6-29-2022

Insights to Protein Pathogenicity from the Lens of Protein Evolution

Janelle Nunez-Castilla
jnune033@fiu.edu

FLORIDA INTERNATIONAL UNIVERSITY

Miami, Florida

INSIGHTS TO PROTEIN PATHOGENICITY FROM THE LENS OF PROTEIN EVOLUTION

A dissertation submitted in partial fulfillment of

the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

BIOLOGY

by

Janelle Nunez-Castilla

2022

To: Dean Michael R. Heithaus
College of Arts, Sciences and Education

This dissertation, written by Janelle Nunez-Castilla, and entitled Insights to Protein Pathogenicity from the Lens of Protein Evolution, having been approved in respect to style and intellectual content, is referred to you for judgment.

We have read this dissertation and recommend that it be approved.

_____
Prem Chapagain

_____
Timothy Collins

_____
Matthew DeGennaro

_____
Wensong Wu

_____
Jessica Siltberg-Liberles, Major Professor

Date of Defense: June 29, 2022

The dissertation of Janelle Nunez-Castilla is approved.

_____
Dean Michael R. Heithaus
College of Arts, Sciences and Education

_____
Andrés G. Gil
Vice President for Research and Economic Development
and Dean of the University Graduate School

Florida International University, 2022

COPYRIGHT PAGE

Nunez-Castilla, J., Stebliankin, V., Baral, P., Balbin, C.A., Sobhan, M., Cickovski, T., Mondal, A.M., Narasimhan, G., Chapagain, P., Mathee, K., and Siltberg-Liberles, J. (2022). Potential autoimmunity resulting from molecular mimicry between SARS-CoV-2 Spike and human proteins. *Viruses, 14*(7), 1415. https://doi.org/10.3390/v14071415

ALL OTHER WORK

DEDICATION

A mis abuelos, por todos los sacrificios que hicieron con coraje y amor. Espero haber valido la pena. Los amo.

ACKNOWLEDGMENTS

This dissertation would not have been possible without the help of my advisor, Dr. Jessica Siltberg-Liberles. So many years ago, you accepted me in your lab as a volunteer, having never done bioinformatics a day in my life, and you have made me the scientist I am today. You've met my struggles with incredible patience, kindness, and understanding. Thank you for always supporting me in my interests. My life changed and was made better because of you, and I am eternally grateful. I could not have done this with anyone else and I would not have wanted to.

Many thanks to my committee members, Dr. Timothy Collins, Dr. Prem Chapagain, Dr. Matthew DeGennaro, and Dr. Wensong Wu. You have each guided and supported me throughout my studies. You've also challenged my thinking with your different perspectives, which I definitely needed. Thank you for the ways you've shaped me as a scientist.

A big thank you to the COVID-informatics team at Florida International University. Working with you all was my first venture into collaborating with people in a field different from my own, which has proven to be an invaluable experience. Thank you for giving me the opportunity to learn from you.

I would also like to thank all the members of the Siltberg-Liberles lab, past and present. I would especially like to thank my co-authors Dr. Helena Gomes Dos Santos, Dr. Joseph Ahrens, Jordon Rahaman, and Christian Balbin for their contributions to the work presented in this dissertation and their years of friendship and camaraderie. I have had so much fun doing (and sometimes complaining about) science with you.

Thank you to the Instructional & Research Computing Center (IRCC) at Florida International University for providing HPC computing resources that have made my work possible. Thank you especially to Dr. Cassian D'Cunha for your support in getting my scripts to work when I was already at my wits' end.

While a PhD is largely about research, I'd be remiss to leave out all the help I've gotten with teaching, which formed a great part of my graduate experience. To Thomas Pitzer, thank you for believing in me as the Head Teaching Assistant for General Biology II labs and for allowing me to implement an exercise to introduce these students to bioinformatics. Thank you also to Roberto Pereira for showing me the ropes while I had that position and guiding me through the many conflicts that came my way. My biggest thank you goes to my friend and mentor Dr. Jaime Mayoral. I have so loved being a part of the Make your Mutant team and co-teaching with you. Your training in the lab allowed me to gain experiences that I would not have otherwise had the opportunity to. You included me in the Design-2-Data team and allowed me to foster connections with awesome scientists across the country. Thank you for welcoming me into your family and always lending me and open ear, a helping hand, and sage advice.

It's been a long road. And, if these Acknowledgments are any indication, it has taken a village. I've left the largest and most special part of my village for the end – my wonderful family and friends. I cannot name you all, but that does not lessen my appreciation. To my grandparents—Abuela Olga, Abuelo Roberto, Abuelo Milo, Abuela Esperanza, and Abuelo Panchitoco—thank you for all the sacrifices you've made so that I could have a chance at the life I live. You've been the greatest examples of what hard work and dedication can bring. I am here because of you. I thank my parents, Ledis

Castilla and Juan Pablo Nunez, for loving me and always supporting my educational pursuits and pushing me to reach higher and strive for more. Thank you to my sister, Monika Nunez-Castilla, for lifting me up and indulging me when I need to vent. Nobody does it better than PLU. Thank you to my dearest friends Jennifer Kramer, Rebecca Luis, and Fiorella Salamanca. You bring me light and levity on my darkest days in ways no one else can. Finally, I thank my husband, Dr. Cesar Gonzalez, for always making me feel understood. You've been there for it all, both bad—frustrations, stress, tears—and good—celebrations, accomplishments, joy. Your unwavering love and support have paved this long and bumpy road.

Last, I would like to acknowledge the sources of funding that have supported a portion of the work presented herein, namely the NIGMS RISE Fellowship (Summer 2016 through Summer 2017) and the National Science Foundation Grant No. 2037374 (Fall 2020 through Spring 2021).

ABSTRACT OF THE DISSERTATION

INSIGHTS TO PROTEIN PATHOGENICITY FROM THE LENS OF PROTEIN EVOLUTION

by

Janelle Nunez-Castilla

Florida International University, 2022

Miami, Florida

Professor Jessica Siltberg-Liberles, Major Professor

As protein sequences evolve, differences in selective constraints may lead to outcomes ranging from sequence conservation to structural and functional divergence. Evolutionary protein family analysis can illuminate which protein regions are likely to diverge or remain conserved in sequence, structure, and function. Moreover, nonsynonymous mutations in pathogens may result in the emergence of protein regions that affect the behavior of pathogenic proteins within a host and host response. I aimed to gain insight on pathogenic proteins from cancer and viruses using an evolutionary perspective. First, I examined p53, a conformationally flexible, multifunctional protein mutated in ~50% of human cancers. Multifunctional proteins may experience rapid sequence divergence given trade-offs between functions, while proteins with important functions may be more constrained. How, then, does a protein like p53 evolve? I assessed the evolutionary dynamics of structural and regulatory properties in the p53 family, revealing paralog-specific patterns of functional divergence. I also studied flaviviruses, like Dengue and Zika virus, whose conformational flexibility contributes to antibody-dependent enhancement (ADE). ADE has long complicated vaccine development for

these viruses, making antiviral drug development an attractive alternative. I identified fitness-critical sites conserved in sequence and structure in the proteome of flaviviruses with the potential to act as broadly neutralizing antiviral drug target sites. I later developed Epitopedia, a computational method for epitope-based prediction of molecular mimicry. Molecular mimicry occurs when regions of antigenic proteins resemble protein regions from the host or other pathogens, leading to antibody cross-reactivity at these sites which can result in autoimmunity or have a protective effect. I applied Epitopedia to the antigenic Spike protein from SARS-CoV-2, the causative agent of COVID-19. Molecular mimicry may explain the varied symptoms and outcomes seen in COVID-19 patients. I found instances of molecular mimicry in Spike associated with COVID-19-related blood-clotting disorders and cardiac disease, with implications on disease treatment and vaccine design.

TABLE OF CONTENTS

CHAPTER PAGE

PREFACE 1

I. INTRODUCTION 3
LITERATURE CITED 13

II. FUNCTIONAL DIVERSIFICATION AFTER GENE DUPLICATION: PARALOG SPECIFIC REGIONS OF STRUCTURAL DISORDER AND PHOSPHORYLATION IN P53, P63 AND P73 20
ABSTRACT 21
INTRODUCTION 21
RESULTS 24
DISCUSSION 47
METHODS 53
ACKNOWLEDGMENTS 59
LITERATURE CITED 60
Appendices 66

III. EXPLORING EVOLUTIONARY CONSTRAINTS IN THE PROTEOMES OF ZIKA, DENGUE, AND OTHER FLAVIVIRUSES TO FIND FITNESS-CRITICAL SITES 84
ABSTRACT 85
INTRODUCTION 85
METHODS 90
RESULTS 97
DISCUSSION 110
ACKNOWLEDGMENTS 117
LITERATURE CITED 117
Appendices 126

IV. EPITOPEDIA: IDENTIFYING MOLECULAR MIMICRY BETWEEN PATHOGENS AND KNOWN IMMUNE EPITOPES 128
ABSTRACT 129
INTRODUCTION 129
EPITOPEDIA IMPLEMENTATION 131
EPITOPEDIA DEMONSTRATION 139
PENTAPEPTIDE STRUCTURAL SPACE ANALYSIS 141
CONCLUSION 148
ACKNOWLEDGMENTS 150
DATA AND SOFTWARE AVAILABILITY 150
LITERATURE CITED 150
Appendices 154

V. POTENTIAL AUTOIMMUNITY RESULTING FROM MOLECULAR MIMICRY BETWEEN SARS-COV-2 SPIKE AND HUMAN PROTEINS 164
ABSTRACT....165
INTRODUCTION 165
METHODS 168
RESULTS AND DISCUSSION 175
CONCLUSION....193
ACKNOWLEDGMENTS 196
LITERATURE CITED 197
Appendices....205

VI. CONCLUSIONS AND FUTURE DIRECTIONS....223
LITERATURE CITED 235

VITA 241

LIST OF TABLES

TABLE PAGE

CHAPTER III

1 Flavivirus protein function....98
2 Summary of target sites 102
3 Conservation of target sites in ZIKV, DENV, and WNV strains 105
4 Target sites of the WNV clade with sites identified as having significant evolutionary rate-shifts across clades 108

CHAPTER IV

1 Median RMSD values resulting from RMSD distribution for structural space analysis of pentapeptide pairs of various identity levels and secondary structural categories shown in Figure 5 144
2 Comparisons across pentapeptide identity levels for the same structural class for the no filter and 30% filter datasets, respectively. Significantly different comparisons are shown in blue 145
3 Comparisons between pentapeptide identity levels for the same structural class for the 30% filter vs the no filter dataset. Significantly different comparisons are shown in blue 147

CHAPTER V

1 3D-mimics found for SARS-CoV-2 Spike 176
2 Human AF-3D-mimics for SARS-CoV-2 Spike 177

LIST OF FIGURES

FIGURE PAGE

CHAPTER II

1 p53 origins. (A) Overview of the p53 family phylogeny including 74 representative species across Metazoa and in choanoflagellates, built based on their p53 DBD domains. For the invertebrate part of the tree, support values at the nodes indicate posterior probabilities. Nodes with posterior probability <0.5 are unresolved. For detailed support values and for the vertebrate clade, see Appendix 1. (B) Pfam domain architectures showing the multidomain context in which the p53 DBDs are found. (C) Heat map representation of the disorder propensities predicted by IUPred [15] based on the full-length proteins. Rows correspond to protein sequences and columns to alignment sites; the color gradient from blue to white to red mirrors the disorder propensity gradient from low (blue) to high (red), with white being the boundary between order and disorder (alignment gaps are colored in grey) 26

2 Disorder propensity across the p53 family in vertebrates. (A) Cartoon representation of the p53 family DNA-based phylogeny is shown (p53 clade, grey; p63 clade, blue; p73 clade, green). The p53, p63, and p73 clades contain 101, 102, and 98 sequences, respectively, ranging from shark to human. Horizontal width represents sequence divergence. (B) The profiles of disorder propensity predicted by IUPred [15] are plotted per site according to the multiple sequence alignment. Profiles colored by clade (i) and by species according to the color guide for sequences in the p53 clade (ii), p73 clade (iii), and p63 clade (iv). The cut-off applied to assign structural disorder (≥0.4) or order (<0.4) is marked by the red line. (C) Boxplots showing the fraction of predicted structural disorder for the 301 vertebrate proteins and for the p53 DBD domain for the same vertebrates and for 47 invertebrates separately (all differences in means are statistically significant based on non-parametric tests with p-values <0.05 with the exception of p53-p63 disorder fractions in full length proteins where p-value = 0.25).....32

3 Graphical representation of sequence-based predictions in vertebrates. Heat maps for structural traits plotted in the order of the DNA-based phylogenetic tree context, showing taxa names as boxes colored according to the color guide in Fig 2. The heat maps are showing sequence-based predictions mapped to their corresponding residue sites on the multiple sequence alignment (gaps in the alignment are colored grey): (A) continuous structural disorder propensities by IUPred [15] colored according to the gradient in Fig 1, (B) secondary structure predictions by PSIPRED [24] displaying loop (white), alpha helix (purple) and beta strand (yellow), and (C) sites predicted to be phosphorylated by NetPhos [25] using a 0.75 cut-off (red). Above the

heat maps, normalized evolutionary rates per site are shown for amino acid sequence (SEQ) in green [26] vs binary traits [27] of disorder-order transitions (DOT) in orange (upper left), secondary structure elements-loop transitions (SLT) in blue (upper center), and phosphorylation transitions (PT) in pink (upper right). All evolutionary rates were normalized with a mean of zero and standard deviation of 1 (negative rates for slow evolving sites and positive rates for fast evolving sites). Grey shaded areas delimitate Pfam domain regions. For greater detail on the p53 clade, see Appendix 6 34

4 Accumulated evolutionary rates per site in vertebrates. Accumulated evolutionary rates per site, (A) for the p53 family, (B-D) per clade, p53, p73, p63. SEQ, DOT, SLT, and PT colored according to Fig 3. Light pink shaded areas delimitate Pfam domain regions. Grey shaded areas have at least 10% gaps. One site with accumulated value >10 is marked with a dot 38

5 Distribution of rapid evolutionary rates per region for sites with <10% gaps in vertebrates. The number of sites with above average rates are shown, (A) for the p53 family, (B-D) per clade p53, p73, and p63. SEQ, DOT, SLT, and PT colored according to Fig 3. In addition, the number of sites with all rates below average (ALL_slow: light blue) and all rates above average (ALL_fast: brown) are shown. The numbers below each region label correspond to the total number of sites kept in that region after filtering out all sites with at least 10% gaps 39

6 Three dimensional context of disorder-order transitions (DOT) and structural disorder conservation in vertebrates. DOT and disorder fraction (gaps included) per site are shown mapped onto representative PDB structures for TAD (PDB code 3dac [29]), p53 DBD (PDB code 4hje [30]), and OD domains (PDB code 1olg [31] for p53 and 4a9z [*To be Published*] for p63/p73); (A) DOT, and (E) disorder fraction for the p53 family showing, from left to right, TAD binding interface with MDM2, p53 DBD domains in their functional tetrameric state binding DNA and Zn as cofactor, and ODs in their functional tetrameric state (on top, values were mapped onto a p53 tetramer, and on the bottom values were mapped onto a p63 tetramer); (B-D) DOT and (F-H) disorder fraction per clade p53, p73, and p63 were mapped onto monomeric states. For further information on the ranges of the mapped regions, See Appendix 10. In addition, a p53 DBD domain colored by the rainbow color scheme based on secondary structure succession (from blue to red corresponding to N-terminus and C-terminus, respectively) and mapped onto a string of secondary structure elements is shown inside the box. The same string of secondary structure elements is shown in (F-H) colored by disorder fractions for an easier visualization of the differences across paralogs 42

7 Shared and clade-specific predicted phosphorylation patterns. (A) WebLogos [35] per clade showing 66 alignment positions following a 50% majority rule

of phosphorylation predictions based on a phosphorylation cut-off = 0.75 (NetPhos), gaps included. (B) Phosphorylation predictions mapped onto their alignment sites (numeration based on the full alignment), with scores ranging from 0 (blue) to 1 (red) with 0.5 as the midpoint (white). Gaps are shown in grey. The colored boxes on the left show the distribution of species sorted by the phylogenetic tree following the color scheme as in Fig 2. Shared and clade-specific phosphorylation sites are distributed along domains (yellow shaded areas) and linkers. Sites marked with a circle means p53 clade-specific (black, the phosphorylation site is experimentally validated in PhosphoSite; grey, an adjacent site is experimentally validated to be phosphorylated in PhosphoSite). Sites marked with a star are predicted to be phosphorylated in a p63 or p73 clade-specific manner while p53 has a different experimentally verified posttranslational modification [34].....46

8 Major evolutionary events in the early p53 family. The sequences in Fig 1 are arranged by NIH Common tree taxonomy to show the evolutionary order of events (left). Branches with evidence of gene duplications are marked with a star. Branches with domain loss are marked with a triangle. Branches are not to scale. The protein distribution per species is shown (right). Presence of domains per protein are colored according to the color scheme for domains in Fig 1, with the addition that grey denotes missing domain and white denotes that no additional proteins were detected 49

CHAPTER III

1 Schematic of the flavivirus polyprotein illustrating a the proteins that make up the polyprotein and b the domains that make up the proteins 97

2 Phylogenetic reconstruction of 42 flavivirus polyproteins. Taxa often associate by vector association: no known vector (NKV), tick-borne flaviviruses (TBFV), and mosquito-borne flaviviruses (MBFV). A second NKV group is found within the MBFVs. These viruses have been found to replicate in vitro within *Aedes* spp. cells (Kuno 2007). Nodes indicated with an asterisk have a posterior probability greater than 0.9 but less than 1, all other nodes have a posterior probability of 1.....101

3 Target sites mapped to the RNA-dependent RNA polymerase structure for ZIKV (PDB id: 5TFR (Upadhyay et al. 2017)). Sites shown in purple are shared among 19 MBFVs. Sites in red are ZIKV+DENV clade specific. Sites in blue are WNV clade specific. Residues shown as spheres are exposed to the surface as determined by the PyMOL script findSurfaceResidues.py (Vertrees 2019) using a 2.5 Å² cut-off 103

4 Percent of sites exhibiting significant evolutionary rate-shifts based on a evolutionary rates for individual proteins and b evolutionary rates for the full-length polyprotein. Percent significant rate-shifts are shown in a per-protein context.....107

5 WNV target sites with significant evolutionary rate-shifts shown in a phylogenetic context. Phylogeny on the left is shown re-rooted to recover the three clades used in the rate-shift analysis. Taxa are colored as in Fig. 2. Target sites for the WNV clade are shown on the right in the context of the multiple sequence alignment for the MBFVs. The last row of the alignment, labeled Significant Rate-Shifts, has stars at the sites that were identified to be experiencing significant evolutionary rate-shifts.....109

6 The target sites in the NS3 helicase (DEAD domain) for the ZIKV+DENV clade has the motif HATFT shown in purple. Only two of the five sites (positions 2 and 3, A and T) are surface accessible. However, in ZIKV NS3, these two sites **a** participate in coordinating ssRNA (PDB id: 5GJB (Tian et al. 2016)) and **b** are found in a deep pocket when ssRNA is not bound (PDB id: 5JPS, not published) 113

7 Structural alignment for ZIKV RdRP (PDB id: 5U0C (Zhao et al. 2017)) and HCV RdRP (PDB id: 4WTG (Appleby et al. 2015)). Each of the 8 entities of 5U0C was aligned with the single entity of 4WTG with CATH-SSAP v0.16.2 (Taylor and Orengo 1989; Orengo and Taylor 1996). Entity 1 of 5U0C had the lowest RMSD (4.57 Å) and the highest SSAP score (74.17) with 4WTG. The SSAP alignment for these two entities was used to superpose the structures using CATH-superpose v0.16.2 (Taylor and Orengo 1989). ZIKV is shown in beige, while HCV is shown in gray. Target sites for the MBFVs, ZIKV+DENV clade, and WNV clade are shown mapped onto the ZIKV RdRP and are colored as in Fig. 3. RNA is shown in green. Sofosbuvir is shown in blue. Views of the structural alignment are shown from a the front and b the back. Additionally, ZIKV and HCV RdRPs are shown from c, e the front and d, f the back. g A sequence alignment for the target sites in the ZIKV+DENV clade, the WNV clade, and HCV. ZIKV and WNV are shown as the representative sequences of their clades 115

CHAPTER IV

1 Overview of Epitopedia. Epitopedia is initiated with one or more PDB structures as input. In *Step 1*, a BLASTP search against linear epitope sequences in EPI-SEQ is performed with the corresponding sequence (seqres) from each PDB input as query. In *Step 2*, BLASTP hits that include sequence fragments from the query that do not contain at least 5 consecutive amino acids and where less than 3 amino acids are surface accessible based on the input structure are discarded. For the remaining hits, the PDB fragment is extracted from the input structure. These are considered 1D-mimics. In *Step*

3, structural fragments from the hits from EPI-SEQ that correspond to the 1D-mimics are extracted from PDB structural representatives of the source antigens. In *Step 4* (optional), for hits against epitopes in human source antigens that are not represented in PDB, structural fragments are extracted from AlphaFold models for regions with a certain confidence level (specified by the user). In *Step 5*, TM-align is used to calculate the RMSD of the structural alignment of the BLAST hit fragment or peptide pairs. In *Step 6*, RMSD results for all fragment pairs for all inputs for the run are combined. EpiScore (length of alignment/RMSD) and RMSD histograms are generated, and Z-scores are calculated based on the whole run. A top list of fragment pairs with RMSD ≤ 1Å is created. These fragment pairs are referred to as 3D-mimics.....135

2 Overview of the Epitopedia web interface for 3D-mimics. For each run, (a) information about the run; (b) the mimic and protein in which the mimic was identified; (c) the epitope and its structural representative; (d) identification of the structural representative with MMseqs2; (e) structural comparison of the mimics including EpiScore, EpiScore Z-Score, and RMSD Z-Score; (f) EpiScore distribution for all structurally represented mimics (blue) during the given run including the EpiScore Z-score (grey), with the current mimic in red; (g) RMSD distribution for all structurally represented mimics (blue) during the given run including the RMSD Z-score (grey), with the location of the current mimic in red; (h) link to 3D visualization of the mimic; (i) and while the Best Mimic is shown from the start, additional mimics for the same motif from the same or different proteins but with higher RMSD are included in a dropdown menu.....138

3 Visualization of the mimic pair in 3D. (a) The motif (green) shown in input protein (brown) (b) and in the structural representative protein (blue). (c) The TM-align structural superimposition for the motif in the input protein (brown) and the structural representative (blue). Panels a-c are interactive. (d) The mimic motif is interactive, hovering over a residue in the motif will highlight in panels a-c.....139

4 Epitopedia output overview using PDB 6VXX, chain A as input. For detailed output see example_output folder on the GitHub repository.....140

5 (a) Violin plots of the resulting RMSD distribution from pentapeptide structure analysis. The distributions for the analysis without the 30% parent sequence identity filter are shown in grey while the corresponding distributions for the pentapeptides from the 30% parent sequence identity set are shown in blue. (b) Violin plots showing the distribution of query coverage between the parent sequences for pentapeptide pairs at various identity levels and secondary structure categories. (c) Violin plots showing the distribution of pairwise identity between the parent sequences for pentapeptide pairs at various identity levels and secondary structure categories.....144

CHAPTER V

1 Molecular mimicry with autoimmune potential across SARS-CoV-2 Spike. (**a**) Overview of molecular mimics (solid arrow: 3D-mimic, dashed arrow: AF-3D-mimic) for Spike in the linear sequence showing Spike domains (NTD: N-terminus domain of S1 subunit (green), RBD: receptor binding domain of S1 subunit (orange), CTD: C-terminus domain of S1 subunit (cyan), S2: S2 domain (purple)) as predicted by Pfam [51] based on the NCBI reference sequence (YP:009724390.1). The boundary between S1 and S2 subunits is indicated at S1/S2. (**b**) Surface representation of Spike (PDB id: 6XR8 [18]) colored by subunit (pink, beige, light blue) with residues colored by number of occurrences in a molecular mimic (blue: 1, green: 2, purple: 3, orange: 4 or more). Structural visualization generated with PyMOL [24]. (**c**) The number of occurrences of the sequence motif in human RefSeq Select isoforms arranged in order from the N-terminus to the C-terminus and colored by primary secondary structure element (magenta: α-helix, yellow: β-sheet, blue: coil) based on PDB id 6XR8 chain A.....179

2 The hTPO pathway to induce platelet production. Simplified JAK-STAT signaling pathway in megakaryocytes where hTPO activates the TPO receptor and triggers signaling cascades that stimulate platelet production [60,61]. Created with BioRender.com (accessed on 12 August 2021).....181

