, Todd M Allen, Marcus Altfeld, Steven G Deeks, Florencia Pereyra, Mary Carrington, Bruce D Walker, and Arup K Chakraborty. Effects of thymic selection of the T-cell repertoire on HLA class I-associated control of HIV infection. *Nature*, 465(7296):350–354, May 2010.
- [280] Paul E Chappell, El Kahina Meziane, Michael Harrison, Łukasz Magiera, Clemens Hermann, Laura Mears, Antoni G Wrobel, Charlotte Durant, Lise Lotte Nielsen, Søren Buus, Nicola Ternette, William Mwangi, Colin Butter, Venugopal Nair, Trudy Ahjee, Richard Duggleby, Alejandro Madrigal, Pietro Roversi, Susan M Lea, and Jim Kaufman. Expression levels of MHC class I molecules are inversely correlated with promiscuity of peptide binding. *Elife*, 4:e05345, April 2015.
- [281] International HIV Controllers Study, Florencia Pereyra, Xiaoming Jia, Paul J McLaren, Amalio Telenti, Paul I W de Bakker, Bruce D Walker, Stephan Ripke, Chanson J Brumme, Sara L Pulit, Mary Carrington, Carl M Kadie, Jonathan M Carlson, David Heckerman, Robert R Graham, Robert M Plenge, Steven G Deeks, Lauren Gianniny, Gabriel Crawford, Jordan Sullivan, Elena Gonzalez, Leela Davies, Amy Camargo, Jamie M Moore, Nicole Beattie, Supriya Gupta, Andrew Crenshaw, Noël P Burt, Candace Guiducci, Namrata Gupta, Xiaojiang Gao, Ying Qi, Yuko Yuki, Alicja Piechocka-Trocha, Emily Cutrell, Rachel Rosenberg, Kristin L Moss, Paul Lemay, Jessica O'Leary, Todd Schaefer, Pranshu Verma, Ildiko Toth, Brian Block, Brett Baker, Alissa Rothchild, Jeffrey Lian, Jacqueline Proudfoot, Donna Marie L Alvino, Seanna Vine, Marylyn M Addo, Todd M Allen, Marcus Altfeld, Matthew R Henn, Sylvie Le Gall, Hendrik Streeck, David W Haas, Daniel R Kuritzkes, Gregory K Robbins, Robert W Shafer, Roy M Gulick, Cecilia M Shikuma, Richard Haubrich, Sharon Riddler, Paul E Sax, Eric S Daar, Heather J Ribaudo, Brian Agan, Shanu Agarwal, Richard L Ahern, Brady L Allen, Sherly Altidor, Eric L Altschuler, Sujata Ambardar, Kathryn Anastos, Ben Anderson, Val Anderson, Ushan Andrad, Diana Antoniskis, David Bangsberg, Daniel Barbaro, William Barrie, J Bartczak, Simon Barton, Patricia Basden, Nesli Basgoz, Suzane Bazner, Nicholas C Bellos, Anne M Benson, Judith Berger, Nicole F Bernard, Annette M Bernard, Christopher Birch, Stanley J Bodner, Robert K Bolan, Emilie T Boudreaux, Meg Bradley, James F Braun, Jon E Brndjar, Stephen J Brown, Katherine Brown, Sheldon T Brown, Jedidiah Burack, Larry M Bush, Virginia Cafaro, Omobolaji Campbell, John Campbell, Robert H Carlson, J Kevin Carmichael, Kathleen K Casey, Chris Cavacuiti, Gregory Celestin, Steven T Chambers, Nancy Chez, Lisa M Chirch, Paul J Cimoch, Daniel Cohen, Lillian E Cohn, Brian Conway, David A Cooper, Brian Cornelson, David T Cox, Michael V Cristofano, George Cuchural, Jr, Julie L Czartoski, Joseph M Dahman, Jennifer S Daly, Benjamin T Davis, Kristine Davis, Sheila M Davod,

Edwin DeJesus, Craig A Dietz, Eleanor Dunham, Michael E Dunn, Todd B Ellerin, Joseph J Eron, John J W Fangman, Claire E Farel, Helen Ferlazzo, Sarah Fidler, Anita Fleenor-Ford, Renee Frankel, Kenneth A Freedberg, Neel K French, Jonathan D Fuchs, Jon D Fuller, Jonna Gaberman, Joel E Gallant, Rajesh T Gandhi, Efrain Garcia, Donald Garmon, Joseph C Gathe, Jr, Cyril R Gaultier, Wondwoosen Gebre, Frank D Gilman, Ian Gilson, Paul A Goepfert, Michael S Gottlieb, Claudia Goulston, Richard K Groger, T Douglas Gurley, Stuart Haber, Robin Hardwicke, W David Hardy, P Richard Harrigan, Trevor N Hawkins, Sonya