UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

Advancing Variant Effect Prediction Beyond Protein-Level by Incorporating Systems-Level Architecture

Permalink

https://escholarship.org/uc/item/0q46m3cx

Author

Ozturk, Kivilcim

Publication Date

2022

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA SAN DIEGO

Advancing Variant Effect Prediction Beyond Protein-Level by Incorporating Systems-Level Architecture

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

in

Bioinformatics and Systems Biology

by

Kivilcim Ozturk

Committee in charge:

Professor Hannah Carter, Chair
Professor Prashant Mali, Co-Chair
Professor Vineet Bafna
Professor Olivier Harismendy
Professor Nathan Lewis

2022

Copyright

Kivilcim Ozturk, 2022

All rights reserved.

The Dissertation of Kivilcim Ozturk is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2022

DEDICATION

I dedicate this dissertation to my parents.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First and foremost, I would like to acknowledge Hannah Carter for being the best advisor one can ever ask for. Thank you for your guidance and support as I navigated my way through this PhD. For listening to my ideas and thoughts, for answering my questions, for countless hours of meetings and discussions, and for many amazing lab hangouts and activities. I am very grateful to have had the opportunity to be a part of your lab, and I hope that we continue to work together in the future.

I would like to acknowledge and thank my committee, Prashant Mali, Vineet Bafna, Olivier Harismendy, and Nathan Lewis, for their continues support and guidance throughout my PhD and for providing invaluable feedback on my research. I would like to especially thank Prashant Mali for being a wonderful collaborator.

I would like to acknowledge my friends for making my graduate experience so enjoyable, fun, and enriching. For making me feel that I have a new home here, and a new family. Especially my lab, Michelle, Andrea, Meghana, Clarence, Adam, James, and many other members. Thank you for making graduate work not just bearable, though only just bearable at times, but also exciting. Thank you for all the fun and crazy activities, and for your support and friendship, this has been truly the best lab one can imagine.

Finally, I would like to acknowledge my parents, Alev and Kemal, for their unconditional support and love. Thank you for raising me to be strong and independent, and encouraging me in pursuing my dreams, however difficult they may be. Thank you for always being there when I need. For being caring, loving, fun, and all around amazing. I really couldn't have done it without you.

The introduction, chapter forewords and conclusion include reformatted reprints of the materials as it appears in "The emerging potential for network analysis to inform precision cancer medicine" in Journal of Molecular Biology, 2018 by Kivilcim Ozturk, Michelle Dow, Daniel E. Carlin, Rafael Bejar, and Hannah Carter; and in "Integrating molecular networks with genetic variant interpretation for precision medicine" in Wiley Interdisciplinary Reviews: Systems Biology and Medicine, 2019 by Emidio Capriotti, Kivilcim Ozturk, and Hannah Carter. The dissertation author was a primary author of the first review paper, and a secondary author of the second review paper.

Chapter 1, in full, is a reformatted reprint of the material as it appears in "Predicting functional consequences of mutations using molecular interaction network features" in Human Genetics, 2021 by Kivilcim Ozturk and Hannah Carter. The dissertation author was a primary investigator and author of this paper.

Chapter 2, in full, is a reformatted reprint of the material as it appears in "Elevated neoantigen levels in tumors with somatic mutations in the HLA-A, HLA-B, HLA-C and B2M genes" in BMC Medical Genomics, 2019 by Andrea Castro, Kivilcim Ozturk, Rachel M. Pyke, Su Xian, Maurizio Zanetti, and Hannah Carter. The dissertation author was a primary investigator and author of this paper.

Chapter 3, in full, is a reformatted reprint of the material currently being prepared for submission for publication as "Interface-guided phenotyping of coding variants of transcription factor RUNX1 with SEUSS" by Kivilcim Ozturk, Rebecca Panwala, Jeanna Sheen, Prashant Mali, and Hannah Carter. The dissertation author was a primary investigator and author of this paper.

VITA

2010 Bachelor of Science in Biological Sciences and Bioengineering, Sabanci University

2015 Master of Science in Computer Science and Engineering, Sabanci University

2022 Doctor of Philosophy in Bioinformatics and Systems Biology, University of California San Diego

PUBLICATIONS

Ozturk K, Carter H. Predicting functional consequences of mutations using molecular interaction network features. Hum Genet. 2022;141: 1195–1210.

Chen K*, **Ozturk K***, Liefeld T, Reich M, Mesirov JP, Carter H, Fraley SI. A phenotypically supervised single-cell analysis protocol to study within-cell-type heterogeneity of cultured mammalian cells. STAR Protoc. 2021;2: 100561.

Chen K, **Ozturk K**, Contreras RL, Simon J, McCann S, Chen WJ, Carter H, Fraley SI. Phenotypically supervised single-cell sequencing parses within-cell-type heterogeneity. iScience. 2021;24: 101991.

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

Monzon AM, Carraro M, Chiricosta L, Reggiani F, Han J, **Ozturk K**, Wang Y, Miller M, Bromberg Y, Capriotti E, Savojardo C. Performance of computational methods for the evaluation of pericentriolar material 1 missense variants in CAGI-5. Hum Mutat. 2019;40: 1474–1485.

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

Fong SH, Carlin DE, **Ozturk K**, Ideker T. Strategies for Network GWAS Evaluated Using Classroom Crowd Science. Cell Systems. 2019. p. 414. doi:10.1016/j.cels.2019.09.001

Ozturk K, Carter H. Identifying Driver Interfaces Enriched for Somatic Missense Mutations in Tumors. Methods Mol Biol. 2019;1907: 51–72.

Capriotti E, **Ozturk K**, Carter H. Integrating molecular networks with genetic variant interpretation for precision medicine. Wiley Interdiscip Rev Syst Biol Med. 2019;11: e1443.

Ozturk K, Dow M, Carlin DE, Bejar R, Carter H. The Emerging Potential for Network Analysis to Inform Precision Cancer Medicine. J Mol Biol. 2018;430: 2875–2899.

*These authors contributed equally to this work.

ABSTRACT OF THE DISSERTATION

Advancing Variant Effect Prediction Beyond Protein-Level
by Incorporating Systems-Level Architecture

by

Kivilcim Ozturk

Doctor of Philosophy in Bioinformatics and Systems Biology

University of California San Diego, 2022

Professor Hannah Carter, Chair
Professor Prashant Mali, Co-Chair

Cancer is a complex disease that harbors substantial genetic heterogeneity. Recent advances in sequencing technologies revealed large numbers of somatic mutations across human tumors. However only a small proportion of these mutations are expected to contribute to tumor growth and progression, making determining functional mutations an important challenge in cancer genomics. Missense variants are particularly difficult to understand as they change only a single amino acid in a protein sequence yet can have large and varied effects on

protein activity. Numerous tools have been developed to identify missense variants with putative disease consequences from protein sequence and structure. However, biological function arises through higher order interactions among proteins and molecules within cells, and diseases are often associated with perturbations to protein interactions. Different perturbations can result in different phenotypes, and the level of impact caused by mutations to the underlying molecular interaction network may determine the likelihood of generating a disease phenotype. Thus, in this dissertation, I aim to incorporate systems-level architecture to bridge the gap between genotype and phenotype in cancer by exploring different network-based strategies to study the impact of variants on biological systems. I first integrated protein structure and network information to design variant features that capture orthogonal information to classical amino acid features and showed their potential to improve variant classification within a machine-learning framework. Next, I investigated how patterns of network rewiring of mutations on cancer genes can be informative for unearthing different selective oncogenic pressures. Finally, I examined transcriptomic effects of perturbation of distinct protein interactions as a way to better define the landscape of prospective phenotypes reachable by individual amino acid substitutions. Overall, this body of work demonstrates that variant effect interpretation can be significantly improved by incorporating information about the role of proteins and their molecular interactions within biological systems.

INTRODUCTION

Cancer is a complex disease that harbors substantial genetic heterogeneity. Recent advances in sequencing technologies have significantly reduced the costs of genome sequencing, allowing detection of vast numbers of somatic mutations across human tumors. Studies of the mutational landscape of tumor genomes determined that the majority of detected mutations are passengers that are not involved in the oncogenic process [1–3]; while only a small number of causal driver mutations are expected to contribute to tumor growth and progression [4]. Therefore, determining causal driver mutations and the genes they target has become an important challenge in cancer genomics.

Multiple methods exist to identify candidate driver mutations based on signatures of functional impact using information about protein sequence and structure [2,5,6]. Another approach is to look for signatures of positive selection for mutations at the gene level. Genes can be prioritized based on signatures of positive selection, including elevated mutation rate relative to expectation [7,8] or unexpected clustering within protein sequence or structure [9,10]. Then mutations within such genes can be classified as drivers or passengers according to their predicted effect on the function of the gene. Typically, these tools rely on protein sequence and structure information to predict variant effects at the protein level, and the scores they provide tend to capture coarse grained estimates of impact (e.g., damaging, benign, and tolerated).

Systems biology provides a toolkit for modeling complex biological systems and their constituent interactions. Increasingly, techniques from systems biology such as network analysis are being applied to analyze genomic data in order to better understand disease [11]. Networks have been employed in the study of tumor cohorts in order to gain insights into tumor biology, to identify putative biomarkers and relevant disease subtypes, and to find possible targets for therapy.

In these scenarios, networks are often used to control heterogeneity and to increase statistical power through aggregation. The structures of the networks can be defined from known molecular interactions (e.g., protein–protein interactions), such that the network itself is a model of the biological system. This confers meaning to interactions within the network such that connected nodes are expected to be more functionally related than distant nodes, and traversal of adjacent edges in the network can be loosely interpreted as biological information flow.

Biological functions and cellular behaviors arise from interactions among proteins and other molecules within cells, and diseases are often associated with perturbations to protein interactions. Different perturbations can result in different phenotypes [12], and the level of impact caused by mutations to the underlying molecular interaction network may determine the likelihood of generating a phenotype [13]. Thus, a protein's location within the system provides biological context that may be important for understanding the effects of mutations [14].

Proteins themselves also frequently have multiple functions via complex interaction dynamics with multiple partners. Therefore, different mutations on the same protein may have different effects on its functions [15,16]. While destabilizing mutations at the core of a protein are likely to interfere with all protein activities, mutations on the surface could potentially interfere with specific protein activities while preserving others [16]. It follows that different mutations targeting the same protein might perturb its interactions differently, affecting different pathways that the protein is involved in, resulting in different disease phenotypes [17]. Indeed, analyses have demonstrated an unexpected enrichment of Mendelian mutations [18–20] and somatic mutations [10,21–23] at protein interaction interfaces. Although protein structure-derived features have long been integral to variant effect interpretation, some more recent features capturing 3D location of mutations within key protein regions including local density of mutation and location at interface

regions have emerged [24–27]. While these features begin to capture information about the potential of variants to affect distinct interactions, they do not incorporate context about the importance of specific interactions within the larger interactome.

As networks impact the potential of variants to have a phenotypic effect, and variants cause phenotypic effects with rewiring of molecular networks, it should be possible to use network information to study variants. Thus, in this dissertation, I aim to incorporate systems-level architecture to bridge the gap between genotype and phenotype in cancer by exploring different network-based strategies to study the impact of variants on biological systems. Chapter 1 seeks to assess the potential for artificial intelligence-based methods for variant interpretation to derive new information from molecular interaction data. I achieved this by first integrating protein structure and network information to enable mapping of mutations to network edges to describe their potential to impact biological function. Then I designed features capturing information about proteins and amino acids in the context of their importance to the network architecture and evaluated them within a machine-learning variant classification framework, and found that network-based features capture orthogonal information to classical amino acid sequence/structure-based features, and can improve variant classification. Chapter 2 investigates how patterns of network rewiring of mutations on cancer genes can be informative for unearthing different selective oncogenic pressures. Here, I focus on a distinct pattern of network architecture perturbation in the tumor suppressor B2M subnetwork, and show that the mutational pattern targeting B2M's interaction with HLA-A, HLA-B, and HLA-C proteins could be indicative of immune evasion type pressures. I explored this with a mechanistic investigation of the interaction in terms of mutation burden, antigen binding affinity, and immune infiltration in tumors, which provides insight into how mutations perturbing this interaction enable immune escape. Chapter 3

examines transcriptomic effects of perturbing distinct protein interfaces as a way to better define the landscape of prospective phenotypes reachable by individual amino acid substitutions. I investigated this by employing an interface-guided Perturb-seq style approach to generate mutations at physical interfaces of the transcription factor RUNX1 with the potential to perturb different interactions, and therefore produce transcriptional readouts implicating different aspects of the RUNX1 regulon. I analyzed these readouts to identify functionally distinct groups of RUNX1 mutations, characterize their effects on cellular programs and study the implications for cancer mutations. Overall, my research helps advance variant effect interpretation in cancer beyond protein-level by integrating the architecture of biological systems.

Acknowledgements

The introduction, in part, includes reformatted reprints of the materials as it appears in "The emerging potential for network analysis to inform precision cancer medicine" in Journal of Molecular Biology, 2018 by Kivilcim Ozturk, Michelle Dow, Daniel E. Carlin, Rafael Bejar, and Hannah Carter; and in "Integrating molecular networks with genetic variant interpretation for precision medicine" in Wiley Interdisciplinary Reviews: Systems Biology and Medicine, 2019 by Emidio Capriotti, Kivilcim Ozturk, and Hannah Carter. The dissertation author was a primary author of the first review paper, and a secondary author of the second review paper.

References

1. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW. Cancer Genome Landscapes. Science. 2013;339: 1546–1558.

2. Carter H, Chen S, Isik L, Tyekucheva S, Velculescu VE, Kinzler KW, et al. Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. Cancer Res. 2009;69: 6660–6667.

3. Pleasance ED, Cheetham RK, Stephens PJ, McBride DJ, Humphray SJ, Greenman CD, et al. A comprehensive catalogue of somatic mutations from a human cancer genome. Nature. 2010;463: 191–196.

4. Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, et al. Patterns of somatic mutation in human cancer genomes. Nature. 2007;446: 153–158.

5. Shihab HA, Gough J, Cooper DN, Day INM, Gaunt TR. Predicting the functional consequences of cancer-associated amino acid substitutions. Bioinformatics. 2013. doi:10.1093/bioinformatics/btt182

6. Torkamani A, Schork NJ. Prediction of Cancer Driver Mutations in Protein Kinases. Cancer Res. 2008;68: 1675–1682.

7. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013;499: 214–218.

8. Tokheim CJ, Papadopoulos N, Kinzler KW, Vogelstein B, Karchin R. Evaluating the evaluation of cancer driver genes. Proceedings of the National Academy of Sciences. 2016;113: 14330–14335.

9. Tamborero D, Gonzalez-Perez A, Lopez-Bigas N. OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. Bioinformatics. 2013;29: 2238–2244.

10. Kamburov A, Lawrence MS, Polak P, Leshchiner I, Lage K, Golub TR, et al. Comprehensive assessment of cancer missense mutation clustering in protein structures. Proc Natl Acad Sci U S A. 2015;112: E5486–95.

11. Barabási A-L, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nat Rev Genet. 2011;12: 56–68.

12. Vidal M, Cusick ME, Barabási A-L. Interactome networks and human disease. Cell. 2011;144: 986–998.

13. Capriotti E, Ozturk K, Carter H. Integrating molecular networks with genetic variant interpretation for precision medicine. Wiley Interdiscip Rev Syst Biol Med. 2019;11: e1443.

14. Ozturk K, Dow M, Carlin DE, Bejar R, Carter H. The Emerging Potential for Network Analysis to Inform Precision Cancer Medicine. J Mol Biol. 2018;430: 2875–2899.

15. Sahni N, Yi S, Zhong Q, Jailkhani N, Charloteaux B, Cusick ME, et al. Edgotype: a fundamental link between genotype and phenotype. Curr Opin Genet Dev. 2013;23: 649–657.

16. Zhong Q, Simonis N, Li Q-R, Charloteaux B, Heuze F, Klitgord N, et al. Edgetic perturbation models of human inherited disorders. Mol Syst Biol. 2009;5: 321.

17. Engin HB, Hofree M, Carter H. Identifying mutation specific cancer pathways using a structurally resolved protein interaction network. Pac Symp Biocomput. 2015; 84–95.

18. David A, Razali R, Wass MN, Sternberg MJE. Protein-protein interaction sites are hot spots for disease-associated nonsynonymous SNPs. Hum Mutat. 2012;33: 359–363.

19. Guo Y, Wei X, Das J, Grimson A, Lipkin SM, Clark AG, et al. Dissecting disease inheritance modes in a three-dimensional protein network challenges the "guilt-by-association" principle. Am J Hum Genet. 2013;93: 78–89.

20. Wang X, Wei X, Thijssen B, Das J, Lipkin SM, Yu H. Three-dimensional reconstruction of protein networks provides insight into human genetic disease. Nat Biotechnol. 2012;30: 159–164.

21. Engin HB, Kreisberg JF, Carter H. Structure-Based Analysis Reveals Cancer Missense Mutations Target Protein Interaction Interfaces. Srinivasan N, editor. PLoS One. 2016;11: e0152929.

22. Porta-Pardo E, Garcia-Alonso L, Hrabe T, Dopazo J, Godzik A. A Pan-Cancer Catalogue of Cancer Driver Protein Interaction Interfaces. Nussinov R, editor. PLoS Comput Biol. 2015;11: e1004518.

23. Raimondi F, Singh G, Betts MJ, Apic G, Vukotic R, Andreone P, et al. Insights into cancer severity from biomolecular interaction mechanisms. Sci Rep. 2016;6: 34490.

24. Iqbal S, Pérez-Palma E, Jespersen JB, May P, Hoksza D, Heyne HO, et al. Comprehensive characterization of amino acid positions in protein structures reveals molecular effect of missense variants. Proc Natl Acad Sci U S A. 2020;117: 28201–28211.

25. Laskowski RA, Stephenson JD, Sillitoe I, Orengo CA, Thornton JM. VarSite: Disease variants and protein structure. Protein Sci. 2020;29: 111–119.

26. Tokheim C, Bhattacharya R, Niknafs N, Gygax DM, Kim R, Ryan M, et al. Exome-Scale Discovery of Hotspot Mutation Regions in Human Cancer Using 3D Protein Structure. Cancer Res. 2016;76: 3719–3731.

27. Tokheim C, Karchin R. CHASMplus Reveals the Scope of Somatic Missense Mutations Driving Human Cancers. Cell Syst. 2019;9: 9–23.e8.

CHAPTER 1: Revisiting the impact of coding variants through the lens of biological systems architecture

1.1 Foreword

The analysis of genomic data and its relation to phenotypes is a fundamental step for study of human diseases. Although in the last decades many studies have uncovered genetic variants associated with diseases [1], the complexity of biological systems makes it difficult to evaluate their impact at the level of cellular processes and behaviors. Networks provide a versatile framework for modeling the architecture of biological systems, wherein nodes usually represent biomolecules, and edges represent interactions among them. Graphical modeling then allows quantitative measures to be derived from the network topology for analyzing different aspects of biological systems. The network structure itself can be analyzed, or biological measurement data can be mapped onto network nodes and edges to facilitate integration or interpretation of those measurements to capture organization of the underlying system. To link genotype to phenotype, genetic alterations can be mapped onto their respective proteins to identify PPIs and biochemical pathways that are potentially affected.

Studies of PPI network topology have generated multiple insights linking protein location and connectivity within the network to particular phenotypes. The characteristics of PPI network topology that enable function and robustness of biological systems also create certain kinds of vulnerability. PPI networks tend to have a scale-free topology, such that the number of edges connecting to each node follows a power-law distribution [2], meaning that only a minority of nodes are hubs with large numbers of interaction partners while most nodes participate in only a few interactions. This is thought to render the system robust to random error, since genetic variants at random are more likely to affect a protein with few interactors, and thus cause only a minor

perturbation to the overall topology of the network. However, this leads to vulnerabilities, as mutations affecting a highly connected hub are likely to have a significant impact on the system and can be very disruptive to function [3].

Various network metrics have been developed to describe characteristics of nodes within networks [4], and can be used to identify nodes with key "roles" within the network topology [5]. Centrality measures can capture the importance of a node to information flow in a network. For example, degree centrality, quantifying the number of interaction partners, can be used to designate proteins as hubs or peripheral nodes; while betweenness centrality describes the number of shortest paths that traverse a node, inspecting it as a potential bottleneck (Figure S1.6). Closeness centrality measures the overall closeness to all other nodes in the network; while eigenvector centrality informs centrality of the neighbors by quantifying if a node is connected to other high-degree nodes (Figure S1.6), and the clustering coefficient measures the embeddedness of the immediate network neighborhood of a node. Nodes can also be assigned to modules within the network using community detection algorithms [6,7].

Given the clear evidence that nodes with distinct characteristics in the network support different aspects of the function of biological systems and are under different evolutionary constraints, it makes sense to evaluate the implications for fitness-related phenotypes. Many studies have taken advantage of network measures to examine different classes of genes. Essential genes encode proteins that are required for organismal survival, such that the loss of the gene is lethal. These genes are reported to have higher degrees in the PPI networks [8], higher betweenness centralities [9], and higher clustering coefficients [10,11]. Cancer genes are also found to be highly central by several studies [12–15]. In contrast, Mendelian disease genes are found to be less central than essential and cancer genes [12,13], particularly when essential Mendelian genes are excluded

from the analysis [13]. Interestingly, disease genes associated with dominant disorders are found to have higher degrees than genes associated with recessive disorders [10]. In contrast, gene deletion at the periphery of the network is less frequently associated with an essential or disease phenotype [12,16].

Most genetic variants are not loss of function events, but rather result in more subtle changes to protein sequences, and mutations within the same protein can have very different effects on its function [17]. It has been shown experimentally that most nonsynonymous Mendelian disease mutations generate stable proteins, supporting that mutation effects on specific protein activities rather than absence of protein drives disease phenotypes [18]. Mapping variants to nodes in the network cannot capture such subtle differences in variant effect, however the integration of information about protein structure and functional sites with protein interactions has made it possible to better discriminate variants in some cases by mapping them to network edges. Das et al. [19], Sahni et al. [20] and Zhong et al. [21] introduced the concept of "edgetics" to describe the potential of mutations to perturb distinct interactions in which a protein participates. Under this model, variants mapping to the core have the potential to eliminate all interactions by destabilizing the protein, while variants mapping to interaction interfaces have the potential to perturb specific subsets of interaction.

Studies of edgetic effects require information about the three-dimensional structure of protein complexes, so that amino acid residues can be labeled according to their location in the protein core, on the surface or at an interface between interacting proteins (Figure S1.7). The framework of edgetics thus allows variants to be studied not only in the context of their location in the network, but also according to their direct impact on network topology. Structurally resolved interactome networks, which integrate information about the domains or amino acid residues that

physically interact, are increasingly available to explore the mechanisms by which mutations cause disease at scale [19,22–25]. Multiple studies using structurally resolved networks revealed a statistical excess of known disease mutations at protein interaction interfaces [26–28], with in-frame disease mutations enriched at interfaces relative to truncating mutations [28], confirming the utility of such networks for systematically investigating disease mechanisms.

My research in Chapter 1 aims to assess the potential for artificial intelligence-based methods for variant interpretation to derive new information from molecular interaction data, and shows that network-based features capture orthogonal information to classical amino acid sequence/structure-based features, and can improve variant classification.

1.2 Abstract

Variant interpretation remains a central challenge for precision medicine. Missense variants are particularly difficult to understand as they change only a single amino acid in a protein sequence yet can have large and varied effects on protein activity. Numerous tools have been developed to identify missense variants with putative disease consequences from protein sequence and structure. However, biological function arises through higher order interactions among proteins and molecules within cells. We therefore sought to capture information about the potential of missense mutations to perturb protein interaction networks by integrating protein structure and interaction data. We developed 16 network-based annotations for missense mutations that provide orthogonal information to features classically used to prioritize variants. We then evaluated them in the context of a proven machine-learning framework for variant effect prediction across multiple benchmark datasets to demonstrate their potential to improve variant classification. Interestingly, network features resulted in larger performance gains for classifying somatic mutations than for germline variants, possibly due to different constraints on what mutations are tolerated at the cellular versus organismal level. Our results suggest that modeling variant potential to perturb context-specific interactome networks is a fruitful strategy to advance in silico variant effect prediction.

1.3 Introduction

Advances in high throughput sequencing technologies have resulted in the rapid accumulation of genomic data and allowed profiling of patient genomes in clinical settings. Such studies frequently uncover previously unobserved and uncharacterized genetic variants of ambiguous relevance to health, making variant interpretation an important challenge in precision medicine [29]. Missense mutations are particularly challenging as they only change a single amino acid in a protein sequence yet can have effects spanning no difference to complete loss of function. Numerous methods have been developed to prioritize functional missense variants [30–38]. Typically, these tools rely on protein sequence/structure information to predict variant effects at the protein level, and the scores they provide tend to capture coarse grained estimates of impact (e.g damaging, benign, tolerated).

Biological functions and cellular behaviors arise from interactions among proteins and other molecules within cells, and biological systems evolve to be robust to random error [39]. Diseases are often associated with perturbations to protein interactions, different perturbations can result in different phenotypes [40], and the level of impact caused by mutations to the underlying molecular interaction network may determine the likelihood of generating a phenotype [41]. For example, loss of function mutations were more likely to be tolerated when they affected proteins at the periphery of the interactome [16]. Similarly, variants that otherwise were predicted to have little effect were more likely to be deleterious if they had a large number of interaction partners [42] and de novo missense variants in autism probands with functional Polyphen2 predictions were enriched at protein interfaces of more central proteins relative to similar mutations in control siblings [43]. Thus, a protein's location within the system provides biological context that may be important for understanding the effects of mutations [44].

Within proteins, different mutations may have different effects on protein functions [20,21]. While destabilizing mutations at the core of a protein are likely to interfere with all protein activities, mutations on the surface could potentially interfere with specific protein activities while preserving others [21]. In this way, different mutations targeting the same protein might perturb its interactions differently, affecting different pathways that the protein is involved in, and resulting in different disease phenotypes [45]. Indeed, analyses have demonstrated an unexpected enrichment of Mendelian mutations [26–28], and somatic mutations [46–49] at protein interaction interfaces. Although protein-structure derived features have long been integral to variant classification, some more recent features capturing 3D location of mutations within key protein regions including local density of mutation and location at interface regions have emerged [50–53]. While these features begin to capture information about the potential variants to affect distinct interactions, they do not incorporate context about the importance of specific interactions within the larger interactome.

Based on the above we sought to assess the potential for artificial intelligence based methods for variant interpretation to derive new information from molecular interaction data. We first integrated structure and protein-protein interaction (PPI) networks to enable systematic annotation of proteins according to location and interactions (Figure 1.1a). We mapped various germline variants and somatic mutations to network edges to describe their potential to impact biological function (Figure 1.1b). We then designed features capturing information about proteins and amino acids in the context of their importance to the network architecture and evaluated them within a machine-learning variant classification framework (Figure 1.1c). We found that network-based features capture orthogonal information to classical amino acid (AA) sequence/structure-

based features and can improve variant classification, though they may be more informative for some variant classification tasks than others.

1.4 Results

Disease causing genes are central in PPI networks. The architectures of biological networks can provide important information for understanding the pathogenesis of mutations [44,54]. The scale-free topology of PPI networks suggests that they are more tolerant to random failures, but variants affecting higher degree nodes are more likely to disrupt function [3]. Indeed, when we compared disease genes using a high-confidence human PPI network of experimentally verified interactions from STRING [55], cancer driver [56] and Mendelian disease genes [1] score higher with various centrality measures than other genes (Figure 1.2). This suggests that the network niche of a gene provides information about the potential of an amino acid substitution to create deleterious phenotypes, a relationship that has proven robust to study bias [57]; in our data, only node degree correlates weakly with the number of publications (Pearson r=0.23, Figure S1.1).

Creating a structurally resolved PPI network. While disease mutations target proteins more central in interaction networks (Figure 1.2), protein level descriptors of centrality are not capable of distinguishing the effects of different mutations within proteins. Investigation of residue-specific network perturbations requires mapping of mutations to 3D protein structures and interaction interfaces so that we can model their potential to affect network edges (Figure 1.1b). We constructed a structurally resolved PPI network (called SRNet from here on) comprising 6,230 proteins and 10,615 PPIs using 3D structures and homology models (Figure 1.1a). This network contains annotations for 530,668 interface residues, defined here as the subset of amino acid residues that mediate physical contact between proteins. Otherwise, amino acids are annotated according to location at the surface or core based on relative solvent accessible surface area calculated from protein 3D structures (Methods). SRNet is an updated and extended version of our previous structurally resolved PPI network [46].

Disease mutations frequently target interface or core residues. We further assessed the potential for SRNet to capture information about residue-based network-perturbation by analyzing location of mutations relative to core, surface or interface regions. Similar to the finding in Engin et al. [46], SRNet supports that somatic missense mutations in tumors (obtained from The Cancer Genome Atlas (TCGA) [58]) target surface regions in oncogenes (OR=1.32, p=1.4e-06) and other genes (OR=1.15, p=1.07e-59), but are relatively depleted at surface regions in tumor suppressor genes (OR=0.91, p<0.1) due to a larger proportion of core mutations (Figure 1.3a), consistent with more loss of function mutations in tumor suppressors. However, when focusing only on surface positions, somatic mutations are more likely to be found at interface regions of oncogenes (OR=1.11, p<0.05) and tumor suppressors (OR=1.30, p=7.8e-07) relative to other genes (Figure 1.3b). Analysis of pathogenic germline variants (ClinVar [59]) versus neutral variants (EXAC [60], SwissVar [61], ClinVar [59]) found similar trends. Pathogenic variants were relatively depleted at the surface (OR=0.56, p=1.5e-42), suggesting they were far more likely to affect core regions, whereas neutral variants were biased toward the surface (OR=1.69, p=1e-19) (Figure 1.3c). On protein surfaces, pathogenic variants were more often found at interface regions (OR=5.65, p=2.2e-308) though neutral variants also showed increased odds of affecting an interface (OR=2.87, p=2.6e-115) (Figure 1.3d).

Network-based features for variant classification. As the above analyses support that both protein and amino acid level information derived from networks is informative about disease-association, we hypothesized that network information would be useful for machine-learning-based variant classification. We designed and analyzed 16 features describing network-level effects of mutations, including 7 protein-level features (Figure 1.4a) that estimate the significance of the target protein in the network, and 9 residue-level features (Figures 1.4b-c) quantifying the

potential of individual amino acid positions on the protein to impact network architecture. The residue-level features are based on comparing network measures before and after removing edges in SRNet potentially affected by a mutation (Methods). These 16 features show potential to distinguish between different classes of variants (Figures 1.4a-c) and are not strongly correlated with other classic non-network based amino acid features used for variant classification, such as measures of site-specific conservation (Figure 1.4d), suggesting that they add new and useful information (Figure S1.2).

Utility of network features for classifying cancer driver mutations. To evaluate the benefit of using network features for somatic mutation classification, we trained a Random Forest to predict driver or passenger class labels using different combinations of features. We separately evaluated classifier performance when trained using all 16 network derived features, only the 7 protein-level features or the 9 residue-level features alone, or in combination with 83 amino acid level features obtained from the SNVBox database [62]. As a training set, we used likely-driver and likely-passenger missense mutations from Tokheim et al., which they obtained from TCGA using a semi-supervised approach based on known cancer driver gene annotations and mutation rates with the goal of generating a more balanced training set consisting of both driver and passenger mutations in cancer genes [52]. While passengers greatly outnumber drivers in practice, we constrained the ratio as 1:4 driver vs. passenger mutations for classifier training (Methods). Generalization error was estimated using a 5-fold cross-validation with gene hold out to prevent information leakage and consequent overfitting. We measured performance using the area under the ROC (auROC) and precision recall curve (auPRC) metrics, similar to prior variant effect prediction studies [52]. We note that use of network features limits training and prediction to mutations that can be annotated by SRNet.

For driver classification, protein-level network features performed better than residue-specific features (Figures 1.5a and S1.3a), though performance for residue-level features was better for the top scoring ~20% of drivers (left edge of ROC curve). We note that residue-level features alone classify all surface non-interface mutations as passengers since their feature values should all be the same (there is no change to network centrality measures when no edges are affected). Combining residue-level features with more classic amino acid level features significantly boosts performance over residue-level features alone (Figures 1.5b and S1.3b). Interestingly, network features alone slightly outperform amino acid level features alone, pointing to the extreme centrality of driver genes. As residue-level features are likely to be most informative for mutations at interfaces, we further explored performance for interface mutations only (Figure 1.5c). Here we see that residue-specific network features perform considerably better as they are not hindered by misclassification of surface mutations (Figure 1.5c). Overall, the combination of network-based and amino acid features displays the highest performance (Figures 1.5b-c and S1.3b-c). Notably, precision-recall curves indicate that incorporating both network and classic AA features resulted in a significant drop in false positive predictions relative to either type of feature alone (Figure S1.3a-c).

Incorporating in silico predicted interface residues. The restriction to analysis of mutations for which 3D structural information about interfaces is available is a problematic limitation. In silico prediction of interfaces can be used to augment interface coverage, as done for Interactome INSIDER [63]. To explore whether in silico predicted interfaces could boost mutation coverage without loss of performance, we repeated our analysis on an extended network with both structure-derived and predicted interfaces. This resulted in improved performance overall (Figures 1.5d-f and S1.3d-f), suggesting that improvements to interface features and the ability to train on

a larger set of mutations enabled by higher coverage in the expanded network outweigh the introduction of noise caused by interface prediction error. We also noted a larger gain in precision for network features relative to AA features when using expanding the network (Figure S1.3e). A more stringent comparison considering only proteins shared between the original and the expanded network found similar results (auROC is 0.832 and 0.871 for SRNet and the extended network for the classifier with Network & AA features respectively).

As we obtained our optimal performance using all features with the extended network, we used this classifier to evaluate feature importances. In Random Forest classifiers, feature importances is determined as the mean decrease in impurity when using that feature to split training examples according to class label during classifier training [64]. Fourteen of the top 21 features were network derived, including the top 9 (Figure S1.2). Protein-level network features were more informative than residue-level features, possibly reflecting the limitation of residue-level features to distinguish surface mutations. Simple 3D location annotating mutations to location at core, surface or interface contributed less information, which may reflect its redundancy with other network features.

We further investigated residue-level network features in the extended network by examining cases where the classifier was successful in differentiating between driver and passenger mutations occurring in the same proteins. Since residue-level features only vary within-protein for interface mutations, we looked for cancer genes where both driver and passenger mutations at interfaces were correctly classified. We found 7 cancer genes (EGFR, HRAS, KRAS, TP53, PIK3R1, CTNNB1 and PTEN) that contained both correctly classified interface driver and interface passenger mutations. Focusing on 212 correctly classified interface mutations in these genes, we observed a significant difference in distribution of residue-level features for the driver

and passenger classes (Figure 1.6), further supporting that residue-level network features provide information useful for within gene mutation classification.

Overall performance on benchmark datasets. We next sought to evaluate the improvement obtained from network features on independent studies of cancer mutations. We used our highest performing classifier, trained on cancer mutations that map to the extended network and all network-based and classic amino acid features (Figure 1.5e, Net & AA classifier with auROC=0.880, Figure S1.2). Since no 'gold standard' dataset exists for cancer, we evaluated classifier performance relative to best-in-class methods that do not use network-derived features on 4 external pan-cancer datasets constructed using different approaches: an in vivo screen: Kim et al. [65], an in vitro assay: Ng et al. [66], and 2 literature-derived datasets: MSK-IMPACT and CGC-recurrent, previously described in Tokheim et al. [52]. For each dataset, considering the mutations scored by all methods, classifier performance was evaluated using the area under the ROC (auROC) and PR curves (auPRC), accuracy, F1 score and the Matthews correlation coefficient (MCC).

We assessed the performance of our classifier relative to both cancer-specific: CHASM [67], ParsSNP [68], TransFIC [69], and CanDrA [70], and population-based methods: VEST [71], SIFT [32], PolyPhen [33], CADD [34], ClinPred [72], DANN [73], DEOGEN2 [74], FATHMM [75], LIST-S2 [76], LRT [77], M-CAP [78], MPC [79], MVP [80], MetaLR and MetaSVM [81], MutPred [38], MutationAssessor [82], MutationTaster [83], PROVEAN [84] and REVEL [36] (Figure 1.7). Comparison was based on the set of benchmark mutations scored by all methods, and was on the basis of auROC, auPRC, accuracy, F1 score and Matthew's correlation coefficient (MCC). We note the later three use discrete labels rather than continuous scores. We used provided labels or recommended cutoffs for all methods as possible, and used a cutoff of 0.5 otherwise.

Our classifier performed well on all 5 metrics across the 4 benchmark sets relative to most other methods. It had the highest auPRC (Figure 1.7b), F1 score (Figure 1.7d) and MCC (Figure 1.7e), of all methods on the 4 benchmark sets, and also performed well on auROC and accuracy (Figure 1.7a,c). In most cases, the difference in auROC relative to other methods was deemed significant by the DeLong test. After our method, the next best performing method was ParsSNP, after which there was considerable variation in what methods performed best by various measures. Overall, these results suggest that network-derived features that capture abstract information about the role of proteins in networks and the potential of mutations to perturb this role, are helpful for driver classification across a variety of settings, though the gains over methods trained on classic amino acid based features are modest.

We separately compared our approach to two methods that incorporate interactome related features: SuSPect [42] and CHASMplus [52]. SuSPect includes protein degree as a predictive feature, while CHASMplus now includes a feature indicating the number of interactions affected if a mutation occurs at a protein interaction interface, along with other improvements relative to the original method. Location at an interface was reported as the second most informative feature after a feature describing within protein clustering of observed mutations [52]. We noted improved performance relative to SuSPect and comparable performance to CHASMplus (Figure S1.4). We note that both our method and CHASMplus derive classic AA features from the SNVbox database [62]. We further analyzed mutations that were correctly classified by our method but misclassified by SuSPect or CHASMplus to see whether the network features implemented relative to these methods explained the difference. These mutations were significantly enriched at interface regions compared to surface or core (OR=3.78, p=1.45e-11) and they had significantly different distributions of all network features (Mann-Whitney U test, $p<0.05$) apart from closeness change

(p=0.22), when compared to mutations misclassified by our method but correctly classified by SuSPect or CHASMplus, suggesting that even though some shared features exist between the methods, our classifier better reflects the network rewiring by mutations.

However, it should also be considered that our network-based approach is dependent on the inclusion of proteins in the network and availability of annotations mapping amino acid residues to core, surface or interface residues. This generally results in a smaller training set than other methods, and an inability to score some fraction of mutations. For the benchmark sets evaluated here, 95.77% of Kim et al., 72.03% of Ng et al., 78.82% of MSK-IMPACT, and 74.11% of CGC recurrent dataset mutations could be scored respectively. It is possible that better network and amino acid annotation coverage could further boost performance.

Utility of network features for classifying pathogenic germline variants. We next evaluated whether network features are also useful in the context of germline variation. We previously observed that inherited disease genes were less central than cancer genes and both pathogenic and neutral mutations were enriched at interface positions, though to different extents. We once again trained a Random Forest classifier to prioritize missense mutations that alter protein activity using the 16 network-based features and 83 amino acid descriptors, using a training set composed of pathogenic and neutral variants (described in the section "Disease mutations frequently target interface or core residues").

For germline variants, residue-specific features yielded similar (with SRNet) or higher performance (with extended network) than protein-level features for all mutations (Figures 1.8a and S1.5a), and for interface mutations only (Figures 1.8c and S1.5c). But overall, network features are outperformed by non-network amino acid features (Figures 1.8b and S1.5b) which is the opposite of the case with cancer driver classifier. This is consistent with proteins targeted by

pathogenic germline variants being less central than cancer driver genes. Since proteins harboring germline pathogenic variants have fewer interaction partners, pathogenic variants in the protein core or at interfaces tend not to result in as extreme values of residue-level network features as driver mutations do (Figure 1.4c), despite the observed enrichment for these variants in core and interface regions (Figures 1.3c-d). The similar performance by network features in both SRNet and the extended network (Figure 1.8) suggests that either increased coverage does not improve performance as much, or the noise introduced by interface prediction error counteracts the performance gained by higher coverage in this setting. A stricter comparison considering only shared proteins between networks once again showed similar performance (auROC is 0.835 and 0.849 for SRNet and the extended network for the classifier with Network & AA features respectively). Precision-recall curves show similar results (Figure S1.5).