3 Structural mimicry between a TQLPP motif in SARS-CoV-2 Spike and an antibody binding epitope in thrombopoietin. (**a**) Pairwise sequence alignment for the TQLPP motif in the epitope for human thrombopoietin (hTPO, IEDB Epitope ID: 920946) and Spike, amino acids colored by Taylor [65] for sites with ≥ 50% conservation in the amino acid property [66]. The region of molecular mimicry is highlighted in the red dashed box. Surface representation of Spike from (**b**) the top and (**c**) the side, with Spike trimer (PDB id: 6XR8 [23]) colored by subunit (pink, beige, light blue) and red indicating the location of the TQLPP epitope fragment, illustrating the surface accessibility of TQLPP and highlighting the location of RBD (dashed oval) and NTD (dashed circle). (**d**) Surface representation shown for hTPO (gray, PDB id: 1V7M [62]) and its TN1 antibody (blue) with the TQLPP motif (red) at the interface. (**e**) TM-align generated structural alignment for TQLPP in Spike (beige) and hTPO (gray), with RMSD = 0.61 Å. (**f**) Violin plots of RMSD values resulting from the comparison of the TQLPP region in 20 Spike trimer structures (60 chains) vs TQLPP in two hTPO structures (PDB ids: 1V7M and 1V7N, chain X for both [49]). Statistical analysis with Mann-Whitney U reveals no statistical significance between the sets. Box plots, bounded by the 1st and 3rd quartiles, show median value (horizontal solid bold line), vertical lines (whiskers) represent 1.5 × IQR, while outliers are marked as black points. For further details, see Methods. Alignment representations were generated with Jalview 2.11.2.2 [66] and structural visualizations were generated with PyMOL 2.5.0 [30] 183

4 Binding of SARS-CoV-2 Spike to TN1 Fab antibody. Equilibrated structure (1 ns) of the modeled TN1 Fab antibody (blue, PDB id: 1V7M) complexed with Spike trimer model (pink, beige, light blue) shown from (a) the side and (b) the top, with TQLPP shown as red spheres. (**c**) The Spike NTD (beige) and TN1 Fab complex used for MD simulations (200 ns), with adjacent glycans at N17 and N74 highlighted in purple. The representative amino acids contributing to hydrogen bonds (dashed lines) during the last 50 ns of simulations for the (d) hTPO-TN1 and (**e**) Spike-TN1 complexes are highlighted as cyan sticks. (f) Violin plot showing the distribution of the MaSIF binding score values for randomly selected patch pairs (blue), the interacting region of Spike-antibody (yellow) and hTPO-TN1 (gray) complexes, and for modeled Spike-TN1 complexes across 40 Spike configurations (red). Statistical analysis with Mann-Whitney U shows that all pairwise comparisons except for Spike-Ab and hTPO-TN1 are significantly different after Bonferroni correction (Appendix 12). Box plots, bounded by the 1st and 3rd quartiles, show median value (horizontal solid bold line), vertical lines (whiskers) represent 1.5 × IQR, while outliers are marked as black points. For further details, see Methods. Structural visualizations were generated with PyMOL 2.5.0 [30] and VMD 1.9.3 [31] 187

5 Predicted interaction patches between TN1 Fab antibody (PDB id: 1V7N) and the TQLPP motif. The best (lowest) binding score is shown for Spike (PDB id: 7LQV, chain A, beige), hTPO (PDB id: 1V7N, chain X, gray), NEK10 (Uniprot: Q6ZWH5, pink), ALG12 (Uniprot: Q9BV10, purple), and FCRL4 (Uniprot: Q96PJ5, light blue). For all, red indicates the TQLPP motif and dark blue dots represent the surface points included in the predicted MaSIF patches 188

6 Structural mimicry between an ELDKY motif in SARS-CoV-2 Spike and epitopes in 6 other proteins. (a) Sequence alignment between SARS-CoV-2 Spike and the epitopes containing the 3D-mimicry motif for human kynureninase (hKYNU, IEDB Epitope ID: 1007556), respiratory syncytial virus fusion F0 glycoprotein (RSV F0, IEDB Epitope ID: 1087776), human cytoplasmic FMR1-interacting protein 1 (hCYFIP1, IEDB Epitope ID: 1346528), human tight junction-associated protein 1 (hTJAP1, IEDB Epitope ID: 1016424), human keratin type I cytoskeletal 18 (hKRT18, IEDB Epitope ID: 1331545), and human tropomyosin alpha-3 (hTPM3, IEDB Epitope ID: 938472). Residues in the molecular mimicry motifs are colored by Taylor [65]. The extended molecular mimicry region is highlighted by the orange dashed box. (b) Surface representation of Spike (PDB id: 6XR8) colored by subunit (beige, pink, light blue) with ELDKY motif indicated in red. Surface representation of proteins (gray) with full or partial 3D-mimics of the ELDKY motif (red): (c) hKYNU (PDB id: 2HZP), (d) RSV F0 (PDB id: 6EAE), (e) hCYFIP1 (PDB id: 4N78), (f) hTJAP1 (Uniprot: Q5JTD0), (g)

hKRT18 (Uniprot: P05783), (h) hTPM3 (Uniprot: P06753). Alignment representations were generated with Jalview 2.11.2.2 [66] and structural visualizations were generated with PyMOL 2.5.0 [30] 190

PREFACE

The following chapters have been submitted for publication and are formatted according to journal specifications:

CHAPTER II:

Dos Santos, H.G.*, Nunez-Castilla, J.*, and Siltberg-Liberles, J. (2016). Functional diversification after gene duplication: paralog specific regions of structural disorder and phosphorylation in p53, p63, and p73. *PLoS ONE, 11*(3), e0151961. https://doi.org/10.1371/journal.pone.0151961

**Co-first authors*

CHAPTER III:

Nunez-Castilla J., Rahaman, J., Ahrens, J.B., Balbin, C.A., and Siltberg-Liberles, J. (2020). Exploring evolutionary constraints in the proteomes of Zika, Dengue, and other Flaviviruses to find fitness-critical sites. *Journal of Molecular Evolution, 88*(4), 399-414. https://doi.org/10.1007/s00239-020-09941-5

CHAPTER IV:

Balbin, C.A.*, Nunez-Castilla, J.*, Stebliankin, V., Baral, P., Sobhan, M., Cickovski, T., Mondal, A.M., Narasimhan, G., Chapagain, P., Mathee, K. and Siltberg-Liberles, J.

Epitopedia: Identifying molecular mimicry between pathogens and known immune epitopes. Under revision for *ImmunoInformatics*.

**Co-first authors*

CHAPTER V:

Nunez-Castilla, J., Stebliankin, V., Baral, P., Balbin, C.A., Sobhan, M., Cickovski, T., Mondal, A.M., Narasimhan, G., Chapagain, P., Mathee, K., and Siltberg-Liberles, J. (2022). Potential autoimmunity resulting from molecular mimicry between SARS-CoV-2 Spike and human proteins. *Viruses*, *14*(7), 1415. https://doi.org/10.3390/v14071415

CHAPTER I

INTRODUCTION

Proteins diverge with time and can evolve at different rates within the same species. For instance, proteins with more critical functions tend to be more conserved than other proteins (Zhang & Yang, 2015). Moreover, residues within the same protein can evolve at different rates depending on a site's structural and functional constraints (Echave et al., 2016), which in turn depend on the environment and intensity of selection (Wollenberg Valero, 2020). Significant site-specific rate-shifts between clades in a phylogeny may be indicative of functional divergence (Gaucher et al., 2002) and can reveal determinants of specificity (Penn et al., 2008). In unrelated organisms, such as viral pathogens and their hosts, convergent evolution of pathogenic proteins can result in short linear motif mimics with the potential to rewire host protein interaction networks (Chemes et al., 2015). Some motif mimics in the pathogen may correspond to immunogenic epitopes in the host (Sarmady et al., 2011) and may trigger a host autoimmune response (Cusick et al., 2012).

The classic paradigm has been that protein structure determines protein function. It would appear to follow, then, that intrinsically disordered proteins lacking a well-defined structure should be similarly lacking in function. However, conformationally flexible proteins are able to sample protein structure space (their conformational ensemble) and can thus adopt multiple functions, more accurately described by the "protein structure-function continuum" model (Uversky, 2019). Conformational flexibility, together with protein promiscuity, promotes evolvability by facilitating functional and structural divergence (Tokuriki & Tawfik, 2009). Further, disordered regions have been found to diverge more rapidly than ordered regions (Brown et al., 2002). However, secondary structure appears to play a role here based on large scale

predictions across eukaryotic protein families, as disordered sites prone to secondary structure have been found to be more conserved than ordered sites also within secondary structure (Ahrens et al., 2016). Sites predicted to be both disordered and within secondary structure elements may in fact be molecular recognition features (MoRFs), which undergo real-time disorder-to-order transitions upon binding (Mohan et al., 2006). MoRFs are known to promote interactions with multiple partners and are a contributor to functional promiscuity (Cumberworth et al., 2013).

Redundancy generated by whole-genome and small-scale gene duplication allows for functional diversification. The most common outcome following a duplication event is that one copy retains the original function while the other copy is lost through pseudogenization (Lynch & Conery, 2000). Alternatively, duplicated genes may be retained through neofunctionalization, where one gene retains the original function while the duplicate may explore novel functions (Ohno, 1970), and subfunctionalization, where the original function is divided between the two copies (Force et al., 1999). Retention of duplicates is affected by gene stoichiometry (dosage effects), such that duplicates from whole-genome duplication are retained at higher rates than those from small-scale duplications (Hughes & Liberles, 2008). The interaction promiscuity of intrinsically disordered proteins is thought to contribute to their sensitivity to dosage effects (Vavouri et al., 2009). On the other hand, multiple interaction partners provide opportunities for subfunctionalization and duplicate retention (Hughes & Liberles, 2008). Additionally, rewiring regions of intrinsic disorder between paralogs (related by duplication) can also result in functional divergence (Ahrens et al., 2017).

The tumor suppressor protein p53 is often described as the Guardian of the Genome. Furthermore, the most common genetic alteration found in human cancers is inactivation of p53, with roughly 50% of cancers having a mutated p53 (Soussi & Béroud, 2001). p53 is an intrinsically disordered transcription factor that functions as a hub protein (Collavin et al., 2010) with roles in maintaining genome integrity, regulating the cell cycle, inducing apoptosis, and more (Bai & Zhu, 2006). Its paralogs, p63 and p73, are also hubs considered to be tumor suppressor proteins although they have clearer roles in developmental processes (Collavin et al., 2010). Intrinsically disordered regions are often enriched in post-translational modifications (Pejaver et al., 2014) and regulate the formation of specific interactions (Uversky et al., 2008), as is the case for the conformationally flexible p53 (Oldfield et al., 2008). Amino acid substitutions may alter the conformational and functional ensemble of a disordered protein like p53, which can have results ranging from no functional effect to gain or loss of function. For a multifunctional protein, the fitness equation can be balanced in a variety of ways if these substitutions lead to the improvement of some functions but the impairment of others. Thus, there may be expansion of a nearly-neutral network that allows for rapid sequence divergence (Wagner, 2008). Alternatively, the nearly-neutral network may be narrowed by the fragility of a protein with many important functions, resulting in slow sequence divergence (Assis & Kondrashov, 2014). To provide insight on the evolution of the conformationally flexible, multifunctional p53 protein and potential functional divergence in its family, I explored the evolutionary dynamics of functional domains, intrinsic disorder, secondary structure, and phosphorylation across p53 and its paralogs.

The flavivirus family includes viruses such as West Nile Virus (WNV), Dengue Virus (DENV), and Zika Virus (ZIKV), the last of which caused an epidemic in the Americas from 2015-2016. WNV is transmitted by a *Culex* spp. vector while DENV and ZIKV are both transmitted by an *Aedes* spp. vector. In phylogenies built for individual proteins, ZIKV is sometimes found to share a more recent common ancestor with WNV than with DENV (Ortiz et al., 2013). Furthermore, there have been conflicting reports on ZIKV's potential to be transmitted by a *Culex* spp. vector (Guedes et al., 2017; Lourenço-de-Oliveira et al., 2018). Antibody-dependent enhancement (ADE) is a phenomenon whereby a prior infection with a closely related virus or serotype can worsen a subsequent viral infection due to insufficient neutralization by existing antibodies. ADE has been observed between the four DENV serotypes (Dejnirattisai et al., 2016; Priyamvada et al., 2016), DENV and ZIKV (Dejnirattisai et al., 2016; Stettler et al., 2016), and WNV and ZIKV (Bardina et al., 2017). ADE has complicated ZIVK (Almeida et al., 2018) and DENV vaccine development (Shukla et al., 2020) with vaccine-enhanced DENV infections having been observed (Hadinegoro et al., 2015), suggesting vaccination efforts may be counterproductive (Ferguson et al., 2016). For DENV, changes in pH have been found to induce functionally relevant conformational transitions in the Envelope protein (Stiasny et al., 2011), and this conformational flexibility has been implicated in altered antibody binding affinity (Kuhn et al., 2015). Given the ADE among closely related flaviviruses and the difficulties facing vaccine development against DENV and ZIKV, efforts may be better concentrated on the development of broadly neutralizing antiviral drugs that consider the evolutionary context of flaviviruses and avoid conformationally flexible regions. To better understand where ZIKV fits in the

flavivirus phylogeny and to explore the potential of broadly neutralizing antiviral drug targets as an alternative to ADE-hindered vaccine efforts, I performed a study to identify fitness-critical sites across ZIKV+DENV and WNV clades.

One member from another group of viruses has recently been the cause of a years-long pandemic. The coronavirus SARS-CoV-2 is the causative agent of COVID-19, a disease whose typical symptoms include fever, cough, shortness of breath (Guan et al., 2020; Wang et al., 2020), and loss of taste or smell (Dawson et al., 2021). As of early July 2022, 554 million cases of COVID-19 have been reported worldwide (World Health Organization, 2022). Recent studies indicate that an estimated one third are asymptomatic (Sah et al., 2021), although many COVID-19 infected individuals experience a variety of complications including liver injury (Tian & Ye, 2020), kidney injury (Han & Ye, 2021), and cardiovascular complications (Long et al., 2020). Similar cardiovascular complications have been observed following vaccination against SARS-CoV-2 as well (Greinacher et al., 2021; Helms et al., 2021; Patone et al., 2021; Schultz et al., 2021). One of the main antigenic proteins in SARS-CoV-2 is Spike (Voss et al., 2021), a protein that protrudes from the viral surface and enables entry into host cells (Shang et al., 2020). Spike is the primary antigenic component in the vaccines against SARS-CoV-2. So, while symptom severity following SARS-CoV-2 infection and vaccination is currently not well understood, molecular mimicry between SARS-CoV-2 Spike and other antigenic or human proteins may offer an explanation. Molecular mimicry refers to shared regions of high molecular similarity in unrelated proteins that allow them to perform similar interactions with other proteins (Cusick et al., 2012). If the mimicry occurs between an antigenic protein and a human protein, cross-reactive antibodies may be produced which

can lead to an autoimmune response (Getts et al., 2013). For example, SARS-CoV-2 Spike may present protein regions that closely resemble epitopes (antigenic protein regions that elicit an immune response) from human proteins, triggering the production of cross-reactive antibodies that erroneously target self-proteins, exacerbating disease symptoms. On the other hand, molecular mimicry may result in heterologous immunity, where prior exposure to a pathogen can result in protective immunity against a different pathogen sharing a mimicry region (Agrawal, 2019). Here, if an individual has been previously exposed to one such pathogen, then upon infection with SARS-CoV-2, molecular mimicry may trigger an immune response sufficient to prevent symptom onset. The study of molecular mimicry can provide insight on disease pathogenesis, improve therapeutic treatment, and inform vaccine design. As such, a means to predict molecular mimicry of known immune epitopes would be of great importance to the broader scientific and medical community. I developed a program, Epitopedia, that predicts molecular mimicry between unrelated proteins. Epitopedia was applied to the Spike protein from SARS-CoV-2 to identify potential regions of molecular mimicry.

The general aim of my doctoral study is to gain insights concerning the evolution of pathogenic proteins from cancer and viruses. I investigate functional divergence in the p53 tumor suppressor protein family. In doing so, I illuminate that p53 is a rapidly evolving protein that appears to still be exploring its function, perhaps explaining why it is so often found mutated in various human cancers. I also identify evolutionarily constrained sites in the flavivirus proteome with the potential to act as sites for broadly neutralizing antiviral drugs. I present Epitopedia, a novel and broadly accessible computational pipeline for the prediction of molecular mimicry from known immune

epitopes. I then apply Epitopedia to the SARS-CoV-2 Spike protein to understand if molecular mimicry could provide an explanation for the variety of disease severity seen in individuals who experience COVID-19. I provide a brief outline of the work performed in the upcoming chapters below.

Functional Diversification After Gene Duplication: Paralog Specific Regions of Structural Disorder and Phosphorylation in p53, p63, and p73

In the second chapter of this dissertation, I investigate functional divergence in the p53 protein family. For nearly 300 vertebrate sequences, I used sequence-based predictors to determine intrinsic disorder, secondary structure, and phosphorylation propensity. I then evaluated the evolutionary dynamics of these structural/functional features in addition to the evolutionary dynamics of the amino acid sequence on a per-site basis. I further assessed the percentage of sites exhibiting rapid evolutionary rates for these four properties across the various domains and linkers found in the p53 protein family. I also mapped the evolutionary dynamics of intrinsic disorder and fraction of intrinsic disorder to structural representatives for the three shared domains in this protein family. Lastly, I evaluated patterns of intrinsic disorder for the p53 DNA-binding domain for a subset of vertebrate sequences compared to invertebrate sequences. Changes in domain composition from invertebrate to vertebrate proteins were also analyzed.

Exploring Functional Constraints in the Proteomes of Zika, Dengue, and Other Flaviviruses to Identify Fitness-Critical Sites

In the third chapter, I identify evolutionarily constrained sites in the flavivirus proteome that are conserved in sequence, structural order, and secondary structure, and present these sites as candidates for broadly neutralizing antiviral drugs. Sites meeting these criteria were considered fitness-critical and are referred to as target sites. I searched for target sites across the full phylogeny of 42 flaviviruses as well as in a clade-specific manner. All target sites were assessed for surface accessibility when possible. I investigated to what extent the identified target sites remained conserved across thousands of Zika virus, Dengue virus, and West Nile virus strains. Further, I determined site-specific evolutionary rates for non-gapped sites in the multiple sequence alignment and analyzed sites experiencing significant rate-shifts between clades as a proxy for functional divergence and determinants of vector specificity.

Epitopedia: Identifying Molecular Mimicry Between Pathogens and Known Immune Epitopes

In the fourth chapter, I present Epitopedia (Balbin et al., 2021), a computational pipeline for the prediction of molecular mimicry of known epitopes. Epitopedia works by taking a structure from the Protein Data Bank (PDB) (Berman et al., 2000) and using the corresponding sequence to BLAST (Altschul et al., 1990) against linear sequence epitopes found in the Immune Epitope Database (Vita et al., 2019). Epitope hits that have at least 5 consecutive amino acids identical to the query are, when possible, further analyzed for structural similarity. Structural representatives for the hits can be identified

from either PDB or AlphaFold2 models of the human proteome (Tunyasuvunakool et al., 2021). In these instances, the corresponding protein region on the query structure is assessed by DSSP (Kabsch & Sander, 1983) to ensure that a minimum of 3 consecutive amino acids are surface accessible and TM-align (Zhang & Skolnick, 2005) is used to calculate the RMSD between the query structure and the hit structure. Hits with an RMSD of at most 1Å are considered candidates for molecular mimicry. Further, I evaluated pentapeptide structural space by comparing the RMSD of pentapeptide pairs from the main secondary structure classes (helix, extended, and coil) at various levels of sequence identity (from 0-100%).

Potential Autoimmunity Resulting from Molecular Mimicry Between SARS-CoV-2 Spike and Human Proteins

Lastly, in the fifth chapter of this dissertation, I use Epitopedia (Balbin et al., 2021) to predict molecular mimicry of the SARS-CoV-2 Spike protein. For all molecular mimicry candidates (see previous section for description), I indirectly evaluated autoimmune potential by assessing how often each molecular mimicry pentapeptide motif was found in the human proteome. I performed more thorough investigations on two mimicry motifs: one from human thrombopoietin, and one found in multiple human proteins and the respiratory syncytial virus glycoprotein. For the mimicry motif in human thrombopoietin, the PDB structure was found bound to an antibody. Thus, the ability of the mimicry motif on SARS-CoV-2 Spike to bind to this antibody was assessed through molecular dynamics simulations and antibody-antigen interface complementarity was assessed with MaSIF-search (Gainza et al., 2019). For the second mimicry motif, I argue

that there is compelling evidence in the literature supporting this motif's potential to produce cardiac disease complications of autoimmune origin in COVID-19 patients.

LITERATURE CITED

Agrawal, B. (2019). Heterologous Immunity: Role in Natural and Vaccine-Induced Resistance to Infections. *Frontiers in Immunology*, *10*, 2631. https://doi.org/10.3389/FIMMU.2019.02631

Ahrens, J. B., Nunez-Castilla, J., & Siltberg-Liberles, J. (2017). Evolution of intrinsic disorder in eukaryotic proteins. *Cellular and Molecular Life Sciences*, *74*, 3163-3174. https://doi.org/10.1007/s00018-017-2559-0

Ahrens, J., Dos Santos, H. G., & Siltberg-Liberles, J. (2016). The Nuanced Interplay of Intrinsic Disorder and Other Structural Properties Driving Protein Evolution. *Molecular Biology and Evolution*, *33*(9), 2248–2256. https://doi.org/10.1093/molbev/msw092

Almeida, R. D. N., Racine, T., Magalhães, K. G., & Kobinger, G. P. (2018). Zika virus vaccines: Challenges and perspectives. *Vaccines*, *6*(3), 62. https://doi.org/10.3390/vaccines6030062

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *245*(3), 403–410.

Assis, R., & Kondrashov, A. S. (2014). Conserved proteins are fragile. *Molecular Biology and Evolution*, *31*(2), 419–424. https://doi.org/10.1093/molbev/mst217

Bai, L., & Zhu, W.-G. (2006). p53: Structure, Function and Therapeutic Applications. *Journal of Cancer Molecules*, *2*(4).

Balbin, C.A., Nunez-Castilla, J., Stebliankin, V., Baral, P., Sobhan, M., Cickovski, T., Mondal, A.M., Narasimhan, G., Chapagain, P., Mathee, K. and Siltberg-Liberles, J. (2021). Epitopedia: identifying molecular mimicry of known immune epitopes. *BioRxiv*. https://doi.org/https://doi.org/10.1101/2021.08.26.457577

Bardina, S. V, Bunduc, P., Tripathi, S., Duehr, J., Frere, J. J., Brown, J. A., Nachbagauer, R., Foster, G. A., Krysztof, D., Tortorella, D., Stramer, S. L., García-Sastre, A., Krammer, F., & Lim, J. K. (2017). Enhancement of Zika virus pathogenesis by preexisting antiflavivirus immunity. *Science*, *356*(6334), 175–180. https://doi.org/10.1126/science.aal4365

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research, 28*(1), 235–242.

Brown, C. J., Takayama, S., Campen, A. M., Vise, P., Marshall, T. W., Oldfield, C. J., Williams, C. J., & Dunker, A. K. (2002). Evolutionary rate heterogeneity in proteins with long disordered regions. *Journal of Molecular Evolution, 55*(1), 104–110.