Heath, Frederick M Hecht, W Keith Henry, Melissa Hladek, Robert P Hoffman, James M Horton, Ricky K Hsu, Gregory D Huhn, Peter Hunt, Mark J Hupert, Mark L Illeman, Hans Jaeger, Robert M Jellinger, Mina John, Jennifer A Johnson, Kristin L Johnson, Heather Johnson, Kay Johnson, Jennifer Joly, Wilbert C Jordan, Carol A Kauffman, Homayoon Khanlou, Robert K Killian, Arthur Y Kim, David D Kim, Clifford A Kinder, Jeffrey T Kirchner, Laura Kogelman, Erna Milunka Kojic, P Todd Korthuis, Wayne Kurisu, Douglas S Kwon, Melissa LaMar, Harry Lampiris, Massimiliano Lanzafame, Michael M Lederman, David M Lee, Jean M L Lee, Marah J Lee, Edward T Y Lee, Janice Lemoine, Jay A Levy, Josep M Llibre, Michael A Liguori, Susan J Little, Anne Y Liu, Alvaro J Lopez, Mono R Loutfy, Dawn Loy, Debbie Y Mohammed, Alan Man, Michael K Mansour, Vincent C Marconi, Martin Markowitz, Rui Marques, Jeffrey N Martin, Harold L Martin, Jr, Kenneth Hugh Mayer, M Juliana McElrath, Theresa A McGhee, Barbara H McGovern, Katherine McGowan, Dawn McIntyre, Gavin X Mcleod, Prema Menezes, Greg Mesa, Craig E Metroka, Dirk Meyer-Olson, Andy O Miller, Kate Montgomery, Karam C Mounzer, Ellen H Nagami, Iris Nagin, Ronald G Nahass, Margret O Nelson, Craig Nielsen, David L Norene, David H O'Connor, Bisola O Ojikutu, Jason Okulicz, Olakunle O Oladehin, Edward C Oldfield, 3rd, Susan A Olender, Mario Ostrowski, William F Owen, Jr, Eunice Pae, Jeffrey Parsonnet, Andrew M Pavlatos, Aaron M Perlmutter, Michael N Pierce, Jonathan M Pincus, Leandro Pisani, Lawrence Jay Price, Laurie Proia, Richard C Prokesch, Heather Calderon Pujet, Moti Ramgopal, Almas Rathod, Michael Rausch, J Ravishankar, Frank S Rhame, Constance Shamuyarira Richards, Douglas D Richman, Berta Rodes, Milagros Rodriguez, Richard C Rose, 3rd, Eric S Rosenberg, Daniel Rosenthal, Polly E Ross, David S Rubin, Elease Rumbaugh, Luis Saenz, Michelle R Salvaggio, William C Sanchez, Veera M Sanjana, Steven Santiago, Wolfgang Schmidt, Hanneke Schuitemaker, Philip M Sestak, Peter Shalit, William Shay, Vivian N Shirvani, Vanessa I Silebi, James M Sizemore, Jr, Paul R Skolnik, Marcia Sokol-Anderson, James M Sosman, Paul Stabile, Jack T Stapleton, Sheree Starrett, Francine Stein, Hans-Jurgen Stellbrink, F Lisa Sterman, Valerie E Stone, David R Stone, Giuseppe Tambussi, Randy A Taplitz, Ellen M Tedaldi, Amalio Telenti, William Theisen, Richard Torres, Lorraine Tosiello, Cecile Tremblay, Marc A Tribble, Phuong D Trinh, Alice Tsao, Peggy Ueda, Anthony Vaccaro, Emilia Valadas, Thanos J Vanig, Isabel Vecino, Vilma M Vega, Wenoah Veikley, Barbara H Wade, Charles Walworth, Chingchai Wanidworanun, Douglas J Ward, Daniel A Warner, Robert D Weber, Duncan Webster, Steve Weis, David A Wheeler, David J White, Ed Wilkins, Alan Winston, Clifford G Wlodaver, Angelique van't Wout, David P Wright, Otto O Yang, David L Yurdin, Brandon W Zabukovic, Kimon C Zachary, Beth Zeeman, and Meng Zhao. The major genetic determinants of HIV-1 control affect HLA class I peptide presentation. *Science*, 330(6010):1551–1557, December 2010.

[282] Yonghong Zhang, Yanchun Peng, Huiping Yan, Keyi Xu, Masumichi Saito, Hao Wu,

- Xinyue Chen, Srinika Ranasinghe, Nozomi Kuse, Tim Powell, Yan Zhao, Weihua Li, Xin Zhang, Xia Feng, Ning Li, Aleksandra Leligdowicz, Xiaoning Xu, Mina John, Masafumi Takiguchi, Andrew McMichael, Sarah Rowland-Jones, and Tao Dong. Multilayered defense in HLA-B51-associated HIV viral control. *J. Immunol.*, 187(2):684–691, July 2011.