1.5 Discussion

Understanding the functional consequences of protein coding variants remains a challenging task. Machine learning methods developed thus far to predict whether a mutation is likely to impair protein activity or cause a pathogenic phenotype have largely been protein-centric, however a growing body of work points to perturbation of the interactome as a major determinant of pathogenicity [12,18,20,26–28,40,42,43,46–49,85–89]. Such studies of variant distribution in biological systems have provided insights as to how molecular interaction networks evolve to ensure robustness or vulnerability to genetic variation [41]. It is increasingly apparent that the role of proteins within molecular networks is a key determinant of the potential of variants to exert deleterious effects [16,42,43]. Motivated by these studies, we investigate here how network-derived features can capture novel information about variant effects that is not already present in the classical amino acid features used by most variant classification methods, and show that combining both sets of features improves classifier performance.

Our approach relies on a structurally resolved PPI network that allows variants to be characterized according to their potential to affect network architecture by mapping them to their location on protein structures and protein-interaction interfaces. These mappings are used to capture the potential of variant positions to perturb information flow through the network. We developed protein-level features to capture the relative importance of a protein within the network, and residue-level features to capture the potential of mutations to alter the architecture of the network. Though protein-level network features are shared by all variants in a protein, they nonetheless can interact with other amino acid level features to support classification; in all cases, combining both protein and residue-level network features with classic amino acid features outperformed combining only residue-level network features with classic amino acid features.

Residue-level features were helpful for distinguishing between variants within proteins, however because we designed them to capture the potential of mutations to alter the network architecture, all surface non-interface variants received the same value for these features. We also did not consider the possibility that surface mutations could generate a new edge in the network. Such mutations could be more common in cancer, where missense mutations have been reported to alter binding specificities for kinases and their substrates thereby remodeling network architectures [90].

Though network features show fairly different distributions for different classes of variants (Figure 1.4) and are orthogonal to the features typically used for variant classification, the best classifier combining both feature types show only modest gains over classifiers that use only classic amino acid features. This may result from more limited availability of training set mutations due to the requirement for structure and interface information to estimate network feature values. This requirement also constrained the coverage of benchmark set mutations that could be classified, though the values remained generally high, and over 70% in the worst case. Performance generally improved when we included predicted protein interactions from Interactome INSIDER [63], suggesting that in silico approaches may be an effective strategy to boost performance until more complete experimentally derived interaction maps are available.

Network features were more informative in the context of somatic mutations than when classifying inherited variants, though performance gains were observed in both cases. This is perhaps expected since inherited disease genes tended to be less central in PPI networks than cancer genes (Figure 1.2), and location at an interface by itself was less discriminatory in the germline setting (Figure 1.3d). We speculate that these differences may arise from different selective pressures acting on somatic versus germline variation. Because development at the

organismal level is likely dependent on the integrity of molecular interaction networks, both pathogenic and neutral variants may be more constrained by network architecture; whereas in cancer, where selection operates at the cellular level and is predominantly positive, mutations may be better tolerated and be more advantageous in central network positions. We note also, however, that training sets for the driver versus passenger classification problems tend to be more gene-centric leading to concerns over whether cancer mutation classifiers distinguish primarily between genes [91]. Although we made an effort to include both drivers and passengers in each driver gene to mitigate this, it may still be reflected in the higher utility of protein-level network features for driver classification.

In conclusion, our study suggests that information about molecular interaction networks can be incorporated into machine-learning based variant interpretation frameworks. This opens future directions for the development of novel features capturing network information. Since networks can be constructed to model cell-type and condition specificity [92], it may be possible to build classifiers that can capture context-specific effects of variants. Furthermore, as studies have shown different interfaces are associated with different protein activities, network-based features could make it possible for machine-learning methods to provide more insight into the potential for mutations within a protein to have distinct functional consequences. We anticipate such advances will boost the utility of variant classification tools for precision medicine applications.

1.6 Materials and Methods

Data and code are available at https://github.com/cartercompbio/NetFeatures.

Source of protein interaction data. To analyze disease gene centrality, we obtained a human PPI network of 12,811 proteins that are involved in 97,376 experimentally verified undirected interactions with a confidence score higher than 0.4 from STRING v11.0 [55].

Disease genes. A list of 125 high confidence cancer genes consisting of 54 oncogenes and 71 tumor suppressor genes was obtained from Vogelstein et al. [56]. We also obtained a list of 4524 Mendelian genes from the OMIM database [93]. These genes were used to evaluate disease gene centrality (Figure 1.2).

Collecting structural protein and protein interaction data. We obtained human protein interaction data (complete set) from Interactome3D [24], which contains a collection of a highly reliable set of experimentally identified human PPIs. We collected experimental co-crystal 3D structures for 5865 of these interactions from the Protein Data Bank (PDB) [94] and homology models for 5768 additional interactions [24] making a total of 11,633 interactions between 6807 proteins with structural protein and interaction data.

Creating a structurally resolved PPI network. Amino acid residues were annotated as participating in a protein interaction interface based on KFC2 [95] scores, and we removed interactions containing fewer than 5 interface residues on either partner. Additionally, we calculated relative solvent accessible surface areas (RSA) using NACCESS [96] for all residues in each protein structure. Residues with RSA<5% and RSA>15% were designated as core and surface residues respectively. Residues with RSAs between these thresholds were excluded from further analysis due to ambiguity. When multiple PDB chains were available for the same protein, we used the consensus designation as the final label. The mapping of PDB residue positions onto

UniProt residue positions was performed via PDBSWS web server [97]. After this mapping, we created a structurally resolved PPI network (named SRNet) of 6230 proteins and 10,615 undirected protein-protein interactions with a total of 530,668 interface residues. To extend coverage of the structurally resolved network, we defined an extended network based on the "High Confidence" dataset of Interactome INSIDER [63], a human PPI network of 14,445 proteins with 110,206 undirected interactions containing in silico interface residue predictions in addition to those derived from 3D structures.

Source of somatic mutation data. To investigate structural location of cancer mutations on proteins (Figure 1.3), we mapped more than 1.4 million somatic missense mutations from TCGA [58] onto the structurally resolved PPI network using structural annotations. Only mutations mapping to canonical proteins were used. After this mapping, we identified a total of 56,667 interface residues of 5005 proteins that are involved in 9235 interactions as mutated.

Training set for cancer mutation prediction. We collected a set of cancer missense mutations designated as likely-driver (n=2051) and likely-passenger (n=623,992) from Tokheim et al. [52]. Of these, 961 driver mutations from 32 genes and 28,043 passenger mutations from 2986 genes mapped to SRNet for a total of 29,004 mutations. All 32 genes with driver mutations also contain passenger mutations. To handle the driver vs. passenger mutation count imbalance in the training set by maintaining an approximate 1:4 driver vs. passenger mutation ratio similar to Carter et al. [67] while not overrepresenting particular genes, we limited the number of passenger mutations for each gene to 16 (median per gene driver mutation count) and collected 4626 passenger mutations at random across all genes with passenger mutations. In the extended network, we mapped 1513 driver mutations from 52 genes and 118,777 passenger mutations from 4478 genes for a total of 120,290 mutations. Thirty-eight of 52 genes with driver mutations also contain

passenger mutations. To maintain an approximate 1:4 driver vs. passenger ratio as described above, we limited the number of passenger mutations for each gene to 17 (median per gene driver mutation count) and collected 6549 passenger mutations at random across all genes with passenger mutations.

Training set for pathogenic variant prediction. We collected 5608 'pathogenic' variants from ClinVar [59], and 3418 neutral variants including 'common' variants (allele frequency>1%) from EXAC [60], variants with 'polymorphism' classification from SwissVar [61] and 'benign' variants from ClinVar [59], that map to SRNet, totaling 9026 missense variants. We also collected 21,819 pathogenic, and 35,522 neutral variants from the same databases that map to the extended network, totaling 57,341 missense variants.

Features. We designed 16 network-based features to quantify the potential impact of a mutation to the underlying network architecture, comprising 7 protein-level and 9 residue-level features. The 7 protein-level features are degree, betweenness, closeness, eigenvector and load centralities, clustering coefficient, and pagerank of the proteins within the PPI network. They are computed using the NetworkX package of Python and they aim to characterize the centrality of a protein in the network based on measures such as the number of nodes it is directly connected to (degree), the amount of shortest paths it is involved in (betweenness and load), the overall closeness to all other nodes (closeness), its embeddedness (clustering coefficient), and the centrality of its neighbors (eigenvector and pagerank). 9 residue-level features describe mutation 3D location (core, interface, surface) on the protein and changes in centrality of the protein within the PPI network resulting from mapping the mutation to network edges. Core mutations are assumed to affect all edges in the network, while interface mutations are mapped to corresponding edges in the network, and surface mutations retain all edges. Each interface mutation causes the

removal of all edges that they are mapped to. The remaining 8 residue-level features are based on this description of how the mutation perturbs the network by capturing degree change, betweenness change, closeness change, eigenvector change, clustering coefficient change, load change, pagerank change and percent degree change. Non-network related amino acid based features (n=83) obtained from the SNVBox database [62] describe substitution effects on amino acid biophysical properties, evolutionary conservation of variant sites, local sequence biases, and site-specific functional annotations. Pearson correlation coefficient was used to evaluate feature correlations. Our proposed classifier (used in Figure 1.7) uses a total of 99 features consisting of all 16 network-based features (7 protein-level and 9 residue-level features) and 83 non-network related amino acid based features. The importance of each feature is computed as the normalized total reduction of the criterion brought by that feature (mean decrease in impurity), also known as the Gini importance (Figure S1.2).

The number of PubMed studies featuring each gene was obtained from the NCBI database (https://ftp.ncbi.nih.gov/gene/DATA/gene2pubmed.gz) for all genes in SRNet with NCBI (Entrez) gene IDs. This was used to assess the potential for protein-level features to be affected by study bias.

Classifier training. We trained a Random Forest classifier (n_estimators=1000, max_features='sqrt') on the training set using the scikit-learn Python package. To avoid classifier overfitting, we performed prediction using a 5-fold gene hold out cross-validation by dividing the training set into 5 random folds for cross-validation while ensuring a balanced number of disease and neutral mutations across the folds. All mutations occurring in the same gene were kept within the same fold. The classifier score represents the percentage of decision trees that classify a mutation as a disease mutation (driver or pathogenic). Receiver Operator Characteristic (ROC)

and precision- recall curves were constructed from the classifier scores and the AUC statistic was used as a measure of classifier performance. To compare the performance of different features for identifying disease mutations, we trained different classifiers on different sets of features: all 16 network-based features (Net), dividing network-based features into 7 protein-level features (Prot) and 9 residue-level features (Res), 83 non-network amino acid (AA) features, 83 amino acid features combined with 9 residue-level features (AA & Res) or 83 amino acid features combined with all 16 network features (AA & Net). Training our proposed classifier (used in Figure 1.7) on 8062 cancer mutations (1513 driver and 6549 passenger mutations mapping to ExtNet) using all 99 features takes ~5.15 seconds using a Jupyter Notebook on a quad Intel Xeon E5-4650 v4 cpu with a total of 56/112 cores/threads and 512GB of RAM. Prediction of 100,000 mutations takes ~2.27 seconds.

Benchmark datasets. We obtained 4 pan-cancer benchmark sets of missense mutations consisting of an in vivo screen: Kim et al. [65], an in vitro assay: Ng et al. [66], and 2 literature-derived datasets: MSK-IMPACT and CGC-recurrent from Tokheim et al. [52]. The in vivo screen contains 71 mutations selected based on their presence in sequenced human tumors and screened in mice to assess oncogenicity and then labeled as 'functional' or 'neutral' based on their abundance [65]. The in vitro assay consists of 747 mutations from a growth factor dependent cell viability assay annotated as 'activating' for increased cell viability, or as 'neutral' for the remaining, with the assumption that a mutation yielding higher cell viability indicates driverness [66]. The MSK-IMPACT dataset is composed of mutations from approximately 10,000 tumors [98] on 414 cancer-related genes (MSK-IMPACT gene panel) labeled as positive class if annotated as 'oncogenic' or 'likely oncogenic' in OncoKB [99], or as negative class if not. The CGC-

recurrent dataset consists of TCGA mutations annotated as positive class if recurrent in a set of curated likely driver genes from the Cancer Gene Census [100], or as negative class if not.

Comparison to other methods. Performance was compared to 24 state-of-the-art methods that do not use network-based information, 4 cancer-focused methods: CHASM [67], ParsSNP [68], TransFIC [69], and CanDrA [70], and 20 population-based methods: VEST [71], SIFT [32], PolyPhen [33], CADD [34], ClinPred [72], DANN [73], DEOGEN2 [74], FATHMM (inherited-disease version) [75], LIST-S2 [76], LRT [77], M-CAP [78], MPC [79], MVP [80], MetaLR and MetaSVM [81], MutPred [38], MutationAssessor [82], MutationTaster [83], PROVEAN [84] and REVEL [36]. We obtained prediction scores for the mutations in the 4 benchmark sets described above for 20 of the methods (VEST, SIFT, PolyPhen, CADD, ClinPred, DANN, DEOGEN2, FATHMM, LIST-S2, LRT, M-CAP, MPC, MVP, MetaLR, MetaSVM, MutPred, MutationAssessor, MutationTaster, PROVEAN and REVEL) from the dbNSFP database (version 4.1a) [31] via the Ensembl Variant Effect Predictor (VEP) [101], and for 4 additional methods (CHASM, ParsSNP, CanDrA and TransFIC) from Tokheim et al. [52]. We also obtained scores on benchmark datasets for 2 additional methods that use network features: SuSPect [42] from (www.sbg.bio.ic.ac.uk/suspect) and CHASMplus from Tokheim et al. [52] (Figure S1.4).

Classifier performance was compared using the area under the ROC (auROC) and PR curves (auPRC), accuracy, F1 score and the Matthews correlation coefficient (MCC). Only the mutations scored by all methods were considered for comparison. Significance of difference of auROC measures are evaluated by DeLong test. auROC and auPRC values were computed using the predicted scores; while accuracy, F1 score and MCC were estimated based on the predicted labels (positive vs. negative). For label assignments, we used the provided labels from dbNSFP for SIFT, PolyPhen, ClinPred, DEOGEN2, FATHMM, LIST-S2, LRT, M-CAP, MetaLR,

MetaSVM, MutationAssessor, MutationTaster and PROVEAN. Methods that did not provide labels directly typically provided a score between 0 and 1 (except CADD and MPC), and we ensured that a higher score indicated a more damaging mutation. Where specified we used recommended score cutoffs (0.1 for ParsSNP, 0.7 for MVP, and 50 for SuSPect) for label assignments when evaluating F1 score, accuracy and MCC results. When no threshold was suggested (or if the suggestion was 0.5), we used a cutoff of 0.5 (our classifier Network&AA, CHASM, VEST, CanDrA, TransFIC, CADD, DANN, MPC, MutPred, REVEL, and CHASMplus). It is important to note that while auROC and auPRC results are independent of the predicted class labels; accuracy, F1 score and MCC results are dependent on the labels and the threshold used for their assignment; therefore assuming a cutoff of 0.5 could underestimate accuracy, F1 and MCC for some methods.

Statistical analysis. Distributions are compared using a Mann-Whitney U test. Correlations are evaluated using the Pearson correlation coefficient. Odds ratios are calculated using Fisher's exact test. auROC scores are compared using the DeLong test.

1.7 Figures

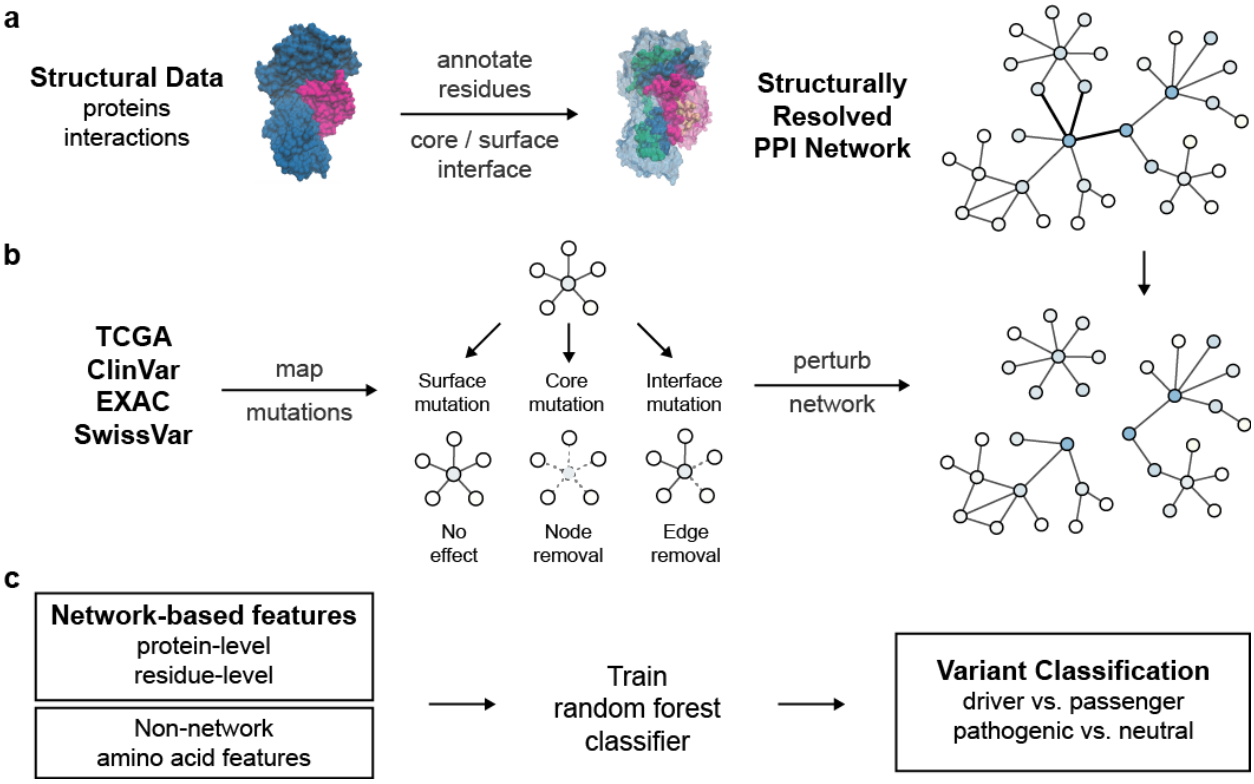


Figure 1.1. Overview of the method. (a) Constructing a structurally resolved PPI network. **(b)** Mapping mutations to perturbed network architectures. **(c)** Designing protein-level and residue-level network-based features and using a machine learning framework to evaluate their potential for variant classification alone and in combination with classic non-network amino acid features.

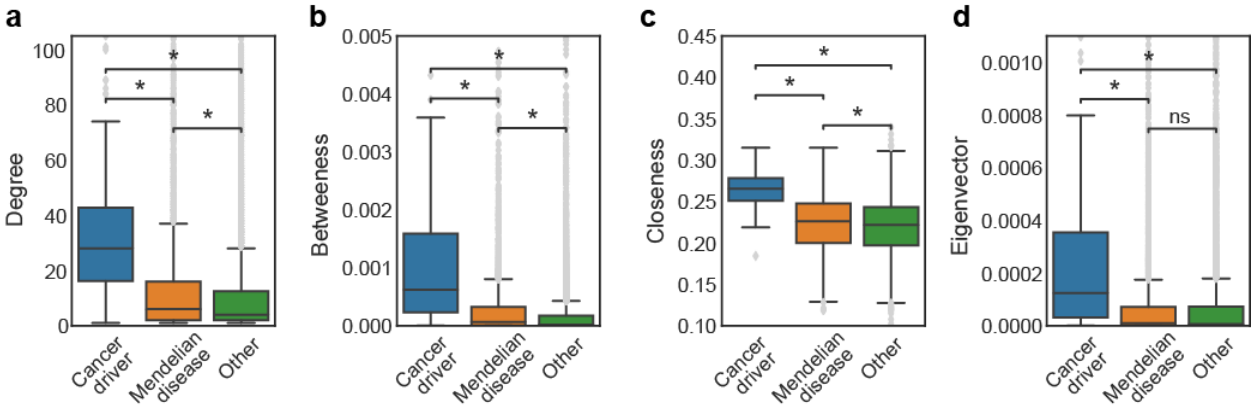


Figure 1.2. Disease genes are central in PPI networks. Boxplots showing distributions of **(a)** degree, **(b)** betweenness, **(c)** closeness, and **(d)** eigenvector centralities of cancer driver, Mendelian disease and other genes (Mann-Whitney U test with Bonferroni correction; *p<1e-04).

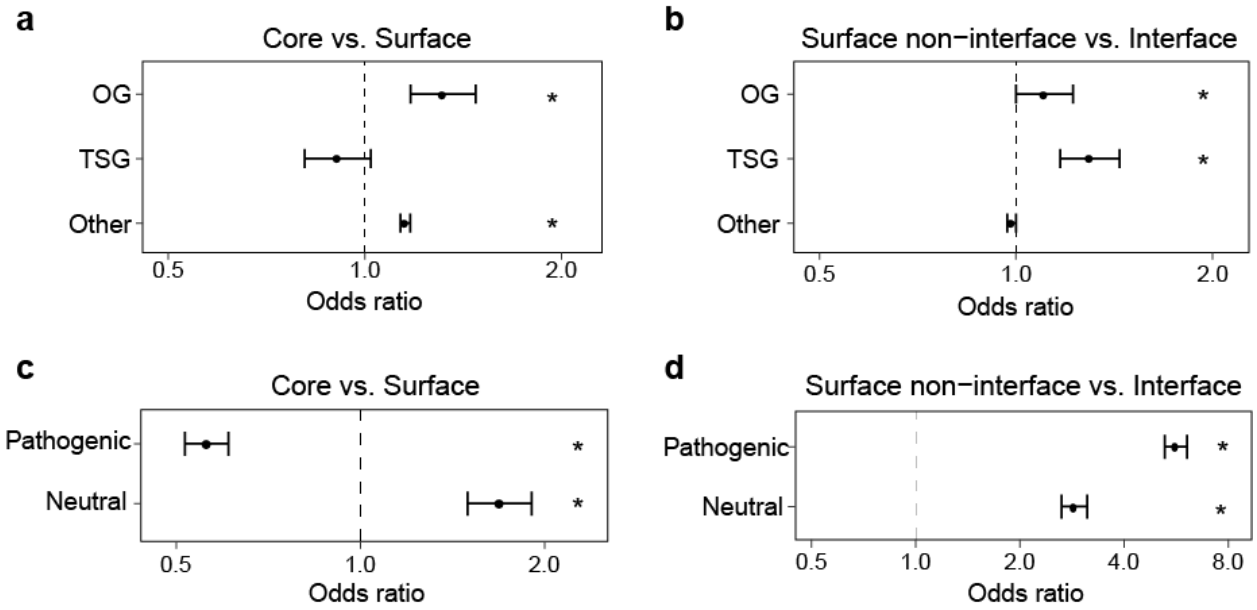
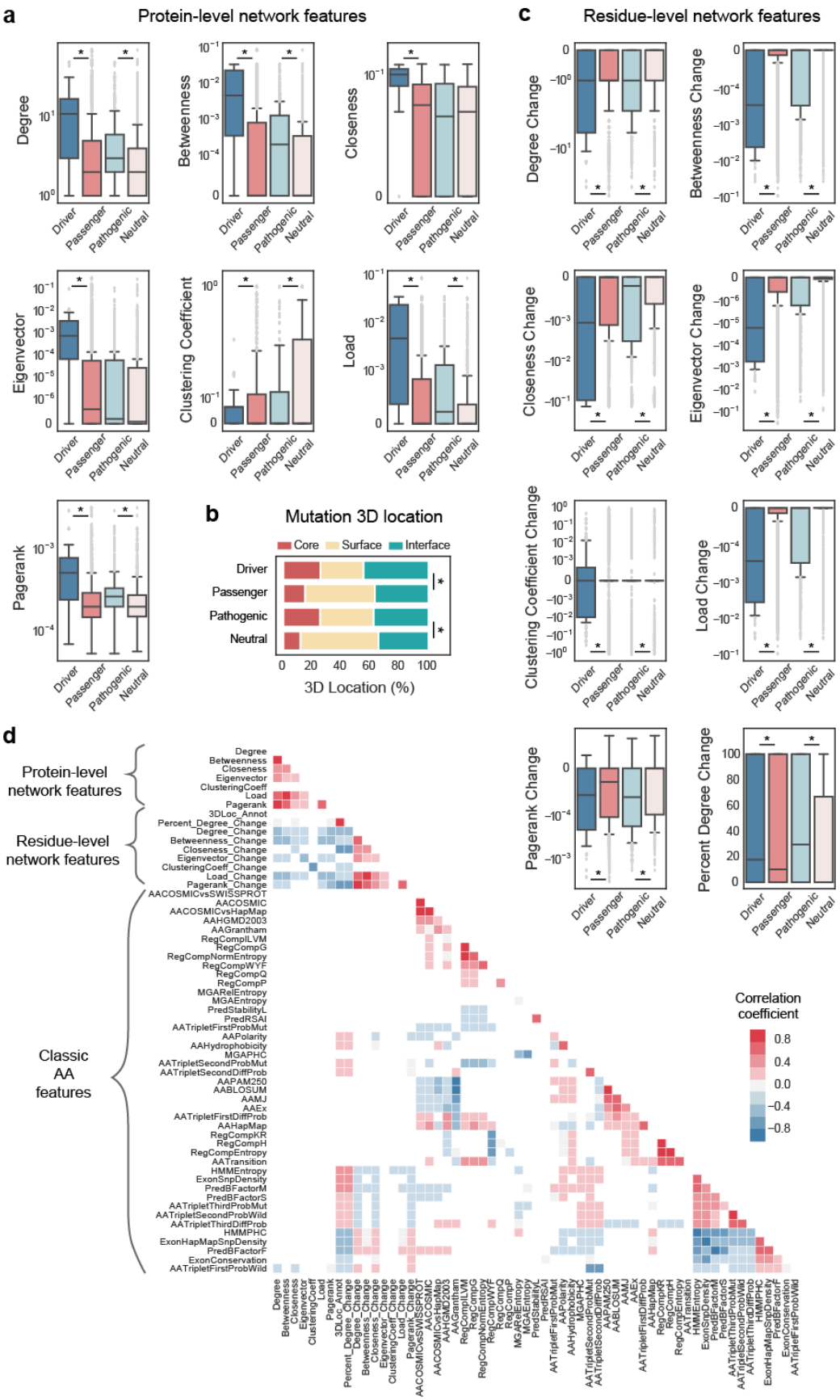


Figure 1.3. Analysis of structural location of missense disease mutations. Odds ratios (OR) and 95% confidence intervals using Fisher's exact test are shown (*p<0.05). **(a-b)** Comparison of somatic mutations in oncogenes (OG), tumor suppressor genes (TSG), and other genes located at **(a)** core vs. surface residues, **(b)** surface non-interface vs. surface interface residues. **(c-d)** Comparison of pathogenic and neutral variants located at **(c)** core vs. surface residues, **(d)** surface non-interface vs. surface interface residues. For **(a)** and **(c)**, an OR>1 means more mutations/variants were found at the surface. For **(b)** and **(d)** an OR>1 means more mutations/variants were found at interfaces.

Figure 1.4. Distribution of network-based features. Distribution of network-based features for driver vs. passenger mutations and pathogenic vs. neutral variants in SRNet (Mann-Whitney U test; *p<1e-04). **(a)** Boxplots showing distribution of protein-level features (degree, betweenness, closeness, eigenvector, clustering coefficient, load and pagerank), **(b)** a stacked bar plot showing percent distribution of 3D locations (core, interface, surface) of mutations, and **(c)** boxplots showing distribution of residue-level network features (degree change, betweenness change, closeness change, eigenvector change, clustering coefficient change, load change, pagerank change and percent degree change). **(d)** Heatmap displaying Pearson correlation coefficients of network-based and classic non-network based amino acid features. Only features that have at least one correlation coefficient higher than 0.3 and only values above 0.1 are shown. Classic amino acid features are ordered based on hierarchical clustering of correlation values.


a
Protein-level network features
Degree
Betweenness
Closeness
Eigenvector
Clustering Coefficient
Load
Pagerank
Driver
Passenger
Pathogenic
Neutral
b
Mutation 3D location
Core
Surface
Interface
3D Location (%)
c
Residue-level network features
Degree Change
Betweenness Change
Closeness Change
Eigenvector Change
Clustering Coefficient Change
Load Change
Pagerank Change
Percent Degree Change
d
Protein-level network features
Residue-level network features
Classic AA features
Correlation coefficient

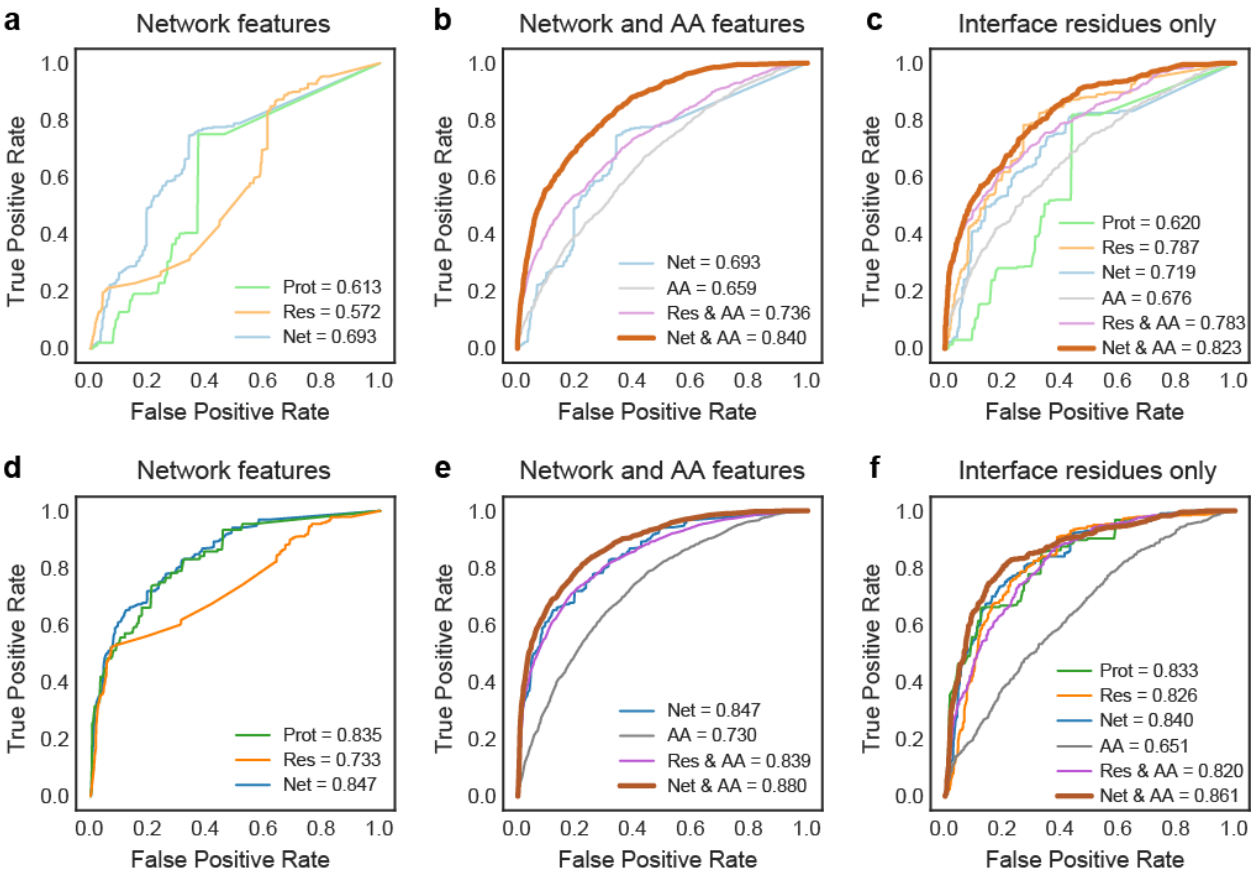


Figure 1.5. Classifier performance in identifying cancer mutations using SRNet vs. the extended network. ROC curves for identifying cancer mutations using **(a-b-c)** SRNet, and **(d-e-f)** the extended network, with **(a-d)** protein-level network features (Prot), residue-level network features (Res), and all network features (Net = Prot + Res); with **(b-e)** all network features, amino acid features (AA), residue-level network and amino acid features (Res & AA), and all network and amino acid features (Net & AA). **(c-f)** ROC curves for identifying cancer mutations targeting interface residues only, using all above-mentioned features. ROC curves using Net & AA features are bold. Performance is measured using auROC scores.

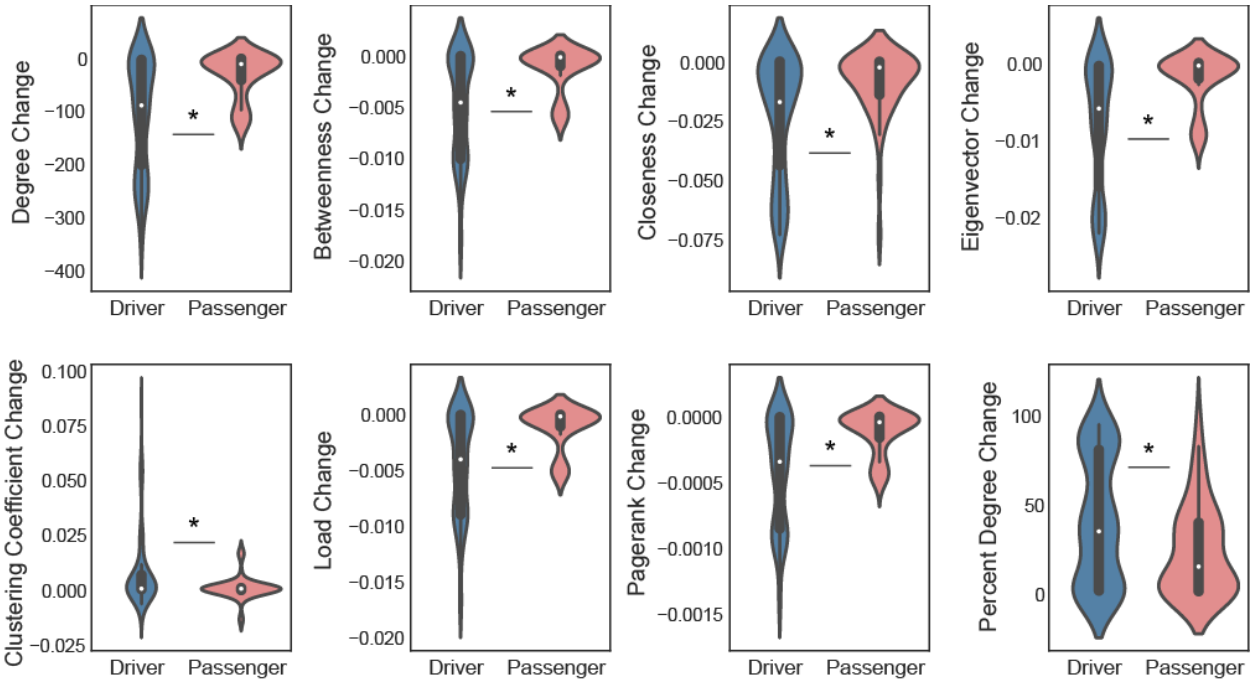
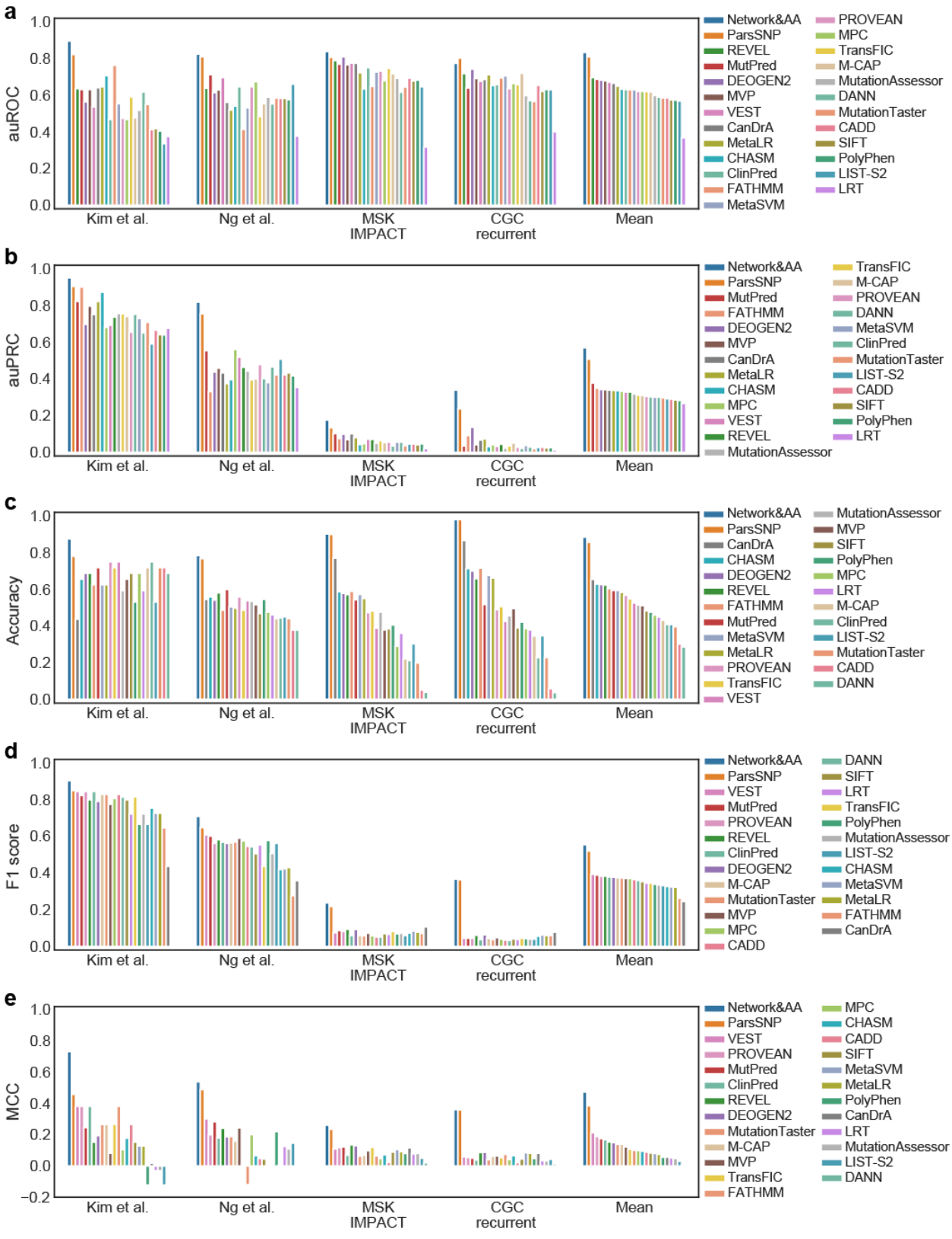


Figure 1.6. Distribution of residue-level network features. Distribution of residue-level network features for correctly labeled driver and passenger mutations occurring in the same proteins (Mann-Whitney U test; *p<0.05).