Chemes, L. B., de Prat-Gay, G., & Sánchez, I. E. (2015). Convergent evolution and mimicry of protein linear motifs in host-pathogen interactions. *Current Opinion in Structural Biology, 32*, 91–101. https://doi.org/10.1016/j.sbi.2015.03.004

Collavin, L., Lunardi, A., & Del Sal, G. (2010). p53-family proteins and their regulators: hubs and spokes in tumor suppression. *Cell Death and Differentiation, 17*(6), 901–911. https://doi.org/10.1038/cdd.2010.35

Cumberworth, A., Lamour, G., Babu, M. M., & Gsponer, J. (2013). Promiscuity as a functional trait: intrinsically disordered regions as central players of interactomes. *Biochemical Journal, 454*(3), 361–369. https://doi.org/10.1042/BJ20130545

Cusick, M. F., Libbey, J. E., & Fujinami, R. S. (2012). Molecular mimicry as a mechanism of autoimmune disease. *Clinical Reviews in Allergy and Immunology, 42*(1), 102–111. https://doi.org/10.1007/s12016-011-8294-7

Dawson, P., Rabold, E. M., Laws, R. L., Conners, E. E., Gharpure, R., Yin, S., Buono, S. A., Dasu, T., Bhattacharyya, S., Westergaard, R. P., Pray, I. W., Ye, D., Nabity, S. A., Tate, J. E., & Kirking, H. L. (2021). Loss of Taste and Smell as Distinguishing Symptoms of Coronavirus Disease 2019. *Clinical Infectious Diseases, 72*(4), 682–685. https://doi.org/10.1093/cid/ciaa799

Dejnirattisai, W., Supasa, P., Wongwiwat, W., Rouvinski, A., Barba-Spaeth, G., Duangchinda, T., Sakuntabhai, A., Cao-Lormeau, V.-M., Malasit, P., Rey, F. A., Mongkolsapaya, J., & Screaton, G. R. (2016). Dengue virus sero-cross-reactivity drives antibody-dependent enhancement of infection with zika virus. *Nature Immunology, 17*, 1102–1108. https://doi.org/10.1038/ni.3515

Echave, J., Spielman, S. J., & Wilke, C. O. (2016). Causes of evolutionary rate variation among protein sites. *Nature Reviews Genetics, 17*(2), 109–121. https://doi.org/10.1038/nrg.2015.18

Ferguson, N. M., Rodríguez-Barraquer, I., Dorigatti, I., Mier-Y-Teran-Romero, L., Laydon, D. J., & Cummings, D. A. T. (2016). Benefits and risks of the Sanofi-Pasteur dengue vaccine: Modeling optimal deployment. *Science, 353*(6303), 1033–1036. https://doi.org/10.1126/science.aaf9590

Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y. L., & Postlethwait, J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, *151*(4), 1531–1545.

Gainza, P., Sverrisson, F., Monti, F., Rodolà, E., Boscaini, D., Bronstein, M. M., & Correia, B. E. (2019). Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nature Methods*, *17*(2), 184–192. https://doi.org/10.1038/s41592-019-0666-6

Gaucher, E. A., Gu, X., Miyamoto, M. M., & Benner, S. A. (2002). Predicting functional divergence in protein evolution by site-specific rate shifts. *Trends in Biochemical Sciences*, *27*(6), 315–321. https://doi.org/10.1016/S0968-0004(02)02094-7

Getts, D. R., Chastain, E. M., Terry, R. L., & Miller, S. D. (2013). Virus infection, antiviral immunity, and autoimmunity. *Immunological Reviews*, *255*(1), 197–209. https://doi.org/10.1111/IMR.12091

Greinacher, A., Thiele, T., Warkentin, T. E., Weisser, K., Kyrle, P. A., & Eichinger, S. (2021). Thrombotic Thrombocytopenia after ChAdOx1 nCov-19 Vaccination. *New England Journal of Medicine*, *384*(22), 2092–2101. https://doi.org/10.1056/nejmoa2104840

Guan, W., Ni, Z., Hu, Y., Liang, W., Ou, C., He, J., Liu, L., Shan, H., Lei, C., Hui, D. S. C., Du, B., Li, L., Zeng, G., Yuen, K.-Y., Chen, R., Tang, C., Wang, T., Chen, P., Xiang, J., … Zhong, N. (2020). Clinical Characteristics of Coronavirus Disease 2019 in China. *New England Journal of Medicine*, *382*(18), 1708–1720. https://doi.org/10.1056/NEJMoa2002032

Guedes, D. R., Paiva, M. H., Donato, M. M., Barbosa, P. P., Krokovsky, L., Rocha, S. W. dos S., Saraiva, K. LA, Crespo, M. M., Rezende, T. M., Wallau, G. L., Barbosa, R. M., Oliveira, C. M., Melo-Santos, M. A., Pena, L., Cordeiro, M. T., Franca, R. F. de O., Oliveira, A. L. de, Peixoto, C. A., Leal, W. S., & Ayres, C. F. (2017). Zika virus replication in the mosquito Culex quinquefasciatus in Brazil. *Emerging Microbes & Infections*, *6*(8), e69. https://doi.org/10.1038/emi.2017.59

Hadinegoro, S. R., Arredondo-García, J. L., Capeding, M. R., Deseda, C., Chotpitayasunondh, T., Dietze, R., Hj Muhammad Ismail, H. I., Reynales, H., Limkittikul, K., Rivera-Medina, D. M., Tran, H. N., Bouckenooghe, A., Chansinghakul, D., Cortés, M., Fanouillere, K., Forrat, R., Frago, C., Gailhardou, S., Jackson, N., … Saville, M. (2015). Efficacy and Long-Term Safety of a Dengue Vaccine in Regions of Endemic Disease. *New England Journal of Medicine*, *373*(13), 1195–1206. https://doi.org/10.1056/NEJMoa1506223

Han, X., & Ye, Q. (2021). Kidney involvement in COVID-19 and its treatments. *Journal of Medical Virology*, *93*(3), 1387–1395. https://doi.org/10.1002/jmv.26653

Helms, J. M., Ansteatt, K. T., Roberts, J. C., Kamatam, S., Foong, K. S., Labayog, J. M. S., & Tarantino, M. D. (2021). Severe, refractory immune thrombocytopenia occurring after sars-cov-2 vaccine. *Journal of Blood Medicine*, *12*, 221–224. https://doi.org/10.2147/JBM.S307047

Hughes, T., & Liberles, D. A. (2008). Whole-genome duplications in the ancestral vertebrate are detectable in the distribution of gene family sizes of tetrapod species. *Journal of Molecular Evolution*, *67*(4), 343–357. https://doi.org/10.1007/s00239-008-9145-x

Kabsch, W., & Sander, C. (1983). Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, *22*(12), 2577–2637. https://doi.org/10.1002/bip.360221211

Kuhn, R. J., Dowd, K. A., Beth Post, C., & Pierson, T. C. (2015). Shake, rattle, and roll: Impact of the dynamics of flavivirus particles on their interactions with the host. *Virology*, *479–480*, 508–517. https://doi.org/10.1016/j.virol.2015.03.025

Long, B., Brady, W. J., Koyfman, A., & Gottlieb, M. (2020). Cardiovascular complications in COVID-19. *American Journal of Emergency Medicine*, *38*(7), 1504–1507. https://doi.org/10.1016/j.ajem.2020.04.048

Lourenço-de-Oliveira, R., Marques, J. T., Sreenu, V. B., Atyame Nten, C., Aguiar, E. R. G. R., Varjak, M., Kohl, A., & Failloux, A.-B. (2018). Culex quinquefasciatus mosquitoes do not support replication of Zika virus. *Journal of General Virology*, *99*(2), 258–264. https://doi.org/10.1099/jgv.0.000949

Lynch, M., & Conery, J. S. (2000). The evolutionary fate and consequences of duplicate genes. *Science*, *290*, 1151–1155.

Mohan, A., Oldfield, C. J., Radivojac, P., Vacic, V., Cortese, M. C., & Dunker, A. K. (2006). Analysis of molecular recognition features (MoRFs). *Journal of Molecular Biology*, *362*(5), 1043–1059. https://doi.org/10.1016/j.mb.2006.07.087

Ohno, S. (1970). *Evolution by gene duplication*. Springer-Verlag.

Oldfield, C. J., Meng, J., Yang, J. Y., Yang, M. Q., Uversky, V. N., & Dunker, A. K. (2008). Flexible nets: disorder and induced fit in the associations of p53 and 14-3-3 with their partners. *BMC Genomics*, *9*(S1). https://doi.org/10.1186/1471-2164-9-S1-S1

Ortiz, J. F., MacDonald, M. L., Masterson, P., Uversky, V. N., & Siltberg-Liberles, J. (2013). Rapid evolutionary dynamics of structural disorder as a potential driving force for biological divergence in flaviviruses. *Genome Biology and Evolution*, *5*(3), 504–513. https://doi.org/10.1093/gbe/evt026

Patone, M., Mei, X. W., Handunnetthi, L., Dixon, S., Zaccardi, F., Shankar-Hari, M., Watkinson, P., Khunti, K., Harnden, A., Coupland, C. A. C., Channon, K. M., Mills, N. L., Sheikh, A., & Hippisley-Cox, J. (2021). Risks of myocarditis, pericarditis, and cardiac arrhythmias associated with COVID-19 vaccination or SARS-CoV-2 infection. *Nature Medicine*, *28*, 1–13. https://doi.org/10.1038/s41591-021-01630-0

Pejaver, V., Hsu, W. L., Xin, F., Dunker, A. K., Uversky, V. N., & Radivojac, P. (2014). The structural and functional signatures of proteins that undergo multiple events of post-translational modification. *Protein Science*, *23*(8). https://doi.org/10.1002/pro.2494

Penn, O., Stern, A., Rubinstein, N. D., Dutheil, J., Bacharach, E., Galtier, N., & Pupko, T. (2008). Evolutionary Modeling of Rate Shifts Reveals Specificity Determinants in HIV-1 Subtypes. *PLoS Computational Biology*, *4*(11), e1000214. https://doi.org/10.1371/journal.pcbi.1000214

Priyamvada, L., Quicke, K. M., Hudson, W. H., Onlamoon, N., Sewatanon, J., Edupuganti, S., Pattanapanyasat, K., Chokephaibulkit, K., Mulligan, M. J., Wilson, P. C., Ahmed, R., Suthar, M. S., & Wrammert, J. (2016). Human antibody responses after dengue virus infection are highly cross-reactive to Zika virus. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(28), 7852–7857. https://doi.org/10.1073/pnas.1607931113

Sah, P., Fitzpatrick, M. C., Zimmer, C. F., Abdollahi, E., Juden-Kelly, L., Moghadas, S. M., Singer, B. H., & Galvani, A. P. (2021). Asymptomatic SARS-CoV-2 infection: A systematic review and meta-analysis. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(34). https://doi.org/10.1073/PNAS.2109229118/-/DCSUPPLEMENTAL

Sarmady, M., Dampier, W., & Tozeren, A. (2011). Sequence- and Interactome-Based Prediction of Viral Protein Hotspots Targeting Host Proteins: A Case Study for HIV Nef. *PLoS ONE*, *6*(6), e20735. https://doi.org/10.1371/journal.pone.0020735

Schultz, N. H., Sørvoll, I. H., Michelsen, A. E., Munthe, L. A., Lund-Johansen, F., Ahlen, M. T., Wiedmann, M., Aamodt, A.-H., Skattør, T. H., Tjønnfjord, G. E., & Holme, P. A. (2021). Thrombosis and Thrombocytopenia after ChAdOx1 nCoV-19 Vaccination. *New England Journal of Medicine*, *384*(22), 2124–2130. https://doi.org/10.1056/NEJMoa2104882

Shang, J., Wan, Y., Luo, C., Ye, G., Geng, Q., Auerbach, A., & Li, F. (2020). Cell entry mechanisms of SARS-CoV-2. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(21), 11727–11734. https://doi.org/10.1073/pnas.2003138117

Shukla, R., Ramasamy, V., Shanmugam, R. K., Ahuja, R., & Khanna, N. (2020). Antibody-Dependent Enhancement: A Challenge for Developing a Safe Dengue Vaccine. *Frontiers in Cellular and Infection Microbiology, 10*. https://doi.org/10.3389/fcimb.2020.572681

Soussi, T., & Béroud, C. (2001). Assessing TP53 status in human tumours to evaluate clinical outcome. *Nature Reviews Cancer, 1*(3), 233–240. https://doi.org/10.1038/35106009

Stettler, K., Beltramello, M., Espinosa, D. A., Graham, V., Cassotta, A., Bianchi, S., Vanzetta, F., Minola, A., Jaconi, S., Mele, F., Foglierini, M., Pedotti, M., Simonelli, L., Dowall, S., Atkinson, B., Percivalle, E., Simmons, C. P., Varani, L., Blum, J., … Corti, D. (2016). Specificity, cross-reactivity, and function of antibodies elicited by Zika virus infection. *Science, 353*(6301), 823–826. https://doi.org/10.1126/science.aaf8505

Stiasny, K., Fritz, R., Pangerl, K., & Heinz, F. X. (2011). Molecular mechanisms of flavivirus membrane fusion. *Amino Acids, 41*(5), 1159–1163. https://doi.org/10.1007/s00726-009-0370-4

Tian, D., & Ye, Q. (2020). Hepatic complications of COVID-19 and its treatment. *Journal of Medical Virology, 92*(10), 1818–1824. https://doi.org/10.1002/jmv.26036

Tokuriki, N., & Tawfik, D. S. (2009). Protein dynamism and evolvability. *Science, 324*(5924), 203–207. https://doi.org/10.1126/science.1169375

Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Žídek, A., Bridgland, A., Cowie, A., Meyer, C., Laydon, A., Velankar, S., Kleywegt, G. J., Bateman, A., Evans, R., Pritzel, A., Figurnov, M., Ronneberger, O., Bates, R., Kohl, S. A. A., … Hassabis, D. (2021). Highly accurate protein structure prediction for the human proteome. *Nature, 596*, 590–596. https://doi.org/10.1038/s41586-021-03828-1

Uversky, V. N. (2019). Protein intrinsic disorder and structure-function continuum. *Progress in Molecular Biology and Translational Science, 166*, 1–17. https://doi.org/10.1016/bs.pmbts.2019.05.003

Uversky, V. N., Oldfield, C. J., & Dunker, A. K. (2008). Intrinsically disordered proteins in human diseases: Introducing the D(2) concept. *Annual Review of Biophysics, 37*, 215–246. https://doi.org/10.1146/annurev.biophys.37.032807.125924

Vavouri, T., Semple, J. I., Garcia-Verdugo, R., & Lehner, B. (2009). Intrinsic Protein Disorder and Interaction Promiscuity Are Widely Associated with Dosage Sensitivity. *Cell, 138*(1), 198–208. https://doi.org/10.1016/j.cell.2009.04.029

Vita, R., Mahajan, S., Overton, J. A., Dhanda, S. K., Martini, S., Cantrell, J. R., Wheeler, D. K., Sette, A., & Peters, B. (2019). The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Research*, *47*(D1), D339–D343. https://doi.org/10.1093/nar/gky1006

Voss, C., Esmail, S., Liu, X., Knauer, M. J., Ackloo, S., Kaneko, T., Lowes, L., Stogios, P., Seitova, A., Hutchinson, A., Yusifov, F., Skarina, T., Evdokimova, E., Loppnau, P., Ghiabi, P., Haijan, T., Zhong, S., Abdoh, H., Hedley, B. D., … Li, S. S. C. (2021). Epitope-specific antibody responses differentiate COVID-19 outcomes and variants of concern. *JCI Insight*, *6*(13). https://doi.org/10.1172/jci.insight.148855

Wagner, A. (2008). Neutralism and selectionism: a network-based reconciliation. *Nature Reviews Genetics*, *9*(12), 965–974. https://doi.org/10.1038/nrg2473

Wang, D., Hu, B., Hu, C., Zhu, F., Liu, X., Zhang, J., Wang, B., Xiang, H., Cheng, Z., Xiong, Y., Zhao, Y., Li, Y., Wang, X., & Peng, Z. (2020). Clinical Characteristics of 138 Hospitalized Patients with 2019 Novel Coronavirus-Infected Pneumonia in Wuhan, China. *Journal of the American Medical Association*, *323*(11), 1061–1069. https://doi.org/10.1001/jama.2020.1585

WHO Coronavirus (COVID-19) Dashboard | *WHO Coronavirus (COVID-19) Dashboard With Vaccination Data*. (2022). https://covid19.who.int/

Wollenberg Valero, K. C. (2020). Aligning functional network constraint to evolutionary outcomes. *BMC Evolutionary Biology*, *20*(1), 1–14. https://doi.org/10.1186/S12862-020-01613-8/FIGURES/5

Zhang, J., & Yang, J. R. (2015). Determinants of the rate of protein sequence evolution. *Nature Reviews Genetics*, *16*(7), 409–420. https://doi.org/10.1038/nrg3950

Zhang, Y., & Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research*, *33*(7), 2302–2309. https://doi.org/10.1093/NAR/GKI524

CHAPTER II

FUNCTIONAL DIVERSIFICATION AFTER GENE DUPLICATION: PARALOG SPECIFIC REGIONS OF STRUCTURAL DISORDER AND PHOSPHORYLATION IN P53, P63 AND P73

ABSTRACT

Conformational and functional flexibility promote protein evolvability. High evolvability allows related proteins to functionally diverge and perhaps to neostructuralize. p53 is a multifunctional protein frequently referred to as the Guardian of the Genome–a hub for e.g. incoming and outgoing signals in apoptosis and DNA repair. p53 has been found to be structurally disordered, an extreme form of conformational flexibility. Here, p53, and its paralogs p63 and p73, were studied for further insights into the evolutionary dynamics of structural disorder, secondary structure, and phosphorylation. This study is focused on the post gene duplication phase for the p53 family in vertebrates, but also visits the origin of the protein family and the early domain loss and gain events. Functional divergence, measured by rapid evolutionary dynamics of protein domains, structural properties, and phosphorylation propensity, is inferred across vertebrate p53 proteins, in p63 and p73 from fish, and between the three paralogs. In particular, structurally disordered regions are redistributed among paralogs, but within clades redistribution of structural disorder also appears to be an ongoing process. Despite its deemed importance as the Guardian of the Genome, p53 is indeed a protein with high evolvability as seen not only in rearranged structural disorder, but also in fluctuating domain sequence signatures among lineages.

INTRODUCTION

Proteins are dynamic, with a natural tendency to rearrange their conformational ensembles in response to the local environment [1]. Conformational flexibility is associated with functional promiscuity and together they promote evolvability [2].

Evolvability offers a route to functional and structural divergence among related proteins, allowing related proteins to functionally diversify and perhaps to neostructuralize [3] and could manifest as a fold transition, a domain change, or a change in conformational flexibility. Conformational flexibility is enabled through the interplay between amino acid residues in proteins and the degree of flexibility depends on the nature of the amino acids. Similarly, structurally disordered protein regions are conformationally flexible. It follows that if the property of structural disorder is not evolutionarily conserved for homologous sites in a protein family, conformational and functional divergence may be inferred.

Recognized as the Guardian of the Genome, yet infamous for its frequent implication in cancer; p53 is a versatile protein, known to perform numerous functions from DNA binding as a transcription factor to a regulator of apoptosis and beyond [4]. With potential to interact with multiple proteins, p53 has been coined a hub, forming an epicenter of incoming and outgoing signals, such as post-translational modifications and interactions with other biomolecules [5]. Conformational flexibility enables p53 to form specific interactions in a regulated fashion [6]. Consequently, a majority of p53's interactions are mediated through structurally disordered regions, which are often enriched in post-translational modifications regulating biomolecular interactions, and p53 is no exception [7]. Many of the structurally disordered regions transition to order upon binding [7], while others may endure a shift in the population of the p53 conformational ensemble [8]. Not only is structural disorder essential for p53's broad functionality, it is accompanied by a complex fitness equation to be considered for every amino acid

substitution in this protein. It was recently reported that the structurally disordered regions in the p53 family were highly diversified in amino acid sequence [9].

For every amino acid substitution, the conformational and functional ensemble may be altered, with plausible scenarios ranging from no change to gain-or-loss of function. While globular protein domains must fold to function, structurally disordered regions may be less constrained, challenging the common concept of structure being more conserved than sequence. Many possibilities to balance the fitness equation exist if some functions are benefitted and others slightly impaired. This could result in an expanded nearly-neutral network that would allow rapid sequence divergence [10]. However, for a protein with many extremely important functions, fragility may narrow the nearly-neutral network ultimately resulting in slow sequence divergence [11]. When a multifunctional, structurally disordered protein like p53 accumulates substitutions on evolutionary time scales, does its functional ensemble diverge? The complexity of this question is apparent; structurally disordered proteins are frequently not found to have their complete structural ensemble experimentally determined, and changes in multifunctionality, as seen for a protein hub, are difficult to conclusively deduce experimentally on evolutionary time scales. Here, we take an evolutionary approach informed by linear predictions to investigate the evolutionary dynamics of structural disorder, secondary structure, functional domains, and phosphorylation, in addition to amino acid substitutions, to gain further insights into the functional ensemble and its potential divergence in the p53 family.

RESULTS

Origins

Reported sightings of a p53 protein and perhaps even a p63/p73 protein in choanoflagellates and invertebrates, suggest that the evolutionary record of p53 predates the beginning of the animal lineage, Metazoa [12]. Thus, a representative p53 family phylogeny including a selection of species ranging from choanoflagellates to primates was constructed for the p53 DNA binding domain (p53 DBD) (Fig 1A). The phylogeny confirms that proteins containing the p53 DBD are found across Metazoa and in choanoflagellates (Fig 1A). In addition to p53 DBD, choanoflagellates and annelids also contain oligomerization domains (ODs) and Sterile Alpha Motif domains (SAMs), while molluscs contain the transactivation domain (TAD), p53 DBD, OD, and SAM. Considering that the same four domain combination is recovered in early chordates, this indicates that this four domain cassette was present prior to the emergence of Ecdysozoa including arthropods (Fig 1B). In the ecdysozoan lineage the p53 ancestor has rapidly diverged and at times regions have been lost, resulting in weak or obliterated traces of the other domains. In hemichordates and early chordates, p53 DBD is found in combinations with OD, TAD and/or SAM. Generally, in non-vertebrates, proteins that not only contain the p53 DBD but additional parts of the four domain cassette tend to cluster, suggesting that more conserved functional sequence motifs may indeed remain within their p53 DBD, compared to the others. Further, cnidarian clusters with the multidomain proteins suggesting that they too may have more of the original functionality left. Noteworthy is that the annelid and mollusc clade, containing *L. gigantea* that comprises the four domain cassette, fall inside the hemichordate and early chordate group. *B. floridae* has two

copies; one (XP_002598770) has the p53 DBD and OD and falls far from all vertebrate p53 domains in this phylogeny, the other (XP_002613954) has the entire four domain cassette. This four domain cassette protein forms the closest outgroup to the entire vertebrate p53 family in this phylogeny and is considered the last common ancestor of all p53, p63 and p73 proteins in vertebrates, in agreement with taxonomy and previous studies [13,14]. In vertebrates, the p53 family consists of two primary clades: one has all p53 proteins, and the other is further split into the p63 and the p73 clades, indicating that p63 and p73 are more similar to each other than to p53.

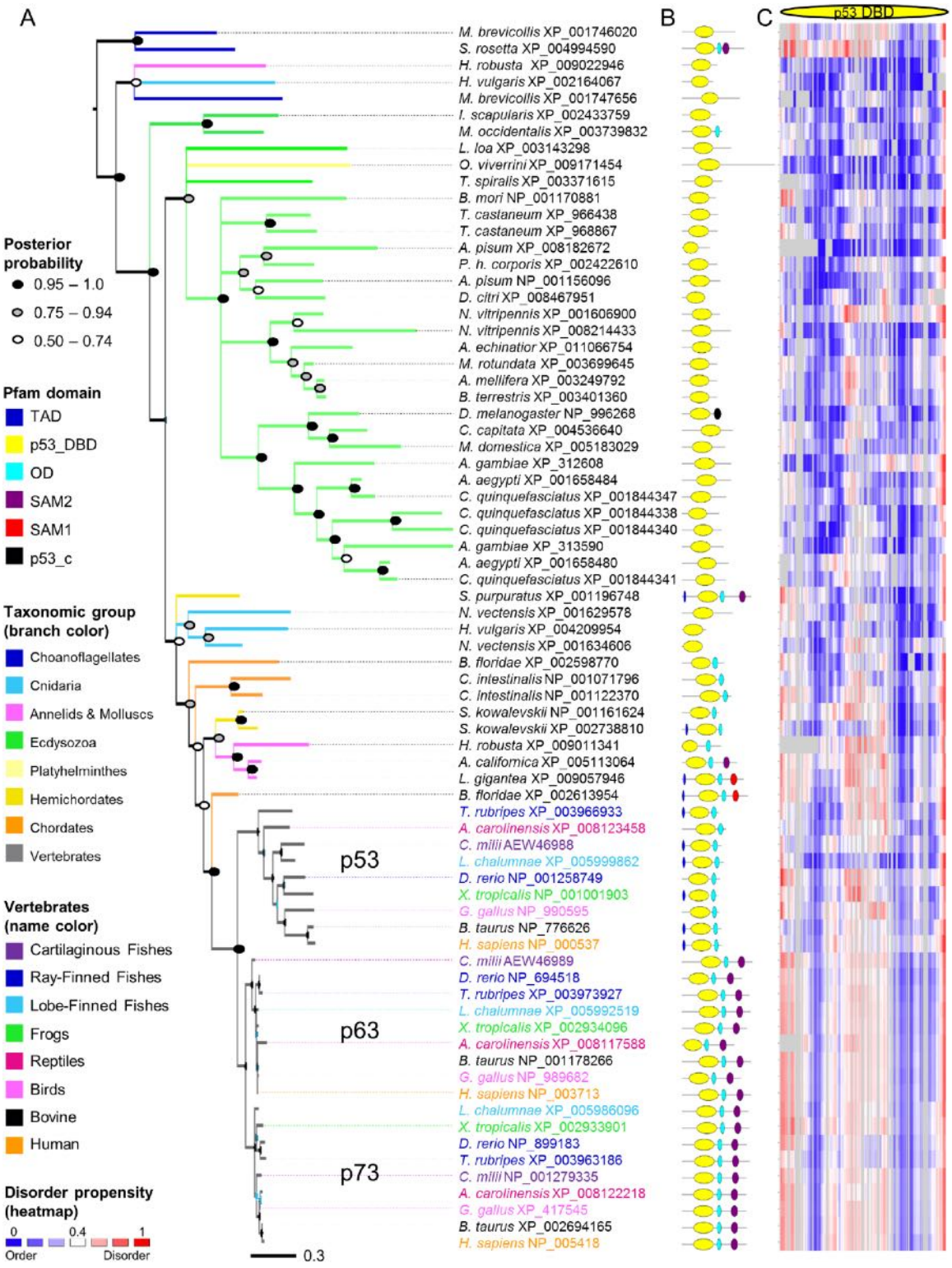


Figure 1. p53 origins. (A) Overview of the p53 family phylogeny including 74 representative species across Metazoa and in choanoflagellates, built based on their p53 DBD domains. For the invertebrate part of the tree, support values at the nodes indicate posterior probabilities. Nodes with posterior probability <0.5 are unresolved. For detailed

support values and for the vertebrate clade, see Appendix 1. (B) Pfam domain architectures showing the multidomain context in which the p53 DBDs are found. (C) Heat map representation of the disorder propensities predicted by IUPred [15] based on the full-length proteins. Rows correspond to protein sequences and columns to alignment sites; the color gradient from blue to white to red mirrors the disorder propensity gradient from low (blue) to high (red), with white being the boundary between order and disorder (alignment gaps are colored in grey).