- [283] Jesse I Mobbs, Patricia T Illing, Nadine L Dudek, Andrew G Brooks, Daniel G Baker, Anthony W Purcell, Jamie Rossjohn, and Julian P Vivian. The molecular basis for peptide repertoire selection in the human leucocyte antigen (HLA) C*06:02 molecule. *J. Biol. Chem.*, 292(42):17203–17215, October 2017.
- [284] D R Madden, J C Gorga, J L Strominger, and D C Wiley. The structure of HLA-B27 reveals nonamer self-peptides bound in an extended conformation. *Nature*, 353(6342):321–325, September 1991.
- [285] Loredana Saveanu, Oliver Carroll, Vivian Lindo, Margarita Del Val, Daniel Lopez, Yves Lepelletier, Fiona Greer, Lutz Schomburg, Doriana Fruci, Gabriele Niedermann, and Peter M van Endert. Concerted peptide trimming by human ERAP1 and ERAP2 aminopeptidase complexes in the endoplasmic reticulum. *Nat. Immunol.*, 6(7):689–697, May 2005.
- [286] Tomo Saric, Shih-Chung Chang, Akira Hattori, Ian A York, Shirley Markant, Kenneth L Rock, Masafumi Tsujimoto, and Alfred L Goldberg. An IFN- γ -induced aminopeptidase in the ER, ERAP1, trims precursors to MHC class I-presented peptides. *Nat. Immunol.*, 3(12):1169–1176, November 2002.
- [287] Ian A York, Shih-Chung Chang, Tomo Saric, Jennifer A Keys, Janice M Favreau, Alfred L Goldberg, and Kenneth L Rock. The ER aminopeptidase ERAP1 enhances or limits antigen presentation by trimming epitopes to 8-9 residues. *Nat. Immunol.*, 3(12):1177–1184, December 2002.
- [288] Gianna Elena Hammer, Federico Gonzalez, Edward James, Hector Nolla, and Nilabh Shastri. In the absence of aminopeptidase ERAAP, MHC class I molecules present many unstable and highly immunogenic peptides. *Nat. Immunol.*, 8(1):101–108, November 2006.
- [289] Emma Reeves, Christopher J Edwards, Tim Elliott, and Edward James. Naturally occurring ERAP1 haplotypes encode functionally distinct alleles with fine substrate specificity. *J. Immunol.*, 191(1):35–43, July 2013.
- [290] Adrian Martín-Esteban, Patricia Gómez-Molina, Alejandro Sanz-Bravo, and José A López de Castro. Combined effects of ankylosing spondylitis-associated ERAP1 polymorphisms outside the catalytic and peptide-binding sites on the processing of natural HLA-B27 ligands. *J. Biol. Chem.*, 289(7):3978–3990, February 2014.
- [291] Genetic Analysis of Psoriasis Consortium & the Wellcome Trust Case Control Consortium 2, Amy Strange, Francesca Capon, Chris C A Spencer, Jo Knight, Michael E Weale, Michael H Allen, Anne Barton, Gavin Band, Céline Bellenguez, Judith G M Bergboer, Jenefer M Blackwell, Elvira Bramon, Suzannah J Bumpstead, Juan P Casas, Michael J Cork, Aiden

- Corvin, Panos Deloukas, Alexander Diltthey, Audrey Duncanson, Sarah Edkins, Xavier Estivill, Oliver Fitzgerald, Colin Freeman, Emiliano Giardina, Emma Gray, Angelika Hofer, Ulrike Hüffmeier, Sarah E Hunt, Alan D Irvine, Janusz Jankowski, Brian Kirby, Cordelia Langford, Jesús Lascorz, Joyce Leman, Stephen Leslie, Lotus Mallbris, Hugh S Markus, Christopher G Mathew, W H Irwin McLean, Ross McManus, Rotraut Mössner, Loukas Moutsianas, Asa T Naluai, Frank O Nestle, Giuseppe Novelli, Alexandros Onoufriadis, Colin N A Palmer, Carlo Perricone, Matti Pirinen, Robert Plomin, Simon C Potter, Ramon M Pujol, Anna Rautanen, Eva Riveira-Munoz, Anthony W Ryan, Wolfgang Salmhofer, Lena Samuelsson, Stephen J Sawcer, Joost Schalkwijk, Catherine H Smith, Mona Ståhle, Zhan Su, Rachid Tazi-Ahnini, Heiko Traupe, Ananth C Viswanathan, Richard B Warren, Wolfgang Weger, Katarina Wolk, Nicholas Wood, Jane Worthington, Helen S Young, Patrick L J M Zeeuwen, Adrian Hayday, A David Burden, Christopher E M Griffiths, Juha Kere, André Reis, Gilean McVean, David M Evans, Matthew A Brown, Jonathan N Barker, Leena Peltonen, Peter Donnelly, and Richard C Trembath. A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between HLA-C and ERAP1. *Nat. Genet.*, 42(11):985–990, November 2010.