Figure 1.7. Comparison of classifier performance on benchmark datasets relative to established methods. Bar plots depict **(a)** the area under the ROC (auROC) and **(b)** the area under the PR curve (auPRC) scores, **(c)** accuracy, **(d)** F1 score, and **(e)** the Matthews correlation coefficient (MCC) results for each method. Mean category displays the mean of scores of each method across datasets. Methods are ordered based on their mean scores. All panels use the same color scheme.


a
auROC
1.0
0.8
0.6
0.4
0.2
0.0
Kim et al.
Ng et al.
MSK IMPACT
CGC recurrent
Mean
Network&AA
ParsSNP
REVEL
MutPred
DEOGEN2
MVP
VEST
CanDrA
MetaLR
CHASM
ClinPred
FATHMM
MetaSVM
PROVEAN
MPC
TransFIC
M-CAP
MutationAssessor
DANN
MutationTaster
CADD
SIFT
PolyPhen
LIST-S2
LRT
b
auPRC
1.0
0.8
0.6
0.4
0.2
0.0
Kim et al.
Ng et al.
MSK IMPACT
CGC recurrent
Mean
Network&AA
ParsSNP
MutPred
FATHMM
DEOGEN2
MVP
CanDrA
MetaLR
CHASM
MPC
VEST
REVEL
MutationAssessor
TransFIC
M-CAP
PROVEAN
DANN
MetaSVM
ClinPred
MutationTaster
LIST-S2
CADD
SIFT
PolyPhen
LRT
c
Accuracy
1.0
0.8
0.6
0.4
0.2
0.0
Kim et al.
Ng et al.
MSK IMPACT
CGC recurrent
Mean
Network&AA
ParsSNP
CanDrA
CHASM
DEOGEN2
REVEL
FATHMM
MutPred
MetaSVM
MetaLR
PROVEAN
TransFIC
VEST
MutationAssessor
MVP
SIFT
PolyPhen
MPC
LRT
M-CAP
ClinPred
LIST-S2
MutationTaster
CADD
DANN
d
F1 score
1.0
0.8
0.6
0.4
0.2
0.0
Kim et al.
Ng et al.
MSK IMPACT
CGC recurrent
Mean
Network&AA
ParsSNP
VEST
MutPred
PROVEAN
REVEL
ClinPred
DEOGEN2
M-CAP
MutationTaster
MVP
MPC
CADD
DANN
SIFT
LRT
TransFIC
PolyPhen
MutationAssessor
LIST-S2
CHASM
MetaSVM
MetaLR
FATHMM
CanDrA
e
MCC
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
Kim et al.
Ng et al.
MSK IMPACT
CGC recurrent
Mean
Network&AA
ParsSNP
VEST
PROVEAN
MutPred
ClinPred
REVEL
DEOGEN2
MutationTaster
M-CAP
MVP
TransFIC
FATHMM
MPC
CHASM
CADD
SIFT
MetaSVM
MetaLR
PolyPhen
CanDrA
LRT
MutationAssessor
LIST-S2
DANN

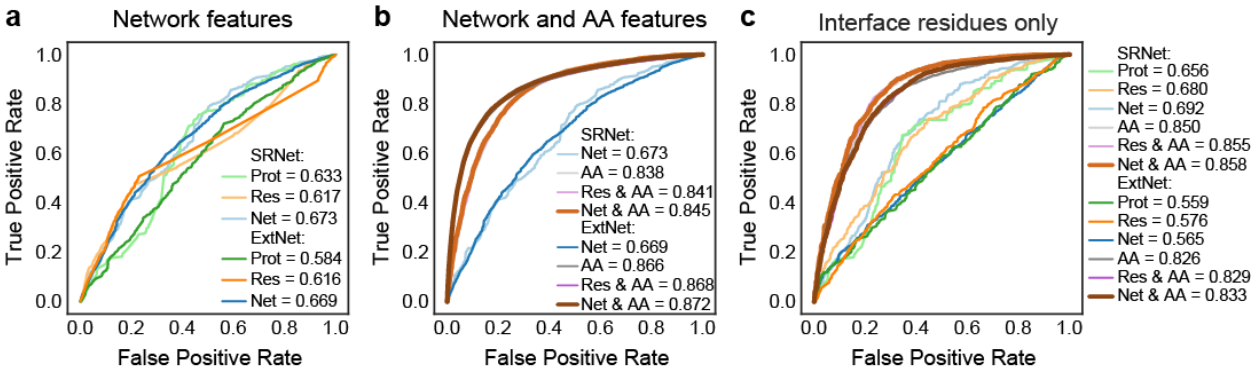


Figure 1.8. Classifier performances for predicting pathogenic vs. neutral variants using SRNet vs. the extended network (ExtNet). ROC curves for identifying variants with **(a)** protein-level network features (Prot), residue-level network features (Res), and all network features (Net = Prot + Res); with **(b)** all network features, amino acid features (AA), residue-level network and amino acid features (Res & AA), and all network and amino acid (Net & AA) features. **(c)** ROC curves for identifying variants targeting interface residues only using all above-mentioned features. ROC curves using Net & AA features are bold. Performance is measured using auROC scores.

1.8 Supplemental Data, Tables, and Figures

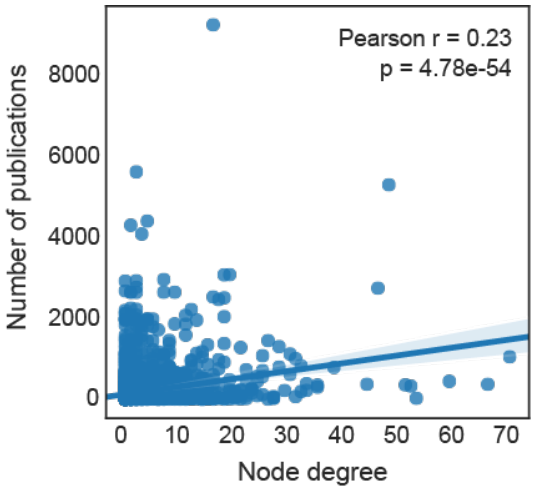
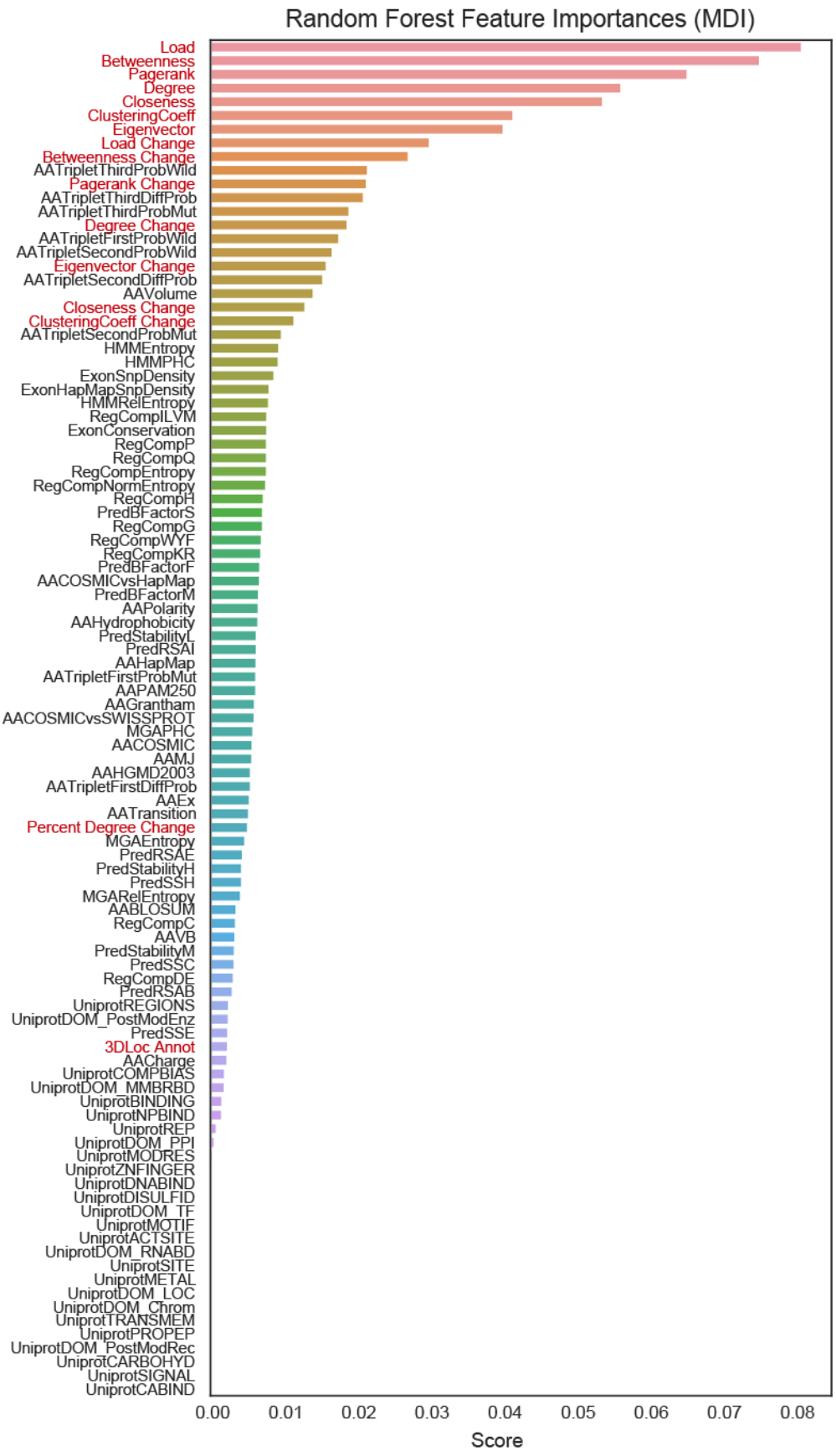


Figure S1.1. Study bias analysis. Correlation of node degree (the number of interacting partners) of proteins in SRNet with the number of PubMed publications they appear in (Pearson r=0.23, p=4.78e-54).

Figure S1.2. Random forest feature importances. The importance of a feature is computed as mean decrease in impurity (MDI), also known as the Gini importance. The network-based features and the amino acid features are colored red and black respectively.


Random Forest Feature Importances (MDI)
Load
Betweenness
Pagerank
Degree
Closeness
ClusteringCoeff
Eigenvector
Load Change
Betweenness Change
AATripletThirdProbWild
Pagerank Change
AATripletThirdDiffProb
AATripletThirdProbMut
Degree Change
AATripletFirstProbWild
AATripletSecondProbWild
Eigenvector Change
AATripletSecondDiffProb
AAVolume
Closeness Change
ClusteringCoeff Change
AATripletSecondProbMut
HMMEntropy
HMMPHC
ExonSnpDensity
ExonHapMapSnpDensity
HMMRelEntropy
RegCompILVM
ExonConservation
RegCompP
RegCompQ
RegCompEntropy
RegCompNormEntropy
RegCompH
PredBFactorS
RegCompG
RegCompWYF
RegCompKR
PredBFactorF
AACOSMICvsHapMap
PredBFactorM
AAPolarity
AAHydrophobicity
PredStabilityL
PredRSAI
AAHapMap
AATripletFirstProbMut
AAPAM250
AAGrantham
AACOSMICvsSWISSPROT
MGAPHC
AACOSMIC
AAMJ
AAHGMD2003
AATripletFirstDiffProb
AAEx
AATransition
Percent Degree Change
MGAEntropy
PredRSAE
PredStabilityH
PredSSH
MGARelEntropy
AABLOSUM
RegCompC
AAVB
PredStabilityM
PredSSC
RegCompDE
PredRSAB
UniprotREGIONS
UniprotDOM_PostModEnz
PredSSE
3DLoc Annot
AACharge
UniprotCOMPBIAS
UniprotDOM_MMBRBD
UniprotBINDING
UniprotNPBIND
UniprotREP
UniprotDOM_PPI
UniprotMODRES
UniprotZNFINGER
UniprotDNABIND
UniprotDISULFID
UniprotDOM_TF
UniprotMOTIF
UniprotACTSITE
UniprotDOM_RNABD
UniprotSITE
UniprotMETAL
UniprotDOM_LOC
UniprotDOM_Chrom
UniprotTRANSMEM
UniprotPROPEP
UniprotDOM_PostModRec
UniprotCARBOHYD
UniprotSIGNAL
UniprotCABIND
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
Score

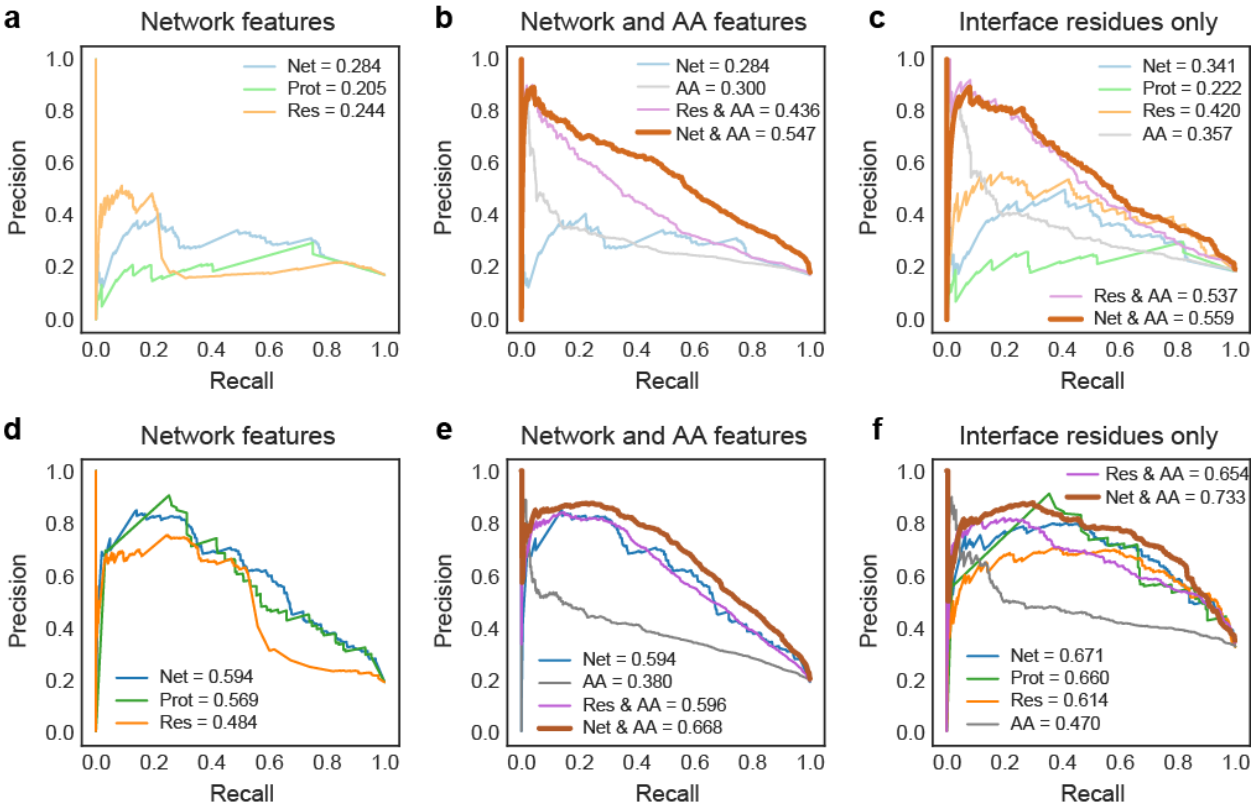


Figure S1.3. Classifier performance using precision-recall in identifying cancer mutations using SRNet vs. the extended network. Precision-recall (PR) curves for identifying cancer mutations using **(a-b-c)** SRNet, and **(d-e-f)** the extended network, with **(a-d)** protein-level network features (Prot), residue-level network features (Res), and all network features (Net = Prot + Res); with **(b-e)** all network features, amino acid features (AA), residue-level network and amino acid features (Res & AA), and all network and amino acid features (Net & AA). **(c-f)** PR curves for identifying cancer mutations targeting interface residues only, using all above-mentioned features. PR curves using Net & AA features are bold. Performance is measured using area under the PR curves.

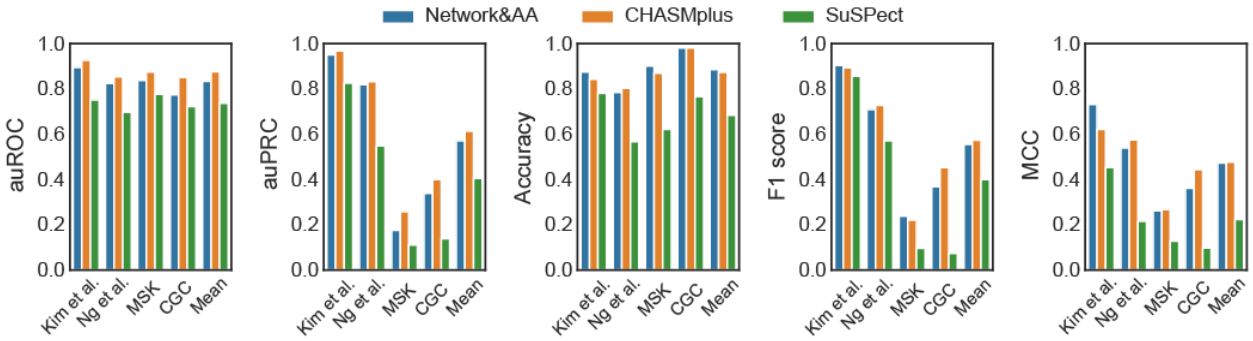


Figure S1.4. Comparison of classifier performance on benchmark datasets against existing methods that use network features. Bar plots depict the area under the ROC (auROC) and the area under the PR curve (auPRC) scores, accuracy, F1 score, and the Matthews correlation coefficient (MCC) results for each method. Mean category displays the mean of scores of each method across datasets. All panels use the same color scheme. A SuSpect cutoff of 50 (as recommended in Yates et al. [42]), and 0.5 for CHASMplus and the Network&AA classifier were used to assign labels for assessing accuracy, F1 score and MCC.

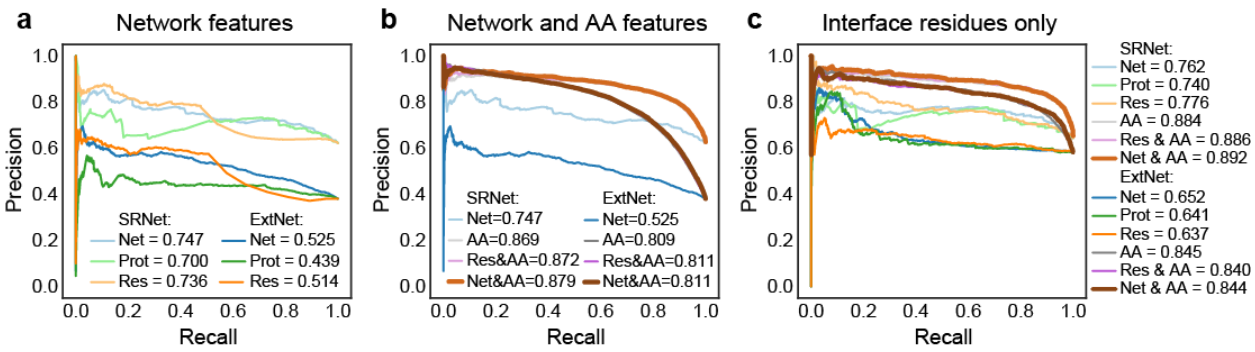


Figure S1.5. Classifier performance using precision-recall for predicting pathogenic vs. neutral variants using SRNet vs. the extended network (ExtNet). Precision-recall (PR) curves for identifying variants with **(a)** protein-level network features (Prot), residue-level network features (Res), and all network features (Net = Prot + Res); with **(b)** all network features, amino acid features (AA), residue-level network and amino acid features (Res & AA), and all network and amino acid (Net & AA) features. **(c)** PR curves for identifying variants targeting interface residues only using all above-mentioned features. PR curves using Net & AA features are bold. Performance is measured using area under the PR curves.

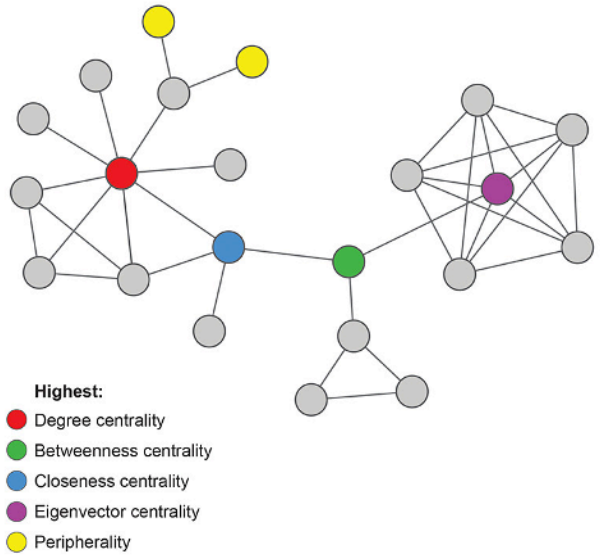


Figure S1.6. Exploring network topology as a determinant of gene-phenotype relationships. Nodes can be described with respect to particular characteristics in the network, including high degree hubs (red), nodes at the periphery (yellow) and nodes with the highest centrality according to four popular measures of centrality.

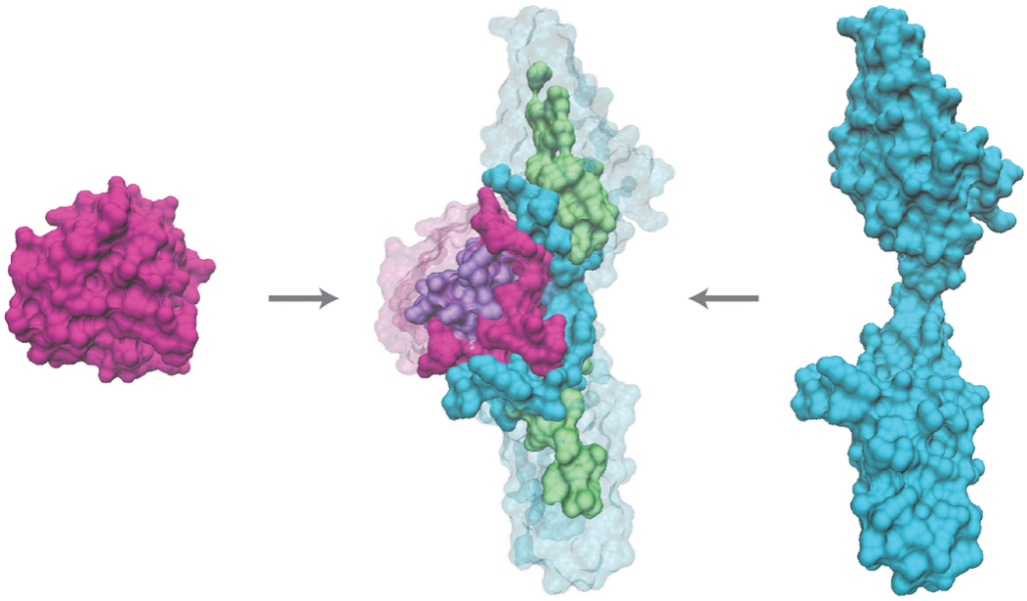


Figure S1.7. Mapping amino acid position to potential to interfere with protein interactions. Protein structures of FGF2 and FGFR1 are shown on the left and right respectively, and as a complex in the center (PDB: 1cvs). In the complex, residues are colored according to location in the protein core (purple and green), at the interface (pink and blue) or at the surface outside of the interface (transparent pink and blue) on the two proteins respectively.

1.9 Author Contributions

Original concept and project supervision by H.C. Project planning, design and method development by K.O. and H.C. Data acquisition, processing, and analysis by K.O. Preparation of manuscript by K.O. and H.C.

1.10 Acknowledgements

This work was supported by SDCSB/CCMI Systems Biology training grant (GM085764 and CA209891) to K.O. and NIH grant DP5 OD017937 and CIFAR award FL-000655 to H.C. NIH grant 2P41GM103504-11 provided access to computational resources.

Chapter 1, in full, is a reformatted reprint of the material as it appears in "Predicting functional consequences of mutations using molecular interaction network features" in Human Genetics, 2021 by Kivilcim Ozturk and Hannah Carter. The dissertation author was a primary investigator and author of this paper.

1.11 References

1. Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. Nucleic Acids Res. 2015;43: D789–98.

2. Barabasi AL, Albert R. Emergence of scaling in random networks. Science. 1999;286: 509–512.

3. Albert R, Jeong H, Barabasi AL. Error and attack tolerance of complex networks. Nature. 2000;406: 378–382.

4. Newman M. Networks: An Introduction. OUP Oxford; 2010.

5. Guimerà R, Amaral LAN. Cartography of complex networks: modules and universal roles. J Stat Mech. 2005;2005: nihpa35573.

6. Leung IXY, Hui P, Liò P, Crowcroft J. Towards real-time community detection in large networks. Phys Rev E Stat Nonlin Soft Matter Phys. 2009;79: 066107.

7. Newman MEJ. Modularity and community structure in networks. Proceedings of the National Academy of Sciences. 2006. pp. 8577–8582. doi:10.1073/pnas.0601602103

8. Jeong H, Mason SP, Barabási A-L, Oltvai ZN. Lethality and centrality in protein networks. Nature. 2001. pp. 41–42. doi:10.1038/35075138

9. Yu H, Kim PM, Sprecher E, Trifonov V, Gerstein M. The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. PLoS Comput Biol. 2007;3: e59.

10. Feldman I, Rzhetsky A, Vitkup D. Network properties of genes harboring inherited disease mutations. Proc Natl Acad Sci U S A. 2008;105: 4323–4328.

11. Said MR, Begley TJ, Oppenheim AV, Lauffenburger DA, Samson LD. Global network analysis of phenotypic effects: protein networks and toxicity modulation in Saccharomyces cerevisiae. Proc Natl Acad Sci U S A. 2004;101: 18006–18011.

12. Garcia-Alonso L, Jiménez-Almazán J, Carbonell-Caballero J, Vela-Boza A, Santoyo-López J, Antiñolo G, et al. The role of the interactome in the maintenance of deleterious variability in human populations. Mol Syst Biol. 2014;10: 752.

13. Goh K-I, Cusick ME, Valle D, Childs B, Vidal M, Barabási A-L. The human disease network. Proc Natl Acad Sci U S A. 2007;104: 8685–8690.

14. Sun J, Zhao Z. A comparative study of cancer proteins in the human protein-protein interaction network. BMC Genomics. 2010. p. S5. doi:10.1186/1471-2164-11-s3-s5

15. Jonsson PF, Bates PA. Global topological features of cancer proteins in the human interactome. Bioinformatics. 2006. pp. 2291–2297. doi:10.1093/bioinformatics/btl390

16. Khurana E, Fu Y, Chen J, Gerstein M. Interpretation of genomic variants using a unified biological network approach. Rzhetsky A, editor. PLoS Comput Biol. 2013;9: e1002886.

17. Taipale M. Disruption of protein function by pathogenic mutations: common and uncommon mechanisms. Biochem Cell Biol. 2019;97: 46–57.

18. Sahni N, Yi S, Taipale M, Fuxman Bass JI, Coulombe-Huntington J, Yang F, et al. Widespread macromolecular interaction perturbations in human genetic disorders. Cell. 2015;161: 647–660.

19. Das J, Fragoza R, Lee HR, Cordero NA, Guo Y, Meyer MJ, et al. Exploring mechanisms of human disease through structurally resolved protein interactome networks. Mol Biosyst. 2014;10: 9–17.

20. Sahni N, Yi S, Zhong Q, Jailkhani N, Charloteaux B, Cusick ME, et al. Edgotype: a fundamental link between genotype and phenotype. Curr Opin Genet Dev. 2013;23: 649–657.

21. Zhong Q, Simonis N, Li Q-R, Charloteaux B, Heuze F, Klitgord N, et al. Edgetic perturbation models of human inherited disorders. Mol Syst Biol. 2009;5: 321.

22. Betts MJ, Lu Q, Jiang Y, Drusko A, Wichmann O, Utz M, et al. Mechismo: predicting the mechanistic impact of mutations and modifications on molecular interactions. Nucleic Acids Res. 2015;43: e10.

23. Meyer MJ, Das J, Wang X, Yu H. INstruct: a database of high-quality 3D structurally resolved protein interactome networks. Bioinformatics. 2013;29: 1577–1579.

24. Mosca R, Céol A, Aloy P. Interactome3D: adding structural details to protein networks. Nat Methods. 2013;10: 47–53.

25. Vázquez M, Valencia A, Pons T. Structure-PPi: a module for the annotation of cancer-related single-nucleotide variants at protein–protein interfaces: Fig. 1. Bioinformatics. 2015. pp. 2397–2399. doi:10.1093/bioinformatics/btv142

26. David A, Razali R, Wass MN, Sternberg MJE. Protein-protein interaction sites are hot spots for disease-associated nonsynonymous SNPs. Hum Mutat. 2012;33: 359–363.

27. Guo Y, Wei X, Das J, Grimson A, Lipkin SM, Clark AG, et al. Dissecting disease inheritance modes in a three-dimensional protein network challenges the “guilt-by-association” principle. Am J Hum Genet. 2013;93: 78–89.

28. Wang X, Wei X, Thijssen B, Das J, Lipkin SM, Yu H. Three-dimensional reconstruction of protein networks provides insight into human genetic disease. Nat Biotechnol. 2012;30: 159–164.

29. Fernald GH, Capriotti E, Daneshjou R, Karczewski KJ, Altman RB. Bioinformatics challenges for personalized medicine. Bioinformatics. 2011. pp. 2323–2323. doi:10.1093/bioinformatics/btr408

30. Cooper GM, Shendure J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nature Reviews Genetics. 2011. pp. 628–640. doi:10.1038/nrg3046

31. Liu X, Li C, Mou C, Dong Y, Tu Y. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. Genome Med. 2020;12: 103.

32. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003;31: 3812–3814.

33. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7: 248–249.

34. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014;46: 310–315.

35. Hecht M, Bromberg Y, Rost B. Better prediction of functional effects for sequence variants. BMC Genomics. 2015;16 Suppl 8: S1.

36. Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, et al. REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. Am J Hum Genet. 2016;99: 877–885.

37. Ponzoni L, Peñaherrera DA, Oltvai ZN, Bahar I. Rhapsody: predicting the pathogenicity of human missense variants. Bioinformatics. 2020;36: 3084–3092.

38. Pejaver V, Urresti J, Lugo-Martinez J, Pagel KA, Lin GN, Nam H-J, et al. Inferring the molecular and phenotypic impact of amino acid variants with MutPred2. Nat Commun. 2020;11: 5918.

39. Félix M-A, Barkoulas M. Pervasive robustness in biological systems. Nat Rev Genet. 2015;16: 483–496.

40. Vidal M, Cusick ME, Barabási A-L. Interactome Networks and Human Disease. Cell. 2011. pp. 986–998. doi:10.1016/j.cell.2011.02.016

41. Capriotti E, Ozturk K, Carter H. Integrating molecular networks with genetic variant interpretation for precision medicine. Wiley Interdiscip Rev Syst Biol Med. 2019;11: e1443.

42. Yates CM, Filippis I, Kelley LA, Sternberg MJE. SuSPect: enhanced prediction of single amino acid variant (SAV) phenotype using network features. J Mol Biol. 2014;426: 2692–2701.

43. Chen S, Fragoza R, Klei L, Liu Y, Wang J, Roeder K, et al. An interactome perturbation framework prioritizes damaging missense mutations for developmental disorders. Nat Genet. 2018;50: 1032–1040.

44. Ozturk K, Dow M, Carlin DE, Bejar R, Carter H. The Emerging Potential for Network Analysis to Inform Precision Cancer Medicine. J Mol Biol. 2018;430: 2875–2899.

45. Engin HB, Hofree M, Carter H. Identifying mutation specific cancer pathways using a structurally resolved protein interaction network. Pac Symp Biocomput. 2015; 84–95.

46. Engin HB, Kreisberg JF, Carter H. Structure-Based Analysis Reveals Cancer Missense Mutations Target Protein Interaction Interfaces. Srinivasan N, editor. PLoS One. 2016;11: e0152929.

47. Kamburov A, Lawrence MS, Polak P, Leshchiner I, Lage K, Golub TR, et al. Comprehensive assessment of cancer missense mutation clustering in protein structures. Proc Natl Acad Sci U S A. 2015;112: E5486–E5495.

48. Porta-Pardo E, Garcia-Alonso L, Hrabe T, Dopazo J, Godzik A. A Pan-Cancer Catalogue of Cancer Driver Protein Interaction Interfaces. Nussinov R, editor. PLoS Comput Biol. 2015;11: e1004518.

49. Raimondi F, Singh G, Betts MJ, Apic G, Vukotic R, Andreone P, et al. Insights into cancer severity from biomolecular interaction mechanisms. Sci Rep. 2016;6: 34490.

50. Iqbal S, Pérez-Palma E, Jespersen JB, May P, Hoksza D, Heyne HO, et al. Comprehensive characterization of amino acid positions in protein structures reveals molecular effect of missense variants. Proc Natl Acad Sci U S A. 2020;117: 28201–28211.

51. Laskowski RA, Stephenson JD, Sillitoe I, Orengo CA, Thornton JM. VarSite: Disease variants and protein structure. Protein Sci. 2020;29: 111–119.

52. Tokheim C, Karchin R. CHASMplus Reveals the Scope of Somatic Missense Mutations Driving Human Cancers. Cell Syst. 2019;9: 9–23.e8.

53. Tokheim C, Bhattacharya R, Niknafs N, Gygax DM, Kim R, Ryan M, et al. Exome-Scale Discovery of Hotspot Mutation Regions in Human Cancer Using 3D Protein Structure. Cancer Res. 2016;76: 3719–3731.

54. Barabási A-L, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nat Rev Genet. 2011;12: 56–68.

55. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 2015;43: D447–52.

56. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW. Cancer Genome Landscapes. Science. 2013;339: 1546–1558.

57. Vinayagam A, Gibson TE, Lee H-J, Yilmazel B, Roesel C, Hu Y, et al. Controllability analysis of the directed human protein interaction network identifies disease genes and drug targets. Proc Natl Acad Sci U S A. 2016;113: 4976–4981.

58. Collins FS, Barker AD. Mapping the cancer genome. Pinpointing the genes involved in cancer will help chart a new course across the complex landscape of human malignancies. Sci Am. 2007;296: 50–57.

59. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: Improving access to variant interpretations and supporting evidence. Nucleic Acids Res. 2018. doi:10.1093/nar/gkx1153

60. Lek M, Karczewski K, Minikel E, Samocha K, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. bioRxiv. 2015. Available: http://www.biorxiv.org/content/early/2015/10/30/030338

61. Mottaz A, David FPA, Veuthey AL, Yip YL. Easy retrieval of single amino-acid polymorphisms and phenotype information using SwissVar. Bioinformatics. 2010. doi:10.1093/bioinformatics/btq028

62. Wong WC, Kim D, Carter H, Diekhans M, Ryan MC, Karchin R. CHASM and SNVBox: Toolkit for detecting biologically important single nucleotide mutations in cancer. Bioinformatics. 2011. doi:10.1093/bioinformatics/btr357

63. Meyer MJ, Beltrán JF, Liang S, Fragoza R, Rumack A, Liang J, et al. Interactome INSIDER: a structural interactome browser for genomic studies. Nat Methods. 2018;15: 107–114.

64. Breiman L. Random forests. Mach Learn. 2001. doi:10.1023/A:1010933404324

65. Kim E, Ilic N, Shrestha Y, Zou L, Kamburov A, Zhu C, et al. Systematic functional interrogation of rare cancer variants identifies oncogenic alleles. Cancer Discov. 2016. doi:10.1158/2159-8290.CD-16-0160

66. Ng PKS, Li J, Jeong KJ, Shao S, Chen H, Tsang YH, et al. Systematic Functional Annotation of Somatic Mutations in Cancer. Cancer Cell. 2018. doi:10.1016/j.ccell.2018.01.021

67. Carter H, Chen S, Isik L, Tyekucheva S, Velculescu VE, Kinzler KW, et al. Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. Cancer Res. 2009;69: 6660–6667.

68. Kumar RD, Swamidass SJ, Bose R. Unsupervised detection of cancer driver mutations with parsimony-guided learning. Nat Genet. 2016;48: 1288–1294.

69. Gonzalez-Perez A, Deu-Pons J, Lopez-Bigas N. Improving the prediction of the functional impact of cancer mutations by baseline tolerance transformation. Genome Med. 2012;4: 89.

70. Mao Y, Chen H, Liang H, Meric-Bernstam F, Mills GB, Chen K. CanDrA: Cancer-specific driver missense mutation annotation with optimized features. PLoS One. 2013. doi:10.1371/journal.pone.0077945

71. Carter H, Douville C, Stenson PD, Cooper DN, Karchin R. Identifying Mendelian disease genes with the Variant Effect Scoring Tool. BMC Genomics. 2013. doi:10.1186/1471-2164-14-S3-S3

72. Alirezaie N, Kernohan KD, Hartley T, Majewski J, Hocking TD. ClinPred: Prediction Tool to Identify Disease-Relevant Nonsynonymous Single-Nucleotide Variants. Am J Hum Genet. 2018;103: 474–483.

73. Quang D, Chen Y, Xie X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics. 2015. pp. 761–763. doi:10.1093/bioinformatics/btu703

74. Raimondi D, Tanyalcin I, Ferté J, Gazzo A, Orlando G, Lenaerts T, et al. DEOGEN2: prediction and interactive visualization of single amino acid variant deleteriousness in human proteins. Nucleic Acids Res. 2017;45: W201–W206.

75. Shihab HA, Gough J, Cooper DN, Day INM, Gaunt TR. Predicting the functional consequences of cancer-associated amino acid substitutions. Bioinformatics. 2013. doi:10.1093/bioinformatics/btt182

76. Malhis N, Jacobson M, Jones SJM, Gsponer J. LIST-S2: taxonomy based sorting of deleterious missense mutations across species. Nucleic Acids Res. 2020;48: W154–W161.

77. Chun S, Fay JC. Identification of deleterious mutations within three human genomes. Genome Res. 2009;19: 1553–1561.

78. Jagadeesh KA, Wenger AM, Berger MJ, Guturu H, Stenson PD, Cooper DN, et al. M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. Nat Genet. 2016;48: 1581–1586.

79. Samocha KE, Kosmicki JA, Karczewski KJ, O'Donnell-Luria AH, Pierce-Hoffman E, MacArthur DG, et al. Regional missense constraint improves variant deleteriousness prediction. doi:10.1101/148353

80. Qi H, Zhang H, Zhao Y, Chen C, Long JJ, Chung WK, et al. MVP predicts the pathogenicity of missense variants by deep learning. Nat Commun. 2021;12: 510.

81. Dong C, Wei P, Jian X, Gibbs R, Boerwinkle E, Wang K, et al. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. Hum Mol Genet. 2015;24: 2125–2137.

82. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res. 2011;39: e118.

83. Schwarz JM, Cooper DN, Schuelke M, Seelow D. MutationTaster2: mutation prediction for the deep-sequencing age. Nat Methods. 2014;11: 361–362.

84. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the functional effect of amino acid substitutions and indels. PLoS One. 2012;7: e46688.

85. Wei X, Das J, Fragoza R, Liang J, Bastos de Oliveira FM, Lee HR, et al. A massively parallel pipeline to clone DNA variants and examine molecular phenotypes of human disease mutations. PLoS Genet. 2014;10: e1004819.

86. David A, Sternberg MJE. The Contribution of Missense Mutations in Core and Rim Residues of Protein–Protein Interfaces to Human Disease. Journal of Molecular Biology. 2015. pp. 2886–2898. doi:10.1016/j.jmb.2015.07.004

87. Nishi H, Nakata J, Kinoshita K. Distribution of single-nucleotide variants on protein-protein interaction sites and its relationship with minor allele frequency. Protein Sci. 2016;25: 316–321.

88. IMEx Consortium Curators, Del-Toro N, Duesbury M, Koch M, Perfetto L, Shrivastava A, et al. Capturing variation impact on molecular interactions in the IMEx Consortium mutations data set. Nat Commun. 2019;10: 10.

89. Piñero J, Berenstein A, Gonzalez-Perez A, Chernomoretz A, Furlong LI. Uncovering disease mechanisms through network biology in the era of Next Generation Sequencing. Sci Rep. 2016;6: 24570.

90. Creixell P, Schoof EM, Simpson CD, Longden J, Miller CJ, Lou HJ, et al. Kinome-wide Decoding of Network-Attacking Mutations Rewiring Cancer Signaling. Cell. 2015;163: 202–217.

91. Raimondi D, Passemiers A, Fariselli P, Moreau Y. Current cancer driver variant predictors learn to recognize driver genes instead of functional variants. BMC Biol. 2021;19: 3.

92. Greene CS, Krishnan A, Wong AK, Ricciotti E, Zelaya RA, Himmelstein DS, et al. Understanding multicellular function and disease with human tissue-specific networks. Nat Genet. 2015;47: 569–576.

93. Amberger JS, Bocchini CA, Scott AF, Hamosh A. OMIM.org: leveraging knowledge across phenotype-gene relationships. Nucleic Acids Res. 2019;47: D1038–D1043.

94. Berman H, Henrick K, Nakamura H. Announcing the worldwide Protein Data Bank. Nat Struct Biol. 2003;10: 980–980.

95. Zhu X, Mitchell JC. KFC2: a knowledge-based hot spot prediction method based on interface solvation, atomic density, and plasticity features. Proteins. 2011;79: 2671–2683.