Vertebrate expansion

The gene duplication pattern resulting in three vertebrate proteins from one ancestral protein is consistent with two whole genome duplications that supposedly occurred at the time of early vertebrates, after the divergence of *B. floridae* but before sharks diverged [13]. To further study the p53 family in vertebrates, a larger vertebrate specific phylogeny was reconstructed. This phylogeny was based on a full-length alignment of 301 sequences with 101, 102, and 98 sequences per p53, p63, and p73 clade, respectively (Appendix 2). The phylogeny shows three specific clades, in agreement with the invertebrate/vertebrate p53 DBD domain tree. Indeed, most vertebrate genomes, from shark to man, seem to encode three genes that belong to the p53 protein family [16], but there are exceptions. Notably, p53 is missing from most of the avian genomes (further discussed below). In addition, there are some lineage-specific small scale duplications of p53. Compared to the ancestral p53 family protein from *B. floridae*, all vertebrate proteins in the p53 family have lost domains, but no domains have been added. Proteins in the p63 and p73 clades overall share the three domain composition of p53 DBD, OD, and SAM. TAD is not identified by Pfam (Appendix 3). In the p53 clade, the evolutionary dynamics of TAD is high. TAD is present in shark, but missing from several ray-finned fish, present in lobe-finned fish and snakes, missing in alligators and birds, and present in most mammals (Appendix 3). For the proteins that lack TAD, the sequence

may remain but the TAD signature is vague. All p53 proteins lack SAM, thus, it was likely lost before sharks diverged. Rarely, SAM is lost from p63 (*P. sinensis* and *B. mutus*) or p73 (*U. maritimus*), and OD is not found in two sequences in the p53 clade. One is after a lineage-specific duplication in *E. edwardii* and the second is from the only bird representative found in data derived from bird genome data, *P. humilis*. Lastly, the N-terminus and linkers between domains are variable in length, and in some cases linkers are even absent.

Birds are not well represented in the p53 clade. Only two bird p53 sequences could be found despite extensive efforts. Notably, the sequence for p53 from *G. gallus* [17] is not found in its whole genome sequence [18,19]. The only avian genome that has remnants of p53 is *P. humilis* [20], although this p53-like sequence only encodes the p53 DBD. *G. gallus* p53 has the p53 DBD and the OD but like many other reptiles, it lacks TAD. Further, these two bird sequences fall outside the reptilian clade as the outgroup to mammals and thus, we cannot conclude that these are the main p53 proteins in *P. humilis* or *G. gallus*. However, given that *G. gallus* and *P. humilis* are distantly related birds and that they fall close to their expected location in the p53 family phylogeny (Appendix 2), it seems plausible that other bird genomes should still encode at least a p53-like protein, but sequencing it from avian genomes appears challenging.

Domain losses or gains between related proteins are strong indications of functional divergence. A domain loss can occur if the sequence diverges beyond recognition or if the region is physically lost [21]. A domain (and a linker) can also appear lost, if different isoforms or partial sequences are considered. Over time, the domain composition of the p53 family has been altered, with high rate of domain loss in

Ecdysozoa where many p53 DBD containing proteins are too short to contain the other domains, but some also have highly divergent OD and SAM domains that no longer generate a significant Pfam domain prediction. In early vertebrates, an ancestral four domain cassette protein was duplicated and subdivided into different proteins, p53, p63, and p73. The p53 clade lost SAM and experienced rapid change in the TAD signature sequence. The p63/p73 clade appears to not change in its current domain organization, but the sequence that once encoded the TAD domain (and may still be present in p63 and p73) has faded beyond recognition, probably prior to the duplication that yielded p63 and p73. Thus, it is possible that a subfunctionalization event followed the first duplication; p53 got most of the TAD domain function, while the p63/p73 ancestor kept the SAM domain.

Sequence divergence: Rate changes at homologous sites

Following the gene duplication resulting in p63 and p73, p63 is much more constrained, manifested by highly conserved sequences among different species, while the p73 clade is less conserved in sequence. The phylogenies based on full-length protein sequence alignments and their corresponding nucleotide sequence alignment reveal that the rate of sequence divergence is greater in the p53 clade (Appendix 2).

A pairwise comparison (based on the full-length protein alignment) between human and shark sequences in the p53, p63, and p73 clades respectively reveal 51.55%, 76.13% and 76.65% sequence identity. Consequently, p63 and p73 are more similar, with 61.81% sequence identity when comparing shark sequences and 59.24% sequence identity when comparing the human sequences. Further, pairwise sequence identity for shark p53 vs. shark p63 and shark p73, reveal 49.02% and 49.86% respectively.

Interestingly, the same comparisons made with the human proteins, p53 vs. p63 and p73, reveal 40.99% and 42.82% pairwise sequence identity, respectively. In summary, the shark p53 family proteins have diverged less than the human counterparts, in accordance with the significantly slower divergence rate found in sharks compared to other vertebrates [22].

Evolutionary dynamics of structural disorder

Highly dependent on conformational flexibility, the proteins in the p53 family are known to vary in stability; p63 is more stable than p73 and the least stable is p53 [23]. Limited studies of p53 proteins from different species show variation in levels of stability also within the p53 clade. Here, we predicted structural disorder propensity as an approximation for conformational flexibility. The disorder profile for the entire p53 family reveals that the predicted disorder propensity per site is highly variable across the entire length of the protein (Fig 2). Dividing the p53 family into the p53, p63, and p73 clades, reveals that the p63 protein is conserved in disorder propensity across the entire protein, while p53 and p73 show multiple regions with varying disorder propensities across their clades (Fig 2). Classifying the sites into either disorder (if the structural disorder propensity is ≥ 0.4) or order (if the structural disorder propensity is <0.4), reveals that, on average, predicted disorder fractions per protein are similar in p53 and p63 clades and higher in p73 clade (means: 0.62 and 0.60 and 0.69, with standard deviations: 0.07, 0.03 and 0.05, respectively). Proteins in the p53 clade show a broader range of disorder, ranging from 0.40 to 0.78 (Fig 2C). However, since p53 has a different domain composition than p63 and p73, comparing only the DBD offers further insights. DBDs in the p53 clade are, on average, predicted to be more ordered than the DBDs in p63 and

p73, with p73 being more disordered than p63 (Fig 2C). The mean and standard deviations are 0.43 (s.d. 0.09), 0.54 (s.d. 0.03) and 0.58 (s.d. 0.08) in p53, p63 and p73 clades (differences in means between them are significant based on non-parametric tests with p-values <0.05). In the p63 and p73 clades, a decrease in the fraction of disorder in DBD domains in ray-finned fish can be observed (Appendix 4). On the contrary, the p53 clade shows the opposite trend, with many ray-finned fish being among the most disordered. It should also be noted that the lobe-finned fish *L. chalumnae* have the most ordered DBD among the entire vertebrate p53 family (Appendix 4). However, also considering the invertebrate p53 DBD, the fractions of disorder in the p53 DBDs are on average smaller than in vertebrates but also more variable within the group (mean 0.23, s.d. 0.16). Single-domain proteins are predicted to be more ordered than those that have contained more of the four domain cassette (Appendix 5).

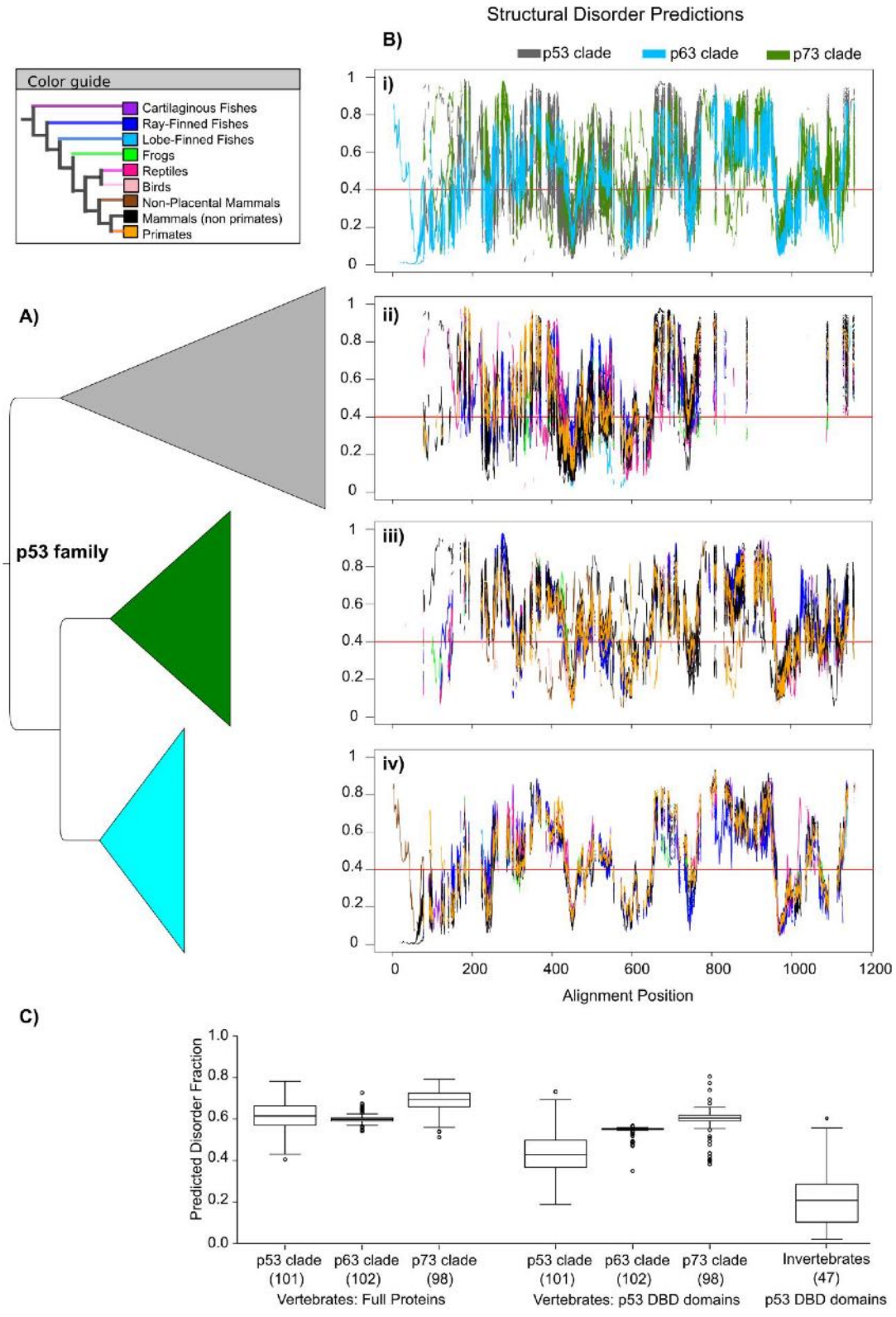


Figure 2. Disorder propensity across the p53 family in vertebrates. (A) Cartoon representation of the p53 family DNA-based phylogeny is shown (p53 clade, grey; p63

clade, blue; p73 clade, green). The p53, p63, and p73 clades contain 101, 102, and 98 sequences, respectively, ranging from shark to human. Horizontal width represents sequence divergence. (B) The profiles of disorder propensity predicted by IUPred [15] are plotted per site according to the multiple sequence alignment. Profiles colored by clade (i) and by species according to the color guide for sequences in the p53 clade (ii), p73 clade (iii), and p63 clade (iv). The cut-off applied to assign structural disorder (≥ 0.4) or order (<0.4) is marked by the red line. (C) Boxplots showing the fraction of predicted structural disorder for the 301 vertebrate proteins and for the p53 DBD domain for the same vertebrates and for 47 invertebrates separately (all differences in means are statistically significant based on non-parametric tests with p-values <0.5 with the exception of p53-p63 disorder fractions in full length proteins where p-value = 0.25).

Although the amount of structural disorder is important for the overall stability of a protein, the location of the disordered and ordered regions, as well as the multidomain context, are crucial. While p63 proteins are consistent for both disorder amount and location across species, the disorder amount and location vary greatly in p53 and p73 proteins from different species, clearly indicating that structural disorder is not conserved here (Fig 2). To address in which regions structural disorder was not conserved, the transition rate of structural disorder-order was examined across the p53 family and in the different clades. The continuous disorder propensity per residue of every protein in the p53 family was mapped onto its corresponding site in the multiple sequence alignment. The resulting heat map, with the sequences arranged corresponding to the phylogenetic tree for the p53 family, reveal interesting patterns of regions that are conserved or changing in disorder propensity (Fig 3A). To further quantify the evolutionary dynamics of structural disorder, the site specific rate of disorder-to-order transition (DOT) was inferred over the phylogeny based on a binary matrix converted from the disorder propensity heat map matrix using the same cut-off as above.

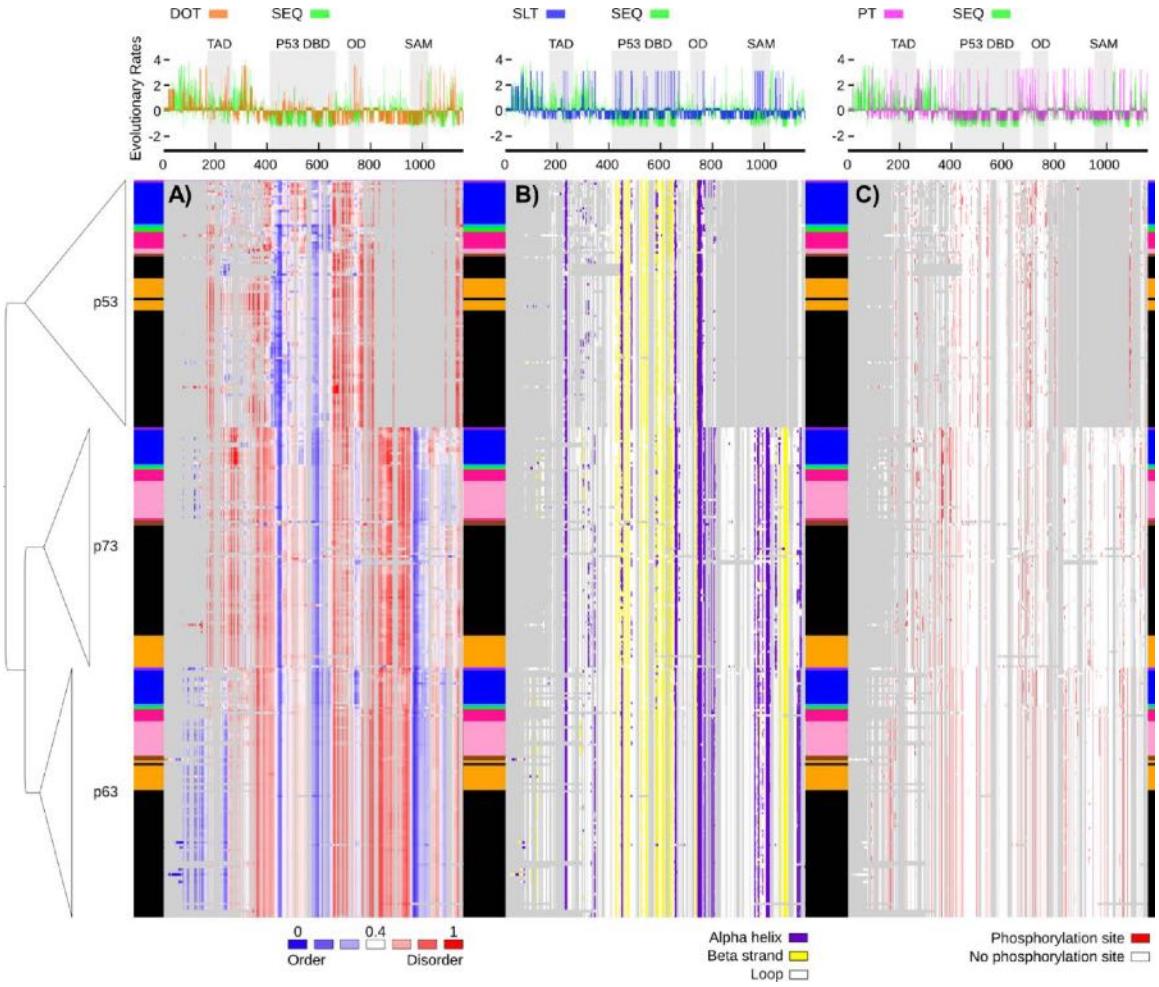


Figure 3. Graphical representation of sequence-based predictions in vertebrates. Heat maps for structural traits plotted in the order of the DNA-based phylogenetic tree context, showing taxa names as boxes colored according to the color guide in Fig 2. The heat maps are showing sequence-based predictions mapped to their corresponding residue sites on the multiple sequence alignment (gaps in the alignment are colored grey): (A) continuous structural disorder propensities by IUPred [15] colored according to the gradient in Fig 1, (B) secondary structure predictions by PSIPRED [24] displaying loop (white), alpha helix (purple) and beta strand (yellow), and (C) sites predicted to be phosphorylated by NetPhos [25] using a 0.75 cut-off (red). Above the heat maps, normalized evolutionary rates per site are shown for amino acid sequence (SEQ) in green [26] vs binary traits [27] of disorder-order transitions (DOT) in orange (upper left), secondary structure elements-loop transitions (SLT) in blue (upper center), and phosphorylation transitions (PT) in pink (upper right). All evolutionary rates were normalized with a mean of zero and standard deviation of 1 (negative rates for slow evolving sites and positive rates for fast evolving sites). Grey shaded areas delimitate Pfam domain regions. For greater detail on the p53 clade, see Appendix 6.

Further, amino acid (sequence) substitution rates per site (SEQ) were inferred (Fig 3). For all rates, throughout this study, positive rates evolve faster than average and negative rates evolve slower than average. DOT is faster than average in most of the p53 spanning region, except in the p53 DBD itself. For the part of the C-terminus that is missing in p53, but before the SAM domain, the sequence is diverging fast, but DOT is slow. Towards the end of SAM and in the C-terminus, p63 and p73 show rapid DOT.

Evolutionary dynamics of secondary structure elements

With a high degree and varying amount of disorder across the p53 family, an analysis of the secondary structure elements propensities was suitable. Mapped in a heat map context, similar to that for the disorder propensity, reveal multiple regions with secondary structure transitions between sequences in the same clade and in a clade-specific manner (Fig 3). To quantify the evolutionary dynamics of secondary structure elements (alpha helix and beta strand) vs. loop across the phylogeny, a binary matrix for these properties was used to infer rates for secondary structure to loop transitions (SLT) (Fig 3). Sites with rapid SLT are found across the entire length of the alignment. Remarkably, the mostly ordered p53 DBD shows several sites with rapid SLT indicating that the structure is fluctuating among species. Also for the seemingly highly similar p63 and p73, like for the DOT, SLT is rapid in the SAM domain.

Evolutionary dynamics of phosphorylation sites

Since phosphorylation frequently modulates the conformations of disordered regions in a regulatory fashion, an analysis of predicted phosphorylation sites was conducted. Here the heat map shows the locations of predicted phosphorylation sites in a binary fashion. Since only Ser, Thr, and Tyr can be phosphorylated, the amount of Ser,

Thr, and Tyr may also be important for how many phosphorylation sites are predicted. However, while there are significant differences in the fraction of Ser, Thr, and Tyr among the different clades (p53, mean 0.17, s.d. 0.01; p63 mean 0.2, s.d. 0.01; p73, mean 0.18, s.d. 0.01) there is no significant difference in the fraction of sites predicted to be phosphorylated when comparing p53, p63 and p73 mean values (p53, mean 0.06, s.d. 0.01; p63, mean 0.06, s.d. 0.01; p73, mean 0.05, s.d. 0.01, significance based on non-parametric tests with p-value < 0.05). In all clades, about 5% of all sites are predicted to be phosphorylated (Appendix 7). To quantify the evolutionary dynamics of phosphorylation sites across the phylogeny, the binary matrix was used to infer rates for presence or absence of phosphorylation sites (PT) (Fig 3). Sites with rapid PT are enriched in the linker regions.

Functional divergence by changes in SEQ, DOT, SLT, and PT rates

Regions that are rapidly changing in disorder, secondary structure, and phosphorylation are likely less important for a conserved function. These rates are calculated for the entire vertebrate p53 family and clade-specific patterns are therefore indistinct. To gain resolution on the clade level, clade-specific rates were estimated (Appendices 8 and 9). Plotting the different rates in an accumulative manner shows that gapped sites indeed have high rates (Fig 4). Since the mere presence of an indel indicates functional change, or perhaps an alternative isoform or a poorly aligned region, our attention is directed to the sites that have less than 10% gaps (Fig 4). For these sites, quantifying the number of sites with rapid DOT, SLT, PT, and SEQ, plus the number of sites that are always fast or always slow for each linker and domain region across the alignment informs which traits are diverging in the different regions (Fig 5A).

Considering the p53 family level, the greater fraction of rapid DOT is found in TAD, the greater fraction of rapid SEQ is found in L1, and the greater fraction of rapid PT is found in L2. The greater fraction of rapid SLT is in L3, however, since the proteins in the p53 clade are shorter than the proteins in the p63 and p73 clades, comparisons beyond the OD domain should be made between p63 and p73 only. Considering the p53 clade (Fig 5B), TAD still has the greater fraction of rapid DOT, and L2 is still high in PT, and L1 in SEQ, but SLT is rather slow. In the p73 clade (Fig 5C), the C-terminus has the greater fraction of rapid DOT, but even the OD domain has almost half of the sites undergoing rapid DOT. SEQ is rather rapid in all linkers, and SLT is rapid for >40% of the 66 sites in the SAM domain. In the p63 clade (Fig 5D), few sites are rapid. In this clade, we note many regions with >50% of sites with all rates slow. OD from p63 and p73 have similar patterns, but more sites are rapid in SLT and PT for p63. The pattern for the OD in p53 is different. Further comparing the C-terminus of p63 to the C-terminus of p73, p63 is more constrained.