- [292] Lam C Tsoi, Sarah L Spain, Jo Knight, Eva Ellinghaus, Philip E Stuart, Francesca Capon, Jun Ding, Yanming Li, Trilokraj Tejasvi, Johann E Gudjonsson, Hyun M Kang, Michael H Allen, Ross McManus, Giuseppe Novelli, Lena Samuelsson, Joost Schalkwijk, Mona Ståhle, A David Burden, Catherine H Smith, Michael J Cork, Xavier Estivill, Anne M Bowcock, Gerald G Krueger, Wolfgang Weger, Jane Worthington, Rachid Tazi-Ahnini, Frank O Nestle, Adrian Hayday, Per Hoffmann, Juliane Winkelmann, Cisca Wijmenga, Cordelia Langford, Sarah Edkins, Robert Andrews, Hannah Blackburn, Amy Strange, Gavin Band, Richard D Pearson, Damjan Vukcevic, Chris C A Spencer, Panos Deloukas, Ulrich Mrowietz, Stefan Schreiber, Stephan Weidinger, Sulev Koks, Külli Kingo, Tonu Esko, Andres Metspalu, Henry W Lim, John J Voorhees, Michael Weichenthal, H Erich Wichmann, Vinod Chandran, Cheryl F Rosen, Proton Rahman, Dafna D Gladman, Christopher E M Griffiths, Andre Reis, Juha Kere, Collaborative Association Study of Psoriasis (CASP), Genetic Analysis of Psoriasis Consortium, Psoriasis Association Genetics Extension, Wellcome Trust Case Control Consortium 2, Rajan P Nair, Andre Franke, Jonathan N W N Barker, Goncalo R Abecasis, James T Elder, and Richard C Trembath. Identification of 15 new psoriasis susceptibility loci highlights the role of innate immunity. *Nat. Genet.*, 44(12):1341–1348, December 2012.
- [293] Xianyong Yin, Hui Qi Low, Ling Wang, Yonghong Li, Eva Ellinghaus, Jiali Han, Xavier Estivill, Liangdan Sun, Xianbo Zuo, Changbing Shen, Caihong Zhu, Anping Zhang, Fabio Sanchez, Leonid Padyukov, Joseph J Catanese, Gerald G Krueger, Kristina Callis Duffin, Sören Mucha, Michael Weichenthal, Stephan Weidinger, Wolfgang Lieb, Jia Nee Foo, Yi Li, Karseng Sim, Herty Liany, Ishak Irwan, Yikying Teo, Colin T S Theng, Rashmi Gupta, Anne Bowcock, Philip L De Jager, Abrar A Qureshi, Paul I W de Bakker, Mark Seielstad, Wilson Liao, Mona Ståhle, Andre Franke, Xuejun Zhang, and Jianjun Liu. Genome-wide meta-analysis identifies multiple novel associations and ethnic heterogeneity of psoriasis susceptibility. *Nat. Commun.*, 6(1):1–11, April 2015.
- [294] Amy L Cummings, Jaklin Gukasyan, Henry Y Lu, Tristan Grogan, Gemalene Sunga,

- Charlene M Fares, Nicholas Hornstein, Jesse Zaretsky, James Carroll, Benjamin Bachrach, Wisdom O Akingbemi, Debory Li, Zorawar Noor, Aaron Lisberg, Jonathan W Goldman, David Elashoff, Alex A T Bui, Antoni Ribas, Steven M Dubinett, Maura Rossetti, and Edward B Garon. Mutational landscape influences immunotherapy outcomes among patients with non-small-cell lung cancer with human leukocyte antigen supertype B44. *Nature Cancer*, 1(12):1167–1175, November 2020.
- [295] M DiBrino, K C Parker, D H Margulies, J Shiloach, R V Turner, W E Biddison, and J E Coligan. Identification of the peptide binding motif for HLA-B44, one of the most common HLA-B alleles in the caucasian population. *Biochemistry*, 34(32):10130–10138, August 1995.
- [296] D A Brewerton, F D Hart, A Nicholls, M Caffrey, D C James, and R D Sturrock. Ankylosing spondylitis and HL-a 27. *Lancet*, 1(7809):904–907, April 1973.
- [297] D A Brewerton, M Caffrey, A Nicholls, D Walters, and D C James. HL-a 27 and arthropathies associated with ulcerative colitis and psoriasis. *Lancet*, 1(7864):956–958, May 1974.
- [298] Lai-Shan Tam, Jieruo Gu, and David Yu. Pathogenesis of ankylosing spondylitis. *Nat. Rev. Rheumatol.*, 6(7):399–405, July 2010.