96. Hubbard SJ, Thornton JM. "NACCESS", Computer Program, Department of Biochemistry and Molecular Biology, University College London. 1993.

97. Martin ACR. Mapping PDB chains to UniProtKB entries. Bioinformatics. 2005;21: 4297–4301.

98. Zehir A, Benayed R, Shah RH, Syed A, Middha S, Kim HR, et al. Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. Nat Med. 2017;23: 703–713.

99. Chakravarty D, Gao J, Phillips SM, Kundra R, Zhang H, Wang J, et al. OncoKB: A Precision Oncology Knowledge Base. JCO Precis Oncol. 2017;2017. doi:10.1200/PO.17.00011

100. Forbes SA, Beare D, Boutselakis H, Bamford S, Bindal N, Tate J, et al. COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Res. 2017;45: D777–D783.

101. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol. 2016;17: 122.

CHAPTER 2: Investigating oncogenic selective signatures on networks: A case study of the B2M subnetwork

2.1 Foreword

The concept of edgetics has provided mechanistic insights for complex diseases including cancer. Proteins frequently have multiple functions via complex interaction dynamics with multiple partners, and targeting of different interactions can generate different phenotypes. Several studies have shown enrichment of somatic mutations at protein interaction interfaces in tumor genomes [1–4], suggesting that perturbations of protein interactions frequently contribute to tumor development. Indeed, my work described in Chapter 1, demonstrated that somatic mutations selectively target interfaces of cancer driver genes.

While this holds true for most cancer genes, our systematic analysis of somatic mutations in relation to the underlying PPI networks revealed a distinct case in tumor suppressor beta-2-microglobulin (B2M), where we observed an uncharacteristic enrichment of mutations at B2M interaction partner binding interfaces [1] (Figure S2.10A). Whereas for most cancer genes, mutations occur preferentially on the cancer gene itself (Figure S2.10B); genes encoding B2M binding partners showed almost as many somatic mutations as B2M (Figure S2.10A). We specifically observed recurring mutations targeting the interface regions of partners HLA-A and HLA-B, where they interact with B2M. This differential pattern of network architecture perturbation is suggestive of different selective oncogenic pressures being at play in the interaction of the tumor suppressor B2M with HLA-A and HLA-B proteins.

B2M, together with the HLA-A, HLA-B, or HLA-C proteins, makes up the major histocompatibility complex class I (MHC-I) molecule, which is responsible for displaying intracellular peptides to T cells, allowing the immune system to recognize and destroy infected or

cancerous cells. Within the complex, a highly polymorphic HLA-encoded alpha chain binds the peptide, and B2M acts as a stabilizing scaffold, making it essential for MHC-I complex formation and peptide presentation. HLA mutations have been implicated as a mechanism of immune evasion during tumorigenesis, and B2M is a known tumor suppressor gene. However, the implications of somatic HLA and B2M mutations have not been fully explored in the context of antigen presentation via the MHC-I molecule during tumor development.

My research in Chapter 2 aims to uncover the selective oncogenic pressures causing the distinct pattern of mutation accumulation in the B2M subnetwork via a mechanistic investigation of the B2M and HLA protein interaction. Given B2M's role as a central component of MHC-I, we show that mutations affecting B2M's interaction with HLA-A, HLA-B, and HLA-C could facilitate immune evasion by altering the availability of MHC-I molecules with distinct specificities, thus affecting presentation of specific peptides to the immune system. To gain a better understanding of the role of somatic mutations affecting B2M and its partners in immune evasion, we examine their effect on mutation burden, antigen binding affinity, immune infiltration, and cytotoxicity in tumors. Thus, my work in Chapter 2 indicates that differential patterns of network rewiring of mutations can uncover genes under distinct selective pressures in cancer, providing insight into their functional roles in carcinogenesis.

2.2 Abstract

The major histocompatibility complex class I (MHC-I) molecule is a protein complex that displays intracellular peptides to T cells, allowing the immune system to recognize and destroy infected or cancerous cells. MHC-I is composed of a highly polymorphic HLA-encoded alpha chain that binds the peptide and a Beta-2-microglobulin (B2M) protein that acts as a stabilizing scaffold. HLA mutations have been implicated as a mechanism of immune evasion during tumorigenesis, and B2M is considered a tumor suppressor gene. However, the implications of somatic HLA and B2M mutations have not been fully explored in the context of antigen presentation via the MHC-I molecule during tumor development. To understand the effect that B2M and HLA MHC-I molecule mutations have on mutagenesis, we analyzed the accumulation of mutations in patients from The Cancer Genome Atlas according to their MHC-I molecule mutation status. Somatic B2M and HLA mutations in microsatellite stable tumors were associated with higher overall mutation burden and a larger fraction of HLA-binding neoantigens when compared to B2M and HLA wild type tumors. B2M and HLA mutations were highly enriched in patients with microsatellite instability. B2M mutations tended to occur relatively early during patients' respective tumor development, whereas HLA mutations were either early or late events. In addition, B2M and HLA mutated patients had higher levels of immune infiltration by natural killer and CD8+ T cells and higher levels of cytotoxicity. Our findings add to a growing body of evidence that somatic B2M and HLA mutations are a mechanism of immune evasion by demonstrating that such mutations are associated with a higher load of neoantigens that should be presented via MHC-I.

2.3 Introduction

Immune evasion is one of the hallmark traits characteristic of cancer cells [5]. The near universal requirement for tumor cells to evade immune elimination implicates the immune system as a major selective force acting on developing tumors. When a tumor cell successfully evades the immune system, the mutations harbored within can persist and propagate as the cell divides.

In humans, the HLA-A, HLA-B, and HLA-C genes encode major histocompatibility complex class I (MHC-I) molecules, which display intracellular peptides on the cell surface for inspection by CD8+ T cells. These T cells have the potential to recognize the MHC-I-peptide complex and become activated cytotoxic T cells (CTLs). Cancer immunotherapies that target CTL activation rely on clinical selection of appropriate neoantigens, mutated peptides specific to tumor cells, to stimulate a response [6]. Although these cancer immunotherapies are of high interest to patients and clinicians, they have not yet shown widespread clinical success [7].

The MHC-I molecule is composed of a highly polymorphic HLA encoded alpha chain and a beta-2-microglobulin (B2M) protein that acts as a stabilizing scaffold. B2M is essential for MHC-I complex formation and peptide presentation. B2M mutations and loss of heterozygosity (LOH) are linked to decreased MHC class I expression and decreased patient survival [8,9]. In addition to HLA-A, HLA-B, and HLA-C, B2M binds to other immune proteins including CD1, FCGRT, HFE, HLA-E, HLA-G LILRB, and MR1. Somatic mutations in HLA-A and HLA-B have also been shown to be under positive selection during tumorigenesis and are more frequent when tumor immune cell infiltration and cytotoxicity are high [10]. Importantly, MHC-I molecule presence on the cell surface can provide an inhibitory signal to natural killer (NK) cell mediated effector functions [11]. In addition to classical HLA molecules HLA-A, HLA-B, HLA-C, and HLA-G, nonclassical HLA-E acts as a ligand to inhibitory receptors on NK cells. Thus, both

presence and antigen presentation function of MHC-I molecules contribute to anti-tumor immunity.

A recent study found that an individual's HLA genotype can facilitate immune evasion and shape the landscape of a patient's acquired mutations [12]. Somatic mutations generating peptides with low affinity for an individual's respective HLA alleles were likely to evade immune detection and persist in the tumor. Somatic LOH in the human leukocyte antigen (HLA) locus is thought to impair immune surveillance and was reported to occur in 40% of non-small-cell lung cancers. The authors found that HLA LOH was significantly associated with a high mutational burden and cancer-specific neoantigens generated from these mutations were biased to bind to the lost HLA allele [13]. Thus, immune evasion may depend on an individual's unique HLA genotype and the specificity of neoantigens for particular HLA alleles.

We previously observed an uncharacteristic enrichment of somatic mutations at B2M interaction partner binding interfaces [1]. Whereas for most cancer genes, mutations occurred preferentially on the cancer gene itself, genes encoding B2M binding partners showed almost as many somatic mutations as B2M (Figure 2.1A). Given B2M's role as a central component of MHC-I, we hypothesized that mutations affecting B2M's interaction with HLA-A, HLA-B, and HLA-C could facilitate immune evasion by altering the availability of MHC-I molecules with distinct specificities, thus affecting presentation of specific peptides to the immune system (Figure 2.1B). To gain a better understanding of the role of somatic mutations affecting B2M and its partners in immune evasion, we examined their effect on mutation burden, antigen binding affinity, immune infiltration, and cytotoxicity in tumors sequenced by The Cancer Genome Atlas (TCGA).

2.4 Results

HLA and B2M mutations in TCGA. B2M mutation calls were obtained directly from the MAF files provided by TCGA. Because the HLA locus is highly polymorphic, mutation calls against the reference genome are unreliable. Instead, we ran Polysolver [10] to simultaneously call patient-specific HLA types and detect somatic mutations affecting a patient's HLA alleles. Out of 10,428 TCGA patients that had the necessary whole exome sequencing data, only 579 patients had an HLA mutation and 125 patients had B2M mutations. Most of these mutations were nonsynonymous (Figure 2.2A). To determine whether nonsynonymous mutations occurred at amino acid residues with the potential to interfere with formation of the MHC-I molecule, experimental 3D structures for the B2M-HLA complex were obtained from the Protein Data Bank [14] and used to annotate amino acid residue location at protein core, surface or at the physical interface between B2M and HLA encoded proteins (Methods).

Mutations on HLA proteins, particularly HLA-A, showed a biased distribution with several recurrent hotspots (Figure 2.2B). Mutations were most concentrated in the α3 domain that mediates interaction with the T cell receptor (TCR) (206 mutations, 40.63% of total; OR=2.04, $p < 2.58e-09$), and included multiple recurrent hotspots. Fifty-one mutations (10.06%) were observed in the transmembrane domain including additional hotspots. Although mutations were observed throughout the α1 and α2 domains that form the peptide binding groove, they tended to be less recurrent (88 mutations, 17.36% for both α1 and α2). This may reflect the much larger heterogeneity of this region across HLA alleles.

Recurrent hotspot mutations often targeted interface and core regions on HLA-A, while they targeted core and surface regions on HLA-B, and surface regions on HLA-C (Figure 2.2B). Since there are many alleles for each HLA protein, we used the consensus of residue annotations

across different alleles to annotate each HLA protein (Figure S2.1). Even though the annotations for most frequently mutated residues were in agreement between different HLA alleles, there were some exceptions, including residue 231 on HLA-A. Although residue 231 (R231) on HLA-A was annotated as surface based on the consensus across HLA-A alleles, the residue is located very close to the interface region (Figure 2.2B) and in fact was predicted as an interface residue on 2 of the 6 HLA-A allele structures analyzed. Additionally, although residue 209 (R209) on HLA-A and HLA-B proteins was annotated as ambiguous due to its intermediate value of relative solvent accessible surface area (RSA) for most HLA-A/B structures analyzed, the average RSA across structures is close to the threshold for core annotation (7.17), and R209 was indeed annotated as core in some of them. Overall, the distribution of HLA mutations for the three proteins was consistent with the previous report by Shukla et al. [10], though the current analysis incorporates an overall larger number of samples. Mutations in B2M were largely loss of function (Figure 2.2A) and more broadly distributed (Figure 2.2C), as expected for a tumor suppressor gene, though several positions were also recurrently mutated.

Expected effects of B2M versus HLA mutations on MHC-I composition. Since B2M is an essential component of all MHC-I molecules, loss of B2M should equally impact MHC-I molecules derived from different HLA alleles. The B2M interface with HLA alleles is shared across the different alleles (Figure S2.2), so mutations at this interface are also likely to affect all variants of an individual's MHC-I molecule, although complexes involving B2M and binding partners that use alternative interfaces should not be affected. In contrast, loss of function or interface mutations affecting a specific HLA allele would only affect the MHC-I molecules derived from that allele. Thus, we speculate that B2M mutations are likely to reduce the total amount of

MHC-I molecules presenting antigens on the tumor cell surface, while HLA mutations would impact which mutations could be presented as neoantigens.

Mutations in MHC-I proteins are associated with increased mutation burden. We hypothesized that both B2M and HLA mutations would affect MHC-I presentation of mutations. Mice with total lack of B2M express little if any cell surface MHC-I and lack cytotoxic CD8+ T cells [15,16]. In human lung cancers, an association was found between higher somatic mutation burden and HLA loss of heterozygosity [13]. If somatic mutations to HLA and B2M similarly impair antigen presentation, we would expect to see an increased mutation burden when comparing to unmutated patients.

We first analyzed 9,055 TCGA patients across 31 solid tumor types that had both exome and RNA sequencing data (Figure 2.3A), removing patients that had synonymous B2M or HLA mutations. We then performed a cancer-specific analysis of 3,514 patients across 8 solid tumor types with at least 5 somatic B2M and HLA mutations (Figure 2.3B, Figure S2.3A). To determine whether somatic mutations to B2M and HLA were associated with an overall higher mutation burden, we compared the total number of expressed nonsynonymous mutations in patients with and without nonsynonymous somatic B2M or HLA mutations. Overall, we observed that both patients with a B2M and an HLA mutation had significantly higher tumor mutation burdens (Mann Whitney test, B2M $p < 1.1e-20$ and HLA $p < 1.1e-30$) than patients without (Figure 2.3A). Pan-cancer, B2M mutated patients also had significantly higher mutation burdens than HLA mutated patients (Mann Whitney test, $p < 0.0028$). There were approximately equal numbers of early stage (I & II) and late stage (III & IV) tumors in these three groups (Figure S2.4). We repeated the pan-cancer mutational burden analysis with Cancer Cell Line Encyclopedia (CCLE) data for 25 B2M-mutated cell lines, 114 HLA cell lines, and 1,381 non-mutated cell lines, and observed the same

trend: cell lines with B2M and HLA mutations had significantly higher overall mutational burden than cell lines without (Figure S2.5). When we analyzed tumors by tissue type, we observed that certain cancers (stomach adenocarcinoma, endometrial cancer, colorectal cancer, lung adenocarcinoma, and cervical cancer) also had significantly higher mutational burden in mutated patients (Figure 2.3B). Stomach, uterine and colorectal cancers have documented high rates of microsatellite instability (MSI), thus we evaluated whether B2M and HLA mutations were biased to occur in high MSI tumors. Using MSI annotations available for 10,415 patients from Kautto et al. [17], we found a significant bias for B2M and HLA mutations to occur in patients with MSI (Fisher's exact test; B2M OR=14.66, $p < 8.7e-24$; HLA OR=6.28, $p < 2.0e-36$). To rule out the possibility that MSI was solely driving our results, we reanalyzed the mutational burden between B2M and HLA mutated and unmutated patients, this time retaining only 8,668 microsatellite stable (MSS) patients. Interestingly, we found similar trends in elevated mutational burden associated with B2M and HLA mutation (Figure 2.3C, 2.3D, Figure S2.3B), and consequently focused on MSS patients only in the subsequent analyses. Thus, even in MSS tumors, B2M and HLA mutations are associated with an increased nonsynonymous mutational burden.

Mutations in MHC-I proteins are associated with increased binding neoantigen counts. To obtain more evidence as to whether the elevated mutation counts observed in HLA and B2M mutated patients were a result of the mutation, or vice versa, we compared the fraction of mutations likely to generate neoantigens across MSS patients with and without B2M and HLA mutations. We speculated that if B2M and HLA mutations are an artifact of higher mutation rates, the proportion of mutations that generate neoantigens should not differ relative to patients without such mutations. However, if these mutations truly facilitate immune escape, neoantigens should be enriched among the observed mutations.

Using HLA allele genotypes called by Polysolver [10], we calculated patient-specific MHC-I presentation scores for all expressed mutations observed in each patient's tumor [12,18]. We previously demonstrated that these affinity-based presentation scores, called PHBR-I scores, can distinguish peptides found in complex with MHC-I on the cell surface in mass spectrometry experiments from random peptides simulated from the human proteome, supporting that affinity is a reasonable proxy for cell surface presentation [12]. Indeed, when we looked at the fraction of expressed mutations considered to be neoantigens at various PHBR-I cutoffs, we found that at any given cutoff, a higher fraction of mutations represented neoantigens in both B2M and HLA mutated patients (Figures 2.4A, B). This corresponded to overall higher numbers of neoantigens in B2M and HLA mutant tumors (Figure S2.6). The higher overall proportion of neoantigens is consistent with both somatic B2M and HLA mutations impairing presentation of neoantigens for immune surveillance.

Assessing bias in neoantigen affinities in patients with mutant HLA alleles. McGranahan et al. reported that in lung cancer, subclones that had lost a particular HLA allele tended to accumulate mutations with higher affinity for the lost allele, suggesting that such mutations were no longer subject to immunoediting [13]. We therefore sought to assess whether mutations accumulating in tumors with HLA mutations showed a bias in affinity toward the affected HLA allele. We first evaluated whether the number of mutant-allele specific mutations in these patients was higher than the average number of mutations specific to each of the other alleles (Figure 2.4C). We observed several patients for which the number of mutant-allele specific mutations was indeed higher (Figure 2.4C; red lines). We note that the current study design differs from the study by McGranahan et al. in that we do not have subclone-specific sequencing data, and thus can not determine which mutations occurred in the same cell population as the mutated

HLA allele. We also did not consider allele-specific deletion events, and thus the assumption that the other 5 HLA alleles are intact may be incorrect for some patients.

Timing of somatic mutations in MHC-I proteins. To better understand B2M and HLA mutation timelines, we analyzed the tumor allelic fraction of expressed mutations for all patients. Early clonal mutations are present in a larger fraction of cancer cells than later subclonal mutations and are, therefore, expected to be present in a higher fraction of the reads generated from that site during tumor sequencing. Although this assumption can be complicated by sampling bias and genomic instability of tumors, we nonetheless expect that somatic point mutations with higher read support will in general have occurred at earlier time points than those with lower read support. Since each individual's tumor is unique, we quantified B2M and HLA mutations in terms of their allelic fraction percentile relative to other mutations observed in the same tumor (Figure 2.4D). Interestingly, B2M mutations tended to be present at higher percentiles than most HLA mutations, suggesting that B2M mutations might occur earlier in tumor development and affect a higher proportion of tumor cells. Most HLA mutations had low percentiles, suggesting these were late, subclonal events, while a subset had high percentiles and likely occurred early during tumor development in those individuals. This observation agrees with the previous report by McGranahan et al. that found HLA loss in lung cancer to be predominantly subclonal with a few observations of clonal loss noted. Patients with MSI tended to have HLA mutations with higher variant allele fraction (VAF) (Fisher's exact test, OR=73.3, $p < 8.1e\text{-}16$). These findings remained even when we considered only mutations in regions unaffected by copy number changes which can confound VAF estimates (Figure S2.7). Interestingly, we found that tumors with early HLA mutations had significantly higher levels of neoantigens predicted to specifically bind to the mutated allele than tumors with late HLA mutations (Figure 2.4E). When we evaluated the bias in

specificity of neoantigens for the mutated allele in patients with early HLA loss, we found a significant difference in the number of binding neoantigens between the mutated HLA allele and average of unmutated HLA alleles (Figure 2.4F). We conclude that somatic B2M and HLA mutations are associated with an overall higher burden of neoantigens, supporting the notion that these mutations facilitate tumor immune escape.

Correlation of B2M versus HLA mutation with immune cell infiltration and cytotoxicity. Effective antigen presentation via MHC-I is associated with CD8+ T cell driven cytotoxicity. Furthermore, cell surface MHC-I molecules deliver an inhibitory signal to natural killer (NK) cells. Thus, changes to cell surface presentation of neoantigens by MHC-I due to mutations in B2M and HLA may be reflected in immune cell infiltration levels and levels of cytotoxicity. We quantified immune cell infiltration from tumor RNA sequencing data using Cibersort [19] and levels of cytotoxicity using the score proposed by Rooney et al. [20]. While Shukla et al. previously evaluated immune infiltrates and cytotoxicity in the context of somatic HLA mutations, to our knowledge B2M mutations have not previously been analyzed in this context [10].

CD8+ T cell levels were elevated in tumors with HLA mutations, both pan-cancer (Figure 2.5A) and in several tumor types (Figure 2.5B, Figure S2.8A). A possible explanation is that CD8+ T cells are primed in secondary lymphoid organs and travel to the tumor where they accumulate due to the lack of the corresponding MHC-I molecule / peptide complex. NK cell levels were elevated in tumors with B2M mutations pan-cancer (Figure 2.5C), however the levels were not significantly different in any given tumor type (Figure 2.5D, Figure S2.8B). Loss of B2M resulting in reduced cell surface MHC-I molecules should reduce the ability of tumor cells to inhibit NK cell driven cytotoxicity, however it is unclear whether this would affect NK cell levels in the tumor.

Cytotoxicity was elevated in both HLA and B2M mutant tumors pan-cancer (Figure 2.5E) and in several tumor types (Figure 2.5F, Figure S2.8C). These trends are consistent with the idea that mutations are a mechanism of escape from immune surveillance, as previously suggested by Shukla et al. for HLA mutations [10].

2.5 Discussion

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

Our analysis builds on work by Shukla et al. that first applied Polysolver to evaluate patterns of HLA mutation across tumors and showed that such mutations occurred preferentially in tumors with high mutation burden and under strong pressure by the immune system as evidenced by high levels of CD8+ T cell infiltration [10]. Here we further analyze patterns of mutation in tumors with HLA mutations, incorporating information about which mutations are likely to be presented by MHC-I molecules derived from patient-specific HLA alleles, and comparing to tumors with B2M mutations or with unaltered MHC-I. Our analysis supports a model where B2M mutations reduce the overall levels of cell surface MHC-I molecules while HLA mutations perturb the overall composition of the MHC-I complex landscape, both providing escape from immune surveillance. Our findings are consistent with those of McGranahan et al. who reported that somatic loss of heterozygosity in the HLA locus was a common mechanism of immune evasion, and that loss of a specific HLA allele could render a subset of neoantigens within the tumor ineffective at generating an immune response upon checkpoint inhibition [13]. While both B2M and HLA mutated patients showed elevated mutation rates, we observed differences in how

neoantigens accumulated in these tumors, with B2M mutant tumors harboring the most neoantigens and tumors with intact MHC-I molecules harboring the least.

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

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

affinity bias of new mutations for the mutated HLA allele. Future studies should address these shortcomings.

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

2.6 Materials and Methods

Data. All available whole exome sequencing (WXS) data as of 5/3/2018 was downloaded from The Cancer Genome Atlas (TCGA) database via their Genomic Data Commons (GDC) client. Both .bam files and auxiliary .bai files were downloaded. All available somatic mutation data as of 5/21/2018 was downloaded from the TCGA database in the form of TCGA project mutation annotation files (MAF). Clinical data were also obtained from the GDC (downloaded on 4/25/2017).

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

visualize protein 3D structures (Figure 2.2B). Exon information for HLA proteins is obtained from the IMGT/HLA database (v3.34; https://www.ebi.ac.uk/ipd/imgt/hla/) [28].

HLA typing and mutation calling. HLA genotyping and mutation calling was performed for HLA-A, HLA-B, and HLA-C genes, which encode the human MHC-I complex. We extracted scripts from the Broad Institute's Polysolver Docker container (https://software.broadinstitute.org/cancer/cga/polysolver_run). We verified that the majority of Polysolver's HLA calls were consistent with that of xHLA [29] (Figure S2.9) and therefore used all available Polysolver results. B2M mutations were taken directly from the TCGA MAF files. Patients with somatic B2M or HLA mutations were grouped for subsequent analysis and compared to patients that had neither. We found that only 13 patients had both B2M and HLA mutations.

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

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

Determining regions with CNVs. Regions affected by copy number variants were determined from TCGA affymetrix SNP6 data by using 0.1 thresholds as the cutoff in either direction. Thus, any region that has a log2 fold change larger than 0.1 or smaller than -0.1 is defined as a position with copy number variation [30]. For the Figure S2.7 we excluded any mutations that occurred in regions with copy number variation.

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

Antigen affinity. We used the netMHCpan4.0 tool [18,31] to obtain mutation affinity scores for all patient HLA alleles. To determine whether a mutation would be effectively bound as a neoantigen to the MHC-I complex, we binarized affinity scores: mutations with scores <= 2 we considered binding, and mutations with scores > 2 we considered non-binding [12,18]. We then took the harmonic mean of the best ranking neoantigen to calculate the Patient Harmonic-mean Best Rank (PHBR) score [12]. To evaluate differences in fraction of binding neoantigens at various presentation score (PHBR-I score) cutoffs, we plotted the empirical cumulative distribution function (ECDF) using the median fraction of neoantigens generated from expressed mutations across patients. The Kolmogorov-Smirnov test was used to determine whether the distribution of neoantigen fractions was significantly different for each group (Figure 2.4A). To determine if the number of neoantigens was significantly different between mutated and control patients at a particular PHBR-I score threshold, we calculated p-values using an unpaired Mann Whitney test for pan-cancer comparisons (Figure 2.4B). To test the significance of the number of neoantigens between mutated and unmutated HLA alleles, we used a paired Wilcoxon test (Figures 2.4C, 2.4F). The Kolmogorov-Smirnov, Mann Whitney, and Wilcoxon tests implemented in the scipy.stats Python package were used for these analyses.

Allelic fraction analysis. For Polysolver-determined HLA mutations, we obtained the tumor allelic fraction ("tumor_f") from the Mutect output files generated by Polysolver. For all other mutations we calculated tumor allelic fraction from tumor alternate allele reads ("t_alt_count") and tumor read depth ("t_depth") from TCGA MAF files. B2M and HLA mutations were further annotated according to their percentile within the ranked list of mutations in the tumor where they were observed. To determine if the distributions of patients with B2M and HLA mutations were significantly different than patients without these mutations, we used an unpaired Mann Whitney statistical test from the scipy.stats Python package.

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

Other statistical considerations. Where appropriate, p-values were adjusted for multiple comparisons using the Benjamini-hochberg method [33].

2.7 Figures

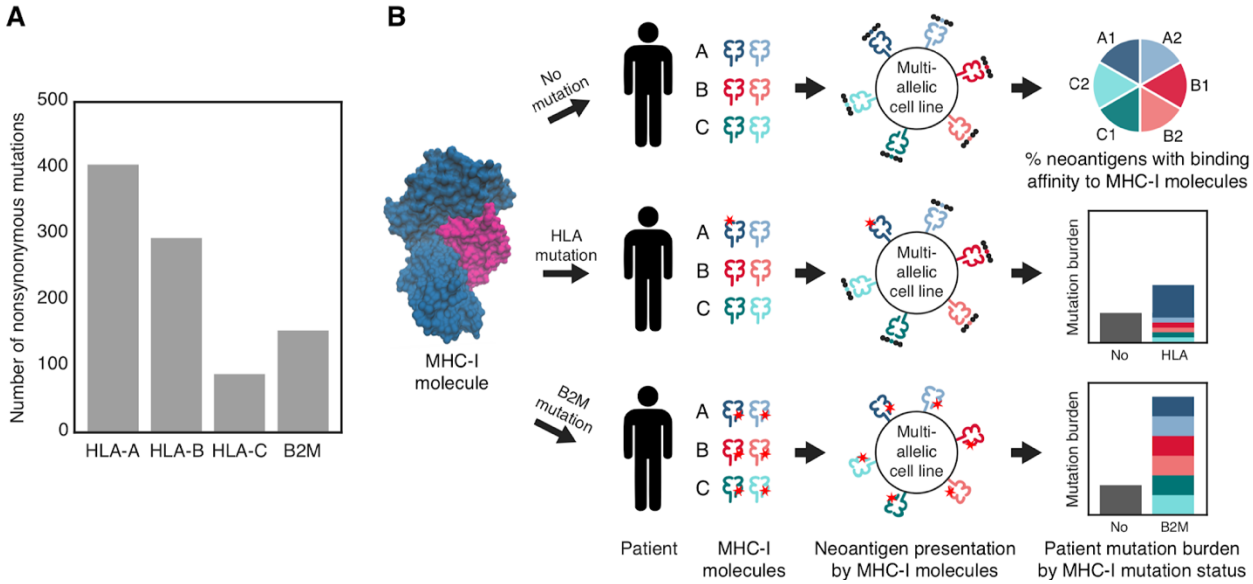
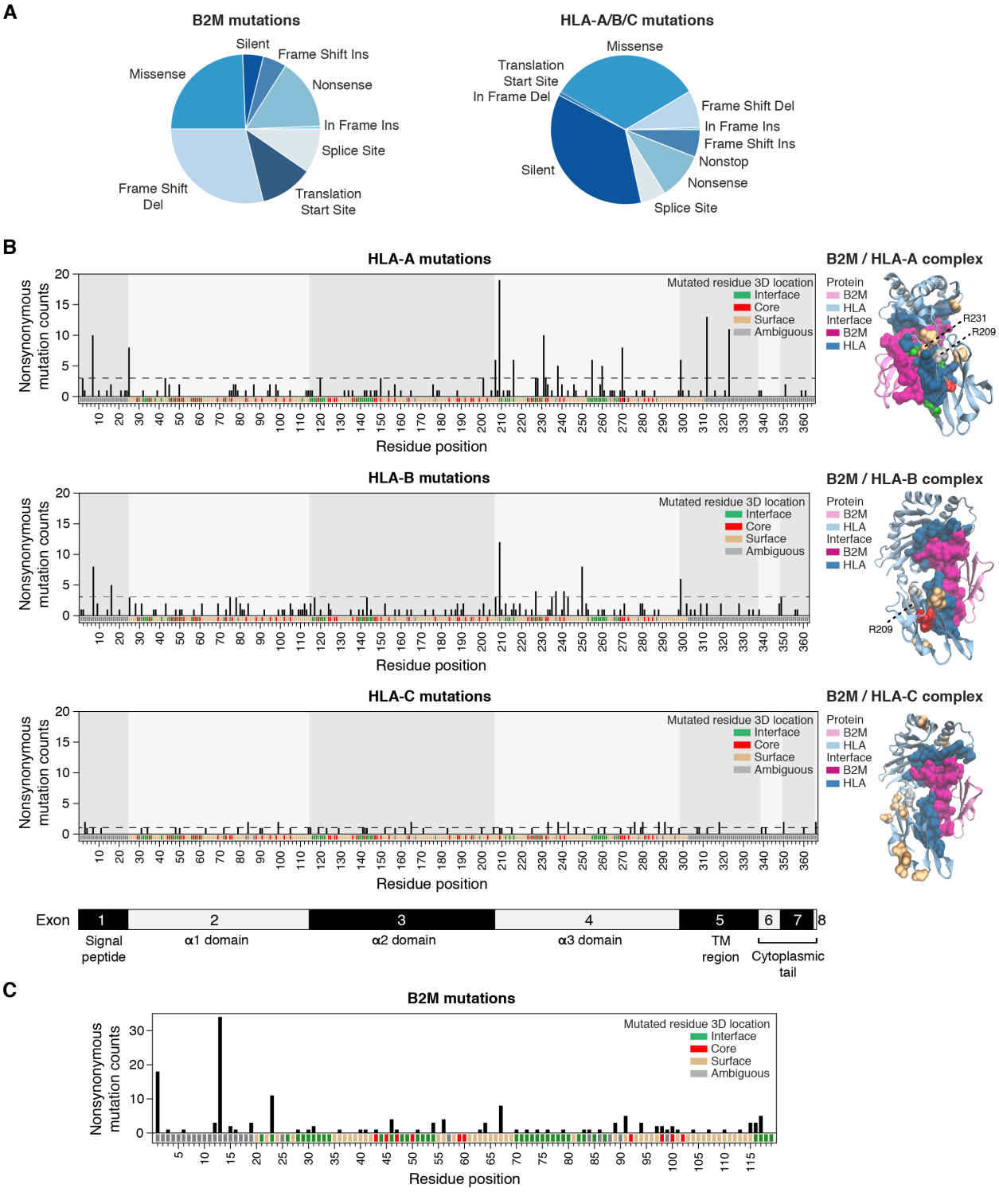


Figure 2.1. Somatic mutations affecting components of the MHC-I molecule. **(A)** The total number of nonsynonymous mutations targeting the genes encoding the components of the MHC-I complex, B2M and HLA-A, HLA-B or HLA-C proteins, across all TCGA patients. The HLA mutation counts were obtained via Polysolver. **(B)** Schematic representation of the effects of mutations that alter the cell surface composition of MHC-I. An HLA mutation will affect a specific MHC-I molecule, whereas a B2M mutation will affect all MHC-I molecules; both mutations can increase the mutation burden of the patient. In the case of an HLA mutation, the patient mutation burden should include more neoantigens with affinity for the mutated HLA allele. The MHC-I molecule displayed is composed of B2M (pink) and HLA-A (blue) proteins (PDB: 3bo8).

Figure 2.2. Mutational analysis of MHC-I complex. **(A)** Pie charts displaying percentages of types of mutation for the B2M protein; and for the combined HLA-A, HLA-B and HLA-C proteins, respectively, across all TCGA patients. **(B)** Distribution of nonsynonymous mutation counts, obtained from Polysolver, for HLA-A, HLA-B, and HLA-C proteins, across functional domains. The corresponding functional domains of HLA proteins are shown at the bottom. The UniProt sequential residue numbering scheme is used for residue numbering, which requires subtraction of the signal peptide (24 residues) for mapping to the IMGT/HLA residue numbering scheme. On the right, 3D crystal structures of MHC-I complex are displayed as B2M and HLA-A complex (PDB: 3bo8), as B2M and HLA-B complex (PDB: 3b3i), and as B2M and HLA-C complex (PDB: 4nt6). Purple ribbons indicate B2M protein, while blue ribbons indicate the HLA proteins. The highlighted purple and blue residues correspond to the interface regions of B2M and HLA proteins, respectively. Hotspot mutations for HLA proteins (frequency>3 for HLA-A, frequency>3 for HLA-B, and frequency>1 for HLA-C) are highlighted as green, red, tan and gray indicating interface, core, surface, and ambiguous residues, respectively. **(C)** Distribution of nonsynonymous mutation counts across the entire B2M protein. On the bottom of the plot, all amino acid residues of B2M protein are colored based on their 3D location: interface, core, surface, or ambiguous.


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

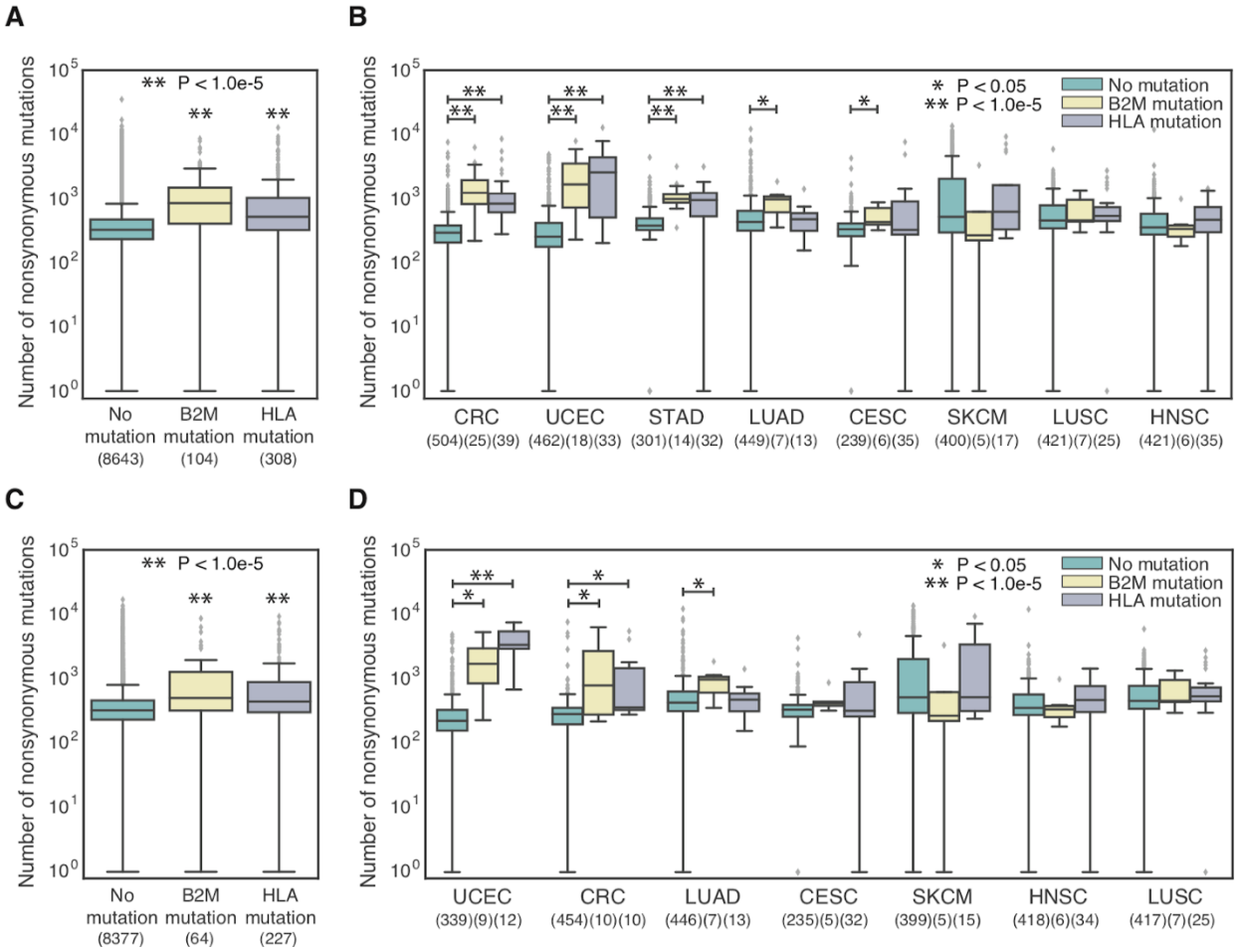


Figure 2.3. Increased mutational burden is related to mutations in MHC-I. (A) and **(C)** Boxplots showing the total number of expressed nonsynonymous mutations of TCGA patients who acquired a mutation in their B2M protein or in one of their HLA alleles versus patients who did not acquire any B2M or HLA mutation, **(A)** for all patients, and **(C)** for only MSS patients. Sample sizes for each patient group are written under their name. **(B)** and **(D)** Boxplots showing total number of expressed nonsynonymous mutations for TCGA patients with or without B2M or HLA mutations, **(B)** for all patients, and **(D)** for only MSS patients. Patients are divided by tumor type and only the tumor types with at least 5 mutated patients are shown. P-values are adjusted for multiple comparisons using the Benjamini–Hochberg procedure. Sample sizes for each patient group are written under the tissue name.