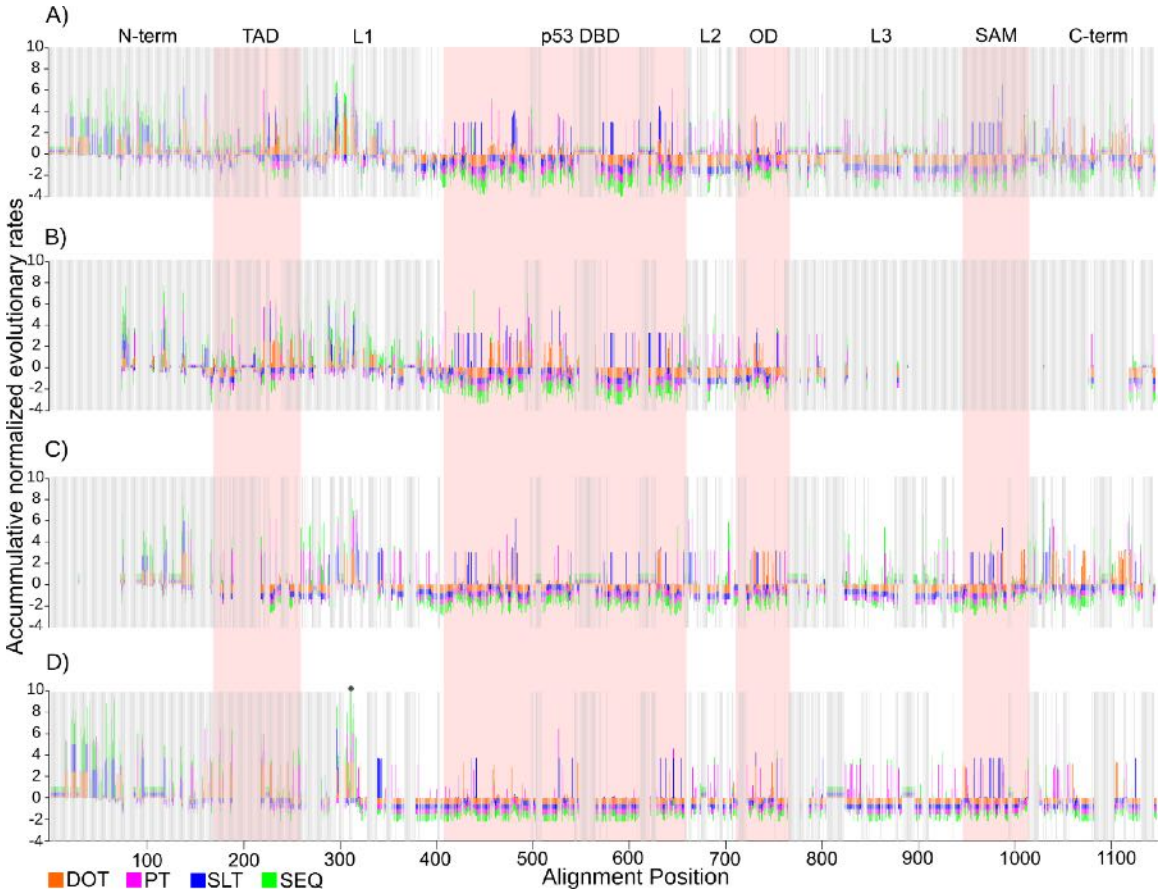


Figure 4. Accumulated evolutionary rates per site in vertebrates. Accumulated evolutionary rates per site, (A) for the p53 family, (B-D) per clade, p53, p73, p63. SEQ, DOT, SLT, and PT colored according to Fig 3. Light pink shaded areas delimitate Pfam domain regions. Grey shaded areas have at least 10% gaps. One site with accumulated value >10 is marked with a dot.

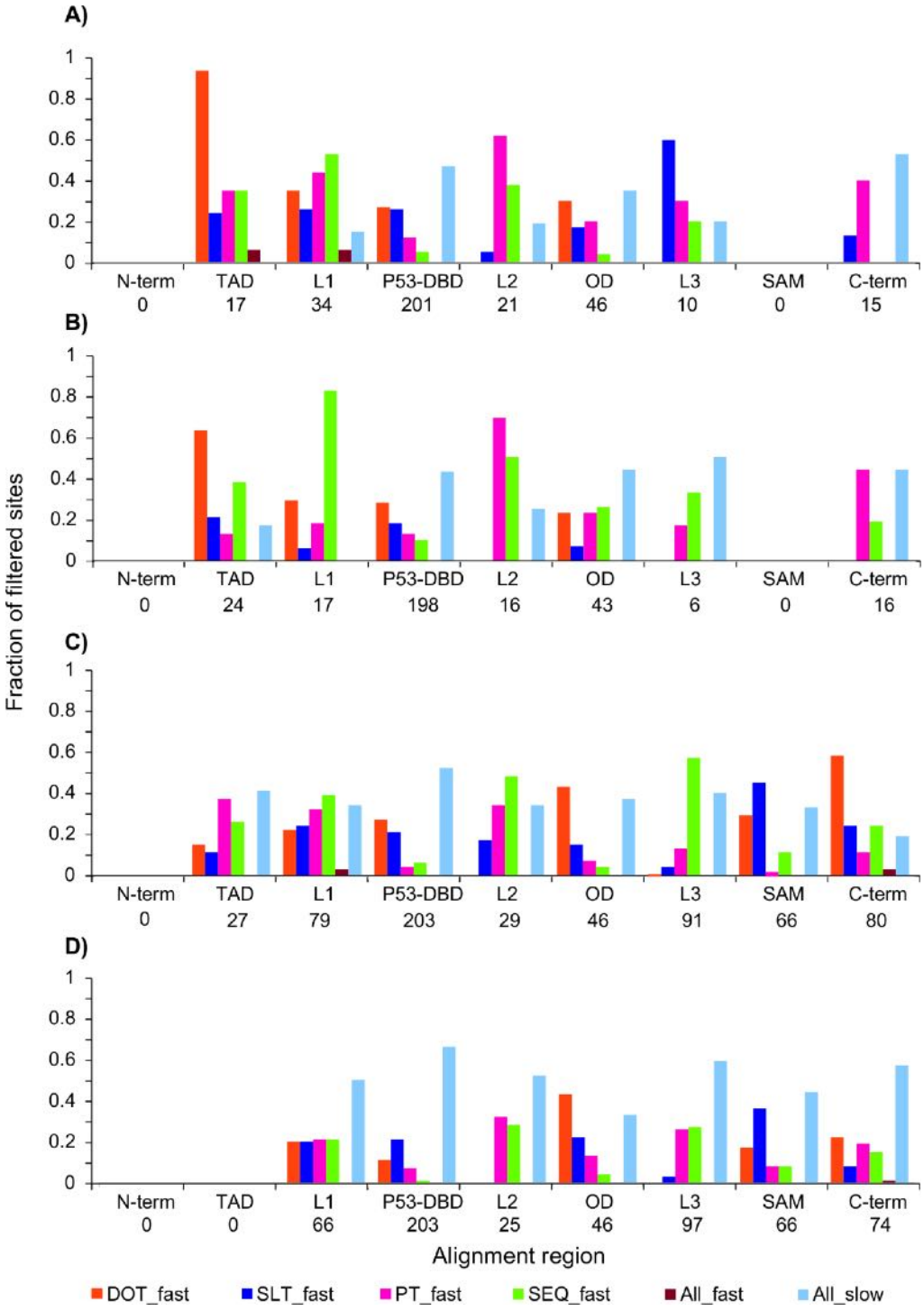


Figure 5. Distribution of rapid evolutionary rates per region for sites with <10% gaps in vertebrates. The number of sites with above average rates are shown, (A) for the p53 family, (B-D) per clade p53, p73, and p63. SEQ, DOT, SLT, and PT colored according to Fig 3. In addition, the number of sites with all rates below average (ALL_slow: light blue) and all rates above average (ALL_fast: brown) are shown. The

numbers below each region label correspond to the total number of sites kept in that region after filtering out all sites with at least 10% gaps.

Structural changes in regions important for molecular interactions

All prediction methods applied are intentionally based on linear sequences and not on 3D structures since the repertoire of 3D structures, although quite impressive for the p53 family, may only provide a limited set of snapshots of the conformational ensemble in which these proteins exist. However, the structural context is valuable and site specific DOT as well as the site specific fractions of predicted disorder were mapped onto structures for TAD, p53 DBD, and OD (all structures used were from human p53 or human p63, and only sites present in the PDB structure were mapped) (Fig 6).

For TAD, the MDM2 binding site is shown (Fig 6A and 6E). Here, moderate DOT is observed for the p53 family. On the clade level, the p53 clade shows rapid DOT (Fig 6B), p73 shows slow DOT (Fig 6C) and p63 has sites with a mixture of slow and rapid DOT (Fig 6D). For disorder conservation in TAD, on the p53 family level and on the p53 clade level intermediate conservation of disorder is observed (Fig 6E and 6F). p73 shows high conservation of disorder (Fig 6G) and p63 shows low conservation of disorder (Fig 6H).

For p53 DBD, the tetrameric state with DNA bound is displayed for the p53 family (Fig 6A and 6E), but for each individual clade, only one of the monomers is shown (Fig 6B-6D and 6F-6H). In general, the region involved in forming the DNA binding p53 DBD dimer and in coordinating Zn as cofactor, has rapid DOT in the p53 clade, as shown in the left circle (Fig 6B). Here, p63 and p73 have slower DOT (Fig 6C and 6D) and conserved disorder (Fig 6G and 6H), while p53 has less conserved disorder (Fig 6F). The p53 clade has rapid DOT at the end of beta strand 4 (B4) and the following

loop (Fig 6F, right circle). The end of the same beta strand shows rapid DOT in p73, while p63 has rapid DOT in the loop. Further, for a second beta strand (B1) in the right circle, p53 is ordered while both p63 and p73 are disordered. Lastly, one of the long beta strands (B10) in the main beta sheet has conserved disorder in p53 while p63 and p73 have conserved order.

For OD, the two different tetrameric states are displayed for the p53 family (Fig 6A and 6E), but for each individual clade, only one of the monomers is shown (Fig 6B-6D and 6F-6H). Earlier studies of the tetramerization in p53 vs. p63 and p73 revealed that the latter two require an additional alpha helix at the C-terminus of OD in order to form stable tetramers and that heterotetramers between p63 and p73, but not p53, can form [28]. Thus, different PDB structures were used to map the functional tetrameric states for p53 and p63/p73, respectively. On the p53 family level, the area around the central horizontal axis and the ends have rapid DOT, while the rest has intermediate DOT. In the p53 clade, DOT is slow except around the horizontal axis (Fig 6B). For p63 and p73, DOT is rapid, perhaps with a slower tendency at the horizontal axis (Fig 6D and 6C). For disorder conservation in OD, p53 has conserved disorder, with slightly less conservation around the horizontal axis (Fig 6F-6H). In p73, sites are more conserved in disorder or lack of disorder, but some sites are not conserved in either property. In p63, most sites are conserved in either disorder or complete lack of disorder.

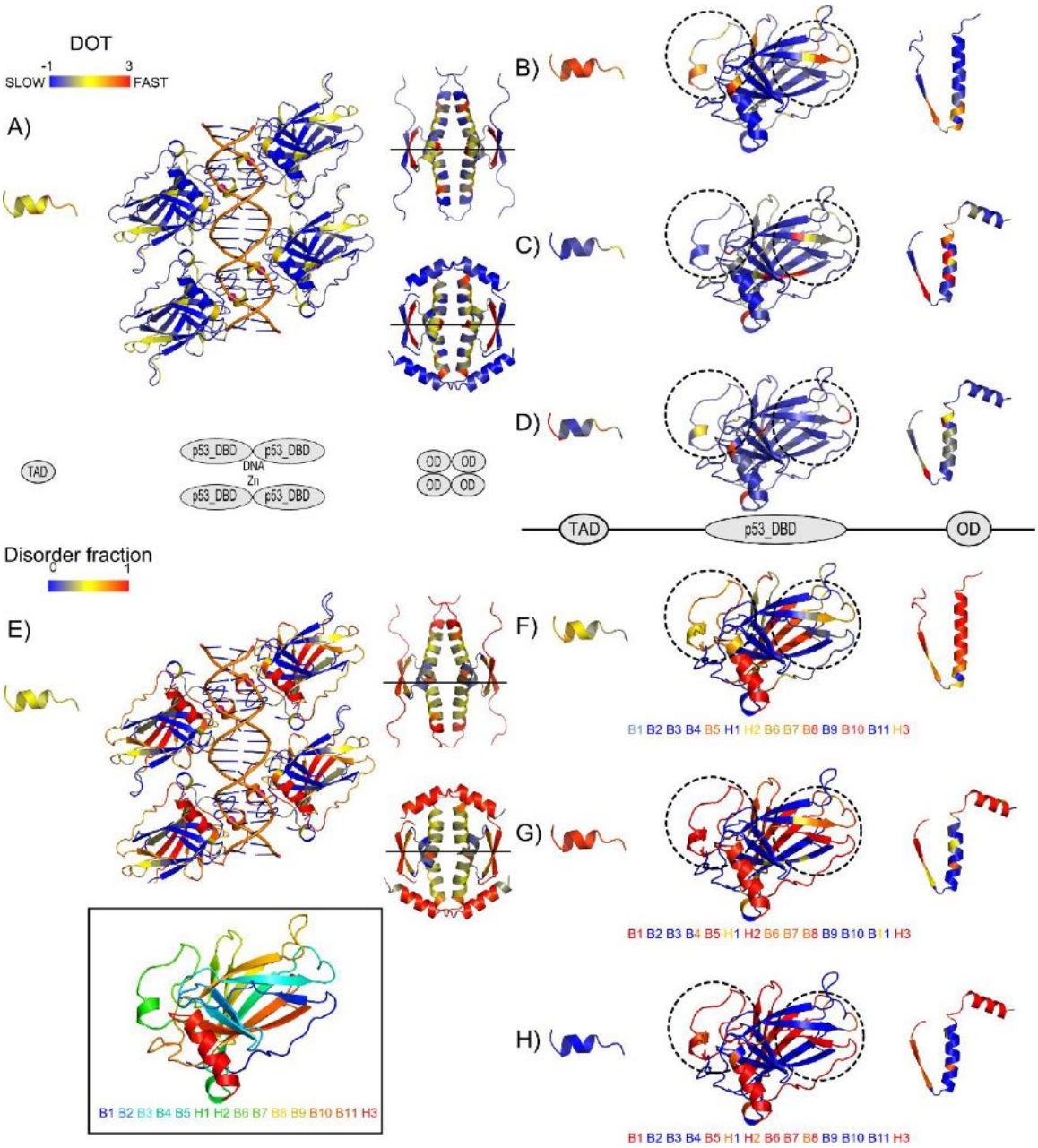


Figure 6. Three dimensional context of disorder-order transitions (DOT) and structural disorder conservation in vertebrates. DOT and disorder fraction (gaps included) per site are shown mapped onto representative PDB structures for TAD (PDB code 3dac [29]), p53 DBD (PDB code 4hje [30]), and OD domains (PDB code 1olg [31] for p53 and 4a9z [*To be Published*] for p63/p73); (A) DOT, and (E) disorder fraction for the p53 family showing, from left to right, TAD binding interface with MDM2, p53 DBD domains in their functional tetrameric state binding DNA and Zn as cofactor, and ODs in their functional tetrameric state (on top, values were mapped onto a p53 tetramer, and on the bottom values were mapped onto a p63 tetramer); (B-D) DOT and (F-H) disorder fraction per clade p53, p73, and p63 were mapped onto monomeric states. For further information on the ranges of the mapped regions, see Appendix 10. In addition, a p53

DBD domain colored by the rainbow color scheme based on secondary structure succession (from blue to red corresponding to N-terminus and C-terminus, respectively) and mapped onto a string of secondary structure elements is shown inside the box. The same string of secondary structure elements is shown in (F-H) colored by disorder fractions for an easier visualization of the differences across paralogs.

Diverging regulation through phosphorylation

To investigate if phosphorylation may be one of the mechanisms utilized to differentiate the regulatory pathways of p53, p63, and p73 from each other, shared and clade-specific phosphorylation sites were identified using a 50% majority rule either within a clade or across the entire p53 family. In total, 66 phosphorylation sites were identified (Appendix 11). Of these 66 sites, only two sites were predicted to be phosphorylated for all three clades. One, and three, sites were shared across p53/p73 and p53/p63, respectively, while eight sites were shared across p63/p73. The remaining 52 sites were clade-specific. In the p53, p63, and p73 clades, respectively, 12, 28, and 12 sites were predicted to be phosphorylated in more than 50% of the sequences for each clade. Since p53 proteins have been extensively studied, many experimental phosphorylation sites are known. For nine out of the 12 p53 clade-specific sites identified here, the NetPhos predictions are in agreement with the experimental data in the PhosphoSite database (as of Dec. 2015) that includes conserved phosphorylation sites for p53 across human, mouse, rat, rabbit and green monkey [32]. For two of the three remaining sites, the adjacent site has been experimentally validated to be phosphorylated. None of the 12 p53 clade-specific sites have been experimentally reported to be phosphorylated in PhosphoSite for either p63 or p73 homologs. For p63 and p73 clade-specific sites, no phosphorylations have been experimentally reported in PhosphoSite for the corresponding site in the p53 homologs, in agreement with the NetPhos predictions.

Indeed, clade-specific positioning of phosphorylation sites in the different clades in the p53 family seem to contribute to their specific regulatory pathways. Further, not only does the phosphorylation site pattern differ between clades, but the p53 family also seem to exploit another strategy for functional diversification through shifts in the type of post-translational modification in homologous sites across paralogs. In particular, for at least three of the p63 and/or p73 clade-specific phosphorylation sites, p53 is also post-translationally modified, but with a different modification (Fig 7 and Appendix 11). Alignment site 253 (TAD region) is predicted to be phosphorylated in the p73 clade (S26 in human p73). This site has Leu in most p63 sequences and Asn in some p53 sequences. For human p53, this site corresponds to Asn30 that has been found to be methylated on the carboxyl by PIMT [33,34]. Similarly, alignment site 498 (p53 DBD region) is predicted to be phosphorylated in the p63 clade (S250 in human p63). This site has Gly in all p73 sequences and Cys in some p53 sequences. For human p53, this site corresponds to Cys182 that has been found to be glutathionylated [34]. Lastly, alignment site 744 (OD region) is predicted to be phosphorylated in the p63 clade (T410 in human p63). This site has Asn in all p73 sequences and Arg in most p53 sequences. For human p53, this site corresponds to Arg337 that is known to be dimethylated [34]. Further, changes in amino acid states with compensatory effects through negatively charged amino acids were observed, e.g. alignment site 225 is phosphorylated in p53 and p63, but has Glu in p73, suggesting that p73 may resemble the phosphorylated state. Also other changes in amino acid among these sites maintain the majority of the physicochemical properties, as in Tyr-Phe transitions, while removing or adding a regulatory switch. Interestingly, some observed transitions are directly involving Ser, Thr or Tyr residues. Phosphorylation

transitions between Ser/Thr (e.g. alignment site 471) are, in general, expected to conserve kinase partner and thus, conserve the regulatory mechanism, while transitions from Ser/Thr phosphorylations to Tyr phosphorylations suggest divergent mechanisms of regulation via different kinases. Alignment site 164 in p63 clade switches from Ser in ray-finned fish to Tyr in the rest of species (with shark as an exception), suggesting divergent regulation in ray-finned fish p63 proteins. Thus, differential regulation within orthologs is implied. Also, alignment site 165 is known to be phosphorylated in human p63 (Tyr36) in PhosphoSite, but this phosphorylation site is missing in all fish, where shark has Cys and the others Phe or Leu.

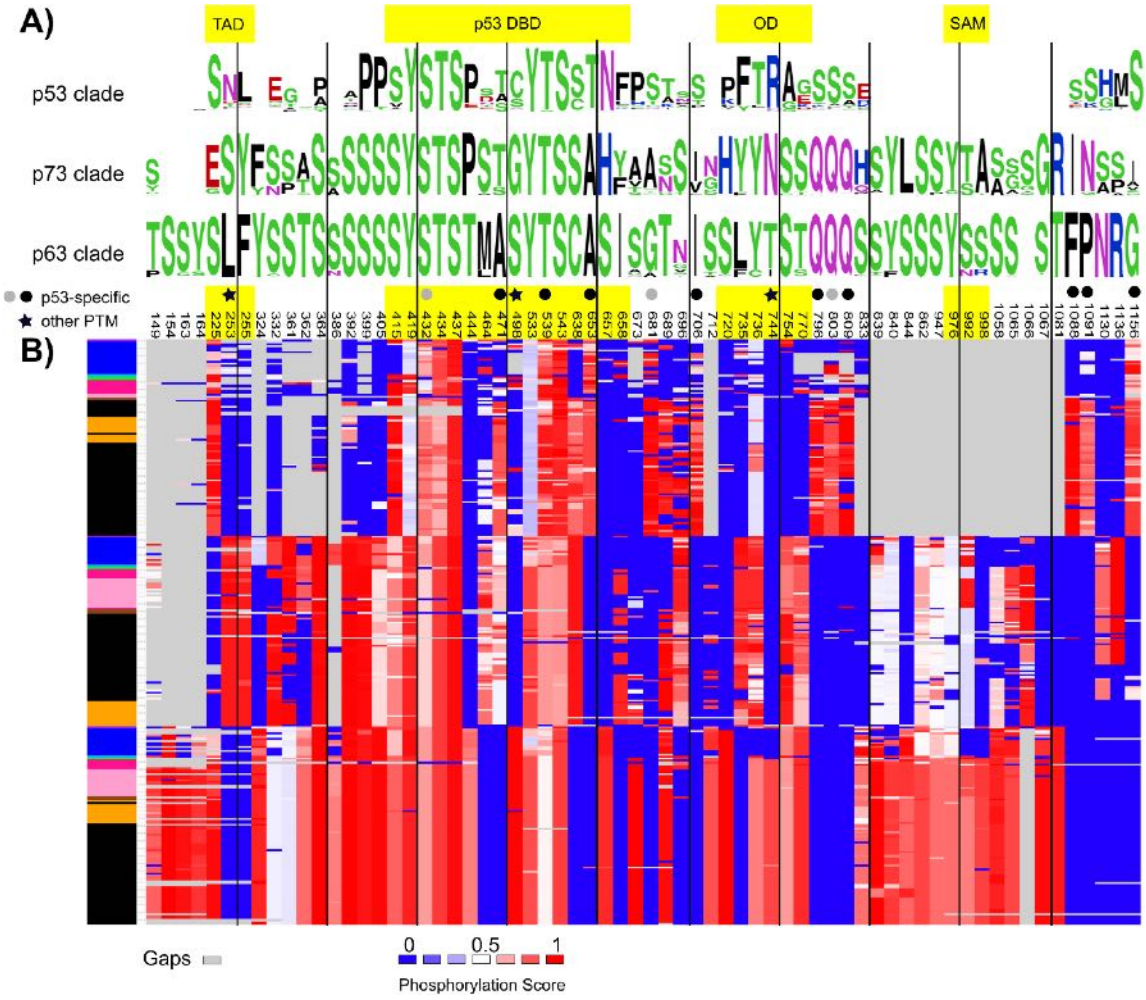


Figure 7. Shared and clade-specific predicted phosphorylation patterns. (A) WebLogos [35] per clade showing 66 alignment positions following a 50% majority rule of phosphorylation predictions based on a phosphorylation cut-off = 0.75 (NetPhos), gaps included. (B) Phosphorylation predictions mapped onto their alignment sites (numeration based on the full alignment), with scores ranging from 0 (blue) to 1 (red) with 0.5 as the midpoint (white). Gaps are shown in grey. The colored boxes on the left show the distribution of species sorted by the phylogenetic tree following the color scheme as in Fig 2. Shared and clade-specific phosphorylation sites are distributed along domains (yellow shaded areas) and linkers. Sites marked with a circle means p53 clade-specific (black, the phosphorylation site is experimentally validated in PhosphoSite; grey, an adjacent site is experimentally validated to be phosphorylated in PhosphoSite). Sites marked with a star are predicted to be phosphorylated in a p63 or p73 clade-specific manner while p53 has a different experimentally verified posttranslational modification [34].

DISCUSSION

Using linear sequence predictors, properties of structural disorder (IUPred), secondary structure (PSIPRED), and phosphorylation sites (NetPhos) have been inferred. It is important to remember that these are predictions and cannot be perfect given that they are (i) independently aiming to predict traits that may depend on each other, (ii) using only the linear sequence context without considering long-range sequence contacts, and (iii) based on experimental data that may not reflect the dynamic nature of a protein sequence, e.g. one PDB structure is merely a snapshot of a conformational ensemble [36]. The accuracy for PSIPRED is >80% compared to actual experimentally determined protein structures [37]. For disordered proteins, fewer proteins are experimentally determined to be disordered. For IUPred, comparing to IDEAL (a small database of disordered proteins [number of proteins = 207]) [38] the accuracy is approximately 85%, but comparing to DisProt (a slightly larger database of disordered proteins [number of proteins = 794]) [39] the accuracy is approximately 62% [40]. However, it has been found that IUPred is more accurate in predicting order vs. disorder for DisProt proteins if the cut-off is set to 0.4 instead of the intended 0.5 [39,41]. In a different study, IUPred predictions of 0.4 were frequently found for disordered residues in partially disordered proteins [42]. Thus, we used the 0.4 cut-off to infer order vs. disorder. The sensitivity reported for NetPhos predictions cover a range from 69–96% [25], partially due to the lack of insufficient data available to train phosphorylation predictors [43]. Still, these are all standard prediction methods, widely used in computational and molecular biology when experimental data is not available.