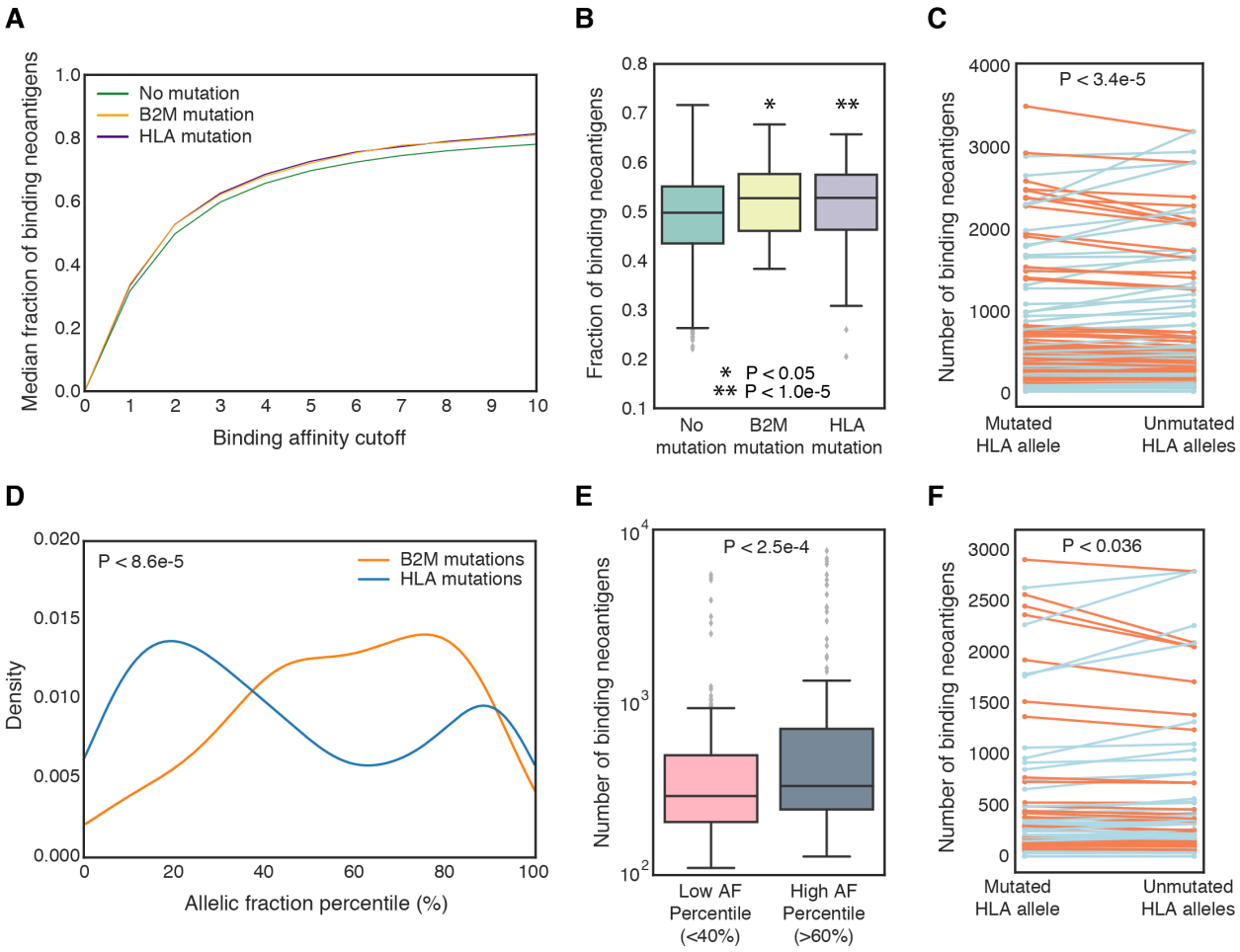


Figure 2.4. Analysis of binding neoantigens to patient HLA alleles. **(A)** Empirical cumulative distribution function showing the proportion of expressed missense and indel mutations labeled as binding neoantigens at different PHBR-I score cutoffs in MSS patients. **(B)** Boxplots comparing the fraction of binding neoantigens in tumors with no B2M or HLA mutation (teal) versus patients with a B2M mutation (yellow) or HLA mutation (purple). A PHBR-I score cutoff of 2 was used to designate a binding neoantigen for this comparison. **(C)** Total number of neoantigens that bind to a patient's mutated HLA allele versus the average number of neoantigens across the unmutated HLA alleles across all cancer types for MSS patients. A red line indicates that there are more neoantigens with binding affinity to the mutated HLA allele than the average across the unmutated HLA alleles; and a blue line depicts the opposite trend. **(D)** Allelic fraction percentile distribution for expressed mutations in MSS patients with B2M and HLA mutations. We used the Kolmogorov-Smirnov statistic to determine whether the two distributions were significantly different. **(E)** Comparison of the number of expressed neoantigens with binding affinity to the patient-specific mutated allele between the low AF percentile (<40%) and the high AF percentile (>60%) HLA mutated patients. Patients with MSI and with mutations in both B2M and HLA genes were excluded. **(F)** Comparison of the total number of neoantigens that bind to a patient-specific mutated HLA allele versus the average number of neoantigens with binding affinity to the five unmutated HLA alleles in patients with high allelic fraction percentile HLA mutations.

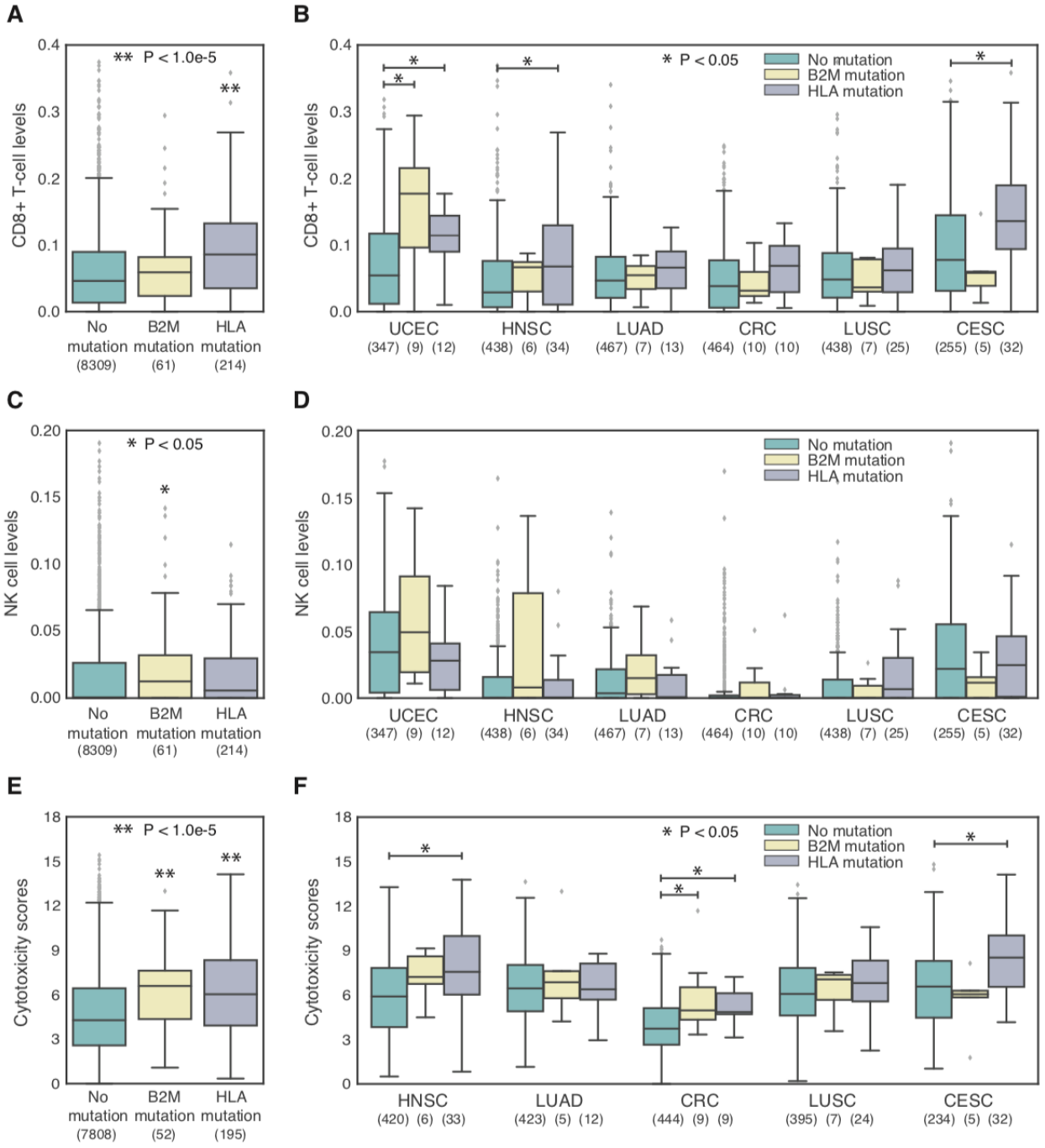


Figure 2.5. Increased NK, CD8+ T-cell and cytotoxicity levels are associated with mutations in MHC-I. (A) and **(C)** and **(E)** Boxplots comparing MSS TCGA patients with or without B2M or HLA mutations, in terms of their **(A)** CD8+ T cell levels, **(C)** natural killer (NK) cell levels, and **(E)** cytotoxicity scores. Sample sizes for each patient group are written under their name. **(B)** and **(D)** and **(F)** Boxplots comparing MSS TCGA patients with or without B2M or HLA mutations, in terms of their **(B)** CD8+ T cell levels, **(D)** natural killer (NK) cell levels, and **(F)** cytotoxicity scores. Patients are divided by tumor type and only the tumor types with at least 5 mutated patients are shown. P-values are adjusted for multiple comparisons. Sample sizes for each patient group are written under the tissue name.

2.8 Supplemental Data, Tables, and Figures

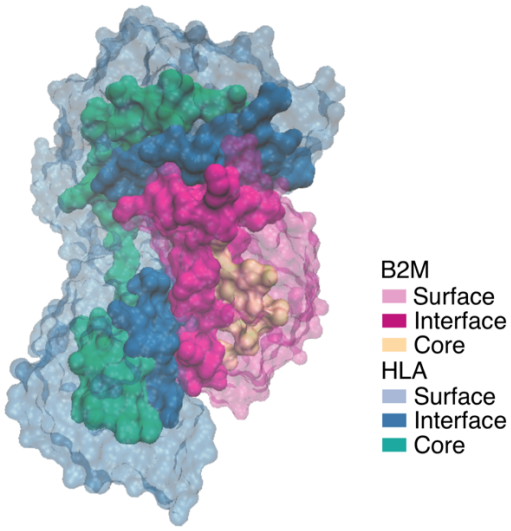


Figure S2.1. MHC-I complex 3D structure. 3D crystal structure of MHC-I complex is displayed as B2M/HLA-A complex (PDB: 3bo8). Interface (blue and violet) and core (green and orange) regions of B2M and HLA-A proteins are highlighted, respectively. Transparent blue and violet regions correspond to the surface regions of B2M and HLA-A proteins, respectively.

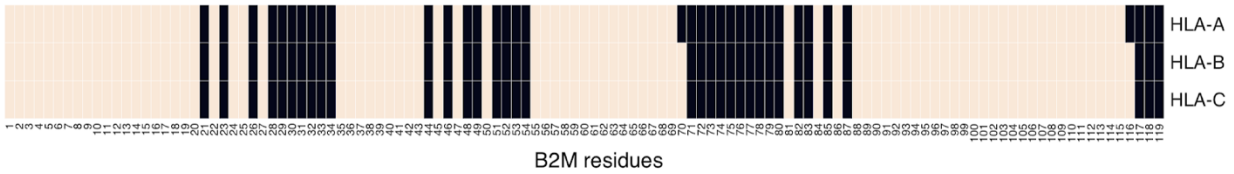


Figure S2.2. B2M interface residue positions for HLA alleles. Residues on B2M that interact with HLA-A, HLA-B, HLA-C proteins are highlighted black.

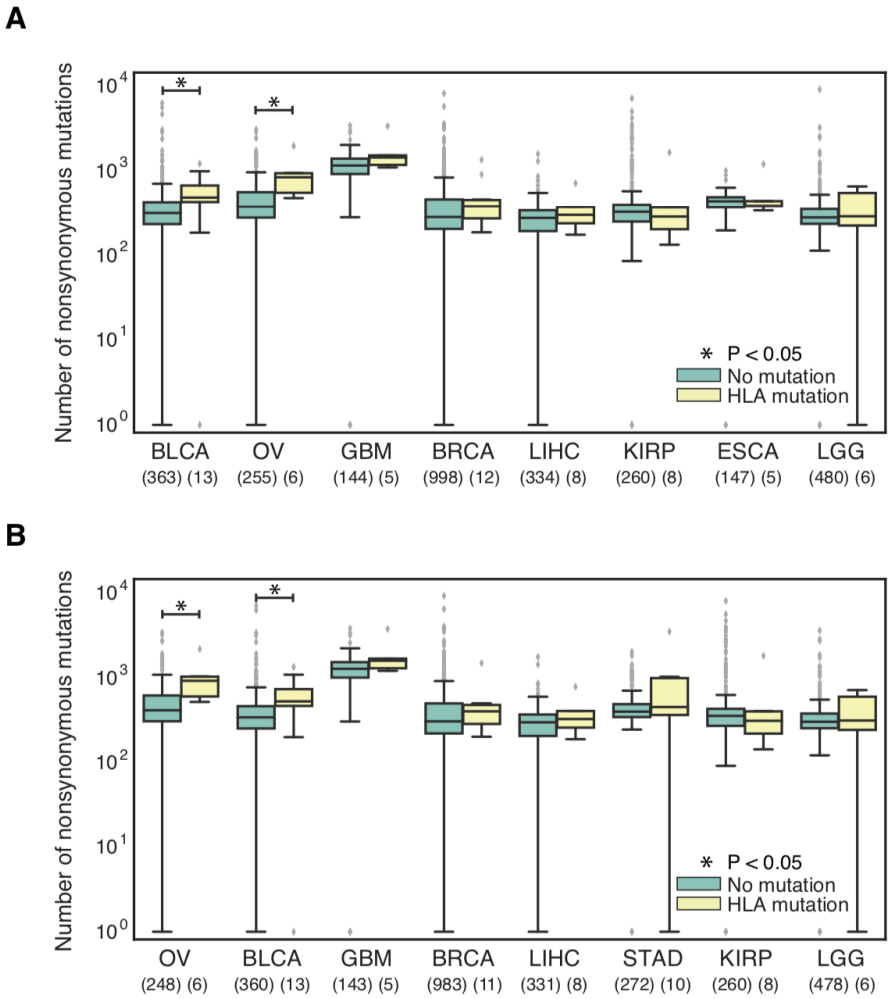


Figure S2.3. Increased mutation burden associated with mutations in HLA, related to Figure 2.3. **(A)** and **(B)** Boxplots showing total number of nonsynonymous mutations for **(A)** MSI and MSS and **(B)** MSS only TCGA patients with or without HLA mutations for additional tissue types not shown in Figures 2.3B, or 2.3D. Patients are divided by tumor type. Only the tumor types containing at least 5 mutated patients and that have not been reported in Figure 2.3 are shown. P-values are adjusted for multiple comparisons using the Benjamini–Hochberg procedure.

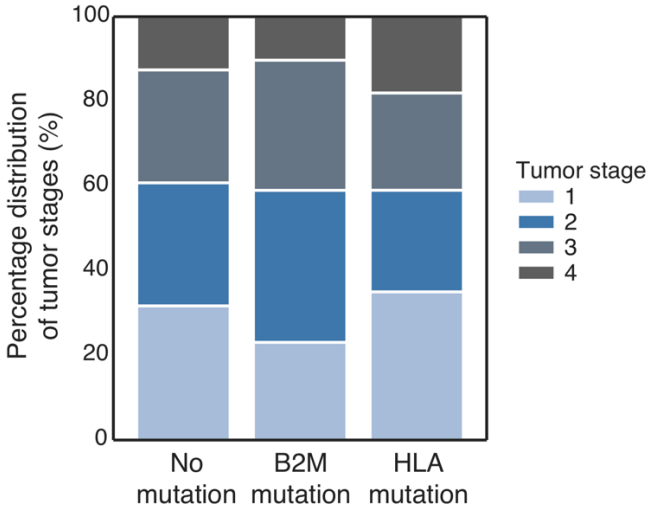


Figure S2.4. Tumor stage analysis for patients with B2M and HLA mutations. Percentage distribution of tumor stages for the patients with or without B2M and HLA mutations.

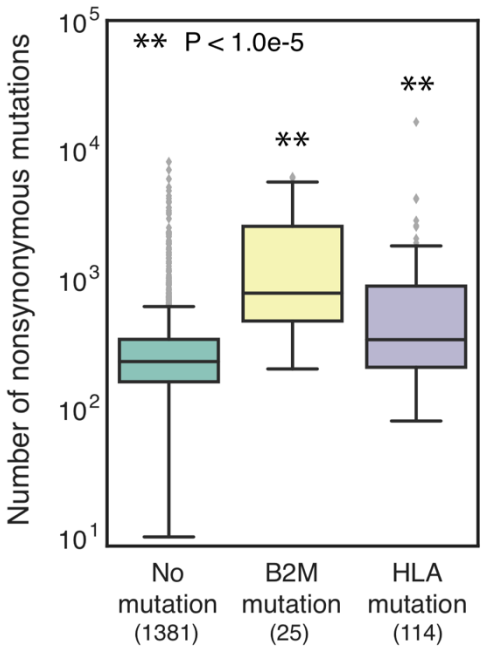


Figure S2.5. Mutation burden in CCLE, related to Figure 2.3. Boxplots showing the total number of nonsynonymous mutations for CCLE cell lines who acquired a B2M or HLA versus cell lines that did not acquire any B2M or HLA mutation. Sample sizes for each group are written under their name.

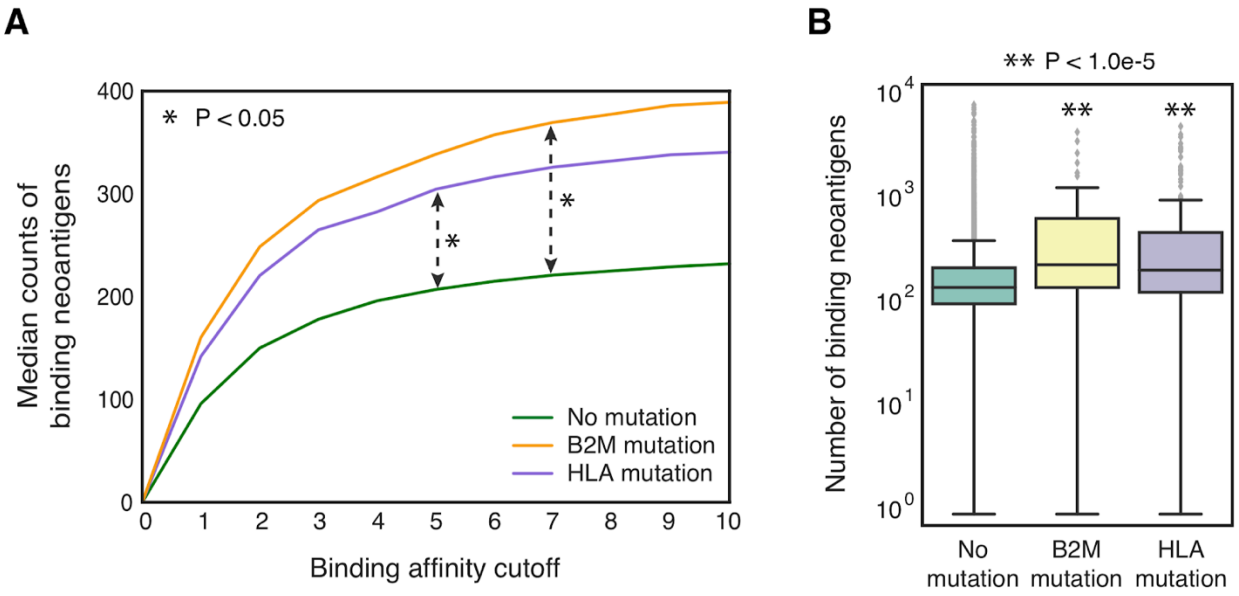


Figure S2.6. Total number of binding neoantigens to patient HLA alleles, related to Figure 2.4. **(A)** Distribution of median total counts of binding neoantigens at different PHBR-I score cutoffs for MSS patients. **(B)** Boxplots comparing the number of neoantigens in MSS patients with no B2M or HLA mutation (teal) versus MSS patients with a B2M mutation (yellow) or an HLA mutation (purple). A PHBR-I score cutoff of 2 was used to designate a binding neoantigen for this comparison.

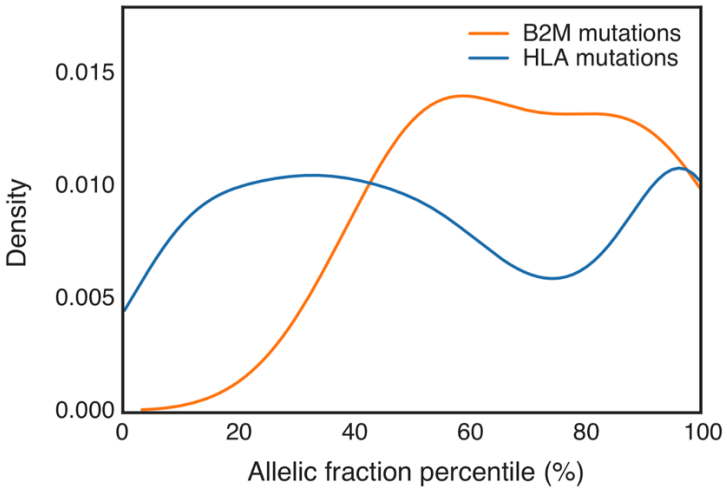


Figure S2.7. **Allelic fraction percentile distribution for patients with B2M and HLA mutations accounting for aneuploidy, related to Figure 2.4**. Allelic fraction percentile distribution for expressed mutations in MSS patients with B2M and HLA mutations, excluding all mutations occurring in regions affected by CNVs. Patients that have both B2M and HLA mutations are excluded.

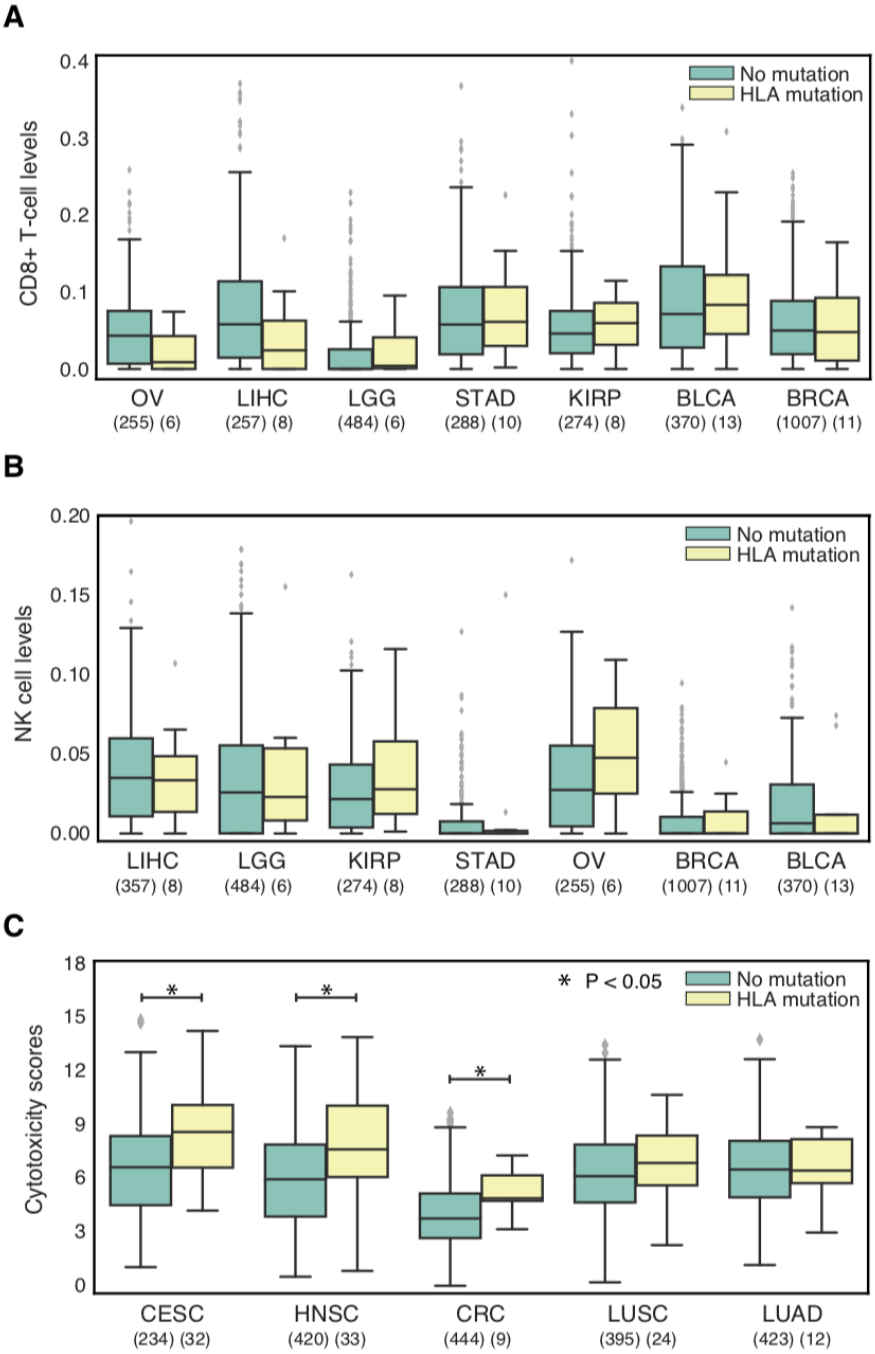


Figure S2.8. NK, CD8+ T-cell and cytotoxicity levels of patients with mutations in HLA, related to Figure 2.5. **(A-B-C)** Boxplots comparing MSS TCGA patients with or without HLA mutations for additional tissue types not shown in Figure 2.5 (B-D-F), in terms of their **(A)** CD8+ T-cell levels, **(B)** NK cell levels, and **(C)** cytotoxicity scores. Patients are divided by tumor type and only the tumor types with at least 5 mutated patients and that have not been reported in Figure 2.5 are shown. P-values are adjusted for multiple comparisons using the Benjamini–Hochberg procedure.

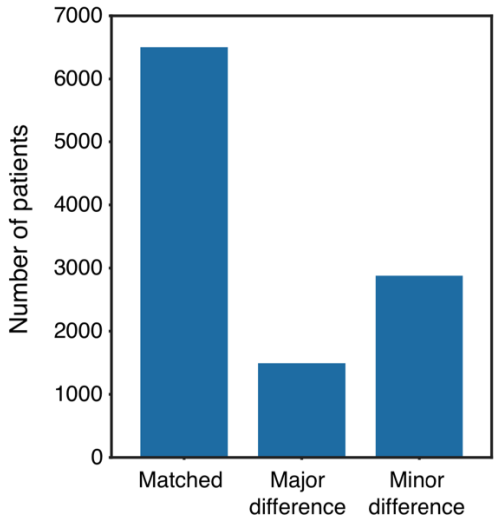


Figure S2.9. HLA allele call comparison between Polysolver and xHLA. Barplot showing matched Polysolver HLA calls and calls with major and minor subtype differences.

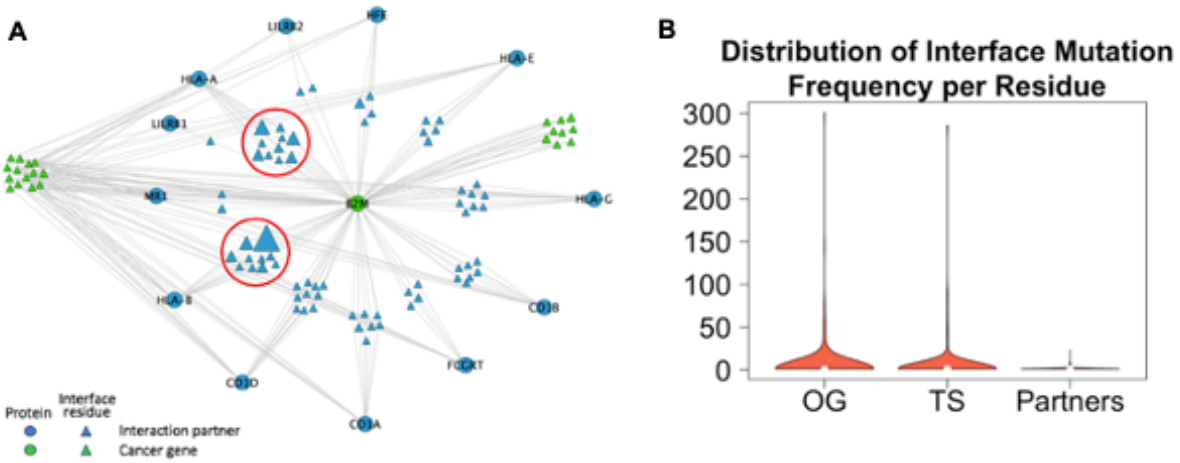


Figure S2.10. Mutational pattern of network architecture perturbation in cancer genes. (A) Mutated structurally resolved network of the tumor suppressor B2M (green circle), its interaction partners (blue circles) and their mutated interface residues (triangles). Green triangles represent mutated interface residues on B2M and blue triangles represent mutated residues on reciprocal interfaces of the associated interaction partner. The size of the triangle corresponds to mutation frequency. **(B)** Violin plots showing the distribution of the number of tumor samples with mutations at a given interface residue.

2.9 Author Contributions

H.C. designed and supervised the research. A.C. and K.O. executed pipelines, analyzed data, and drafted the manuscript. R.M.P. assisted with analysis. S.X. assisted with immune cell infiltration analysis. A.C., K.O., and H.C. designed figures. A.C., K.O., M.Z., and H.C. helped write the manuscript. All authors read and approved the final manuscript.

2.10 Acknowledgements

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

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

Chapter 2, in full, is a reformatted reprint of the material as it appears in "Elevated neoantigen levels in tumors with somatic mutations in the HLA-A, HLA-B, HLA-C and B2M genes" in BMC Medical Genomics, 2019 by Andrea Castro, Kivilcim Ozturk, Rachel M. Pyke, Su Xian, Maurizio Zanetti, and Hannah Carter. The dissertation author was a primary investigator and author of this paper.

2.11 References

1. Engin HB, Kreisberg JF, Carter H. Structure-Based Analysis Reveals Cancer Missense Mutations Target Protein Interaction Interfaces. Srinivasan N, editor. PLoS One. 2016;11: e0152929.

2. Kamburov A, Lawrence MS, Polak P, Leshchiner I, Lage K, Golub TR, et al. Comprehensive assessment of cancer missense mutation clustering in protein structures. Proc Natl Acad Sci U S A. 2015;112: E5486–95.

3. Porta-Pardo E, Garcia-Alonso L, Hrabe T, Dopazo J, Godzik A. A Pan-Cancer Catalogue of Cancer Driver Protein Interaction Interfaces. Nussinov R, editor. PLoS Comput Biol. 2015;11: e1004518.

4. Raimondi F, Singh G, Betts MJ, Apic G, Vukotic R, Andreone P, et al. Insights into cancer severity from biomolecular interaction mechanisms. Sci Rep. 2016;6: 34490.

5. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144: 646–674.

6. Schumacher TN, Schreiber RD. Neoantigens in cancer immunotherapy. Science. 2015;348: 69–74.

7. Rosenberg SA. Raising the bar: the curative potential of human cancer immunotherapy. Sci Transl Med. 2012;4: 127ps8.

8. Sade-Feldman M, Jiao YJ, Chen JH, Rooney MS, Barzily-Rokni M, Eliane J-P, et al. Resistance to checkpoint blockade therapy through inactivation of antigen presentation. Nat Commun. 2017;8: 1136.

9. del Campo AB, Kyte JA, Carretero J, Zinchencko S, Méndez R, González-Aseguinolaza G, et al. Immune escape of cancer cells with beta2-microglobulin loss over the course of metastatic melanoma. International journal of cancer. 2014;134: 102–113.

10. Shukla SA, Rooney MS, Rajasagi M, Tiao G, Dixon PM, Lawrence MS, et al. Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes. Nat Biotechnol. 2015;33: 1152–1158.

11. Kambayashi T, Michaëlsson J, Fahlén L, Chambers BJ, Sentman CL, Kärre K, et al. Purified MHC class I molecules inhibit activated NK cells in a cell-free system in vitro. Eur J Immunol. 2001;31: 869–875.

12. Marty R, Kaabinejadian S, Rossell D, Slifker MJ, van de Haar J, Engin HB, et al. MHC-I Genotype Restricts the Oncogenic Mutational Landscape. Cell. 2017;171: 1272–1283.e15.

13. McGranahan N, Rosenthal R, Hiley CT, Rowan AJ, Watkins TBK, Wilson GA, et al. Allele-Specific HLA Loss and Immune Escape in Lung Cancer Evolution. Cell. 2017;171: 1259–1271.e11.

14. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res. 2000;28: 235–242.

15. Zijlstra M, Bix M, Simister NE, Loring JM, Raulet DH, Jaenisch R. Beta 2-microglobulin deficient mice lack CD4-8+ cytolytic T cells. Nature. 1990;344: 742–746.

16. Koller BH, Marrack P, Kappler JW, Smithies O. Normal development of mice deficient in beta 2M, MHC class I proteins, and CD8+ T cells. Science. 1990;248: 1227–1230.

17. Kautto EA, Bonneville R, Miya J, Yu L, Krook MA, Reeser JW, et al. Performance evaluation for rapid detection of pan-cancer microsatellite instability with MANTIS. Oncotarget. 2017;8: 7452–7463.

18. Hoof I, Peters B, Sidney J, Pedersen LE, Sette A, Lund O, et al. NetMHCpan, a method for MHC class I binding prediction beyond humans. Immunogenetics. 2009;61: 1–13.

19. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods. 2015;12: 453–457.

20. Rooney MS, Shukla SA, Wu CJ, Getz G, Hacohen N. Molecular and genetic properties of tumors associated with local immune cytolytic activity. Cell. 2015;160: 48–61.

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

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

23. Grasso CS, Giannakis M, Wells DK, Hamada T, Mu XJ, Quist M, et al. Genetic Mechanisms of Immune Evasion in Colorectal Cancer. Cancer Discov. 2018;8: 730–749.

24. Hubbard SJ, Thornton JM. NACCESS: Department of Biochemistry and Molecular Biology, University College London. Software available at http://www bioinf manchester ac uk/naccess/nacdownload html. 1993.

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

26. Martin ACR. Mapping PDB chains to UniProtKB entries. Bioinformatics. 2005;21: 4297–4301.

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

28. Robinson J, Halliwell JA, Hayhurst JD, Flicek P, Parham P, Marsh SGE. The IPD and IMGT/HLA database: allele variant databases. Nucleic Acids Res. 2015;43: D423–31.

29. Xie C, Yeo ZX, Wong M, Piper J, Long T, Kirkness EF, et al. Fast and accurate HLA typing from short-read next-generation sequence data with xHLA. Proc Natl Acad Sci U S A. 2017;114: 8059–8064.

30. Beroukhim R, Mermel CH, Porter D, Wei G, Raychaudhuri S, Donovan J, et al. The landscape of somatic copy-number alteration across human cancers. Nature. 2010;463: 899–905.

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

32. Patro R, Mount SM, Kingsford C. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. Nat Biotechnol. 2014;32: 462–464.

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

CHAPTER 3: Interface-guided phenotyping of coding variants of the transcription factor RUNX1 with SEUSS

3.1 Foreword

Understanding the consequences of single amino acid substitutions in driver genes remains an unmet need. High-throughput mutagenesis is emerging as a powerful tool to probe the varying consequences of different amino acid substitutions across the length of a protein or protein domain, however it is currently limited to specific functional readouts such as target protein abundance [1] or functional assays [2–4]. Studying the effects of genetic perturbations on cellular programs and fitness has been challenging using traditional pooled screens. Over the last few years, there has been a surge of interest in Perturb-seq [5] style assays to measure the transcriptional consequences of genetic perturbations ranging from whole gene knockout to amino acid substitutions in single cells [6]. While providing greater function insight, these sequencing-based methods are not yet scalable to exhaustive mutagenesis, necessitating selection of target mutations.

Motivated by the improvement to variant effect interpretation by integration of protein structure and network information in Chapters 1 and 2, I hypothesized that examining the consequences of perturbing distinct protein interactions could provide a useful abstraction of the phenotypic space reachable by individual amino acid substitutions. To explore this hypothesis, in Chapter 3, I employed an interface-guided Perturb-seq style approach to generate mutations at physical interfaces of the transcription factor RUNX1, with the potential to perturb different interactions, and therefore produce transcriptional readouts implicating different aspects of the RUNX1 regulon. I analyzed these readouts to identify functionally distinct groups of RUNX1 mutations, characterize their effects on cellular programs and study the implications for cancer mutations.

3.2 Abstract

Understanding the consequences of single amino acid substitutions in driver genes remains an unmet need. Perturb-seq provides a tool to investigate the effects of individual mutations on cellular programs. Here we deploy SEUSS, a Perturb-seq like approach, to generate mutations at physical interfaces of the RUNX1 Runt domain with proteins and DNA. We measured the impact of more than 100 mutations on RNA profiles in single myelogenous leukemia cells and used the profiles to categorize mutations into three functionally distinct groups: WT-like, LOF-like and hypomorphic. The largest concentration functional mutations (non-WT-like) clustered at the DNA binding site and contained many of the more frequently observed mutations in human cancers. Hypomorphic variants shared characteristics with loss of function variants but had gene expression profiles indicative of a shift toward endothelial cell lineage and glutamate catabolism. DNA accessibility changes were enriched for RUNX1 binding motifs, particularly near differentially expressed genes. Our work demonstrates the potential of targeting protein interaction interfaces to better define the landscape of prospective phenotypes reachable by amino acid substitutions.

3.3 Introduction

Cancer is a disease associated with progressive loss of cell identity and gain of signals promoting inappropriate survival and proliferation. It is now well established that somatic DNA mutations, particularly in oncogenes and tumor suppressors, can drive tumor development and progression [7–10]. It is also increasingly evident that different somatic mutations in the same gene can have different associations with prognosis and therapeutic response. For example, KRAS G13 but not G12 mutant colorectal cancers are sensitive to treatment with cetuximab [11]. In lung cancer, the G13 mutation is associated with shorter overall survival than the G12 mutation [12]. In breast cancer, TP53 mutations within DNA binding motifs are associated with worse prognosis than those outside, but within DNA binding motifs, mutations at codon 179 and the R248W substitution are associated with significantly poorer prognosis than other substitutions [13]. This highlights the need to develop strategies for studying perturbations at a finer scale than gene knock-out or knockdown.

High-throughput mutagenesis is emerging as a powerful tool to probe the varying consequences of different amino acid substitutions across the length of a protein or protein domain, however it is currently limited to specific functional readouts, such as target protein abundance [1] or functional assays [2–4]. Studying the effects of genetic perturbations on cellular programs and fitness has been challenging using traditional pooled screens. Approaches such as ScalablE fUnctional Screening by Sequencing (SEUSS) [14] and Perturb-seq [5] measure the transcriptional consequences of perturbations ranging from whole gene knockout to amino acid substitutions [6] in single cells, the latter making it possible to gain insight into the differences that distinguish mutations at the level of cellular programs relevant to cancer progression. SEUSS has been used to study the consequences of functional domain deletions and hotspot mutations to MYC in human

pluripotent stem cells [14]. Perturb-seq applied to study driver mutations in KRAS and TP53 revealed that these mutations span a continuum of function that is not necessarily predicted by mutation frequency in cancer cohorts [6]. While providing greater function insight, these sequencing-based methods are not yet scalable to exhaustive mutagenesis, necessitating selection of target mutations.

Individual proteins often have multiple functions within cells, mediated through interaction with different binding partners. Somatic mutations in driver genes are overrepresented at interaction interfaces [15–19], suggesting that examining the consequences of perturbing distinct protein interfaces could provide a useful abstraction of the phenotypic space reachable by individual amino acid substitutions. To explore this hypothesis we sought to design a library of mutations perturbing distinct physical interfaces of RUNX1, a transcriptional master regulator of hematopoiesis implicated in multiple cancer types, then use the resulting transcriptional landscape to examine cancer-associated mutations.

Runt-related transcription factor 1 (RUNX1) has been implicated in multiple tumor types such as acute myeloid leukemia and breast cancer [20,21]. The protein encoded by RUNX1 forms the heterodimeric complex core-binding factor (CBF) together with CBFB, and is thought to be involved in the development of normal hematopoiesis [22,23]. CBF binds to the core element of many enhancers and promoters through the Runt domain of RUNX1 [22,24]. Specifically, RUNX members (RUNX1–3) modulate the transcription of their target genes by recognizing the core consensus binding sequence 5-TGTGGT-3, or very rarely, 5-TGCGGT-3, within their regulatory regions via their Runt domain [25]; while CBFB, a non-DNA-binding regulatory subunit in the CBF complex, binds to the Runt domain and leads to increased DNA-binding affinity [26]. A

variety of transcriptional co-regulatory proteins bind RUNX1 via the Runt domain and modulate CBF activity [27].

As RUNX1 is a pioneer transcription factor and master regulator, we hypothesized that mutations affecting its interactions with transcriptional cofactors would manifest as changes to the expression of multiple RUNX1 target genes and thus potentially result in functional diversity. We applied SEUSS to study perturbations to protein interaction interfaces of the RUNX1 Runt domain, which harbors the majority of recurrent cancer mutations. Prior work has estimated that in Perturb-seq style assays, transcriptional profiles of 100-400 single cells should be sufficient to detect variant effects on gene expression, thus we planned for a library size of approximately 120 variants, including wild type (WT) and loss of function (LOF) controls. Thus, we prioritized variants with the potential to perturb distinct interactions, and therefore generate distinct transcriptional readouts implicating different aspects of the RUNX1 regulon. We analyzed these readouts to identify functionally distinct groups of RUNX1 mutations, characterize their effects on cellular programs and study the implications for cancer mutations.

3.4 Results

An interface-guided Perturb-seq assay for coding variant phenotyping of transcription factor RUNX1. We first determined which amino acid substitutions to study. While somatic mutations in RUNX1 span the entire gene, the most recurrent mutations cluster in the Runt domain that binds DNA and includes binding sites for multiple protein partners, including CBFB (Figure 3.1a). We used protein structures and template-based docking [28] to identify amino acid residues of the RUNX1 Runt domain involved in physical interactions with 33 protein partners with structural data (Figure 3.1b). Eighty-three residues within the Runt domain were identified to mediate at least one interaction. These were used to design an open reading frame (ORF) mutation library to assess the impact of perturbation of various RUNX1 interactions.

For each of the 83 residues, we identified amino acid substitutions that would maximally perturb function based on VEST scores [29] and FoldX scores [30]. While VEST is trained to discriminate pathogenic from neutral amino acid substitutions, FoldX estimates the effect of an amino acid substitution on the free energy of a protein. Where possible we also prioritized substitutions observed in tumors from the COSMIC database for these positions [31]. To provide a frame of reference for functional impact, we included WT RUNX1 and LOF controls (RUNX1 replaced with GFP), along with 17 negative controls (expected to be indistinguishable from WT) consisting of 10 silent and 7 predicted neutral (based on VEST scores) mutations, and 10 positive controls (expected to have similar impact to LOF) consisting of 5 truncating and 5 core mutations. In order to evaluate the impact of multiple mutations, we also included 5 mutation combinations, bringing the total to 117 library elements.

We used ScalablE fUnctional Screening by Sequencing (SEUSS) [14] to express the mutant RUNX1 library in K562 cells [32], with and without doxycycline-inducible knockdown of

endogenous RUNX1. Our library was generated from a modified lentiviral vector to contain a hygromycin resistance enzyme gene downstream of the EF1a promoter, followed by a P2A peptide motif, the RUNX1 variant (WT, mutated, or GFP), and a 12 base pair barcode sequence unique to each variant for identification during single cell transcriptome sequencing. The cells were transduced with the pooled variant library at a MOI of ~0.3 to ensure that each cell received a single construct, and were grown with hygromycin to select for cells carrying constructs. At day 7 post transduction, single cell RNA libraries were prepared and sequenced, with the remainder of cells being maintained until day 14 for fitness screening (Figure 3.1c).

We obtained 38,104 cells from sequencing of which 56.3% contains detectable variant barcodes assigned to a single variant only. Four variants (G138V, S145I, P157R, T161I) were excluded due to low cell counts, and one negative control was removed (G143G) due to a frame shift artifact during the mutation library preparation. After additional quality control (QC) filtering, we recovered 20,878 high-quality cells with detectable variant barcodes assigned to a single variant only, covering 112 of 117 assayed variants for downstream analysis (Methods, Figure S3.1).

We next analyzed variants according to their transcriptional profiles. We reasoned that variants with similar effects on RUNX1 targeting should cluster together, while those with distinct functional effects should cluster separately. To provide a frame of reference for functional impact, we compared expression profiles between cells harboring variants and cells carrying the WT or LOF control constructs. Variants were annotated as “wildtype-like” (WT-like) if the induced expression changes were indistinguishable from cells carrying the WT construct, and as “loss-of-function-like” (LOF-like) if indistinguishable from cells with the LOF construct. Any variant that did not have a WT-like expression profile was considered “functional”.

Unsupervised clustering of cells and variants. Single cell gene expression profiles were used to perform unsupervised clustering, which supported 3 clusters of cells (Figure 3.2a). Cluster 1 harbored the majority of cells with the WT construct or negative control mutations expected to be functionally WT (log(OR)=1.69, p<7.39e-19, and log(OR)=2.69, p<6.79e-301, respectively, Figure S3.2), whereas the LOF construct and positive control mutations were most enriched in clusters 2 and 3 (log(OR)=0.28, p<0.01 for LOF, and log(OR)=0.34, p<9.68e-15 for positive controls for cluster 2; and log(OR)=0.87, p<2.22e-15 for LOF, and log(OR)=0.88, p<2.54e-86 for positive controls for cluster 3) (Figure 3.2b, Figure S3.2). Though cells sharing certain variants showed bias toward one side or the other of the UMAP, they did not occupy discrete regions. Thus in order to better quantify differences between variants, we compared their average gene expression profiles.

For each variant, we averaged the expression of each of the top 2000 variable genes across all cells corresponding to the variant. We performed unsupervised clustering of variants based on these profiles, and visualized them in UMAP space, where unsupervised clustering again suggested three groups (Figure 3.2c). Group I included the WT construct and all negative control variants. Group III contained the LOF construct and the majority of positive control variants (8 of 10) expected to result in a truncated or unstable protein (Figure 3.2d). Most of the tested variants also fell into these groups, reflecting expression profiles similar to WT or LOF variants, whereas the separate assignment of 14 variants to cluster II suggested some partial loss of RUNX1 function, distinct from LOF or WT activity. Accordingly, we labeled variants in these groups as WT-like, hypomorphic and LOF-like (Figure 3.2e).

Revisiting the single cell visualization with these labels, cells carrying hypomorphic variants were most prevalent near the boundary between clusters 1 and 2 (Figure 3.2f) and were

most enriched in cluster 2 (log(OR)=0.74, p<9.92e-89) (Figure 3.2g). Cell cluster 1 was highly enriched for cells harboring WT-like variants (log(OR)=3.07, p<2.22e-308), while cells harboring LOF-like variants were enriched both in clusters 2 (log(OR)=0.85, p<5.75e-172) and 3 (log(OR)=2.22, p<2.22e-308) (Figure 3.2g).

To understand whether hypomorphic variants represent partial function, falling between normal and complete loss of function, versus gain of new activity, we further investigated similarity to WT and LOF constructs. We quantified the differences in expression profiles with Hotelling's two-sample T-squared (T2) test, a multivariate generalization of Student's t-test [33] using lower dimensional representation of cells in the principal component space (Methods). Higher scores indicate a higher deviation from the variant being compared, WT or LOF. This statistical analysis revealed that all WT-like variants are indistinguishable from the WT construct via small T2 scores relative to WT (T2WT), but high T2 scores relative to the LOF control (T2LOF) (p<0.05, Figure 3.2h). Similarly, all LOF-like variants are indistinguishable from the LOF construct via small T2LOF scores, but have high T2WT scores (p<0.05, Figure 3.2h). Hypomorphic variants are significantly different from both the WT and LOF controls (p<0.05, Figure 3.2h) suggesting that these variants result in transcriptional changes that are not simply an intermediate between LOF and WT. This result is supported by differential gene expression analysis where 48 of 141 genes that are differentially expressed between hypomorphic variants versus the WT control, are not differentially expressed between LOF versus WT and 107 of 232 genes that are differentially expressed between hypomorphic variants versus the LOF control, are also unique, suggesting gain of new activity.

Our single cell data were derived from the pooling of two biological replicates, each of which was evaluated for cell fitness at day 14 (Methods). Fitness measurements were highly

concordant between replicates (Pearson correlation; r=0.94, p<5.32e-55) (Figure 3.2i), were positively correlated with variant T2WT scores (Pearson correlation; r=0.85, p<1.25e-32), and negatively correlated with T2LOF scores (Pearson correlation; r=-0.77, p<8.40e-24). In general, LOF-like variants tended to associate with increased fitness, consistent with RUNX1 previous reports that reduction or loss of RUNX1 activates cell proliferation [34,35].

Gene expression programs distinguishing RUNX1 variants. We next performed hierarchical clustering of the top 2000 variable genes across 112 RUNX1 Runt domain variants to study transcriptional similarities and differences (Figure 3.3a). The resulting dendrogram once again separated WT-like, LOF-like and hypomorphic variants, and highlighted groups of genes that showed similar expression patterns across groups. While all the dendrograms for all three groups showed additional substructure, we were particularly interested in the hypomorphic variants which appeared to separate into 3 groups, which we referred to as hypomorphic-I, -II and -III, to reflect the progression of gene expression changes observed across the variants (Figure 3.3a).

To better understand how gene expression patterns distinguished these groups, we evaluated functional enrichment of gene clusters associated with each. The top 3 gene clusters showed generally higher expression in LOF-like variants relative to WT, while cluster 4 showed the opposite trend, and clusters 5-10 were more variable (Figure 3.3a, Figure S3.3, Table S3.1). In the hypomorphic group clusters 7 and 9 had lower and higher aggregated mean expression across genes respectively relative to WT and LOF like variant groups. Cluster 4 genes were enriched for immune related functions such as T helper cell lineage commitment and immune cell activation (Figure S3.4), consistent with RUNX1's role in hematopoietic lineage commitment and differentiation [36–38]. LOF variants promoted gene expression programs associated with

hemoglobin levels [36,39], vasculogenesis and ECM regulation in clusters 1-3 respectively. Clusters 5 and 6 were enriched for genes related to the cell cycle. Cluster 8 was enriched for epithelial cell migration, whereas cluster 10, which was highest in WT cells, was enriched for negative regulation of meiotic cell cycle and ER stress, and cluster 7 was enriched for genes involved in secretion, synaptic plasticity and axon guidance (Figure S3.4). Overall, these results suggest that loss of RUNX1 activity resulted in failure to promote differentiation toward immune cell identity and reactivation of developmental programs and shifts in metabolism related to oxidative phosphorylation and hemoglobin [40–42].

The majority of variance in gene expression fell along the WT-like to LOF-like axis (PC1, 31.6% variance explained), both visually (Figure 3.3b), and statistically as it positively correlates with T2WT scores ($r=0.90$, $p<2.2e-16$), and fitness scores ($r=0.93$, $p<2.2e-16$). On the other hand, PCs 2, 3 and 4 (3.4%, 3.1%, and 2.6% variance explained, respectively) seem to correspond to transcriptomic effects more specific to the hypomorphic variants (Figure 3.3b). The most influential genes based on PC loadings were enriched for endothelial cell development and morphogenesis, glutamate catabolism, MAP kinase signaling, cell cycle and mitotic related activities, suggesting these functions most distinguish hypomorphic variants from WT and LOF variants.

Though a non-linear mapping, the first two UMAP dimensions were strongly inversely correlated with PC1 (Figure 3.3b,c). T2WT, T2LOF (Figure 3.3d) and fitness scores (Figure 3.3e) largely agreed with the hierarchical clustering, though variants in hypomorphic II appeared to have both more distinct transcriptional profiles from T2WT and T2LOF as well as higher fitness scores than the the other hypomorphic variants.

LOF-like variants tended to have higher FoldX scores and VEST scores (Figure 3.3f,g). VEST scores were more strongly correlated with T2WT scores and fitness scores than FoldX scores (T2WT: r=0.42, p<4.88e-06 vs r=0.38, p<6.31e-05; fitness r=0.43, p<3.23e-06 vs r=0.27, p<4.06e-03) suggesting that amino acid characteristics associated with pathogenicity are more likely to cause larger deviations from WT gene expression programs.

Kernel density estimates of single cell UMAP embeddings for variant groups demonstrate that the majority of cells belonging to each assigned phenotype (WT-like, hypomorphic-I, -II, -III, and LOF-like) occupy discrete regions in UMAP, though hypomorphic distributions seemed to harbor cells that also overlapped more closely with regions dominated by WT-like and LOF-like cells (Figure 3.3h). This could potentially be due to small differences in expression of mutant construct, stochasticity in measurement of gene expression, or could represent variable penetrance at the cellular level due to buffering built into cellular systems, such as stress response pathways.

LOF-like variants significantly target RUNX1-DNA binding. We first investigated the impact of mutations at the interface of protein-nucleotide interactions of RUNX1. We identified 11 amino acid residues of RUNX1 (R80, R135, R139, R142, G143, K167, T169, V170, D171, R174, R177) as involved in DNA binding based on the 3D crystal structure of the interaction, among which 8 residues are perturbed in our mutation library (R80G, R135G, R139Q, R142S, G143R, T169I, V170M, R174Q) (Figure 3.4a). These 8 mutations were significantly enriched for LOF-like impact vs. WT-like (Fisher's exact test; OR=13.33, p<0.0084) and for functional (LOF-like or hypomorphic) vs. WT-like impact (Fisher's exact test; OR=9.03, p<0.025, Figure 3.4b), suggesting that interruption of DNA-binding is very damaging to the RUNX1 function and the impact can be deduced from the transcriptional profiles of the mutations.

In comparison, for the 19 residues identified as involved in the interaction with CBFB (Figure 3.4a), severity of phenotypic impact of mutations was not highly enriched compared to direct DNA binding. 10 of the mutations had LOF-like or hypomorphic impact while 9 had WT-like consequences (Fisher's exact test; OR=1.27, p<0.79, Figure 3.4b). We saw that functional (LOF-like or hypomorphic) mutations generally had higher VEST and FoldX scores, though not statistically significant (Mann Whitney-U test; VEST: p<0.96; FoldX: p<0.17), which could suggest that mutations resulting in WT-like phenotypes were not as damaging to the protein complex, possibly due to amino acid residues surrounding them that are involved in the interaction are able to maintain enough of the binding.

Comparing coding variant Perturb-seq with recurrence in cancer. In principle, positions with functional variants that improve fitness would be under higher positive selective pressure in tumors and would thus be more frequently mutated across patients. Thus, we tested whether expression-based phenotyping was associated with mutation frequency in cancer (COSMIC [31]). Frequency was weakly positively correlated with fitness (Pearson correlation; r=0.312, p<1.21e-03), but not with T2WT scores (Pearson correlation, r=0.136, p<0.16, Figure 3.4c). We observed a significant enrichment of cancer mutations (frequency>0) (Figure 3.4d) with LOF-like vs. WT-like impact (Fisher's exact test; OR=4.34, p<0.0084, Figure 3.4e), and functional vs. WT-like impact (Fisher's exact test; OR=3.01, p<0.021, Figure 3.4e), meaning that cancer mutations are more likely to be damaging/impactful. This impact is even more significant for cancer mutations with higher prevalence (frequency>2) (Fisher's exact test; OR=15.47, p<0.0035, and OR=12.19, p<0.0059, respectively, Figure 3.4e). Among the 5 most frequently observed mutations (Figure 3.4c), four of them (R174Q, R139Q, R135G and P173S) target amino acid

residues involved in DNA binding and display LOF-like impact; while the remaining one (S114L) targets a residue involved in CBFB binding and is hypomorphic (Figure 3.4d).

Of the RUNX1 mutations shared between our assay and the COSMIC database, the majority of occurrences were in hematopoietic malignancies (n=104), followed by breast cancer (n=10), the urinary tract (n=5) and the large intestine (n=4). Across these four tumor types, approximately 79.6% of observed variants were LOF-like and 14.6% were hypomorphic (Figure 3.4f).

Variant impact on the RUNX1 regulon. Next, we sought to validate the hypomorphic effects variants with bulk RNA sequencing, and investigate whether the variants altered RUNX1 binding and consequently DNA accessibility for regulatory elements associated with differentially expressed genes. We selected 12 variants to study, including the WT and LOF control variants and 9 hypomorphic variants from all three sub-categories: two hypomorphic-I (N82I, P156R), six hypomorphic-II (L62P, G95R, V97D, G100V, V137D, I166S) and one hypomorphic-III (R118G). We also included one LOF-like variant (V159D) predicted to be involved in RUNX1-CBFB binding to further investigate the effects of interruption of this interaction (Table 3.1). Two of the hypomorphic variants (G95R and R118G) were observed in human tumors previously (Figure 3.5a, Table 3.1).

We performed bulk RNA- and ATAC-seq screens for each variant separately in K562 cells grown in doxycycline to induce repression of the endogenous RUNX1, and hygromycin to select for transduced cells. The screen was conducted in three biological replicates for each mutation with greater than 1 million cells in each replicate. At day 7 post transduction, the cells were split into two groups: ~1 million cells to be sequenced to a depth of 30 million reads/sample for bulk

RNA-seq, and 100,000 cells to be sequenced to a depth of 75 million reads/sample for bulk ATAC-seq (Figure 3.1c).

From bulk RNA-seq, we obtained an 18,646 gene by sample expression matrix after QC filtering (Methods). After normalizing the raw counts with respect to library size and removing batch effects between replicates, we averaged gene expression across replicates for each sample. Using the top 2000 variable genes from the scRNAseq analysis, we perform PCA and hierarchical clustering of samples. PC1 once again correlated with the progression of phenotypic effects (WT-like, hypomorphic-I, -II, -III, and LOF-like) (Figure 3.5b), which was also reflected in the gene expression-based hierarchical clustering of the variants (Figure 3.5c) that essentially reproduced the earlier scRNA-seq analysis results (Figure 3.2a). PC2 separated hypomorphic variants from WT- and LOF-like ones (Figure 3.5b), with more specific distinctions obtained by PCs 3 and 4 (Figure 3.5d). We saw similar results when replicates were analyzed separately (Figure S3.5). The similarity between bulk and single cell analysis supports that single cell transcriptomic analysis can reliably identify hypomorphic variants.

From bulk ATAC-seq, we obtained a 101,433 peak by sample count matrix after QC filtering (Methods). 9176 peaks were mapped to promoter regions ((-1 kb, +100 bp) of transcription start sites), among which 2722 had RUNX1 binding motifs, referred to hereafter as RUNX1 promoter peaks. Of these, 207 were found to be differentially accessible between LOF-like samples and WT; while 191 RUNX1 promoter peaks are differentially accessible between hypomorphic samples and WT (Figure 3.5e). Among genes with differentially accessible RUNX1 promoter peaks for LOF-like or hypomorphic variants, increased or decreased accessibility led to an increase or decrease in gene expression for 101 and 70 genes, respectively, in the bulk RNA-seq dataset. There was also significant correlation in the magnitude of change (LOF-like vs. WT:

Pearson correlation r=0.65, p<1.05e-13; and hypomorphic vs. WT: Pearson correlation r=0.52, p<1.89e-06). These genes were generally found to be overrepresented in gene sets that are relevant to RUNX1 activity (e.g. heme Metabolism, reactive oxygen species pathway, oxidative phosphorylation [36,39]), with differentially accessible genes for hypomorphic variants showing enrichment in terms of effects on MYC to the LOF-like variants (e.g. Myc targets V1 and V2) (Figure 3.5f). Tracks of two example genes among these groups, ABT1 [25] for LOF-like and ENSA [25] for hypomorphic variants, are visualized (Figure 3.5g).

3.5 Discussion

In this work we used information about physical contacts between proteins to design a library of amino acid substitutions and profiled their transcriptional consequences at the single cell level using SEUSS. We selected RUNX1 as our target protein for its well-defined role in cancer and its activity as a master regulator of hematopoiesis, reasoning that point mutations in this gene could result in large detectable differences in gene expression. We focused our analysis on the Runt domain which is most enriched for recurrent cancer mutations. We designed mutations based on in silico prediction of occurrence at physical interfaces and potential to perturb protein function. Analysis of the resulting single cell transcriptomic profiles revealed that the majority of mutations we selected had effects similar to loss of function or wild type conditions. Nonetheless, we detected 15 variants that generated unexpected changes to gene expression. Further analysis found that these variants activate MYC signaling and trigger an unfolded protein response.

Intersecting protein interaction-perturbing variants with cancer mutations, we found that the majority of recurrent cancer variants tested resulted in a more general loss of function variant, most likely by preventing RUNX1 from binding DNA altogether. However, the R118G and G95R hypomorphic alleles were also observed in tumors. Loss of RUNX1 DNA binding was associated with higher fitness scores, suggesting that RUNX1 acts as a tumor suppressor in K562 cells.

The role of RUNX1 as an oncogene versus tumor suppressor is still not entirely clear, and may depend on the type of malignancy as well as the other mutations present. Loss of RUNX1 leads to increased susceptibility to AML; point mutations affecting RUNX1 are associated with shorter time to progression from MDS to AML and worse prognosis in AML and CML [43,44]. However, in disease with RUNX1 translocations, a survival dependency on WT RUNX1 has been reported [45]. On the K562 background, loss of function appears to be associated with higher

fitness, suggesting a more tumor suppressive role in this setting (CML with blast crisis), though it is important to note that K562 cells represent disease that developed with WT RUNX1.

Bulk sequencing corroborated the effects found based on single cell sequencing, supporting that our study design was sufficiently powered to detect differences between variants. In UMAP plots, individual mutations were difficult to distinguish without mapping to densities. Even then, it was apparent that individual cells could coincide with WT or LOF mutation regions. It was unclear whether this reflected stochastic differences in construct expression or knockdown of endogenous RUNX1, cell-to-cell differences in read coverage, or bonafide variable penetrance of the variant effect on the phenotype of individual cells.

Limitations of our study include that it was performed exclusively in K562 cells. It is unclear how well the effects will generalize to other leukemic cell lines or non-hematopoietic tumor types. In addition, we focused only on the Runt domain, whereas other domains are also important for RUNX1 interaction with cofactors. Thus, we may not have fully captured the space of possible phenotypes that can be generated by single amino acid substitutions in RUNX1. Furthermore, our epigenetic profiling was limited to DNA accessibility, whereas ChIP-seq would more directly reveal mutation-associated changes to RUNX1 localization. These questions will be the topics of future studies to better understand the role perturb-seq can play in providing exploitable mechanistic insights in cancer.

Understanding the consequences of single amino acid substitutions in driver genes remains an unmet need. New technologies are making it possible to place such variants into the context of cellular programs and fitness. Our study demonstrates the potential of targeting protein interactions to define the space of cellular phenotypes reachable by single amino acid substitutions.

3.6 Materials and Methods

Building of RUNX1 variant library. The gene overexpression vector was generated from a modified lentiviral vector (Addgene #120426). The vector was modified by removal of the mCherry transgene and the hygromycin resistance enzyme gene. The hygromycin resistance enzyme gene was then re-cloned to be immediately downstream of the EF1a promoter, followed by a P2A peptide motif and a NheI restriction site used to clone in the library elements. A 12 base pair barcode sequence was then introduced downstream of the cloning site to identify variants during single cell transcriptome sequencing. To insert the barcode, the backbone was digested with NheI (New England BioLabs), and a pool of 12 base pair long barcodes with flanking sequences compatible with the NheI site was cloned using Gibson assembly. To clone the library elements, the expression vector was digested with NheI for 3 hours at 37 °C. The linearized vector was then purified using a QIAquick PCR Purification Kit (Qiagen).

DNA fragments coding for the library elements were ordered from Twist Bioscience as a site saturation variant library in an arrayed format as linear dsDNA. A fraction of each oligonucleotide was then combined, and the pool was amplified via PCR using KAPA-Hifi (Kapa Biosystems) in 50 mL reactions containing 10 ng of pooled template and 2.5 mL of primers RX1_01 and RX1_02 (10 mM), which include ~30 bp of DNA homologous to the overexpression vector for Gibson assembly cloning. A thermal cycler was used to heat the sample to 95 °C for 3 minutes, then 16 cycles of 98 °C for 20 seconds, 68 °C for 15 seconds, and 72 °C for 45 seconds, followed by a final 5 minute extension at 72 °C. The PCR products were then purified using Agencourt AMPure XP Beads (New England BioLabs) beads at a 0.8:1 bead:PCR reaction ratio. See Table S3.2 for primer sequences.

Gibson assembly was then used to clone the pooled library elements into the vector. For the reaction, 50 ng of the digested vector and 30 ng of the insert were mixed with 5 mL of Gibson Reaction Master Mix (New England BioLabs) in a reaction volume of 10 mL. The Gibson reactions were incubated at 50 ºC for 1 hour and transformed via heat shock into 50 mL of One Shot Stbl3 chemically competent cells (Invitrogen). This was done by incubating the cells with the Gibson on ice for 30 minutes, followed by a 45 second heat shock at 42 °C then 2 minutes on ice, then the addition of 250 mL of SOC media (Thermo Fisher Scientific). The cells were allowed to recover shaking at 37 °C for 1 hour and were then plated on LB-carbenicillin plates. Individual bacterial colonies were picked off of the plate and grown in LB-carbenicillin culture media shaking for 16 hours at 37 °C. After growth, plasmid DNA was isolated via a Qiagen Plasmid Mini Kit. Each colony was Sanger sequenced using the primer RX1_03 to identify the variant, then by the primer RX1_04 to capture the associated barcode. One overexpression vector was created for each variant, each with a single unique barcode associated. After ~30% of the library was cloned, the oligonucleotides for remaining elements were re-pooled and cloned using the above protocol, until the full library was assembled. To generate the combination mutations, the first mutation was created as described above. Subsequent mutations were generated with overlap extension PCR with primers containing the desired mutations.

Lentivirus production. Replication deficient lentiviral particles were produced individually in an arrayed format for each element of the library in HEK293FT cells (Invitrogen) via transient transfection. The HEK293FT cells were grown in DMEM media (Gibco) supplemented with 10% FBS (Gibco) and 1% antibiotic-antimycotic (Thermo Fisher Scientific). One day prior to transfection, HEK293FT were plated in 12 well plates, with one well per element of the library, at ~35% confluency. The day of transfection, the culture medium was removed and

replaced with fresh DMEM plus 10% FBS. Meanwhile, the transfection mix was prepared by mixing 125 mL of Optimem reduced serum media (Life Technologies) with 1.5 mL of lipofectamine 2000 (Life Technologies), 125 ng of pMD2.G plasmid (Addgene #12259), 500 ng of pCMV delta R8.2 plasmid (Addgene #12263), and 375 ng of each plasmid overexpression vector for each library element. The transfection mix was incubated for 30 minutes, then added dropwise to the HEK293FT cells. The viral particles in the supernatant were harvested at 48 and 72 hours post transfection, and the virus for each library element were pooled and filtered with a 0.45 mm filter (Steriflip, Millipore), then concentrated to 1.5 mL using Amicon Ultra-15 centrifugal filters with a cutoff 100,000 NMWL (Millipore). The virus was then mixed, aliquoted and frozen at -80 ºC. For the validation screen, the transfection was performed in 15 cm dishes, one for each of the selected validation mutations, and frozen separately.

Generation of clonal inducible RUNX1 repression cell line. To repress the endogenous RUNX1, the repression vector was generated from a PiggyBac inducible dCas9 construct (Addgene #63800). The vector was modified by removing the inducible transgene, the sequence for the KRAB-dCas9 fusion (Addgene #60954) followed by a P2A sequence then GFP was inserted in its place. The vector was then modified through the insertion of a U6 promoter followed by SaII and AflII cloning sites for insertion of guide RNA sequences, then a guide RNA scaffold. Guides for CRISPRi targeting RUNX1 were chosen from the Dolcetto library set A [46] and ordered via oligonucleotide from IDT. The guides were then cloned into the repression vector after digestion with SaII (New England BioLabs) and AflII (New England BioLabs).

K562 cells (ATCC) were cultured in RPMI 1640 media (Gibco) supplemented with 10% FBS and 1% antibiotic-antimycotic. One day prior to electroporation, the K562 cells were maintained at a concentration of 1 million cells per mL. The day of the electroporation, the cells

were spun down and resuspended at a concentration of 10 million cells per mL. A total of 2 mg of DNA was added to 100 mL of cells containing a 1:2.5 molar ratio of the all-in-one RUNX1 targeting repression vector to the PiggyBac transposase vector (Transposagen). The DNA was then electroporated into the K562 cells using the Ingenio Electroporation Kit (Mirus Bio) and a 4D Nucleofector (Lonza) per the manufacturer's protocol. The cells were recovered for 3 days, then selected for those that received integration by the addition of 1 mg/mL puromycin (Gibco) into the culture media. After 4 days of selection, the cells were split across a 96-well plate into single colonies by serial dilution. Individual colonies were then assessed for their degree of inducible RUNX1 repression, and the clone with the highest repression was selected for use in the screen.

Quantification of RUNX1 expression. To measure RUNX1 repression in the single colonies, each colony was split into two groups and grown in RPMI media supplemented with 10% FBS and 1% antibiotic-antimycotic and 1 mg/mL puromycin. In one of the groups, 1 mg/mL doxycycline (Thermo Fisher Scientific) was added to the media to induce expression of the dCas9-KRAB transgene. Both sets of cells were maintained at 200,000 cells/mL over the course of 3 days after the addition of the doxycycline. On day 3 the cells were pelleted, and RNA was extracted using a Qiagen RNeasy Mini Kit. Complementary DNA (cDNA) was synthesized from the RNA using the Protoscript II First Strand cDNA Synthesis Kit (New England BioLabs) per the manufacturer's protocol, then diluted 1:4 with water. To quantify expression, qPCR was performed on the cDNA using a CFX Connect Real Time PCR Detection System (Bio-Rad). For each sample, two sets of primers were used; a set used to quantify RUNX1 expression (Table S3.2) which was compared to the housekeeping gene GAPDH . The qPCR was carried out in a total volume of 10 mL containing 5 mL of iTaq Universal Sybr Green Master Mix (Bio-Rad), 2 mL of each primer (10 mM), and 1 mL of diluted cDNA. Thermal cycling conditions were 95 °C for 2.5 minutes,

followed by 40 cycles of 95 °C for 10 seconds, then 60 °C for 30 seconds. All samples were run in triplicate, and the RUNX1 expression was determined using the 2-delta delta CT method, by comparing to the GAPDH expression.

Sequencing screening. The K562 clonal cell line previously generated for repression of the endogenous RUNX1 protein was cultured in RPMI media supplemented with 10% FBS and 1% antibiotic-antimycotic. For the single cell RNA sequencing screen, the cells were transduced with the pooled variant library at an MOI of ~0.3 to ensure that each cell received a single construct. The viral transduction was performed by mixing the virus with media containing 8 mg/mL polybrene (Millipore). The cells were suspended in this media at a concentration of 2 million cells per mL and spun at 1000 G for 2 hours at 33 °C in a 12-well plate. The cells were then pelleted and resuspended in fresh media at a concentration of 400,000 cells/mL. 24 hours after transduction, the media was again changed, and the cells were resuspended at 400,000 cells/mL. 48 hours after transduction, the cell culture media was changed to media containing 1 mg/mL puromycin and 200 mg/mL hygromycin (Invitrogen) to select for transduced cells. At that time the cells were also split into two groups, and to one of the groups doxycycline was added daily at a concentration of 1 mg/mL to induce repression of the endogenous RUNX1. Throughout the duration of the screen, the media was changed each day, and the cells were maintained at a concentration of 400,000 cells/mL. The screening was conducted with two biological replicates with greater than 1 million cells in each condition to ensure greater than 1000-fold coverage of the library. At day 7 post transduction, the cells were processed with 10X, with the remainder of cells being maintained until day 14 for fitness screening.

For the bulk RNA sequencing and bulk ATAC sequencing screen, the cells were transduced with the twelve validation mutations separately. 48 hours after transduction, 200

mg/mL hygromycin was used to select for transduced cells and 1 mg/mL doxycycline was used to induce repression of the endogenous RUNX1. The screen was conducted with three biological replicates with greater than 1 million cells in each condition. At day 7 post transduction, the cells were split into two groups, 1 million cells for bulk RNA-seq and 100,000 cells for bulk ATAC-seq.

Single cell RNA sequencing library preparation. scRNA-seq experiments were performed with two replicates per condition (cells with and without doxycycline). Cells were first washed with a solution of PBS (Gibco) with 0.04% BSA (Gibco) by centrifuging the cells for 5 minutes at 300 G then resuspending them in the solution. After the wash, the cells were again centrifuged and resuspended in the same solution. The cells were filtered using a 0.40 mM cell strainer (VWR), and the concentration was determined using a manual hemacytometer (Thermo Fisher Scientific). The cells were then subjected to scRNA seq (10X genomics, chromium single cell 3' v3, with two reactions per replicate) aiming for a target cell recovery of 10,000 cells per library. The single-cell libraries were generated according to manufacturer's protocols with the following conditions: 11 PCR cycles run during cDNA amplification and 10 PCR cycles run during library generation. The libraries were sequenced using Illumina NovaSeq platform. To genotype the cells with the variant, the barcode sequences were amplified off of the cDNA pool generated in the scRNA seq protocol. The barcodes were amplified via PCR using OneTaq 2X Master Mix (New England BioLabs) in 100 mL reactions, each split across 5 PCR tubes (20 mL per tube). For each sample the reactions contained 5 mL of primers RX1_07 and the NEBNext Universal PCR Primer for Illumina (New England BioLabs) (10 mM), 6 mL of cDNA, 50 mL of OneTaq, and the rest filled with water. A thermal cycler was used to heat the sample to 95 °C for 3 minutes, then 20 cycles of 98 °C for 20 seconds, 65 °C for 15 seconds, and 68 °C for 45 seconds,

followed by a final 5 minute extension at 68 °C. The PCR products were purified using AMPure XP Beads beads at a 0.8:1 bead:PCR reaction ratio. The second step of PCR was performed. Subsequently, a NEBNext Ultra RNA Library Prep Kit (New England BioLabs) was used to generate Illumina compatible sequencing libraries; this was done in a 50 mL reaction split across 5 PCR tubes (10 mL per tube) with 20 ng of the first step purified PCR product.

Library fitness screening. A fitness screen was also performed concurrently with the single cell RNA sequencing screen. At days 2, 7, and 14 post-transfection, ~1 million cells were collected, and their genomic DNA was isolated via a Qiagen DNeasy Blood and Tissue Kit. Barcodes corresponding to each library element at each timepoint, and replicate were then amplified from the genomic DNA using OneTaq 2X Master Mix. The sequencing libraries were amplified in 50 mL reactions, each split across 5 PCR tubes (10 mL per tube). For each sample, the reactions contained 2.5 mL of primers A and B (10 mM), 6 mg of gDNA, 25 mL of OneTaq, with the rest filled with water. The thermal cycler was used to heat the sample to 95 ºC for 3 minutes, then 27 cycles of 98 °C for 20 seconds, 65 °C for 15 seconds, and 72 °C for 45 seconds, followed by a final 5 minute extension at 72 °C. The PCR products were purified using AMPure beads at 0.8:1 bead:PCR reaction ratio. NEBNext Multiplexed Oligos for Illumina (New England BioLabs) were then used to index the samples, and the samples were sequenced on an Illumina NovaSeq platform to a depth of 2.5 million reads/sample.

Freezing for bulk RNA-seq. Cells for bulk RNA-seq were pelleted and the media aspirated. They were flash-frozen in liquid nitrogen and stored at -80 ºC.

Freezing for bulk ATAC-seq. Cells for bulk RNA-seq were pelleted in a centrifuge at 1000 G for 5 minutes at 4 ºC, resuspended in cold PBS, and pelleted again. ATAC lysis buffer was made by mixing 100 uL 1M Tris-HCl pH 7.4, 20 uL 5 M NaCl, 30 uL 1M MgCl2, 100 uL 10%

IGEPAL CA-630, and 9.75 mL water. The cells were lysed with the cold ATAC lysis buffer using 100 uL buffer per 100,000 cells and centrifuged at 1000 G for 10 minutes at 4 ºC. The supernatant was removed, and the cells were flash-frozen in liquid nitrogen and stored at -80 ºC.

Bulk RNA sequencing library preparation. Bulk RNA-seq experiments were performed with three replicates per condition. RNA was isolated from the cells using a Qiagen RNeasy Mini Kit according to the manufacturer's protocols. Samples were prepared for bulk RNA-seq using the NEBNext Ultra II RNA Library Prep with Sample Purification Beads Kit (New England Biolabs) according to manufacturer's protocols with the following conditions: 1 ug input RNA, library insert size = 200 nt. The bulk RNA-seq library was sequenced on an Illumina NovaSeq platform to a depth of 30 million reads/sample.

Bulk ATAC sequencing library preparation. Bulk ATAC-seq experiments were performed with three replicates per condition. Tagmentation buffer was prepared with 12.5 uL buffer, 9.75 uL H2O, 0.25 uL digitonin, and 2.5 uL Tn5 enzyme (Illumina) per sample. Each frozen cell pellet sample was resuspended in the tagmentation buffer and incubated at 37 ºC for 45 minutes. 1x volume 40mM EDTA was added to each sample. The tagmented samples were purified using AMPure XP Beads at a 2:1 bead:tagmentation reaction ratio. The samples were incubated with the beads at room temperature for 15 minutes, then placed on a magnetic rack to separate the beads from the supernatant, which was discarded. The beads were washed twice with cold 80% ethanol, and the purified DNA was eluted from the beads using Buffer EB (Qiagen).

The tagmented DNA was dual indexed using i5 and i7 barcodes, giving each sample a unique barcode combination. The DNA and barcodes were added to NEB Hi Fidelity 2x PCR Mix (New England BioLabs) and amplified using the following PCR cycle: 72 ºC for 7 minutes; 98 ºC for 30 seconds; then 10 cycles of 98 ºC for 10 seconds, 63 ºC for 30 seconds, and 72 ºC for 1

minute; and cooling back down to 4 ºC. Double size selection was performed using AMPure XP Beads to select for the size of the final library. First, 0.55x volume AMPure Beads was added to each PCR reaction and incubated at room temperature for 15 minutes. The samples were placed on a magnetic rack and the supernatant transferred to new tubes, to which another 0.65x volume AMPure Beads were added (for a total of 1.2x volume PEG). The samples were incubated at room temperature for 15 minutes, the supernatant was discarded, and the beads were washed twice with cold 80% ethanol. DNA was eluted from the beads using Buffer EB and pooled together to make the final library for sequencing. The bulk ATAC-seq library was sequenced on an Illumina NovaSeq platform to a depth of 75 million reads/sample.

Protein 3D structure analysis. We obtained 61 experimentally verified undirected protein interactions of RUNX1 with a confidence score higher than 0.4 from STRING v9.1 [47]. Experimental 3D co-crystal protein structures for RUNX1-CBFB interaction (PDB: 1ljm, 1e50, 1h9d) were obtained from the Protein Data Bank (PDB) [48], and used to predict amino acid residues of RUNX1 in direct physical contact with CBFB as described in our previous work [49]. The remaining interactions did not have co-crystal structures. Instead, we used in silico template based protein docking on single protein structures with PRISM [28] to identify contact residues. PRISM returned predictions for 32 RUNX1 interaction partners (Figure 3.1b).

Amino acid residues of RUNX1 involved in DNA binding (PDB: 1h9d) were determined using the distance between two non-hydrogen atoms of amino acids and nucleotides, one from the protein and one from the DNA. If the distance was less than 3.5A, we designated those residues as interface residues [17]. Amino acid residues were annotated as core, surface, or intermediate based on their relative solvent accessible surface areas as described in our previous work [49]. VMD [50] was used to visualize protein 3D structures (Figures 3.1a and 3.4a,d).

Selection of variants for library construction. The ORF mutation library consists of 117 elements: 83 single amino acid substitutions at protein interaction interfaces in the RUNX1 Runt domain, 1 WT construct, 1 LOF construct, 17 negative, and 10 positive control mutations, and 5 combinations of two or more interface mutations (Figure 3.3c). Variant effect prediction scores for all possible missense mutations targeting each residue were obtained from VEST [29] and FoldX [30]. Variant frequency in human tumors was determined from the COSMIC database (obtained on 11/7/2022) [31]. For each residue, the most damaging amino acid substitution possible from a single base substitution (the highest VEST or FoldX scored mutation) was chosen to be included in the ORF mutation library, prioritizing cancer mutations where possible, to maximize the possibility of perturbing physical protein interactions. 30 of 83 mutations tested are cancer mutations.

The WT construct consists of WT RUNX1; while the LOF construct contains a green fluorescent protein (GFP) in place of RUNX1. 17 negative control mutations consist of 10 silent and 7 neutral (predicted based on VEST scores) mutations, and are expected to be functionally indistinguishable from the WT construct. 10 positive controls consist of 5 truncating and 5 core mutations, and are expected to have similar impact to the LOF construct. 5 perturbation mutation combinations consists of two or three combinations of perturbation mutations already in the library.

scRNA-seq analysis. The single cell RNA sequencing screen was performed for two conditions: one treated with doxycycline to induce repression of the endogenous RUNX1 (named ‘dox’ group), and the other not treated (named ‘nodox’ group). The screening was conducted with two biological replicates for each condition, and single cell RNA sequencing was performed with two reactions per replicate, making a total of eight libraries: four containing cells treated with

doxycycline (dox) and four not (nodox). Sequencing was run with a target cell recovery of 10,000 cells per library.