By comparing approximately 300 protein sequences from the vertebrate p53 family and an additional ~50 invertebrate p53 DBD domain sequences, we have investigated diverging properties from sequence to structure to regulation in the p53 family. From the invertebrate p53 DBD phylogeny, it appears that p53 DBD sequences primarily form clades based on the domain content of the full-length protein. If the p53 DBD containing proteins from Fig 1 are arranged by species in the order of taxonomy and with focus on their domain composition, a picture of the main evolutionary events of the p53 family emerges (Fig 8). As previously shown, a three domain p53 DBD containing protein is present in choanoflagellates [12]. The shared precursor of this protein and the very first metazoan p53 protein must have had at least three of the four domains found in present day vertebrate p53 family proteins. We observe proteins with all four domains in gastropods, hemichordates, and early chordates. Since these belong to Bilateria, it is clear that the bilaterian ancestor had all four domains. It should also be noted that other species not included here, such as the placozoan, *Trichoplax adhaerens*, have an MDM2 binding site [44]. Although Pfam does not classify this protein to have a TAD domain, the MDM2 binding site indicates that it does, or at least that it used to have a TAD domain. Thus, TAD predates the divergence of Bilateria and Placozoa. Further, TAD and the other non-p53 DBD domains, are frequently lost (Fig 8). In Ecdysozoa, some of these domain losses are due to actual sequence segment loss and others are due to the sequence signature being depleted. Altogether, this clearly suggests that early metazoan, and perhaps even choanoflagellates have p53 family proteins that diverged less than many of the ecdysozoan p53 family proteins that have lost most domains and frequently only consist of the p53 DBD itself. There may be other equally or more remote

p53 DBD proteins in other invertebrates, like e.g. CEP-1 in *Caenorhabditis elegans* [44]. Lineage-specific gene duplications are frequent in invertebrates, but a last common ancestor of all proteins in the vertebrate p53 family is shared with *B. floridae* (Figs 1 and 8).

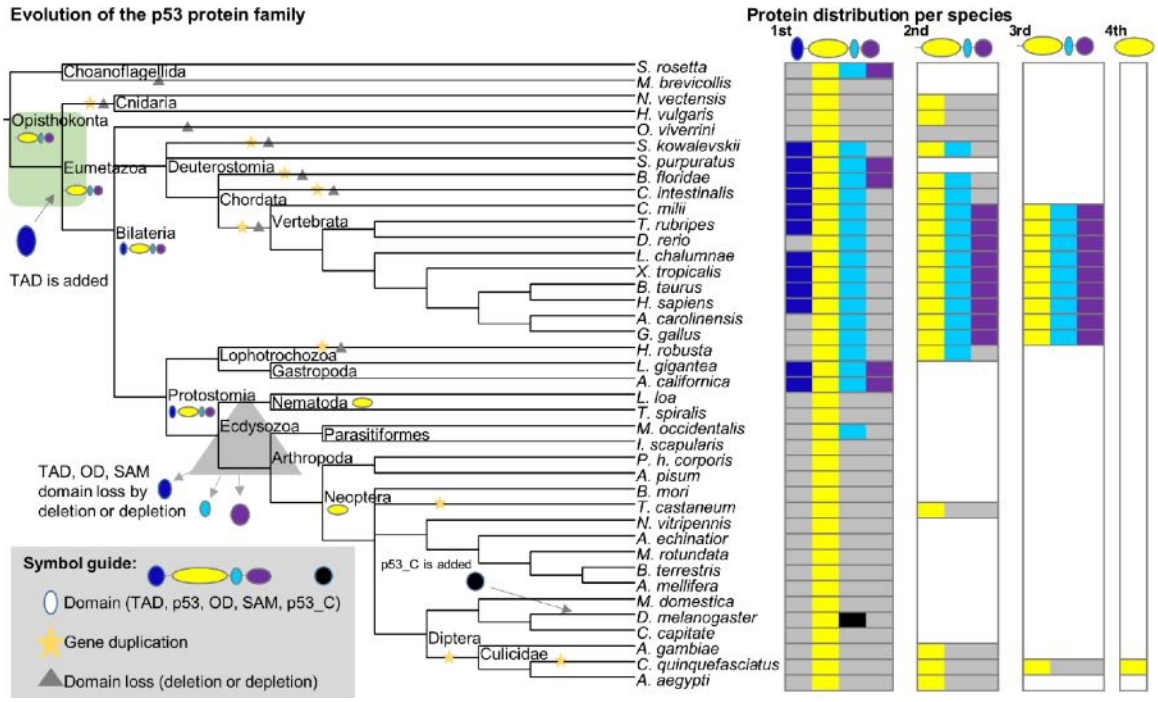


Figure 8. Major evolutionary events in the early p53 family. The sequences in Fig 1 are arranged by NIH Common tree taxonomy to show the evolutionary order of events (left). Branches with evidence of gene duplications are marked with a star. Branches with domain loss are marked with a triangle. Branches are not to scale. The protein distribution per species is shown (right). Presence of domains per protein are colored according to the color scheme for domains in Fig 1, with the addition that grey denotes missing domain and white denotes that no additional proteins were detected.

It is also clear that the p53 DBD is less structurally disordered in single domain invertebrate proteins. In vertebrates, the three paralogs p53, p63, and p73, are diverging at different rates: p63 is highly constrained while p53 is not. Ray-finned fish are demonstrating rapid lineage-specific diversification among all three paralogs. Although this study is mostly focused on the functional domains and their divergence, the inter-

domain linkers vary in length and in disorder/order and secondary structure composition. Linkers are not just flexible spacers but important for controlling the conformational ensemble [45]. The divergence in linker 1 between p53 and p63 and p73 is profound and suggests functional change. TAD is rapidly diverging amongst p53 in different vertebrates, and has already diverged beyond Pfam's domain detection ability in p63 and p73, even if some of TAD's ancestral functionality may have remained. For p53, MDM2 is a critical regulator [44]. When MDM2 binds to key residues F19, W23, and L26 in the human p53 TAD, it can further ubiquitinate p53 on Lys residues throughout the p53 protein marking it for proteasomal degradation (reviewed in [46]). p73 was found to bind MDM2 in the same region, and although binding of MDM2 prevented p73's transcriptional activity, it was not ubiquitinated [47]. Recently, a study found p73 to be ubiquitinated by MDM2 but p73 was not degraded [48]. For p63, the MDM2 interaction is much weaker [49]. Thus, the differential disorder among paralogs in the MDM2 binding region amongst these paralogs suggest and support divergent functional dependence on MDM2. The MDM2 binding region is frequently lost among ray-finned fish p53 proteins, and the TAD Pfam domain in general is not detected in p63 and p73, although the homologous sequence may still be there. Still, remnants of the MDM2 binding site have been found in p53 from early metazoans [44] further supporting that this is an ancestral function.

Additional indications of clade-specific functional divergence emerges from the patterns of phosphorylation. Indeed, functionally relevant phosphorylation transitions were identified and present an interesting picture of how these three paralogs have diversified in the realm of phospho-signaling. Since phosphorylation is performed by

different kinases in response to various signals these seemingly small changes can allow proteins to specialize after a gene duplication. Of the three members of the p53 family, p63 is more constrained to diverge in sequence. The p63 clade has 28 clade-specific predicted phosphorylation sites above 50% conservation, compared to 12 in the p53 and p73 clades alike, suggesting that phosphorylation sites may be lost on the latter two. For at least two of the clade-specific phosphorylation sites in p63, p53 is also post-translationally modified but with a different modification, further enforcing distinct regulatory mechanisms acting on these three paralogs.

Null-mice of p63 or p73 are severely impacted and do not live long while null-mice of p53 survive to adulthood [50], suggesting that p53 is dispensable but p63 and p73 are not. The functional overlap between p53, p63, and p73 is hampered by the complexity of the protein family [51]. p53 presents lineage-specific changes and one can speculate that perhaps p53 is rapidly diversifying in a near-neutral mode due to remaining functional redundancy with p63 and p73.

p53 is a puzzling protein, known to cause and prevent cancer, prevalently mutated, in cancerous and non-cancerous cells [52]. Regardless, it cannot be expected to be functionally conserved amongst invertebrates with different domain composition, nor amongst vertebrates. Interpreting the p53 family from a molecular evolution perspective, p63 and p73 are predominantly responsible for most of the ancient function as indicated by stronger conservation of sequence and the properties here analyzed, but even in these two clades divergent regions suggest ongoing functional divergence. From a systems biology perspective, diversification in phosphorylation alters the signaling and interaction networks in which these different proteins act. From a biophysical perspective, non-

conserved disorder has been interpreted as non-functional [53]. Here non-conserved disorder is found in the DNA binding region of p53, while p63 and p73 both have conserved disorder. This suggests functional diversification of the DNA binding region in p53 causing some species to become ordered in this region, perhaps bypassing a regulatory step of DNA binding regulation. Thus, an alternative interpretation for non-conserved disorder (rapid DOT) could be that it enables or disables fine-tuned signaling, rapid rewiring, or gain and loss of function(s) in a lineage-specific manner, offering a boost to biological diversity. In p53, all scenarios are possible. In the ray-finned fish clade, p53 is rapidly changing compared to the rest of the vertebrates, with many changes from fish to fish in the TAD domain. Also p63 and p73 have ray-finned fish specific changes. For p73, the p53 DBD is more ordered in ray-finned fish than in the rest of the p73 clade. For p63, the OD domain is more ordered in ray-finned fish than in the rest of the p63 clade. Co-evolution is probable. p53 from the lobe-finned fish, *L. chalumnae* has remarkably little disorder. Was the last common ancestor of p53 more ordered than it is today or has disorder been lost in *L. chalumnae*? Given that the rest of the vertebrate p53 family is more disordered, it is likely that *L. chalumae* has lost disorder. Without disorder, is *L. chalumnae*'s p53 still a multifunctional protein, and does it hold clues to critical, non-redundant, p53 functions, perhaps with simplified regulation? Further, what is happening to p53 in the avian genomes?

p53 is an innovative protein. While many proteins simply lose function in response to a mutation, many cancer causing mutations in p53 are thought to cause a gain-of-function [54], perhaps through mutation-driven conformational selection effects [55]. If a mutation can cause a gain-of-function, can controlled experimental conditions

with wt-p53 *in vitro* have similar effects? Some gain-of-function effects seen in cancer mutants may shift the conformational ensemble since structurally disordered proteins are prone to adapt to their environmental conditions (mutation-driven conformational selection [55] vs. allosteric conformational selection [56]). Both of these effects could impact p53 *in vitro*, *in vivo*, and in a tumor cell context.

Inevitably, ongoing functional divergence is present in the p53 family, and especially in the p53 clade. The Guardian of the Genome gives the impression of still exploring its function and does not fit the picture of a resilient Guardian. Perhaps, a more appropriate way to refer to p53 is as a Gambler of the Genome?

METHODS

Sequence retrieval

Three datasets were constructed: (i) the p53 protein family at the whole protein level in vertebrates, (ii) the p53 protein family at the nucleotide level in vertebrates, and (iii) the p53 protein family at the DNA-binding domain level in a representative set of vertebrate sequences and non-vertebrates. For (i), NCBI BLAST [57] was performed using the blastp algorithm with the human p53 protein sequence (NCBI reference sequence: NP_000537.3) against vertebrates in the RefSeq database [58]. To minimize redundancy, only the longest sequence from the same gene was chosen as the representative. Partial or much longer proteins were removed to maintain a high quality multiple sequence alignment. In some instances, sequences from key species missing in the RefSeq database were instead identified by BLAST against the nr database. For (ii), the corresponding nucleotide sequences for the amino acids sequences in (i) were

retrieved from NCBI. For the final dataset (iii), NCBI BLAST was performed using the blastp algorithm with the human p53 protein DNA-binding domain excluding vertebrates in the RefSeq database to get non-vertebrate sequences. Partial proteins with an incomplete p53 DBD were removed to maintain a high quality multiple sequence alignment. To minimize redundancy and to reduce the dataset a selection of sequences was used.

For major vertebrate taxonomic groups, a representative organism with sequence information for all three paralogs in the p53 protein family was selected from (i). Vertebrate organisms included in (iii) were: *Homo sapiens*, *Bos taurus*, *Gallus gallus*, *Anolis carolinensis*, *Xenopus tropicalis*, *Latimeria chalumnae*, *Takifugu rubripes*, *Danio rerio*, and *Callorhinchus milii*. Sequence identifiers for all vertebrate sequences are given in Appendix 12 and protein identifiers are included in the phylogenetic trees that show sequence names.

Phylogenetic reconstruction

Sequences for datasets (i) and (iii) were aligned with MAFFT v7.123–1 [59] using the L-INS-i algorithm for a maximum of 1000 iterations. Sequences in dataset (ii) were aligned using TranslatorX [60] to map corresponding codons to the amino acid alignment from (i). Phylogenetic trees for all datasets were constructed using MrBayes v3.2.2 [61]. For protein based phylogenies [(i) and (iii)], Bayesian MCMC analysis was performed using a mixed amino acid model with gamma distributed rate variation among sites. The nucleotide based phylogeny (ii) was estimated with Bayesian MCMC analysis using a GTR model with gamma distributed rate variation among sites. For all trees, MrBayes ran two simultaneous analyses (each with four chains: three heated and one

cold) for 15 million generations with a sampling frequency of 100 generations. For dataset (i) the best tree was constructed with TBR branch swaps, while for (ii) and (iii) the best trees were constructed with TBR branch swaps disabled. The final average standard deviation of the split frequencies were 0.0060 (max. s.d. 0.051) for dataset (i), 0.0053 (max. s.d. 0.092) for dataset (ii), and 0.0023 (max. s.d. 0.016) for dataset (iii). Consensus trees were built with the default burn-in phase (discarding the first 25% of trees) using the 50% majority rule. The tree from the third dataset was rooted on a branch containing *Monosiga brevicollis* and *Salpingoeca rosetta*. The resulting topology was used to guide rooting the trees from the first two datasets by rooting on the branch containing both p63 and p73 clades and selecting the p53 clade as the outgroup.

Sequence-based predictions

To assess the characterization of the structural properties of the proteins included in our phylogenies, the amino acid sequence of each protein (unaligned sequence) was used as input for different sequence-based predictors in order to predict structural disorder, secondary structure, phosphorylation sites and domain regions. Thereafter, for each prediction method, the predicted value for each residue in each protein sequence was mapped onto its corresponding site in the multiple sequence alignment. This resulted in three matrices for (i) structural disorder prediction, (ii) secondary structure predictions, and (iii) predicted phosphorylation sites. For (i) and (iii), the data predicted was continuous. For (ii), the data had three non-numerical categories. In order to analyze the transitions between order and disorder, between the presence of secondary structure elements and loops, and for presence or absence of phosphorylation sites, all matrices were represented as binary phyletic patterns (as described below). The phyletic patterns

were individually analyzed in their phylogenetic context and transition rates were calculated.

Structural disorder prediction

Structural disorder was predicted using IUPred [15,62] version 1.0 selecting the option for long disordered regions. IUPred was specifically developed for predicting disorder in intrinsically unfolded proteins using estimated energy content. The IUPred prediction generates a disorder propensity for each residue in the protein. The disorder propensities range from 0 (indicating no propensity of being disordered) to 1 (indicating strong propensity of being disordered). While the method was developed to have scores above 0.5 indicating disorder, a cut-off of 0.4 was later demonstrated to give higher accuracy when predicting disorder on proteins from the experimentally verified DisProt database [41,42]. The continuous disorder predictions were mapped onto the multiple sequence alignment, and visualized in a heat map format using iTOL [63]. Further, all sites with IUPred prediction values <0.4 were assigned order and all sites ≥ 0.4 were assigned disorder. This binary matrix was used as a phyletic pattern for analyzing the evolutionary dynamics of structural disorder to order transitions (DOT).

Secondary structure prediction

Secondary structure was predicted using PSIPRED [24,64] version 3.4 with default parameters and the nr database (version March.30.2014), filtered to avoid low complexity regions, coiled-coil regions and transmembrane regions, was selected to generate a sequence profile per protein. PSIPRED is a neural network program which performs an analysis on the sequence profiles obtained from PsiBlast (Position Specific Iterated–BLAST version 2.2.26, blastpgp) [65] converting them to secondary structure

propensities. The three states of secondary structure propensity (alpha helix, beta strand, and loop) were visualized in a heat map. The PSIPRED predictions were converted into binary data: alpha helix/beta strand residues were set to 1 and loop residues were set to 0. This binary matrix was used as a phyletic pattern for analyzing the evolutionary dynamics of secondary structure to loop transitions (SLT).

Phosphorylation site prediction

Phosphorylation sites for Serine, Threonine and Tyrosine residues were predicted using NetPhos [25] version 3.1, an artificial neural network method. Similar to the other predictions, two states were defined: sites with values <0.75 were assigned not phosphorylated or 0 and all sites ≥0.75 were assigned as sites predicted to be phosphorylated or 1. Sites predicted to be phosphorylated were visualized in a heat map. The resulting binary matrix was used for analyzing the evolutionary dynamics of phosphorylation transitions (PT).

Protein domain prediction

Protein domains were predicted based on Pfam [66] version 27 by aligning each sequence to their stored Hidden Markov Model (HMM) profiles using the available batch search scripts. Sites in domains with significant bit scores based on pre-defined gathering thresholds, predicted to be part of a Pfam_A domain (based on the envelope coordinates), were visualized in a heat map.

Evolutionary dynamics of sequence data

Rate4Site [67] was used to estimate the amino acid substitution rates (SEQ) by an empirical Bayesian principle under the Jones, Taylor, and Thornton [68] amino acids substitution model (JTT) using a prior gamma distribution including 16 discrete

categories. Rate4Site estimates the site specific rates considering the topology and branch lengths of the phylogenetic tree. The branch lengths were not optimized as the input trees were obtained by Bayesian inference. Normalized evolutionary rates in Rate4Site are Z-scores, scaled such that the average across all sites is equal to zero and standard deviation is equal to 1. This means that sites showing a normalized evolutionary rate <0 are evolving slower than average, and those with a rate >0 are evolving faster than average.

Evolutionary dynamics of predicted data

To study the gain/loss transitions of structural properties in related proteins along their evolutionary history, a protocol that includes the estimation of evolutionary rates per site based on the phylogenetic trees and the binary matrices generated was adopted. GLOOME software [27] was used to study the evolutionary dynamics of structural disorder (DOT rate; disorder-order transitions), secondary structures (SLT rate; secondary structure-loop transitions), and phosphorylation sites (PT rate; phosphorylation transitions). GLOOME was originally developed to study the gain/loss events across phylogenies. Here GLOOME was applied to analyze trends in binary presence (1) and absence (0) patterns in predicted protein sequence features (disorder vs. no-disorder, secondary structure vs. no secondary structure, phosphorylation site vs. no phosphorylation site) with default equal substitution rates for transitions within the same state (0 to 0, 1 to 1) and default equal rates for substitutions from one state to another (0 to 1, 1 to 0) and a rate distribution of 6 gamma categories. The outputs include the evolutionary rates per alignment site normalized as a Z-score (the same way as for the sequence data in Rate4Site). Lastly, for each of the evolutionary rates calculated (SEQ,

DOT, SLT and PT) for the family and the individual clades, we further analyzed those aligned sites with less than 10% of gaps per alignment position.

Non-parametric tests

Inference methods implemented in R statistical software [69] were used for testing if differences in means across groups are statistically significant (p-value <0.05). According to the Shapiro-Wilk test [70] normality could not be assumed and non-parametric tests were performed. For three or more samples the Kruskal-Wallis test [71] was applied, while the pairwise testing involved the use of the Mann-Whitney U test with Bonferroni correction [72,73].

3D mapping of structural disorder conservation and disorder-to-order transition rates

Conservation, here defined as the fraction of disorder per site from the binary matrices (gaps included), was calculated for the p53 family and the individual clades. Site specific rates and conservation of disorder were mapped onto representative PDB structures for the different domains. Appendix 10 shows the details of the mapped regions. Figures were generated using PyMOL [74].

ACKNOWLEDGMENTS

The authors would like to acknowledge the Instructional & Research Computing Center (IRCC) at Florida International University for providing HPC computing resources that have contributed to the research results reported within this paper, web: http://ircc.fiu.edu.

LITERATURE CITED

1. Gunasekaran K, Ma B, Nussinov R (2004) Is allostery an intrinsic property of all dynamic proteins? Proteins 57: 433–443. pmid:15382234

2. Tokuriki N, Tawfik DS (2009) Protein dynamism and evolvability. Science 324: 203–207. pmid:19359577

3. Siltberg-Liberles J (2011) Evolution of structurally disordered proteins promotes neostructuralization. Mol Biol Evol 28: 59–62. pmid:21037204

4. Vogelstein B, Lane D, Levine AJ (2000) Surfing the p53 network. Nature 408: 307–310. pmid:11099028

5. Collavin L, Lunardi A, Del Sal G (2010) p53-family proteins and their regulators: hubs and spokes in tumor suppression. Cell Death Differ 17: 901–911. pmid:20379196

6. Uversky VN, Oldfield CJ, Dunker AK (2008) Intrinsically disordered proteins in human diseases: Introducing the D(2) concept. 37: 215–246.