Sequencing reads in FASTQ format were aligned using the 10X Genomics Cell Ranger pipeline (version 3.1.0) [51], to the human transcriptome GRCh38 (version GRCh38-3.0.0), resulting in a gene by cell matrix of UMI counts for each library. To assign one or more genotypes to each cell, the plasmid barcode reads were aligned to GRCh38 using BWA, and labeled with its corresponding cell and UMI tags as described in the SEUSS pipeline [14] (https://github.com/yanwu2014/genotyping-matrices).

The UMI count matrices were processed using Seurat (version 4.1.0) [52]. Four dox libraries were merged resulting in 38,104 cells and 20,389 genes, after removal of genes expressed in fewer than 3 cells. 16,653 cells not containing a genotype barcode, or containing more than one, were removed. To filter out low quality cells, we removed cells expressing fewer than 200 genes or more than 5000 genes. We also discarded cells that have over 20% of reads aligned to mitochondrial genes, resulting in 20,878 cells. Four perturbation variants (G138V, S145I, P157R, T161I) were excluded due to low cell counts (less than 10 cells), and one negative control was removed (G143G) due to a frame shift artifact during the mutation library preparation, resulting in 112 remaining variants.

The count matrix was log-normalized with the default scale factor of 10,000 and the top 2000 variable genes were identified to be used for downstream analyses. Mitochondrial or ribosomal genes were not included in the top 2000 variable gene list. We then applied a linear transformation on the count matrix to scale and center expression of each gene. We assigned cell cycle scores to each cell based on its expression of G2/M and S phase markers and applied a linear model to regress out effects of cell cycle heterogeneity.

Unsupervised clustering of single cells. We performed linear dimensionality reduction (PCA) on the scaled data using the top 2000 variable genes. Keeping the first 20 principal components, we clustered cells by first determining the nearest neighbors of each cell in the PCA space, and then by applying a modularity optimization algorithm that iteratively groups cells together with a resolution parameter of 0.3. We used UMAP, a non-linear dimensionality reduction technique (Becht et al., 2018), to visualize the three predicted unsupervised clusters where similar cells are placed together in low-dimensional space (Figure 3.2a,b,f).

We applied Fisher's exact test to evaluate the enrichment or depletion of assigned phenotypes (Figure 3.2g), variant classes, or cell cycle phases (Figure S3.2) in each cluster. A log odds ratio (log(OR))>0 indicates enrichment, while a log(OR)<0 indicates depletion.

Unsupervised clustering of variants. To cluster variants into discrete classes, we used mean expression profiles of each variant. For each variant, we computed the mean expression (log-normalized) of each of the top 2000 variable genes across all cells corresponding to the variant, resulting in an expression vector of size 2000 for each variant, representing its mean expression profile. This generated a count matrix of 112 variants by 2000 genes.

We performed PCA on the resulting count matrix, and keeping the first 20 principal components, we clustered variants by first determining the nearest neighbors of each variant in the PCA space, and then by applying a modularity optimization algorithm that iteratively groups variants together with a resolution parameter of 0.8. We used UMAP to visualize the three predicted unsupervised clusters where similar variants are placed together in low-dimensional space (Figure 3.2c-e). We performed differential expression analysis between clusters using DESeq2 [53].

T2 scores. In order to quantify the extent to which expression profile of a variant deviates from the WT or the LOF variant, we used the Hotelling's two-sample T-squared statistic (T2), a generalization of Student's t-statistic that is used in a two-sample multivariate hypothesis testing [33]. For this comparison, we employed the principal component space, and by considering the top 20 principal components (PC), we compared matrices of cells by 20 PCs for each variant. We used the hotellings2 function from the spm1d python package to compute the test statistics, named here as T2 scores. For each variant, first we compared against cells overexpressing the WT variant (T2 scores (vs. WT)), then we compared against cells overexpressing the LOF variant (T2 scores (vs. LOF)). Higher scores indicate a higher deviation from the variant being compared.

Fitness analysis. To calculate fitness effects from genomic DNA reads, we first aligned reads to mutation barcodes (MagECK [54]) and counted the number of reads corresponding to each mutation for each replicate at each timepoint (days 2, 7 and 14 post transduction), resulting in a mutation by samples read counts matrix. We normalized read counts for each sample by dividing each column by its sum. We then divided read counts of each sample by the counts at day 2 post transduction, and log2 transformed it to obtain a measurement to represent fitness effects for each mutation and sample. We averaged fitness measurements from the two biological replicates taken at day 14 to compute mean fitness scores (Figure 3.3i).

Hierarchical variant clustering. We also hierarchically clustered variants based on Pearson correlation of their mean gene expression profiles. We ordered the leaves of the resulting dendrogram by increasing T2 scores obtained from comparison to the WT variant. To obtain discrete cluster assignments, we cut the hierarchy based on visual inspection, obtaining three main variant clusters that largely agree with WT-like, hypomorphic and LOF-like annotations (only 1 variant difference for the hypomorphic/LOF-like separation). We further cut the hierarchy of the

middle cluster, representing hypomorphic variants, into three sub-clusters: named as hypomorphic-1, hypomorphic-2 and hypomorphic-3.

Hierarchical gene clustering. To determine genes whose expression is impacted by variants, we hierarchically clustered genes based on Manhattan distance between them using mean gene expression profiles of variants, resulting in gene groups with various expression profiles across variant clusters.

Bulk RNA-seq analysis. Sequencing reads in FASTQ format were aligned to the human transcriptome GRCh38 (Gencode v30 - GRCh38.p12) using STAR (version 2.7.1a) [55]. RSEM (version 1.3.1) is used to calculate read counts for each sample and replicate ('rsem-calculate-expression' command), and to generate a gene by sample matrix ('rsem_generate_data_matrix' command) of the raw counts ('expected_count' column). Starting with 57,535 gene features, we removed genes with less than 10 reads in total across all the samples, along with mitochondrial and ribosomal genes, resulting in 18,646 remaining genes.

We first normalized raw counts using the variance stabilizing transformation, which transforms counts on the log2 scale and normalizes with respect to library size [53]. We removed two outlier samples (18B and 60B: replicates B of samples with N82I and V137D mutations, respectively) identified based on expression profiles of top 2000 variable genes (Figure S3.6), and removed batch effects between replicates using the limma package [56]. For visualization purposes, we averaged gene expression across replicates for each sample. In order to validate variant clustering results obtained from scRNA-seq here in the bulk setting, we used top 2000 variable genes obtained from the scRNA-seq analysis, to perform PCA and hierarchical clustering of samples and genes based on Manhattan distance. We ordered the leaves of the resulting sample dendrogram by increasing T2 scores obtained from the scRNA-seq analysis by comparison to the

WT variant. Same analysis is also performed using all replicates instead of their means (Figure S3.5). Differential expression analysis between samples is performed with DESeq2 [53].

Bulk ATAC-seq analysis. Sequencing reads in FASTQ format were aligned to the human transcriptome GRCh38 and processed using the nf-core/atacseq pipeline (version 1.2.2) [57], built using Nextflow (version 22.04.0), in conjunction with Singularity. The command used is 'nextflow run nf-core/atacseq -r master -name "run/name" -profile "singularity" -work-dir "work/directory/path" -params-file "params/file/path" --genome GRCh38', with default parameters. First, fastq files from two ATAC-seq runs were merged with "cat" command for each read (reads 1 and 2 for paired-end data) of each replicate (three biological replicates) of each sample (12 samples); and the pipeline was run on the merged FASTQ files.

Briefly, the pipeline performs adapter trimming using Trim Galore! (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/), read alignment with BWA [58], filtering with SAMtools [59] (e.g. removal of mitochondrial reads), BEDTools [60], BamTools [61], Pysam (https://github.com/pysam-developers/pysam), and picard (https://broadinstitute.github.io/picard/), normalized coverage track generation with BEDTools and bedGraphToBigWig [62], genome-wide enrichment with deepTools [63], peak calling with MACS2 [64] (broad peaks by default), genomic features peak annotation with HOMER [65], consensus peak set creation with BEDTools, counting reads in consensus peaks with featureCounts [66], and quality control and statistics reporting with MultiQC [67]. During the consensus peak set creation, a peak was included only if present in at least two of the three biological replicates for each sample.

Using HOMER [65], we found enriched motifs in the consensus peak set (findMotifsGenome function), and selected the enriched Runt domain motifs (5 motifs total) to

represent RUNX1 DNA binding sites. Then we identified the genomic locations of these motifs on the human transcriptome GRCh38 (scanMotifGenomeWide function). Using the genomic feature annotations (annotatePeaks function) of the peaks in the consensus set, we identified peaks as promoters if located within 1 kbp downstream and 100 bp upstream of the transcription start sites (TSS) of associated genes. We selected promoter peaks also containing RUNX1 motifs, to study genes regulated by RUNX1.

Based on QC results (Figure S3.7), we selected two highest-quality ATAC-seq replicates of each sample for downstream analysis. Using the consensus peak set, we performed differential peak analysis between LOF-like, or hypomorphic samples, versus WT condition with DESeq2 [53]. We selected RUNX1 promoter peaks that are differentially accessible ($p<0.05$) between conditions, and for the corresponding genes, we compared accessibility change to gene expression change by bulk RNA-seq. The coverage tracks are visualized using the IGV genome browser [68].

Gene set overrepresentation analysis. Gene set overrepresentation analysis is performed using the Gene Ontology (GO) biological process terms (2021) or the MSigDB hallmark gene sets (2020), with the EnrichR package [69].

Statistical analysis. Correlations are evaluated using the Pearson correlation coefficient. Odds ratios are calculated using Fisher's exact test. Distributions are compared using a Mann–Whitney U test. Multiple testing correction is applied where applicable.

Data and code availability. Data will be made available via the SRA and code will be shared via Github.

3.7 Figures

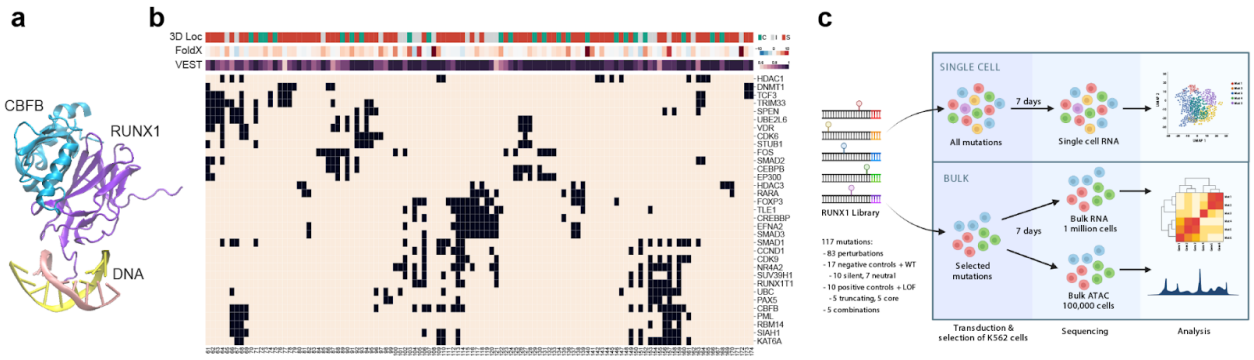


Figure 3.1. An interface-guided Perturb-seq assay for coding variant phenotyping of RUNX1. (a) 3D crystal structure of transcription factor CBF, consisting of RUNX1 Runt domain (purple) and CBFB (blue), interacting with DNA (yellow and pink strands) (PDB: 1h9d). **(b)** Amino acid residue map of the RUNX1 Runt domain. Columns represent amino acid residues, while rows represent interaction partners of RUNX1. At each row, interface residues involved in interaction to the partner are highlighted black. Rows are hierarchically clustered. On top: 3D location annotations of each residue (C: core, I: intermediate, S: surface), followed by VEST and FoldX scores of most damaging mutations targeting the residue. The darker the color, the more damaging (VEST) or destabilizing (FoldX) the mutation is. **(c)** Experimental and computational overview: ORF mutation library design, transduction, single cell RNA-sequencing of all 117 library elements, bulk RNA and ATAC-sequencing of 12 selected library elements, and computational analysis.

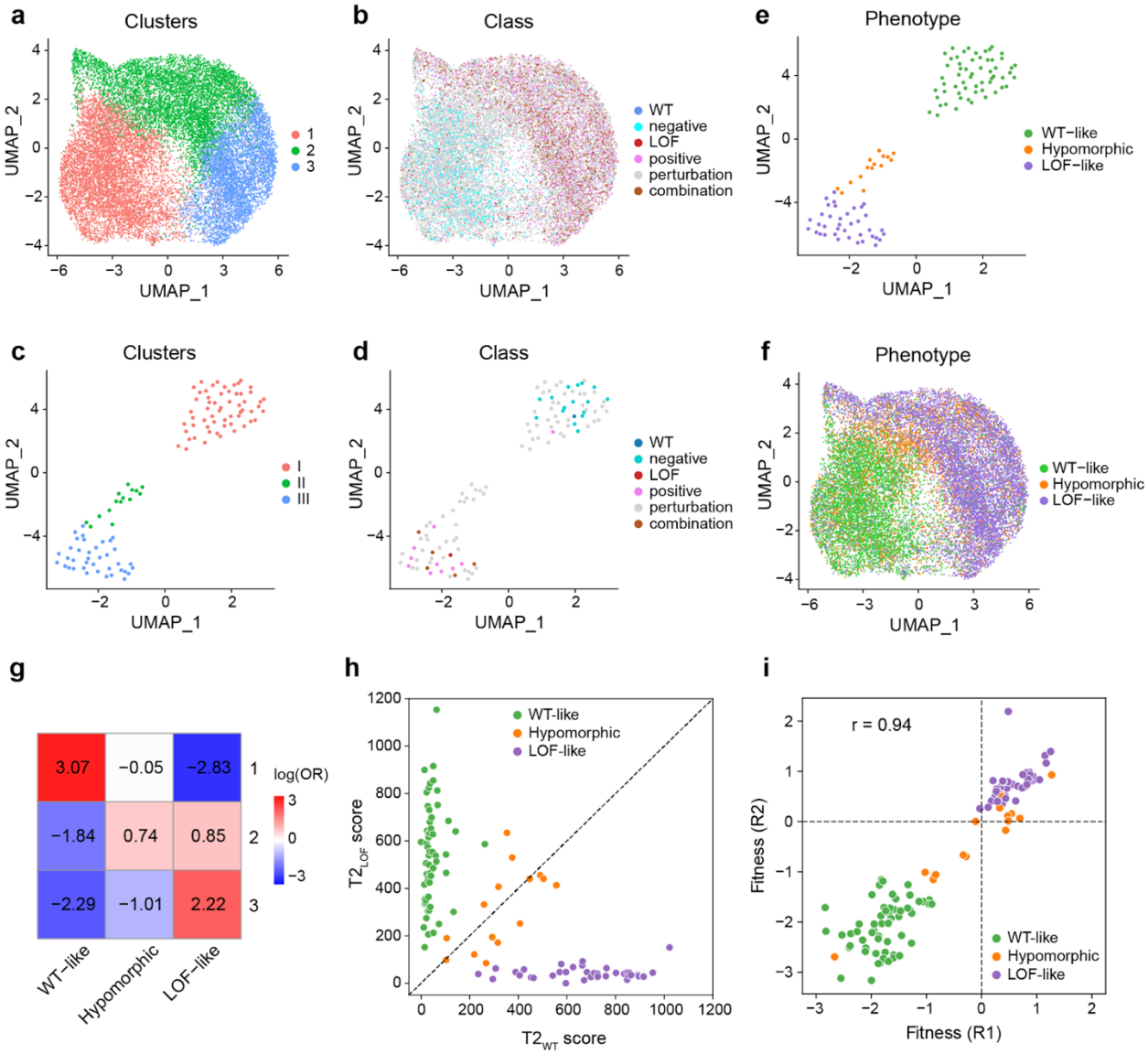
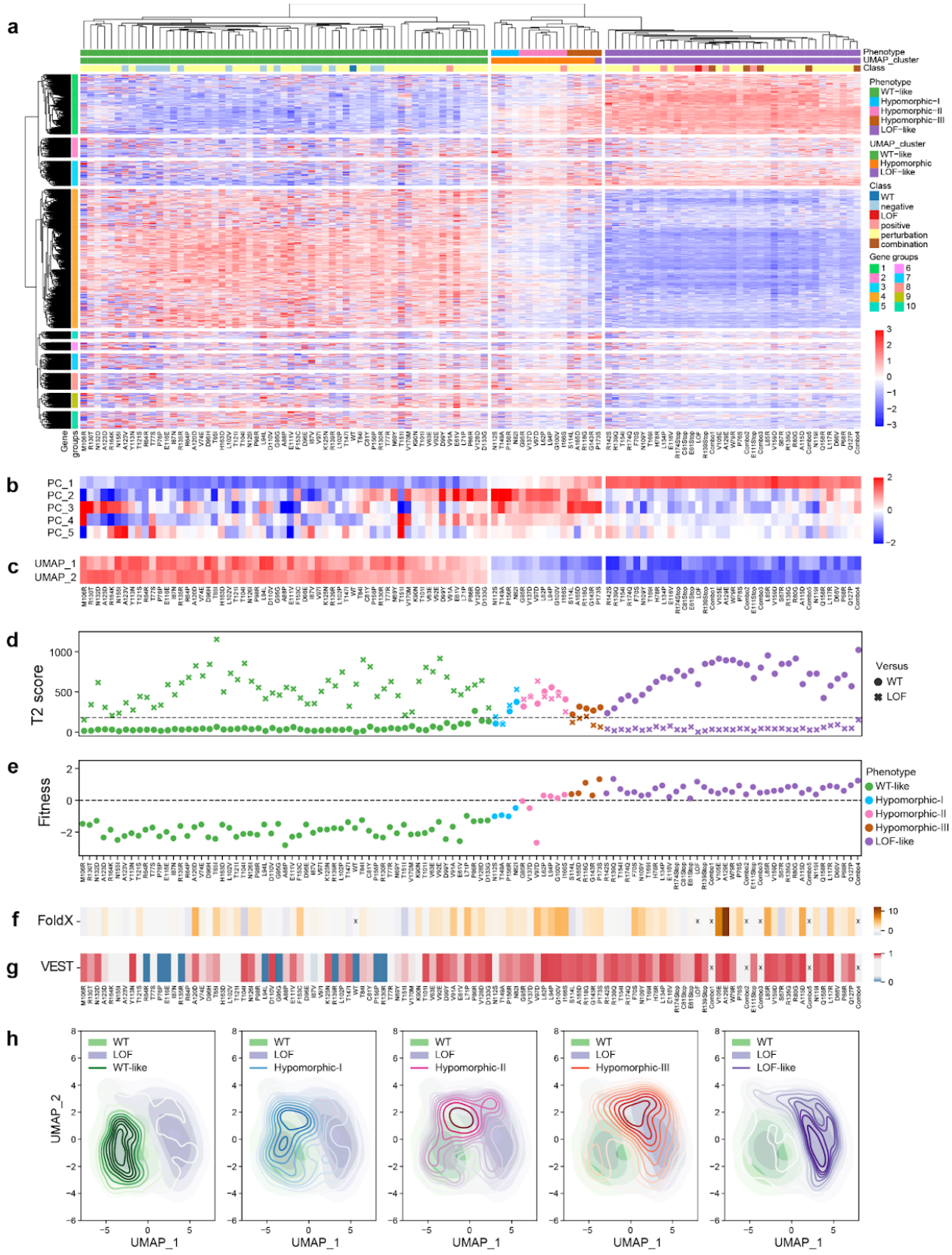


Figure 3.2. Unsupervised analysis of RUNX1 variant transcriptional effects informs WT-like, LOF-like and hypomorphic variants. (a-b) UMAP embedding of single cells colored by **(a)** unsupervised clusters, or **(b)** variant classes, obtained using the top 2000 variable genes. Cell cycle effects are regressed out. **(c-d-e)** UMAP embedding of variants constructed from mean gene expression vectors of the top 2000 variable genes for each variant, colored by **(c)** unsupervised clusters, **(d)** variant classes, or, **(e)** variant functional designations of WT-like, LOF-like, and hypomorphic for unsupervised clusters in **c**. **(f)** UMAP embedding of single cells colored by variant functional designations of WT-like, LOF-like, and hypomorphic obtained from unsupervised cluster designations in **e.** Cell cycle effects are regressed out. **(g)** Cluster enrichment of single cells (unsupervised clusters from **a**) for assigned phenotypes (from **f**) based on log of odds ratios obtained using Fisher's exact test. Positive values indicate enrichment, while negative values indicate depletion. **(h)** T2 scores of each variant when compared to the WT (x-axis) or LOF (y-axis) control, colored by assigned phenotypes. **(i)** Fitness scores of variants computed from two biological replicates, colored by assigned phenotypes (R1: replicate 1, R2: replicate 2).

Figure 3.3. Mapping the phenotypic consequences of RUNX1 interface variants with transcriptomic analysis. (a) Heatmap showing mean expression profiles of all variable genes (rows) in each variant (columns). Genes and variants are hierarchically clustered into ten (row colors) and five clusters (column colors: green: WT-like, blue: hypomorphic-I, pink: hypomorphic-II, brown: hypomorphic-III, and purple: LOF-like), respectively. The leaves of the variant dendrogram are ordered by increasing $T2_{WT}$ scores. Gene expression values are z-scored. **(b)** Top 5 PC, and **(c)** UMAP embeddings of each variant based on their mean gene expression profiles. Rows are scaled to have a mean of zero and unit variance. **(d)** T2 scores of each variant when compared to the WT (circle) or LOF (cross) control, colored by phenotypes. Dotted line highlights value 178.79, the median of $T2_{WT}$ scores of all WT-like and LOF-like variants. **(e)** Mean fitness scores of variants, colored by phenotypes. **(f)** FoldX, and **(g)** VEST scores of each variant. Variants that could not be scored (combination of perturbation mutations, and/or WT and LOF variants) are grayed out and marked with an X. **(h)** Kernel density estimate plots comparing UMAP embedding distributions of single cells belonging to each assigned phenotype (WT-like (green line), hypomorphic-I (blue line), hypomorphic-II (pink line), hypomorphic-III (brown line), and LOF-like (purple line)), to the cells overexpressing the WT (green shade) or LOF variant (purple shade).


a
Phenotype
UMAP_cluster
Class
Gene groups
WT-like
Hypomorphic-I
Hypomorphic-II
Hypomorphic-III
LOF-like
Hypomorphic
WT
negative
LOF
positive
perturbation
combination
b
PC_1
PC_2
PC_3
PC_4
PC_5
c
UMAP_1
UMAP_2
d
T2 score
Versus
WT
LOF
e
Fitness
f
FoldX
g
VEST
h
WT
LOF
WT-like
Hypomorphic-I
Hypomorphic-II
Hypomorphic-III
LOF-like
UMAP_1
UMAP_2

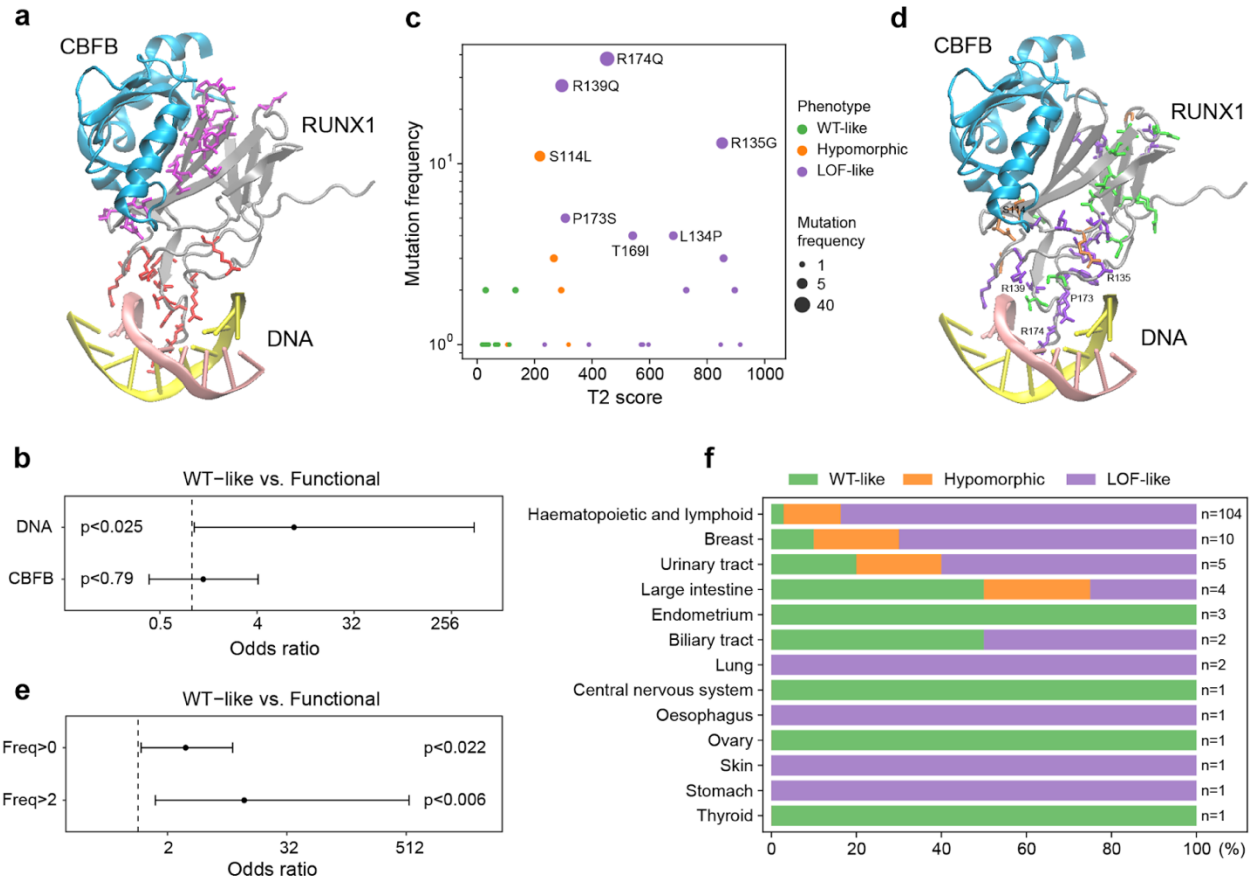


Figure 3.4. Mapping oncogenic variants into the RUNX1 regulatory landscape. (a) 3D crystal structure of transcription factor CBF, consisting of RUNX1 Runt domain (gray) and CBFB (blue), interacting with DNA (yellow and pink strands). Amino acid residues involved in interaction with DNA and CBFB are colored red and purple, respectively (PDB: 1h9d). **(b)** Odds ratios (OR) and 95% confidence intervals using Fisher's exact test. Enrichment or depletion of WT-like vs. functional (LOF-like or hypomorphic) impact variants for DNA or CBFB binding residues of RUNX1. OR>1 means enrichment for functional variants, while OR<1 means depletion. **(c)** Scatter plot of mutations present in cancer in terms of $T2_{WT}$ scores vs. mutation frequency. Variants are colored based on their phenotypic annotations. **(d)** 3D crystal structure of transcription factor CBF, consisting of RUNX1 Runt domain (gray) and CBFB (blue), interacting with DNA (yellow and pink strands). Amino acid residues mutated in cancer are colored based on their phenotypic effects (green: WT-like, orange: hypomorphic, purple:LOF-like). 5 most frequent cancer mutations are annotated (PDB: 1h9d). **(e)** Odds ratios and 95% confidence intervals using Fisher's exact test. Enrichment or depletion of WT-like vs. functional impact variants for cancer mutations based on frequency. OR>1 means enrichment for functional variants, while OR<1 means depletion. **(f)** A stacked bar plot showing percent distribution of phenotypic annotations of variants (WT-like, hypomorphic, and LOF-like) across cancers observed in different primary tissues. Sample size (n) for each tissue is displayed on the right.

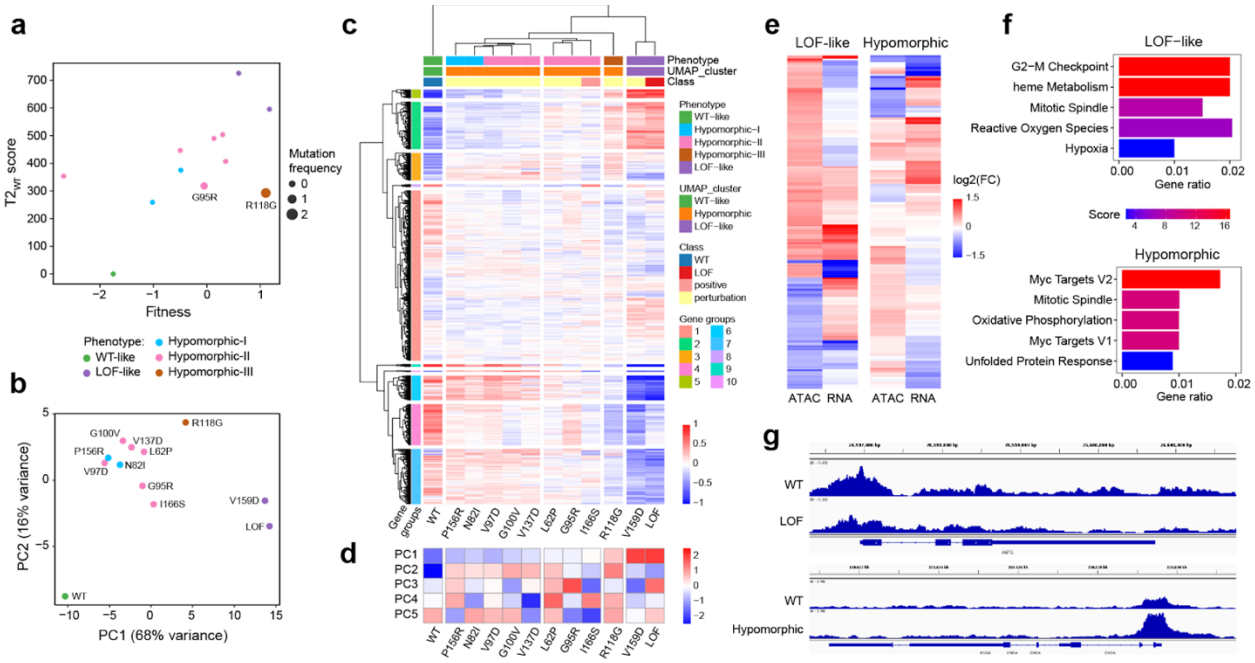


Figure 3.5. Bulk RNA- and ATAC-seq analysis of 12 validation mutations. (a) Overview of 12 validation mutations in terms of $T2_{WT}$ and fitness scores (obtained from the scRNAseq analysis), and mutation frequency in cancer. Cancer mutations are annotated. **(b)** PCA plot of 12 validation mutations, in the bulk RNA-seq setting, using top 2000 variable genes from scRNA-seq. Gene expression is averaged across replicates. **(c)** Hierarchical clustering of samples (rows) and genes (columns) in the bulk RNA-seq setting, using top 2000 genes obtained from scRNA-seq. Gene expression is averaged across replicates. The leaves of the variant dendrogram are ordered by increasing $T2_{WT}$ scores. Gene expression values are z-scored. **(d)** Top 5 PC embeddings of each sample based on mean gene expression across replicates. Rows are scaled to have a mean of zero and unit variance. **(e)** Comparison of accessibility change to gene expression change for RUNX1 promoter peaks that are differentially accessible between LOF-like vs. WT, or hypomorphic vs. WT samples. Rows are hierarchically clustered. **(f)** Gene set overrepresentation (mSigDb hallmark gene set) of overlapping differentially accessible RUNX1 regulated promoters and differentially expressed genes. Top 5 gene sets are displayed. **(g)** Tracks of two example genes: ABT1, differentially accessible in the LOF variant versus WT, and ENSA, differentially accessible in the hypomorphic variants versus WT are visualized using the IGV genome browser. Tracks for each comparison are set at the same range.

3.8 Tables

Table 3.1. 12 validation mutations selected for bulk RNA- and ATAC-sequencing.

Variants	Class	Phenotype	Cancer
L62P	perturbation	Hypomorphic-2	0
N82I	perturbation	Hypomorphic-1	0
G95R	perturbation	Hypomorphic-2	1
V97D	perturbation	Hypomorphic-2	0
G100V	perturbation	Hypomorphic-2	0
R118G	perturbation	Hypomorphic-3	2
V137D	perturbation	Hypomorphic-2	0
P156R	perturbation	Hypomorphic-1	0
V159D	perturbation	LOF-like	0
I166S	positive	Hypomorphic-2	0
LOF	LOF	LOF-like	0
WT	WT	WT-like	0

3.9 Supplemental Data, Tables, and Figures

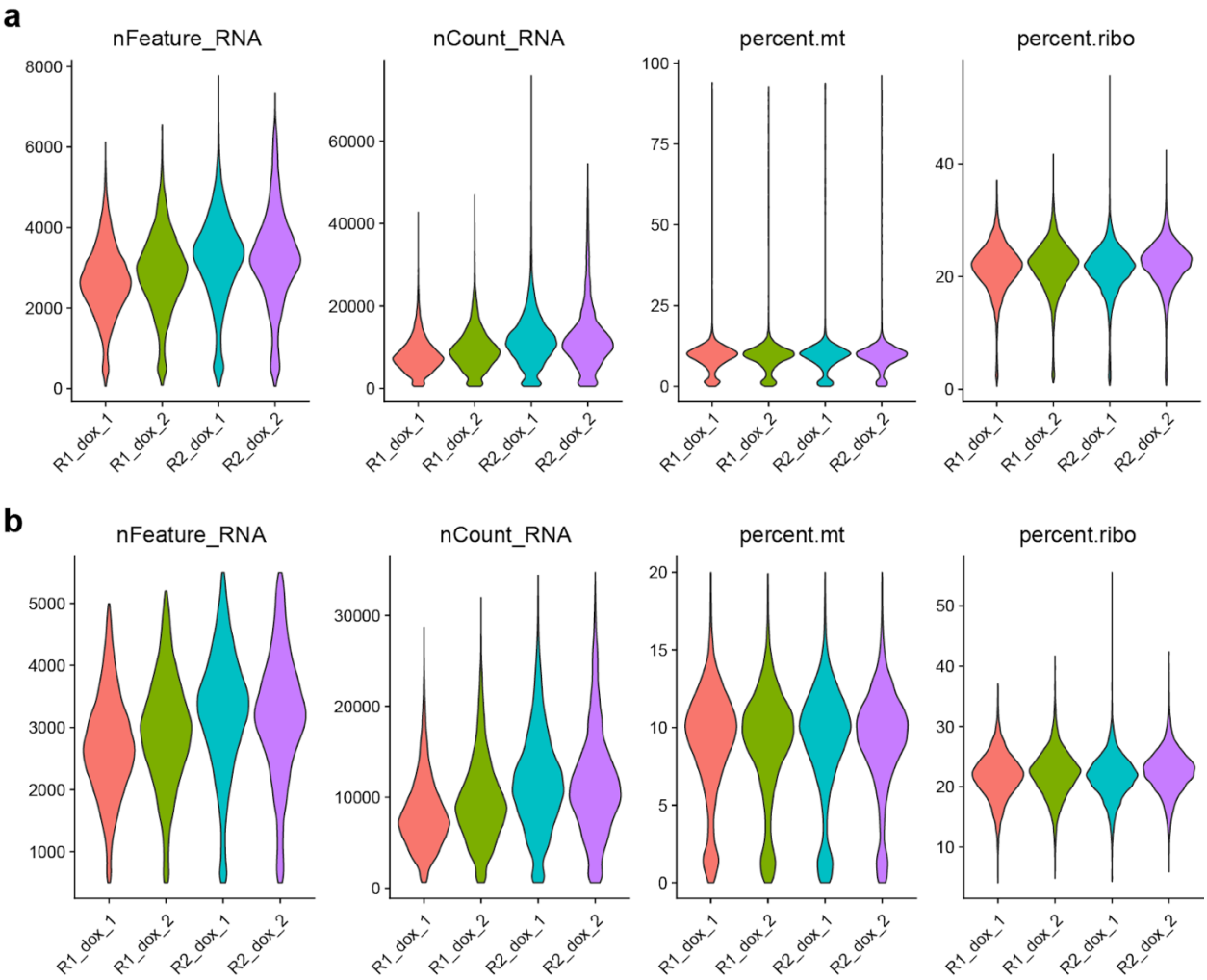


Figure S3.1. Violin plots displaying unique and total gene counts, and percentage of mitochondrial or ribosomal genes. In single cells, **(a)** before, and **(b)** after QC filtering.

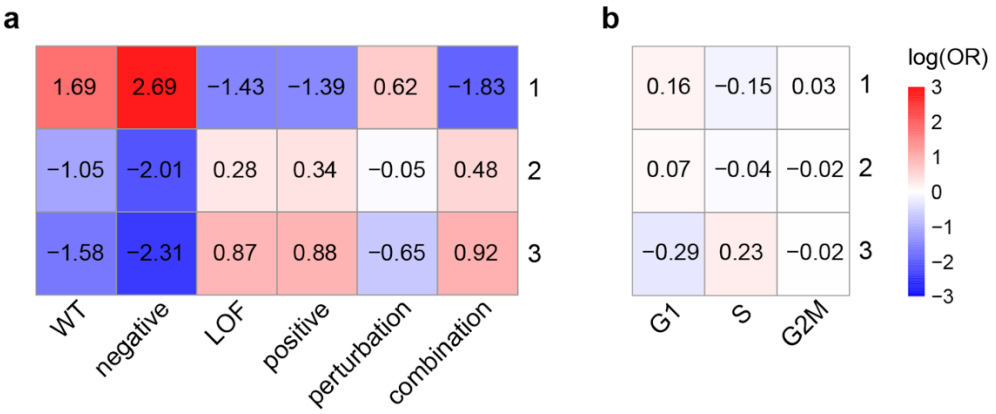


Figure S3.2. Cluster enrichment of single cells for unsupervised clusters from Figure 3.2a. For **(a)** variant classes (Figure 3.2b), and **(b)** cell cycle phases, based on log of odds ratios obtained using Fisher's exact test. Positive values indicate enrichment, while negative values indicate depletion.

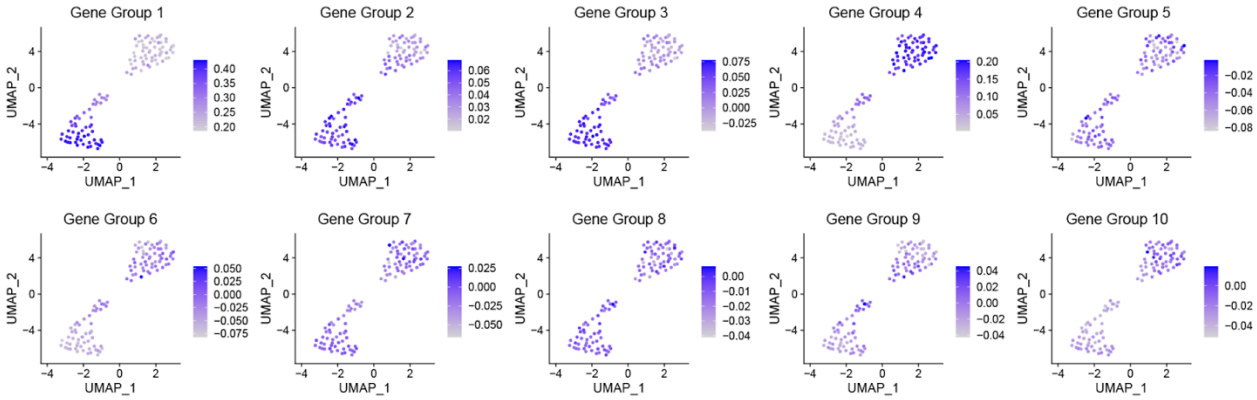


Figure S3.3. Aggregated mean expression of genes for each gene group (Figure 33.a) across cells for each variant.