7. Oldfield CJ, Meng J, Yang JY, Yang MQ, Uversky VN, Dunker AK (2008) Flexible nets: disorder and induced fit in the associations of p53 and 14-3-3 with their partners. BMC Genomics 9 Suppl 1: S1. pmid:18366598

8. Yu Q, Ye W, Wang W, Chen H-F (2013) Global conformational selection and local induced fit for the recognition between intrinsic disordered p53 and CBP. PLoS One 8: e59627. pmid:23555731

9. Xue B, Brown CJ, Dunker AK, Uversky VN (2013) Intrinsically disordered regions of p53 family are highly diversified in evolution. Biochim Biophys Acta 1834: 725–738. pmid:23352836

10. Wagner A (2008) Neutralism and selectionism: a network-based reconciliation. Nat Rev Genet 9: 965–974. pmid:18957969

11. Assis R, Kondrashov AS (2014) Conserved proteins are fragile. Mol Biol Evol 31: 419–424. pmid:24202613

12. Nedelcu AM, Tan C (2007) Early diversification and complex evolutionary history of the p53 tumor suppressor gene family. Dev Genes Evol. 217: 801–806. pmid:17924139

13. Putnam NH, Butts T, Ferrier DEK, Furlong RF, Hellsten U, Kawashima T, et al. (2008) The amphioxus genome and the evolution of the chordate karyotype. Nature 453: 1064–1071. pmid:18563158

14. Berná L, Alvarez-Valin F (2014) Evolutionary genomics of fast evolving tunicates. Genome Biol Evol 6: 1724–1738. pmid:25008364

15. Dosztányi Z, Csizmok V, Tompa P, Simon I (2005) IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics 21: 3433–3434. pmid:15955779

16. Lane DP, Madhumalar A, Lee AP, Tay B-H, Verma C, Brenner S, et al. (2011) Conservation of all three p53 family members and Mdm2 and Mdm4 in the cartilaginous fish. Cell Cycle 10: 4272–4279. pmid:22107961

17. Soussi T, Bègue A, Kress M, Stehelin D, May P (1988) Nucleotide sequence of a cDNA encoding the chicken p53 nuclear oncoprotein. Nucleic Acids Res 16: 11383. pmid:3060861

18. International Chicken Genome Sequencing Consortium. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. (2004) Nature 432: 695–716. pmid:15592404

19. Belyi VA, Ak P, Markert E, Wang H, Hu W, Puzio-Kuter A, et al. (2010) The origins and evolution of the p53 family of genes. Cold Spring Harb Perspect Biol 2: a001198. pmid:20516129

20. Cai Q, Qian X, Lang Y, Luo Y, Xu J, Pan S, et al. (2013) Genome sequence of ground tit *Pseudopodoces humilis* and its adaptation to high altitude. Genome Biol 14: R29. pmid:23537097

21. Bornberg-Bauer E, Beaussart F, Kummerfeld SK, Teichmann SA, Weiner J (2005) The evolution of domain arrangements in proteins and interaction networks. Cell Mol Life Sci 62: 435–445. pmid:15719170

22. Venkatesh B, Lee AP, Ravi V, Maurya AK, Lian MM, Swann JB, et al. (2014) Elephant shark genome provides unique insights into gnathostome evolution. Nature 505: 174–179. pmid:24402279

23. Brandt T, Kaar JL, Fersht AR, Veprintsev DB (2012) Stability of p53 homologs. PLoS One 7: e47889. pmid:23112865

24. McGuffin LJ, Bryson K, Jones DT (2000) The PSIPRED protein structure prediction server. Bioinformatics 16: 404–405. pmid:10869041

25. Blom N, Gammeltoft S, Brunak S (1999) Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. J Mol Biol 294: 1351–1362. pmid:10600390

26. Pupko T, Bell RE, Mayrose I, Glaser F, Ben-Tal N (2002) Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. Bioinformatics 18: S71–S77. pmid:12169533

27. Cohen O, Ashkenazy H, Belinky F, Huchon D, Pupko T (2010) GLOOME: gain loss mapping engine. Bioinformatics 26: 2914–2915. pmid:20876605

28. Joerger AC, Wilcken R, Andreeva A (2014) Tracing the evolution of the p53 tetramerization domain. Structure 22: 1301–1310. pmid:25185827

29. Popowicz GM, Czarna A, Holak TA (2008) Structure of the human Mdmx protein bound to the p53 tumor suppressor transactivation domain. Cell Cycle 7: 2441–2443. pmid:18677113

30. Chen Y, Zhang X, Dantas Machado AC, Ding Y, Chen Z, Qin PZ, et al. (2013) Structure of p53 binding to the BAX response element reveals DNA unwinding and compression to accommodate base-pair insertion. Nucleic Acids Res 41: 8368–8376. pmid:23836939

31. Clore GM, Omichinski JG, Sakaguchi K, Zambrano N, Sakamoto H, Apella E, et al. (1994) High-resolution structure of the oligomerization domain of p53 by multidimensional NMR. Science 265: 386–391. pmid:8023159

32. Hornbeck PV, Zhang B, Murray B, Kornhauser JM, Latham V, Skrzypek E (2015) PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. Nucleic Acids Res 43: D512-D520. pmid:25514926

33. Lee J-C, Kang S-U, Jeon Y, Park JW, You J-S, Ha S-W, et al. (2012) Protein L-isoaspartyl methyltransferase regulates p53 activity. Nat Commun 3: 927. pmid:22735455

34. Nguyen T-A, Menendez D, Resnick MA, Anderson CW. (2014) Mutant TP53 posttranslational modifications: challenges and opportunities. Hum Mutat 35: 738–755. pmid:24395704

35. Crooks GE, Hon G, Chandonia J-M, Brenner SE (2004) WebLogo: a sequence logo generator. Genome Res 14: 1188–1190. pmid:15173120

36. Slabinski L, Jaroszewski L, Rodrigues APC, Rychlewski L, Wilson IA, Lesley SA, et al. (2007) The challenge of protein structure determination--lessons from structural genomics. Protein Sci 16: 2472–2482. pmid:17962404

37. Bryson K, McGuffin LJ, Marsden RL, Ward JJ, Sodhi JS, Jones DT (2005) Protein structure prediction servers at University College London. Nucleic Acids Res 33: W36–W38. pmid:15980489

38. Fukuchi S, Sakamoto S, Nobe Y, Murakami SD, Amemiya T, Hosoda K, et al. (2012) IDEAL: Intrinsically Disordered proteins with Extensive Annotations and Literature. Nucleic Acids Res 40: D507–D511. pmid:22067451

39. Sickmeier M, Hamilton JA, LeGall T, Vacic V, Cortese MS, Tantos A, et al. (2007) DisProt: the Database of Disordered Proteins. Nucleic Acids Res 35: D786–D793. pmid:17145717

40. Di Domenico T, Walsh I, Tosatto SCE (2013) Analysis and consensus of currently available intrinsic protein disorder annotation sources in the MobiDB database. BMC Bioinformatics 14 Suppl 7: S3. pmid:23815411

41. Fuxreiter M, Tompa P, Simon I (2007) Local structural disorder imparts plasticity on linear motifs. Bioinformatics 23: 950–956. pmid:17387114

42. Xue B, Oldfield CJ, Dunker AK, Uversky VN (2009) CDF it all: consensus prediction of intrinsically disordered proteins based on various cumulative distribution functions. FEBS Lett 583: 1469–1474. pmid:19351533

43. Trost B, Kusalik A (2011) Computational prediction of eukaryotic phosphorylation sites. Bioinformatics 27: 2927–2935. pmid:21926126

44. Lane DP, Cheok CF, Brown C, Madhumalar A, Ghadessy FJ, Verma C (2010) Mdm2 and p53 are highly conserved from placozoans to man. Cell Cycle.

45. Ma B, Tsai C-J, Haliloğlu T, Nussinov R (2011) Dynamic allostery: linkers are not merely flexible. Structure 19: 907–917. pmid:21742258

46. Chao CC-K (2015) Mechanisms of p53 degradation. Clin Chim Acta 438: 139–147. pmid:25172038

47. Bálint E, Bates S, Vousden K (1999) Mdm2 binds p73 alpha without targeting degradation. Oncogene 18: 3923–3929. pmid:10435614

48. Wu H, Leng RP (2015). MDM2 mediates p73 ubiquitination: a new molecular mechanism for suppression of p73 function. Oncotarget 6: 21479–21492. pmid:26025930

49. Zdzalik M, Pustelny K, Kedracka-Krok S, Huben K, Pecak A, Wladyka B, et al. (2014) Interaction of regulators Mdm2 and Mdmx with transcription factors p53, p63 and p73. Cell Cycle 9: 4584–4591.

50. Stiewe T (2007) The p53 family in differentiation and tumorigenesis. Nat Rev Cancer 7: 165–816. pmid:17332460

51. Costanzo A, Pediconi N, Narcisi A, Guerrieri F, Belloni L, Fausti F, et al. (2014) TP63 and TP73 in cancer, an unresolved “family” puzzle of complexity, redundancy and hierarchy. FEBS Lett 588: 2590–2599. pmid:24983500

52. Martincorena I, Roshan A, Gerstung M, Ellis P, Van Loo P, McLaren S, et al. (2015) High burden and pervasive positive selection of somatic mutations in normal human skin. Science. 348: 880–886.

53. van der Lee R, Buljan M, Lang B, Weatheritt RJ, Daughdrill GW, Dunker AK, et al. (2014) Classification of intrinsically disordered regions and proteins. Chem Rev 114: 6589–6631. pmid:24773235

54. Brosh R, Rotter V (2009) When mutants gain new powers: news from the mutant p53 field. Nat Rev Cancer 9: 701–713. pmid:19693097

55. Siltberg-Liberles J, Grahnen JA, Liberles DA (2011) The Evolution of Protein Structures and Structural Ensembles Under Functional Constraint. Genes (Basel). 2: 748–762.

56. Nussinov R, Ma B, Tsai C-J (2014) Multiple conformational selection and induced fit events take place in allosteric propagation. Biophys Chem 186: 22–30. pmid:2429303

57. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 245: 403–410.

58. Pruitt KD, Tatusova T, Maglott DR (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 33: D501–D504. pmid:15608248

59. Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 30: 3059–3066. pmid:12136088

60. Abascal F, Zardoya R, Telford MJ (2010) TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. Nucleic Acids Res 38: W7–W13. pmid:20435676

61. Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Hohna S, et al. (2012) MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Softw Syst Evol 61: 539–542.

62. Dosztányi Z, Csizmók V, Tompa P, Simon I. (2005) The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. J Mol Biol 347: 827–839. pmid:15769473

63. Letunic I, Bork P (2007) Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics 23: 127–8. pmid:17050570

64. Jones DT (1999). Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 292: 195–202. pmid:10493868

65. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389–402. pmid:9254696

66. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. (2014) Pfam: the protein families database. Nucleic Acids Res 42: D222–D230. pmid:24288371

67. Mayrose I, Graur D, Ben-Tal N, Pupko T (2004) Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. Mol Biol Evol 21: 1781–1791. pmid:15201400

68. Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 8: 275–282. pmid:1633570

69. The R Core Team (2012) R: A language and environment for statistical computing. Vienna, Austria: R foundation for Statistical Computing, Vienna, Austria

70. Shapiro SS, Wilk MB (1995) An Analysis of Variance Test for Normality (Complete Samples). Biometrika 52: 591–611.

71. Kruskal WH, Wallis WA (1952) Use of Ranks in One-Criterion Variance Analysis. J Am Stat Assoc 47: 583–621.

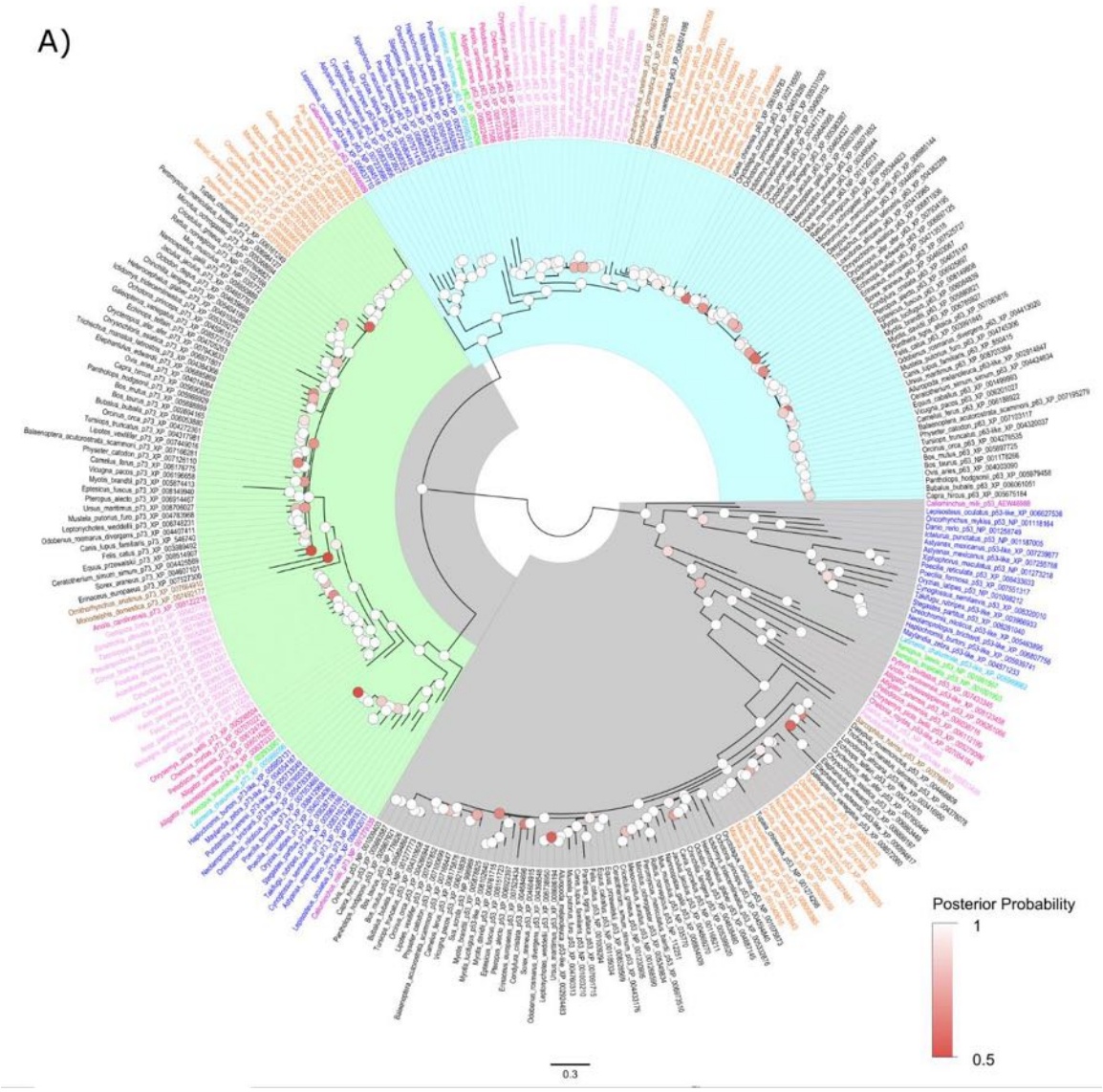
72. Mann HB, Whitney DR (1947) On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. Ann Math Stat 18: 50–60.

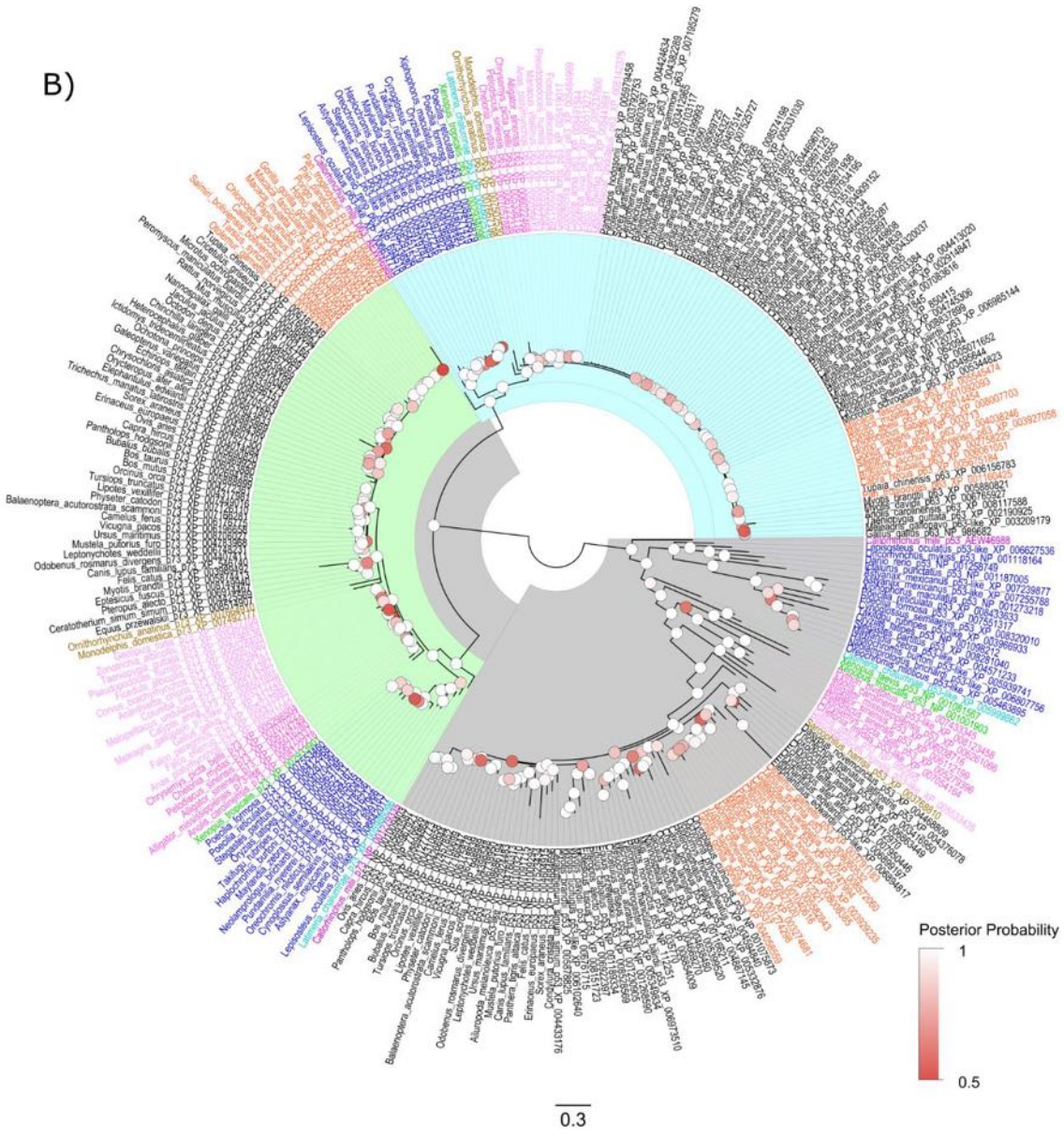
73. Dunn OJ (1961) Multiple Comparisons among Means. J Am Stat Assoc ;56: 52–64.

74. Schrödinger LLC (2014) The PyMOL molecular graphics system, Version 1.7.2. https://sourceforge.net/projects/pymol/

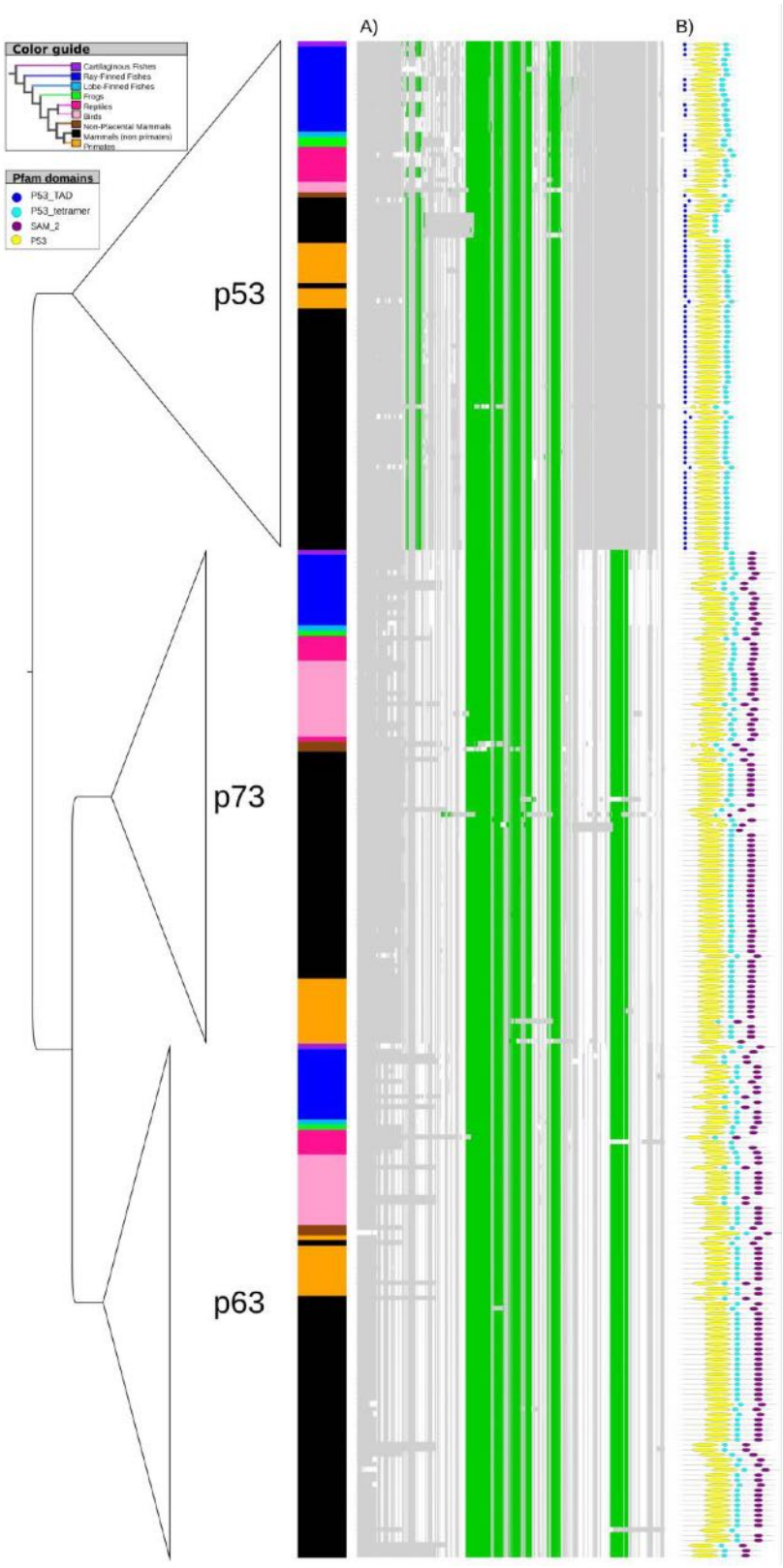
Appendices

Appendix 1. p53 domain phylogeny for Metazoa and Choanoflagellates. Overview of the p53 family phylogeny including 74 representative species across Metazoa and Choanoflagellates, built based on their p53 DBD domains. Support values at the nodes indicate posterior probabilities. Nodes with posterior probability < 0.5 are unresolved.