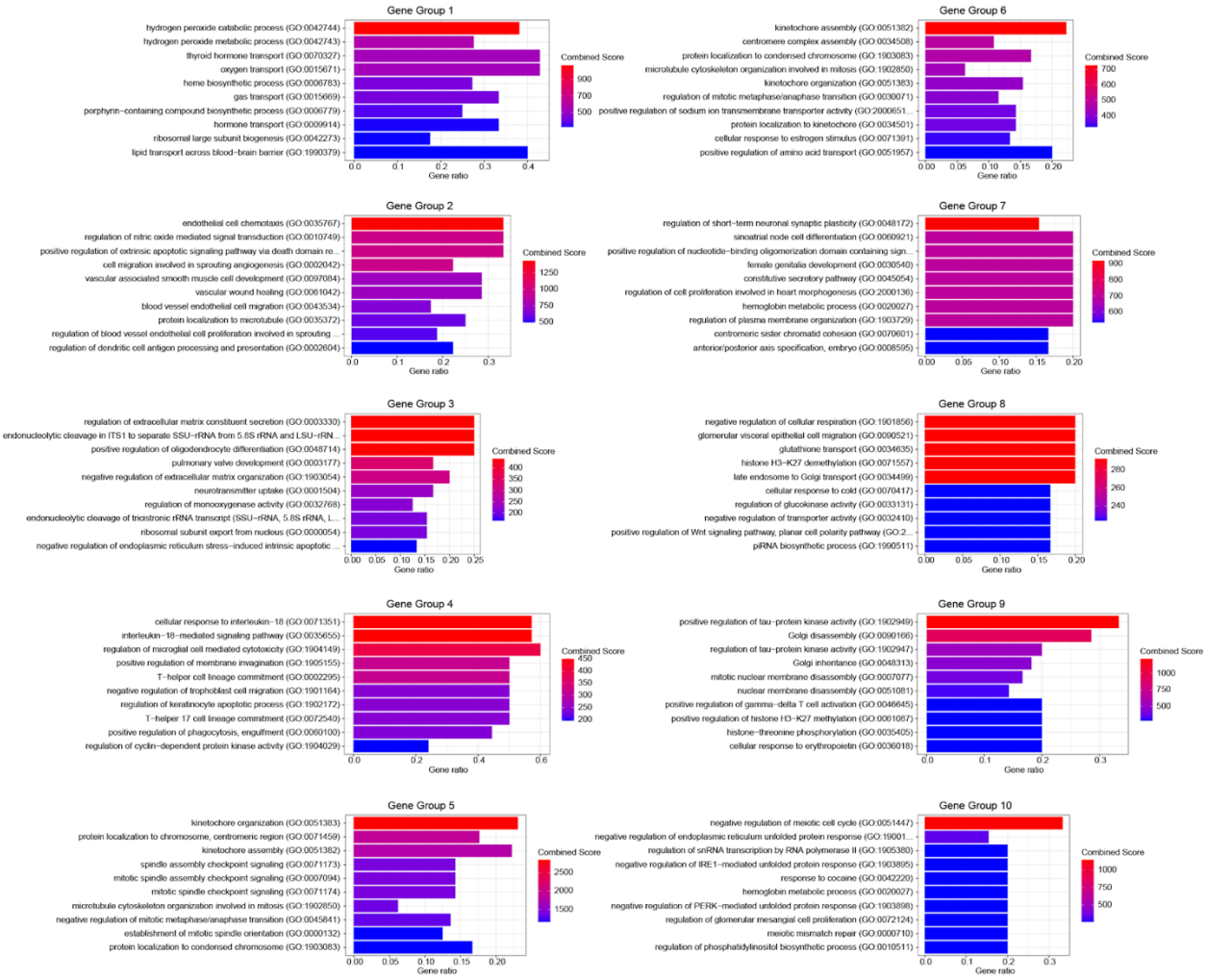


Figure S3.4. Gene set overrepresentation analysis results for GO Biological Process terms for each gene group (Figure 3.3a) displaying top 10 terms.

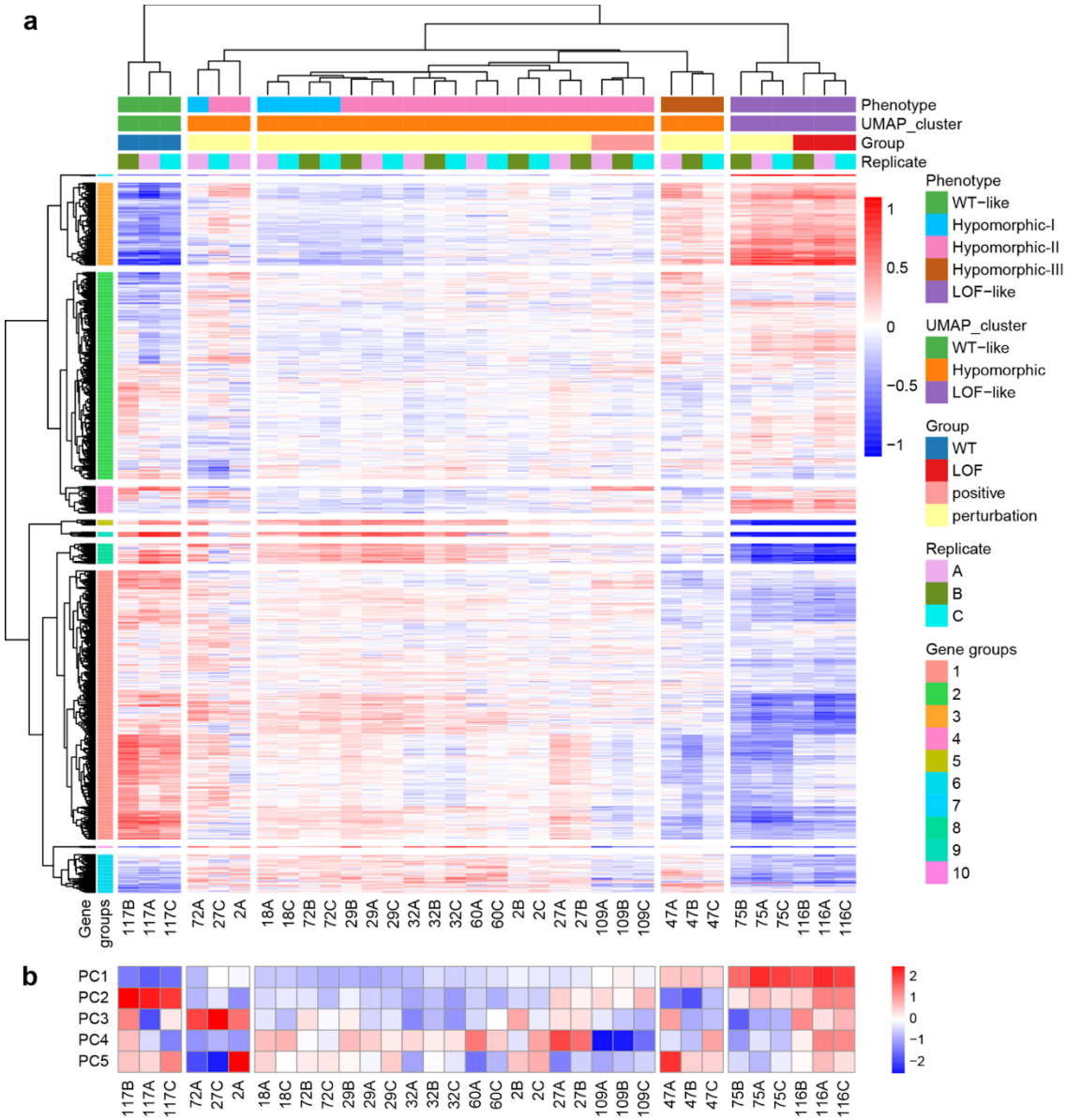


Figure S3.5. Heatmap of gene expression for bulk RNA-seq. (a) Hierarchical clustering of samples (rows) and genes (columns) in the bulk RNA-seq setting, using top 2000 genes obtained from scRNA-seq. All sample replicates are present. The leaves of the variant dendrogram are ordered by increasing $T2_{WT}$ scores. Gene expression values are z-scored. **(b)** Top 5 PC embeddings of each sample based on mean gene expression across replicates. Rows are scaled to have a mean of zero and unit variance.

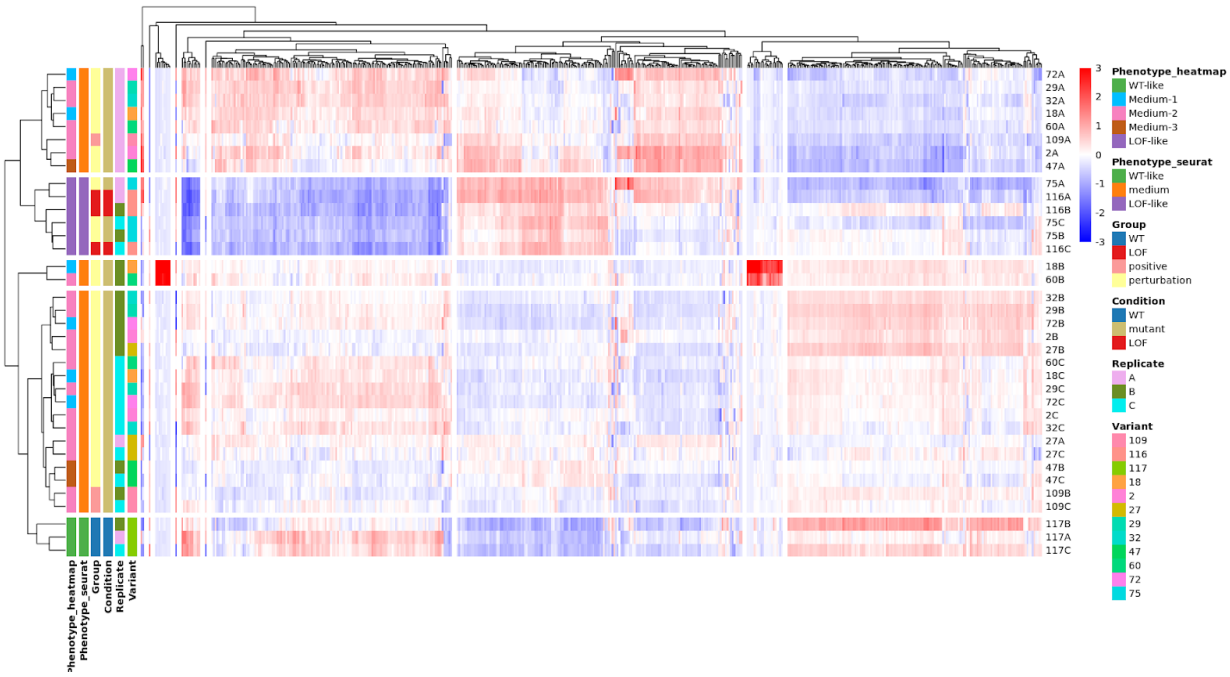


Figure S3.6. Hierarchical clustering of samples identifying two outlier bulk RNA-seq samples. Identifying outlier samples 18B and 60B, before batch effect removal between replicates.

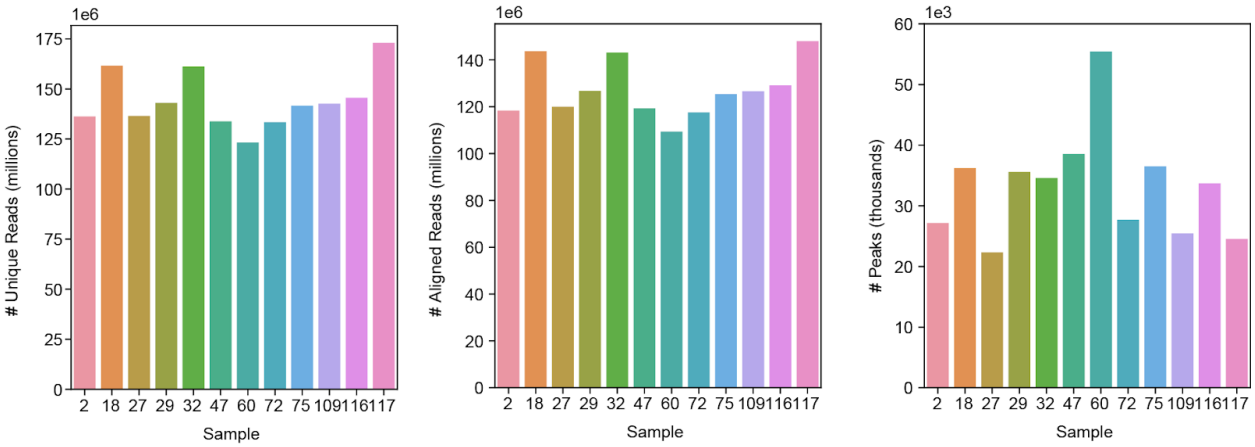


Figure S3.7. ATAC-seq plots. For number of unique reads, aligned reads after filtering (e.g. removal of mitochondrial reads), and number of peaks for each sample across replicates.

Table S3.1. Gene group (Figure 3.3a) scores for each phenotype cluster.

	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group 7	Group 8	Group 9	Group 10
WT-like	0.2210	0.0318	-0.0003	0.1756	-0.0459	-0.0233	-0.0162	-0.0158	-0.0094	-0.0152
Hypomorphic	0.3075	0.0568	0.0474	0.1038	-0.0362	-0.0262	-0.0248	-0.0100	0.0085	-0.0362
LOF-like	0.4028	0.0563	0.0633	0.0346	-0.0401	-0.0566	-0.0098	-0.0112	-0.0151	-0.0307

Table S3.2. Primers.

Name	Description	Sequence
RX1_01	Used to amplify the dsDNA oligo pool of library variants for cloning	GCCGGAGATGTCGAAGAGAATCCTGG ACCGATGCGTATCCCCGTAGATGC
RX1_02	Used to amplify the dsDNA oligo pool of library variants for cloning	ACAGCCAGGAAATAGTTCTAACTTAGCT AGTCAGTAGGGCCTCCACACG
RX1_03	Used in sanger sequencing to identify mutation in the RUNX1 gene to determine variant in library	CTGTGTAGAAGTACTCGCCGATAGTG
RX1_04	Used in sanger sequencing to capture barcode associated with each variant	TCTTGTCTTCGTTGGGAGTG
RX1_05	Used in qPCR to quantify RUNX1 expression	CCACCTACCACAGAGCCATCAA
RX1_06	Used in qPCR to quantify RUNX1 expression.	TTCACTGAGCCGCTCGGAAAAG
RX1_07	Used to amplify the barcodes from cDNA	GACTGGAGTTCAGACGTGTGCTCTTCC GATCTAGAACTATTTCCTGGCTGTTACGCG
GAPDH_F	Used for qPCR of overexpressed peptides.	ACAGTCAGCCGCATCTTCTT
GAPDH_R	Used for qPCR of overexpressed peptides.	ACGACCAAATCCGTTGATC

3.10 Author Contributions

Original concept and project supervision by HC; Project planning, design, and method development by KO and HC; Experiments by RP, JS and PM; Data acquisition, processing and analysis by KO; Preparation of manuscript by KO, RP, JS, PM and HC.

3.11 Acknowledgements

This work was supported by NIH grants DP5OD017937 and U01HG012059 and CIFAR grant FL-000655 to HC and CCMI pilot project funds under NIH Grant U54CA209891. Computational infrastructure support was provided by NIH grant P41GM103504. We thank Adam Klie, Kyle Ford, Jennifer Phuong Nguyen, Agnieszka D'Antonio-Chronowska and Kelly Frazer for many helpful discussions.

Chapter 3, in full, is a reformatted reprint of the material currently being prepared for submission for publication as “Interface-guided phenotyping of coding variants of transcription factor RUNX1 with SEUSS” by Kivilcim Ozturk, Rebecca Panwala, Jeanna Sheen, Prashant Mali, and Hannah Carter. The dissertation author was a primary investigator and author of this paper.

3.12 References

1. Matreyek KA, Starita LM, Stephany JJ, Martin B, Chiasson MA, Gray VE, et al. Multiplex Assessment of Protein Variant Abundance by Massively Parallel Sequencing. doi:10.1101/211011

2. Starita LM, Young DL, Islam M, Kitzman JO, Gullingsrud J, Hause RJ, et al. Massively Parallel Functional Analysis of BRCA1 RING Domain Variants. Genetics. 2015;200: 413–422.

3. Gasperini M, Starita L, Shendure J. The power of multiplexed functional analysis of genetic variants. Nat Protoc. 2016;11: 1782–1787.

4. Fowler DM, Fields S. Deep mutational scanning: a new style of protein science. Nat Methods. 2014;11: 801–807.

5. Dixit A, Parnas O, Li B, Chen J, Fulco CP, Jerby-Arnon L, et al. Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. Cell. 2016;167: 1853–1866.e17.

6. Ursu O, Neal JT, Shea E, Thakore PI, Jerby-Arnon L, Nguyen L, et al. Massively parallel phenotyping of coding variants in cancer with Perturb-seq. Nat Biotechnol. 2022;40: 896–905.

7. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW. Cancer Genome Landscapes. Science. 2013;339: 1546–1558.

8. Stratton MR, Campbell PJ, Andrew Futreal P. The cancer genome. Nature. 2009. pp. 719–724. doi:10.1038/nature07943

9. Weinstein JN, The Cancer Genome Atlas Research Network, Collisson EA, Mills GB, Mills Shaw KR, Ozenberger BA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nature Genetics. 2013. pp. 1113–1120. doi:10.1038/ng.2764

10. Garraway LA, Lander ES. Lessons from the Cancer Genome. Cell. 2013;153: 17–37.

11. Roock WD, De Roock W, Jonker DJ, Di Nicolantonio F, Sartore-Bianchi A, Tu D, et al. Association of KRAS p.G13D Mutation With Outcome in Patients With Chemotherapy-Refractory Metastatic Colorectal Cancer Treated With Cetuximab. JAMA. 2010. p. 1812. doi:10.1001/jama.2010.1535

12. Yu HA, Sima CS, Shen R, Kass S, Gainor J, Shaw A, et al. Prognostic impact of KRAS mutation subtypes in 677 patients with metastatic lung adenocarcinomas. J Thorac Oncol. 2015;10: 431–437.

13. Olivier M, Langerød A, Carrieri P, Bergh J, Klaar S, Eyfjord J, et al. The clinical value of somatic TP53 gene mutations in 1,794 patients with breast cancer. Clin Cancer Res. 2006;12: 1157–1167.

14. Parekh U, Wu Y, Zhao D, Worlikar A, Shah N, Zhang K, et al. Mapping Cellular Reprogramming via Pooled Overexpression Screens with Paired Fitness and Single-Cell RNA-Sequencing Readout. Cell Syst. 2018;7: 548–555.e8.

15. Porta-Pardo E, Garcia-Alonso L, Hrabe T, Dopazo J, Godzik A. A Pan-Cancer Catalogue of Cancer Driver Protein Interaction Interfaces. Nussinov R, editor. PLoS Comput Biol. 2015;11: e1004518.

16. Betts MJ, Lu Q, Jiang Y, Drusko A, Wichmann O, Utz M, et al. Mechismo: predicting the mechanistic impact of mutations and modifications on molecular interactions. Nucleic Acids Res. 2015;43: e10.

17. Engin HB, Kreisberg JF, Carter H. Structure-Based Analysis Reveals Cancer Missense Mutations Target Protein Interaction Interfaces. Srinivasan N, editor. PLoS One. 2016;11: e0152929.

18. Kamburov A, Lawrence MS, Polak P, Leshchiner I, Lage K, Golub TR, et al. Comprehensive assessment of cancer missense mutation clustering in protein structures. Proc Natl Acad Sci U S A. 2015;112: E5486–95.

19. Raimondi F, Singh G, Betts MJ, Apic G, Vukotic R, Andreone P, et al. Insights into cancer severity from biomolecular interaction mechanisms. Sci Rep. 2016;6: 34490.

20. Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J, Tamborero D, Schroeder MP, Jene-Sanz A, et al. IntOGen-mutations identifies cancer drivers across tumor types. Nat Methods. 2013;10: 1081–1082.

21. Janes KA. RUNX1 and its understudied role in breast cancer. Cell Cycle. 2011. doi:10.4161/cc.10.20.18029

22. Bruijn MF de, de Bruijn MF, Speck NA. Core-binding factors in hematopoiesis and immune function. Oncogene. 2004. pp. 4238–4248. doi:10.1038/sj.onc.1207763

23. Ichikawa M, Asai T, Chiba S, Kurokawa M, Ogawa S. Runx1/AML-1 Ranks as a Master Regulator of Adult Hematopoiesis. Cell Cycle. 2004;3: 720–722.

24. Collins A, Littman DR, Taniuchi I. RUNX proteins in transcription factor networks that regulate T-cell lineage choice. Nat Rev Immunol. 2009;9: 106–115.

25. Stelzer G, Rosen N, Plaschkes I, Zimmerman S, Twik M, Fishilevich S, et al. The GeneCards suite: From gene data mining to disease genome sequence analyses. Curr Protoc Bioinformatics. 2016. doi:10.1002/cpbi.5

26. Fukunaga J, Nomura Y, Tanaka Y, Amano R, Tanaka T, Nakamura Y, et al. The Runt domain of AML1 (RUNX1) binds a sequence-conserved RNA motif that mimics a DNA element. RNA. 2013;19: 927–936.

27. Wang Z, Wang P, Li Y, Peng H, Zhu Y, Mohandas N, et al. Interplay between cofactors and transcription factors in hematopoiesis and hematological malignancies. Signal Transduct Target Ther. 2021;6: 24.

28. Baspinar A, Cukuroglu E, Nussinov R, Keskin O, Gursoy A. PRISM: a web server and repository for prediction of protein–protein interactions and modeling their 3D complexes. Nucleic Acids Research. 2014. pp. W285–W289. doi:10.1093/nar/gku397

29. Carter H, Douville C, Stenson PD, Cooper DN, Karchin R. Identifying Mendelian disease genes with the Variant Effect Scoring Tool. BMC Genomics. 2013. doi:10.1186/1471-2164-14-S3-S3

30. Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L. The FoldX web server: An online force field. Nucleic Acids Res. 2005. doi:10.1093/nar/gki387

31. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Research. 2019. pp. D941–D947. doi:10.1093/nar/gky1015

32. Zhou B, Ho SS, Greer SU, Zhu X, Bell JM, Arthur JG, et al. Comprehensive, integrated, and phased whole-genome analysis of the primary ENCODE cell line K562. Genome Res. 2019;29: 472–484.

33. Hotelling H. The Generalization of Student's Ratio. The Annals of Mathematical Statistics. 1931. pp. 360–378. doi:10.1214/aoms/1177732979

34. Tanaka K, Tanaka T, Ogawa S, Kurokawa M, Mitani K, Yazaki Y, et al. Increased expression of AML1 during retinoic-acid-induced differentiation of U937 cells. Biochem Biophys Res Commun. 1995;211: 1023–1030.

35. Yokomizo T, Ogawa M, Osato M, Kanno T, Yoshida H, Fujimoto T, et al. Requirement of Runx1/AML1/PEBP2alphaB for the generation of haematopoietic cells from endothelial cells. Genes Cells. 2001;6: 13–23.

36. Sood R, Kamikubo Y, Liu P. Role of RUNX1 in hematological malignancies. 2017;129(15):2070-2082. Blood. 2018;131: 373.

37. Ot□lora-Ot□lora B, Henr□quez B, L□pez-Kleine L, Rojas A. RUNX family: Oncogenes or tumor suppressors (Review). Oncology Reports. 2019. doi:10.3892/or.2019.7149

38. Kellaway SG, Coleman DJL, Cockerill PN, Raghavan M, Bonifer C. Molecular Basis of Hematological Disease Caused by Inherited or Acquired RUNX1 Mutations. Exp Hematol. 2022;111: 1–12.

39. Bagla S, Regling KA, Wakeling EN, Gadgeel M, Buck S, Zaidi AU, et al. Distinctive phenotypes in two children with novel germline mutations - one with myeloid malignancy and increased fetal hemoglobin. Pediatr Hematol Oncol. 2021;38: 65–79.

40. Kuvardina ON, Herglotz J, Kolodziej S, Kohrs N, Herkt S, Wojcik B, et al. RUNX1 represses the erythroid gene expression program during megakaryocytic differentiation. Blood. 2015;125: 3570–3579.

41. Gerritsen M, Yi G, Tijchon E, Kuster J, Schuringa JJ, Martens JHA, et al. RUNX1 mutations enhance self-renewal and block granulocytic differentiation in human in vitro models and primary AMLs. Blood Advances. 2019. pp. 320–332. doi:10.1182/bloodadvances.2018024422

42. Owens DDG, Anselmi G, Oudelaar AM, Downes DJ, Cavallo A, Harman JR, et al. Dynamic Runx1 chromatin boundaries affect gene expression in hematopoietic development. Nat Commun. 2022;13: 773.

43. Nishimoto N, Arai S, Ichikawa M, Nakagawa M, Goyama S, Kumano K, et al. Loss of AML1/Runx1 accelerates the development of MLL-ENL leukemia through down-regulation of p19ARF. Blood. 2011;118: 2541–2550.

44. Motoda L, Osato M, Yamashita N, Jacob B, Chen LQ, Yanagida M, et al. Runx1 protects hematopoietic stem/progenitor cells from oncogenic insult. Stem Cells. 2007;25: 2976–2986.

45. Goyama S, Schibler J, Cunningham L, Zhang Y, Rao Y, Nishimoto N, et al. Transcription factor RUNX1 promotes survival of acute myeloid leukemia cells. J Clin Invest. 2013;123: 3876–3888.

46. Sanson KR, Hanna RE, Hegde M, Donovan KF, Strand C, Sullender ME, et al. Optimized libraries for CRISPR-Cas9 genetic screens with multiple modalities. Nat Commun. 2018;9: 5416.

47. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 2015;43: D447–52.

48. Berman H, Henrick K, Nakamura H. Announcing the worldwide Protein Data Bank. Nat Struct Biol. 2003;10: 980–980.

49. Ozturk K, Carter H. Predicting functional consequences of mutations using molecular interaction network features. Hum Genet. 2022;141: 1195–1210.

50. Humphrey W, Dalke A, Schulten K. VMD: Visual molecular dynamics. J Mol Graph. 1996. doi:10.1016/0263-7855(96)00018-5

51. Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017;8: 14049.

52. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell. 2015;161: 1202–1214.

53. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15: 550.

54. Li W, Xu H, Xiao T, Cong L, Love MI, Zhang F, et al. MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. Genome Biol. 2014;15: 554.

55. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29: 15–21.

56. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43: e47.

57. Ewels PA, Peltzer A, Fillinger S, Patel H, Alneberg J, Wilm A, et al. The nf-core framework for community-curated bioinformatics pipelines. Nat Biotechnol. 2020;38: 276–278.

58. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25: 1754–1760.

59. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25: 2078–2079.

60. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26: 841–842.

61. Barnett DW, Garrison EK, Quinlan AR, Stromberg MP, Marth GT. BamTools: a C API and toolkit for analyzing and managing BAM files. Bioinformatics. 2011. pp. 1691–1692. doi:10.1093/bioinformatics/btr174

62. Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D. BigWig and BigBed: enabling browsing of large distributed datasets. Bioinformatics. 2010;26: 2204–2207.

63. Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, et al. deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res. 2016;44: W160–5.

64. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008;9: R137.

65. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Molecular Cell. 2010. pp. 576–589. doi:10.1016/j.molcel.2010.05.004

66. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30: 923–930.

67. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016. pp. 3047–3048. doi:10.1093/bioinformatics/btw354

68. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol. 2011;29: 24–26.

69. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. 2016;44: W90–7.

CONCLUSION

Identifying cancer driver mutations among large numbers of passenger mutations is a major challenge in cancer genomics, and a variety of tools have been developed for this task. Typically, these tools rely on protein structure and sequence information to predict variant effects at the protein-level, but biological function happens through interactions among proteins and molecules within cells, creating a need for systems-level information to be incorporated into variant interpretation studies.

There is a growing body of work that points to perturbation of the interactome as a major determinant of pathogenicity [1–18]. Such studies of variant distribution in biological systems have provided insights as to how molecular interaction networks evolve to ensure robustness or vulnerability to genetic variation [19]. It is increasingly apparent that the role of proteins within molecular networks is a key determinant of the potential of variants to exert deleterious effects [10,14,20]. Motivated by these studies, in this dissertation, I aimed to integrate network architecture information to predict variant effects in cancer. First, I designed network-based variant features by combining protein structure and network architecture, and showed that they add new information that is not already present in classical amino acid sequence or structure-based features, and that they can improve variant classification. Next, I demonstrated that distinct patterns of network rewiring of mutations can be indicative of different selective oncogenic pressures, with a case study of B2M and its protein partners HLA-A, HLA-B, and HLA-C, which I found evidence to be under immune evasion type pressures. Finally, I examined transcriptomic effects of distinct protein interaction perturbations as a way to better define the landscape of prospective phenotypes reachable by individual amino acid substitutions, by using an interface-guided Perturb-seq style approach on the transcription factor RUNX1, which is a master regulator of cellular programs and

identity. Overall, this dissertation demonstrates that variant effect interpretation can be significantly improved by incorporating information about the role of proteins and their molecular interactions within biological systems.

One key insight gained through this work is that there was less gain in discriminatory potential for germline variants versus somatic variants, suggesting that network perturbing mutations may not be well tolerated in the germline. This highlights that interface mutations can have large effects on function. In cancer, it was apparent from patterns of enrichment of location of mutations at particular interfaces that positive and purifying selection manifests as patterns on the network. With the rapid growth of human variation databases, it may be possible to study analogous patterns of variation in the germline case, where more purifying selection appears to occur, which could potentially highlight network regions that are important for disease risk in human populations. Somatic variants found in normal tissues may provide additional insights. Finally, the RUNX1 study demonstrated that while most missense variants generated WT-like or non-functional protein, there was some limited potential for interface variants to generate new functions. It is interesting to speculate that the opportunity for new function might be relatively limited for a given protein. Does this generalize, or is there more constraint on neomorphic variants for essential genes than more peripheral genes?

It is also important to consider that my approach relies on a structurally resolved PPI network that allows variants to be characterized according to their potential to affect network architecture by mapping them to their location on protein structures and protein-interaction interfaces. These mappings are used to capture the potential of variant positions to perturb information flow through the network. While use of protein structures and interface information allow analysis of variants by their distinct network perturbation patterns and improve variant effect

interpretation, it also limits availability of training set mutations due to the requirement for structure and interface information to estimate network feature values. It also constrains the coverage of mutations that can be classified.

Additionally, data availability and quality are important considerations for network analyses. PPI networks remain incomplete and may contain many false positive connections. Furthermore, networks assembled from various published experiments may exhibit literature bias; proteins associated with certain phenotypes may be more studied and as a result, may appear more connected in the network, giving the illusion that higher degree nodes are associated with phenotypes of interest [21–23]. Choice of network may thus influence the biological conclusions drawn from network analysis. Indeed, Huang et al. [24] showed that network performance at recovering known disease genes varied according to the disease, and no single network performed best for all diseases. To partially address this issue, several methods for selecting high-quality PPI datasets and score the reliability of an interaction have been developed [25].

Structure is not available for many human proteins, limiting the investigation of variants affecting protein interactions. The extent to which networks themselves are complete also remains poorly understood. Many conclusions have been drawn based on the architecture of the human interactome, however some estimates suggest that at most 20% of interactions have been experimentally measured [26]. In addition, gene expression patterns differ widely across cell types, suggesting that for accurate inference, network architectures need to be cell-type specific. Indeed, disease network modules tend to include genes that are co-expressed in specific tissues [27], and several groups have now constructed tissue-specific networks to study disease variation [28,29], which opens future directions for variant analyses discussed in this work to be revisited in a tissue- or cell-type specific setting. An immediate opportunity would be to integrate tissue specific

expression information to generate tissue-specific network modules, though this will provide only a limited representation. Exhaustively experimentally determining cell-type specific interactions which may change under different stress conditions will be challenging, but possibly a mix of computational modeling and informed experiments could help guide improvements to tissue specific networks. This is indeed the goal of the Impact of Genomic Variants on Function Consortium recently established by the National Human Genome Research Institute (https://igvf.org/).

Finally, network representations are usually static, whereas the biological networks that they represent are dynamic and conditional. Protein interactions often require particular localization or posttranslational modification. Novel technologies such as APEX, a proximity labeling technique recently developed to enable spatially resolved analysis of protein interaction networks [30], may provide a solution to further resolve cell-type specific interactions and subcellular location thereof. Distinguishing between constitutive and transient interactions, and cell-state specific interactions may be important for further understanding the potential of variants to generate relevant phenotypes [31].

New technologies are emerging that can accelerate the pace of interaction profiling and that will create more complete networks and new opportunities for analysis. Next generation sequencing-based interaction screening technologies allow higher throughput screening for binary interactions. Combining such technologies with deep mutational scanning techniques [32] could allow more systematic profiling of the edgetic effects of variants. Indeed, in this work, I employed a Perturb-seq style approach for high-throughput profiling of distinct protein interaction perturbation effects on single cell transcriptional profiles, to better define the landscape of prospective phenotypes reachable by individual amino acid substitutions in a specific cellular

context. This proof of concept sets the stage for future efforts to integrate the transcriptional consequences of variants with context specific networks to further improve predictive modeling, potentially implicating functional effects in specific cell-types or tissues, and provides new insight for both the potential importance of context and approaches to effectively explore this space.

Acknowledgements

The conclusion, in part, includes reformatted reprints of the materials as it appears in "Predicting functional consequences of mutations using molecular interaction network features" in Human Genetics, 2021 by Kivilcim Ozturk and Hannah Carter; and in "Integrating molecular networks with genetic variant interpretation for precision medicine" in Wiley Interdisciplinary Reviews: Systems Biology and Medicine, 2019 by Emidio Capriotti, Kivilcim Ozturk, and Hannah Carter. The dissertation author was a primary author of the first paper, and a secondary author of the second paper.

References

1. Vidal M, Cusick ME, Barabási A-L. Interactome Networks and Human Disease. Cell. 2011. pp. 986–998. doi:10.1016/j.cell.2011.02.016

2. David A, Razali R, Wass MN, Sternberg MJE. Protein-protein interaction sites are hot spots for disease-associated nonsynonymous SNPs. Hum Mutat. 2012;33: 359–363.

3. Sahni N, Yi S, Taipale M, Fuxman Bass JI, Coulombe-Huntington J, Yang F, et al. Widespread macromolecular interaction perturbations in human genetic disorders. Cell. 2015;161: 647–660.

4. Wang X, Wei X, Thijssen B, Das J, Lipkin SM, Yu H. Three-dimensional reconstruction of protein networks provides insight into human genetic disease. Nat Biotechnol. 2012;30: 159–164.

5. Engin HB, Kreisberg JF, Carter H. Structure-Based Analysis Reveals Cancer Missense Mutations Target Protein Interaction Interfaces. Srinivasan N, editor. PLoS One. 2016;11: e0152929.

6. Raimondi F, Singh G, Betts MJ, Apic G, Vukotic R, Andreone P, et al. Insights into cancer severity from biomolecular interaction mechanisms. Sci Rep. 2016;6: 34490.

7. Porta-Pardo E, Garcia-Alonso L, Hrabe T, Dopazo J, Godzik A. A Pan-Cancer Catalogue of Cancer Driver Protein Interaction Interfaces. Nussinov R, editor. PLoS Comput Biol. 2015;11: e1004518.

8. Guo Y, Wei X, Das J, Grimson A, Lipkin SM, Clark AG, et al. Dissecting disease inheritance modes in a three-dimensional protein network challenges the “guilt-by-association” principle. Am J Hum Genet. 2013;93: 78–89.

9. Wei X, Das J, Fragoza R, Liang J, Bastos de Oliveira FM, Lee HR, et al. A massively parallel pipeline to clone DNA variants and examine molecular phenotypes of human disease mutations. PLoS Genet. 2014;10: e1004819.

10. Chen S, Fragoza R, Klei L, Liu Y, Wang J, Roeder K, et al. An interactome perturbation framework prioritizes damaging missense mutations for developmental disorders. Nat Genet. 2018;50: 1032–1040.

11. David A, Sternberg MJE. The Contribution of Missense Mutations in Core and Rim Residues of Protein–Protein Interfaces to Human Disease. Journal of Molecular Biology. 2015. pp. 2886–2898. doi:10.1016/j.jmb.2015.07.004

12. Nishi H, Nakata J, Kinoshita K. Distribution of single-nucleotide variants on protein-protein interaction sites and its relationship with minor allele frequency. Protein Sci. 2016;25: 316–321.

13. Kamburov A, Lawrence MS, Polak P, Leshchiner I, Lage K, Golub TR, et al. Comprehensive assessment of cancer missense mutation clustering in protein structures. Proc Natl Acad Sci U S A. 2015;112: E5486–E5495.

14. Yates CM, Filippis I, Kelley LA, Sternberg MJE. SuSPect: enhanced prediction of single amino acid variant (SAV) phenotype using network features. J Mol Biol. 2014;426: 2692–2701.

15. IMEx Consortium Curators, Del-Toro N, Duesbury M, Koch M, Perfetto L, Shrivastava A, et al. Capturing variation impact on molecular interactions in the IMEx Consortium mutations data set. Nat Commun. 2019;10: 10.

16. Sahni N, Yi S, Zhong Q, Jailkhani N, Charloteaux B, Cusick ME, et al. Edgotype: a fundamental link between genotype and phenotype. Curr Opin Genet Dev. 2013;23: 649–657.

17. Garcia-Alonso L, Jiménez-Almazán J, Carbonell-Caballero J, Vela-Boza A, Santoyo-López J, Antiñolo G, et al. The role of the interactome in the maintenance of deleterious variability in human populations. Mol Syst Biol. 2014;10: 752.

18. Piñero J, Berenstein A, Gonzalez-Perez A, Chernomoretz A, Furlong LI. Uncovering disease mechanisms through network biology in the era of Next Generation Sequencing. Sci Rep. 2016;6: 24570.

19. Capriotti E, Ozturk K, Carter H. Integrating molecular networks with genetic variant interpretation for precision medicine. Wiley Interdiscip Rev Syst Biol Med. 2019;11: e1443.

20. Khurana E, Fu Y, Chen J, Gerstein M. Interpretation of genomic variants using a unified biological network approach. Rzhetsky A, editor. PLoS Comput Biol. 2013;9: e1002886.

21. Chen J, Aronow BJ, Jegga AG. Disease candidate gene identification and prioritization using protein interaction networks. BMC Bioinformatics. 2009;10: 73.

22. Feldman I, Rzhetsky A, Vitkup D. Network properties of genes harboring inherited disease mutations. Proc Natl Acad Sci U S A. 2008;105: 4323–4328.

23. Xu J, Li Y. Discovering disease-genes by topological features in human protein-protein interaction network. Bioinformatics. 2006;22: 2800–2805.

24. Huang JK, Carlin DE, Yu MK, Zhang W, Kreisberg JF, Tamayo P, et al. Systematic Evaluation of Molecular Networks for Discovery of Disease Genes. Cell Syst. 2018;6: 484–495.e5.

25. Peng X, Wang J, Peng W, Wu F-X, Pan Y. Protein-protein interactions: detection, reliability assessment and applications. Brief Bioinform. 2017;18: 798–819.

26. Menche J, Sharma A, Kitsak M, Ghiassian SD, Vidal M, Loscalzo J, et al. Disease networks. Uncovering disease-disease relationships through the incomplete interactome. Science. 2015;347: 1257601.

27. Kitsak M, Sharma A, Menche J, Guney E, Ghiassian SD, Loscalzo J, et al. Tissue Specificity of Human Disease Module. Sci Rep. 2016;6: 35241.

28. Greene CS, Krishnan A, Wong AK, Ricciotti E, Zelaya RA, Himmelstein DS, et al. Understanding multicellular function and disease with human tissue-specific networks. Nat Genet. 2015;47: 569–576.

29. Pierson E, GTEx Consortium D, Koller D, Battle A, Mostafavi S, Ardlie KG, et al. Sharing and Specificity of Co-expression Networks across 35 Human Tissues. Rigoutsos I, editor. PLoS Comput Biol. 2015;11: e1004220.

30. Lobingier BT, Hüttenhain R, Eichel K, Miller KB, Ting AY, von Zastrow M, et al. An Approach to Spatiotemporally Resolve Protein Interaction Networks in Living Cells. Cell. 2017;169: 350–360.e12.

31. Perkins JR, Diboun I, Dessailly BH, Lees JG, Orengo C. Transient protein-protein interactions: structural, functional, and network properties. Structure. 2010;18: 1233–1243.

32. Fowler DM, Fields S. Deep mutational scanning: a new style of protein science. Nat Methods. 2014;11: 801–807.