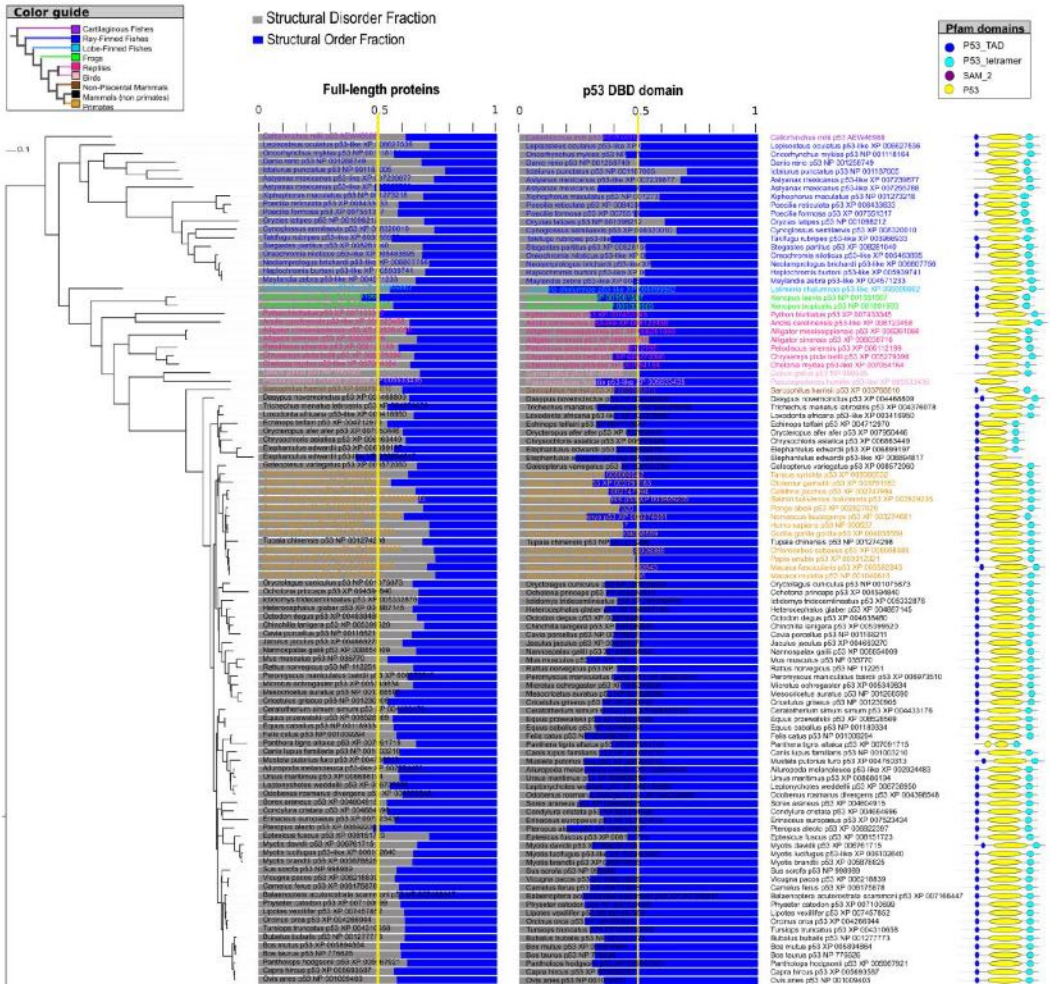

A)
Posterior Probability
1
0.5
0.3

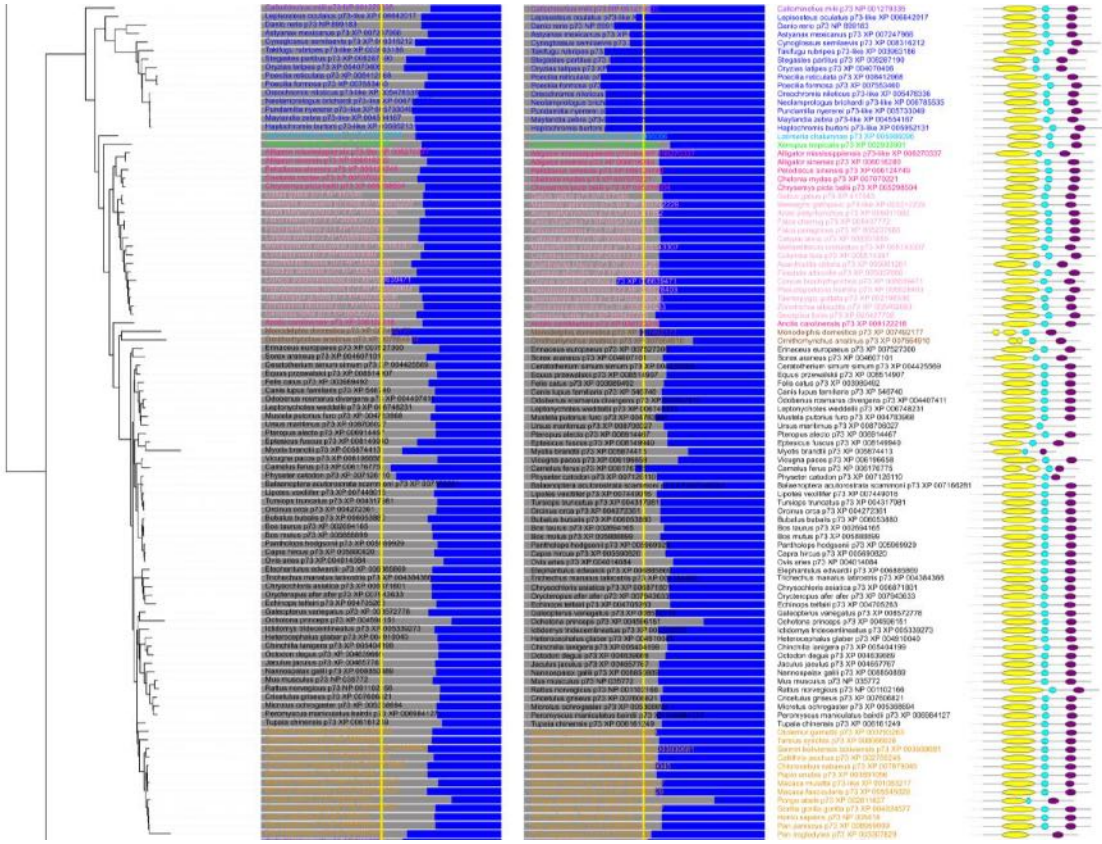


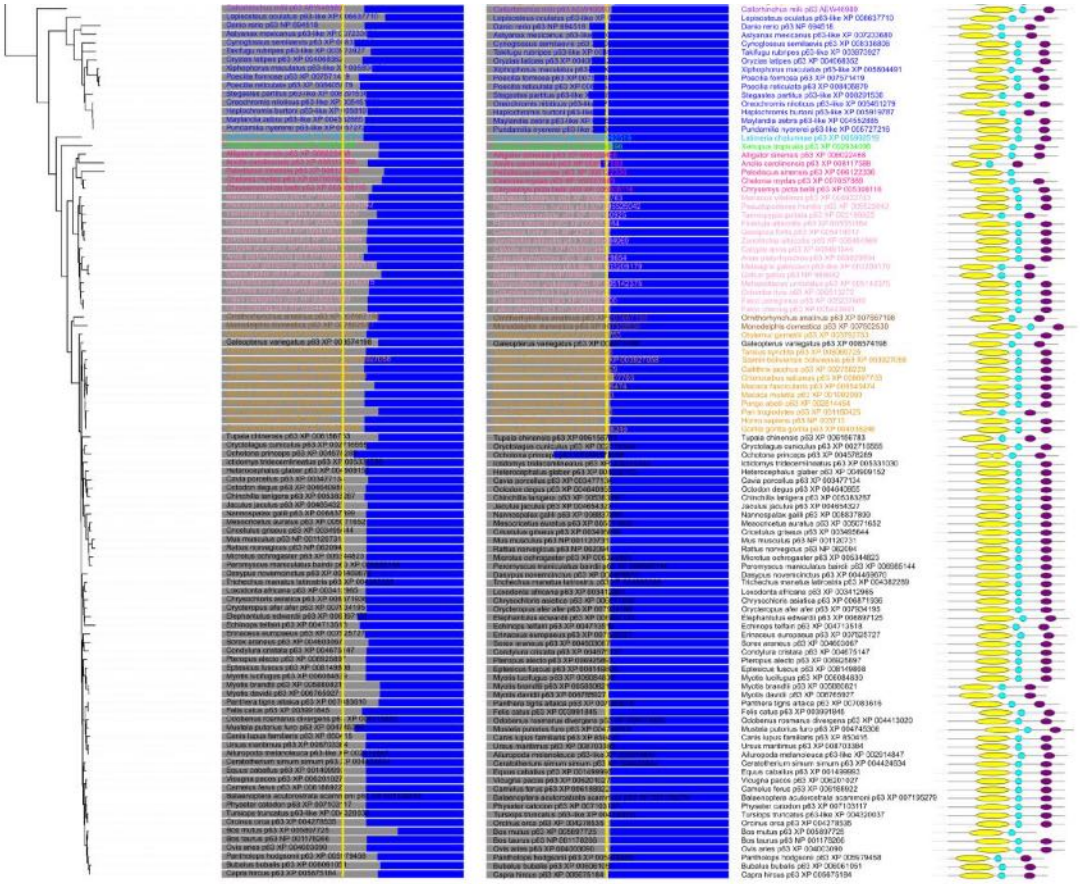
Appendix 2. p53 phylogenies for 301 vertebrate proteins. (A) Circular representations of p53 DNA-based phylogeny and (B) its corresponding full-protein-based phylogeny. These consensus trees were obtained with MrBayes 3.2.2 after sampling trees for 15 million generations with the default burn-in phase (discarding the first 25% of trees) and using the 50% majority rule. Node circles show posterior probabilities ranging from 0.5 in red to 1 in white. Here proteins were colored by clade (p53 in grey, p63 in blue and p73 in green) with tip labels following the color guide from Fig 2. Figure generated with FigTree (http://tree.bio.ed.ac.uk/software/figtree/).


Color guide
Cartilaginous Fishes
Ray-Finned Fishes
Lobe-Finned Fishes
Frogs
Reptiles
Birds
Non-Placental Mammals
Mammals (non primates)
Primates
Pfam domains
P53_TAD
P53_tetramer
SAM_2
P53
p53
p73
p63
A)
B)

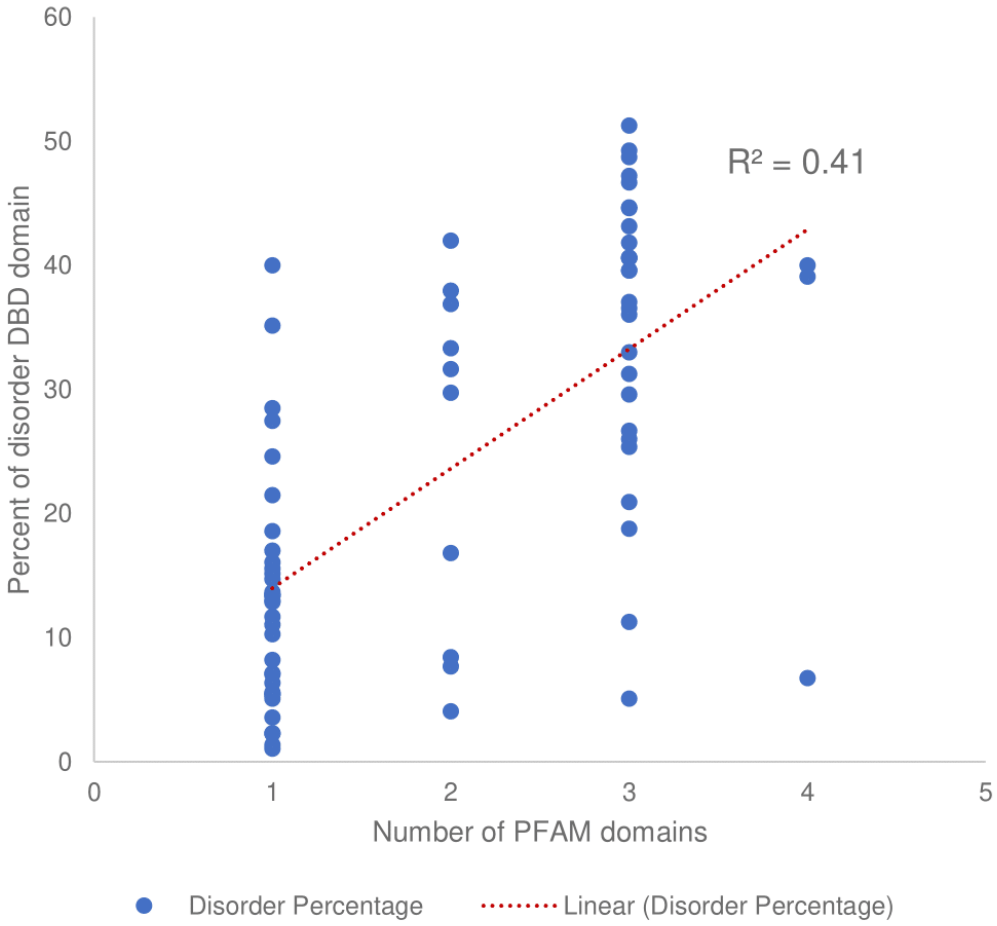
Appendix 3. Domain composition in vertebrate proteins. (A) Heat map showing Pfam domain predictions per protein into their corresponding multiple sequence alignment sites (rows show protein hits; columns show alignment positions; sites that belong to Pfam_A domains are colored, green; linkers between domains, white; gaps in the alignment, grey), all in the context of the p53 DNA-based phylogeny with tip labels colored according to the color guide. (B) In addition, individual domain architectures (labeled and colored as shown in Pfam domains box) were also included to highlight their actual lengths enforcing missing or broken domains. Figure generated with iTOL [63].


Color guide
Cartilaginous Fishes
Ray-Finned Fishes
Lobe-Finned Fishes
Frogs
Reptiles
Birds
Non-Placental Mammals
Mammals (non primates)
Primates
Structural Disorder Fraction
Structural Order Fraction
Full-length proteins
p53 DBD domain
Pfam domains
P53_TAD
P53_tetramer
SAM_2
P53

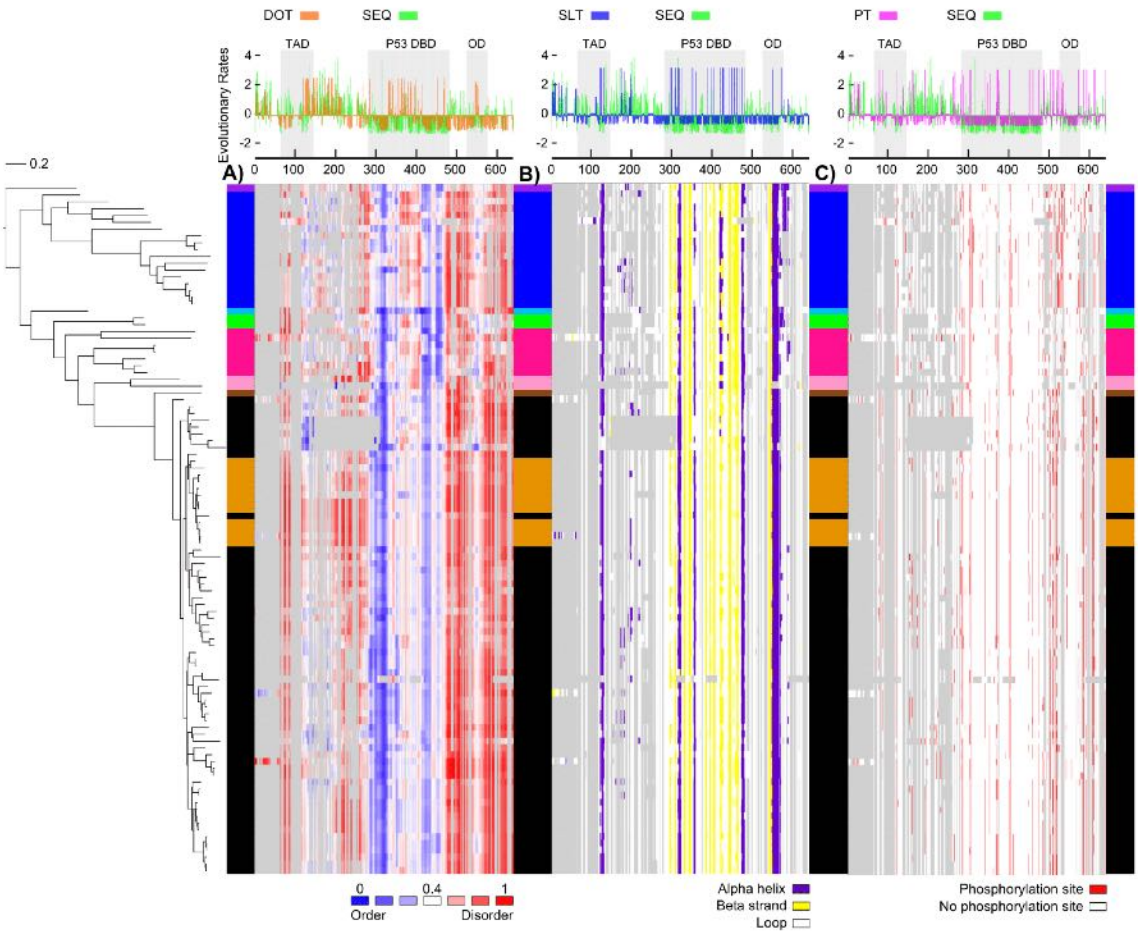




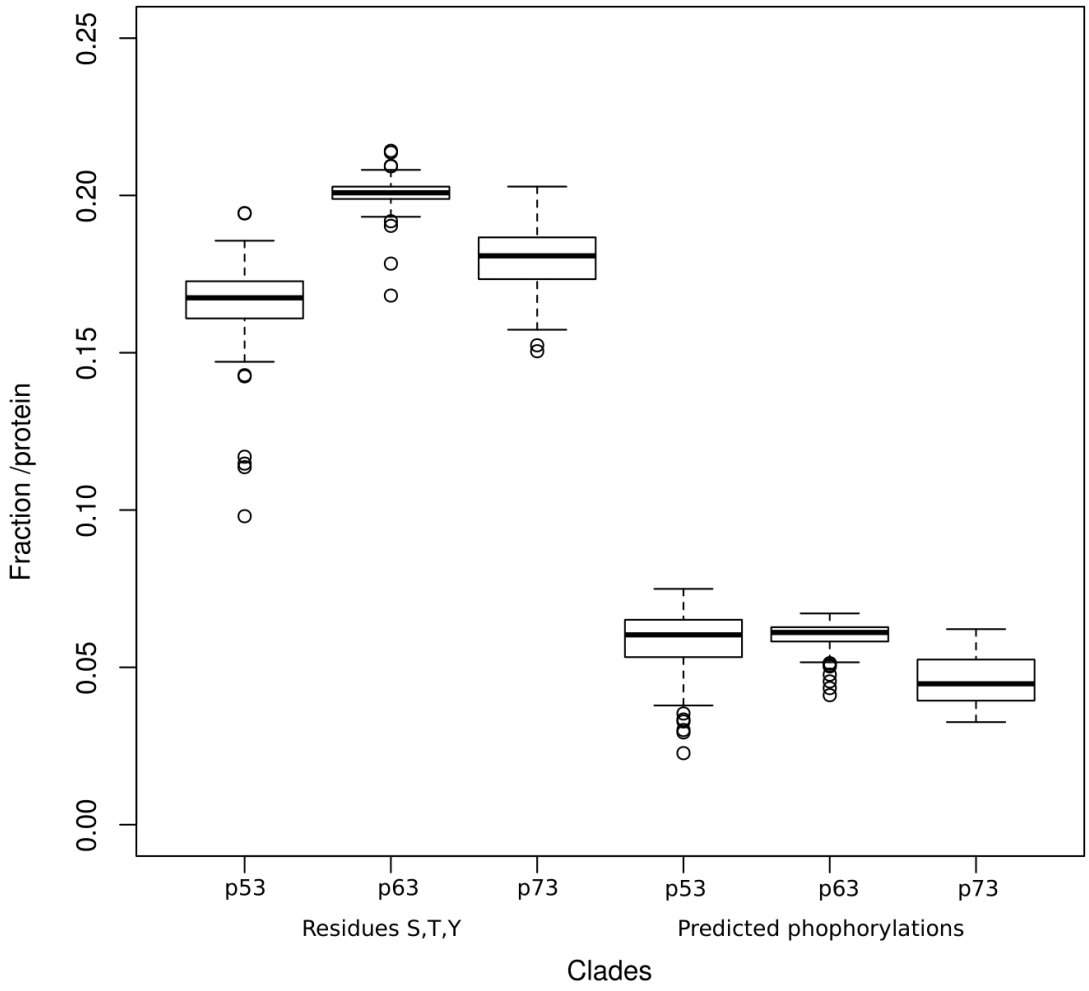
Appendix 4. Structural disorder fractions in vertebrate proteins. Distribution of structural disorder (grey) and order (blue) in full-length proteins and in p53 DBD domains sorted by p53 DNA-based phylogenetic tree with tip labels following the color guide. Furthermore, individual domain architectures (labeled and colored as shown in Pfam domains box) were also included. Figure generated with iTOL [63].



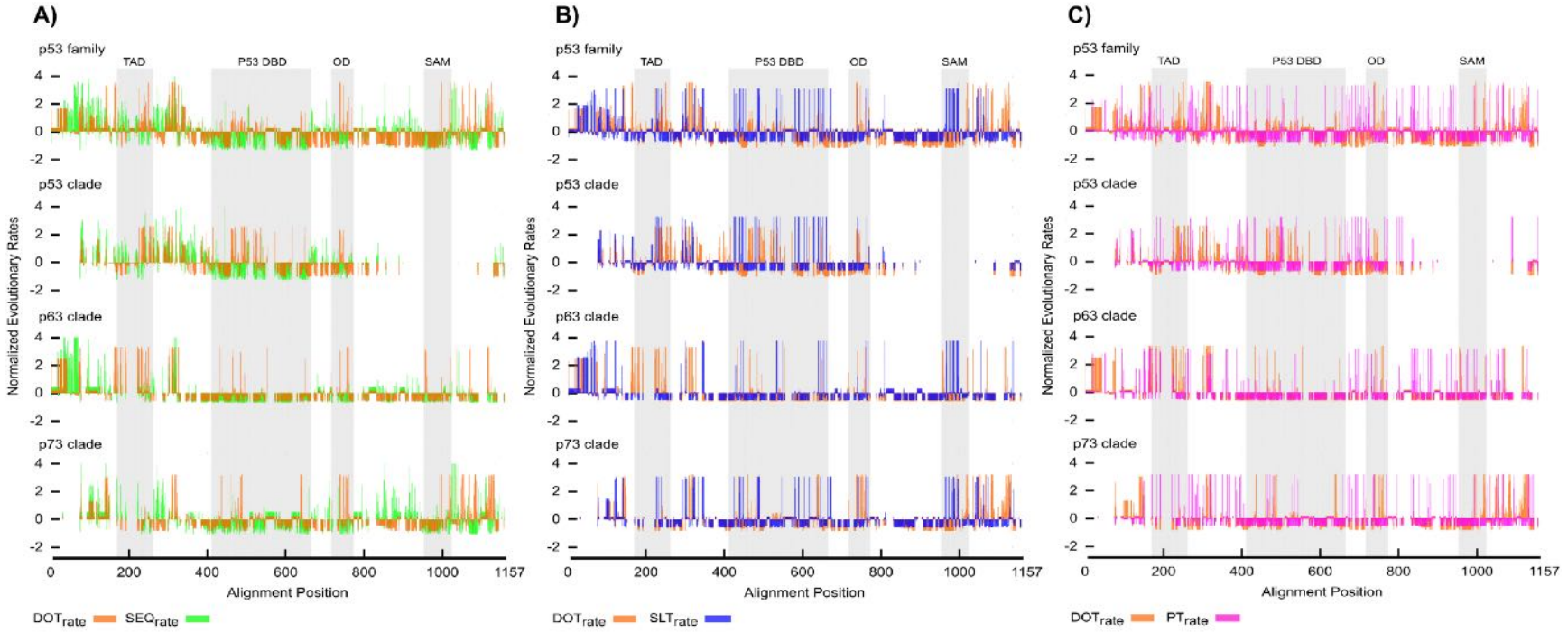
Appendix 5. p53 DBD structural disorder content increases with the number of domains. Scatter plot of the p53 DBD structural disorder percentage vs. the number of Pfam domains per protein from 74 hits, including invertebrates and vertebrates proteins. There is a positive correlation between these two variables (Pearson correlation coefficient R = 0.64, R^2 = 0.41, and p-value < 0.05, concluding that linear correlation different to 0 is statistically significant).



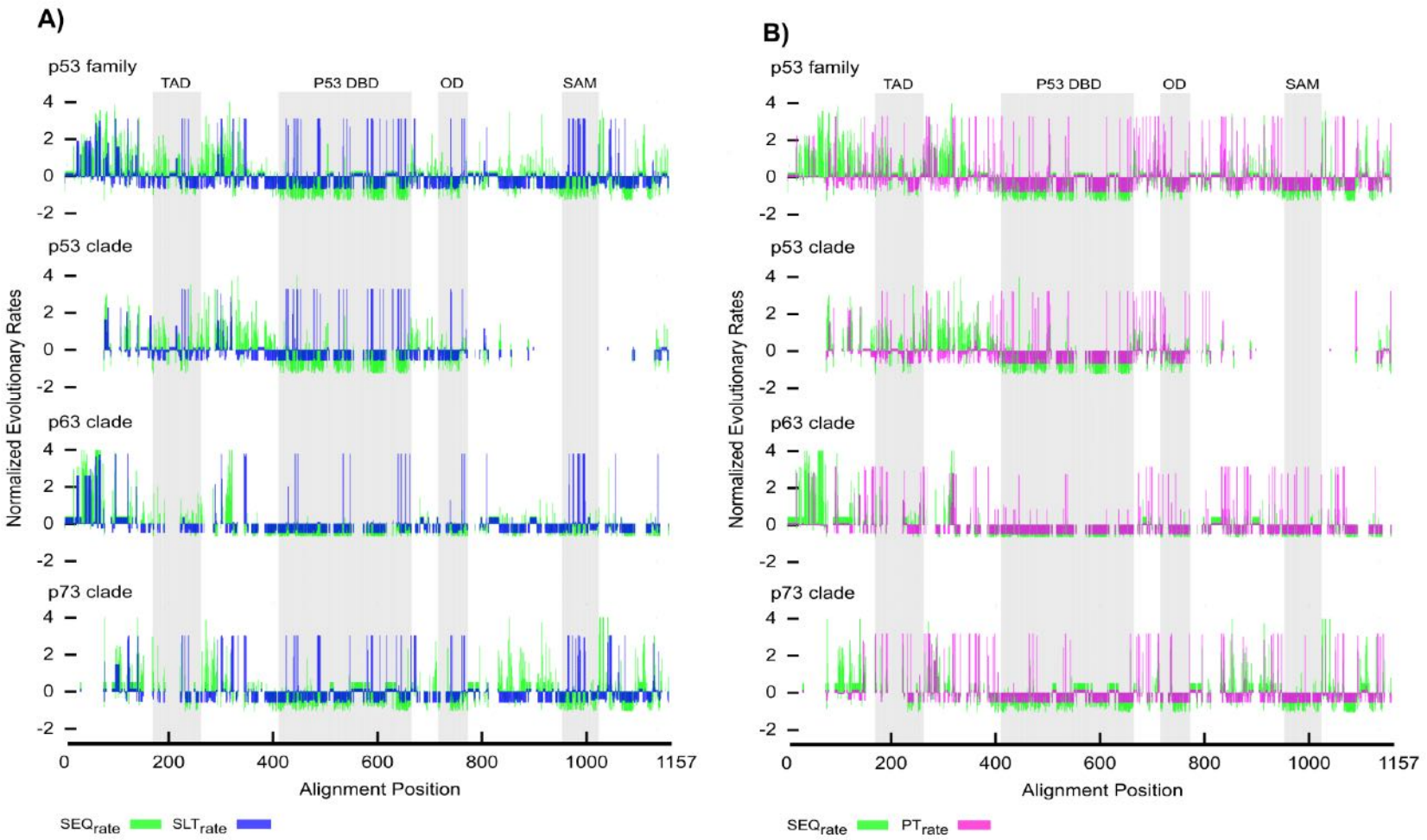
Appendix 6. p53 clade in detail: graphical representation of sequence-based predictions. Heat map for structural traits plotted in the order of the p53 DNA-based phylogenetic tree context, showing p53 protein names as boxes colored according to the color guide in Fig 2. These heat maps are showing sequence-based predictions mapped to their corresponding residue sites in the multiple sequence alignment, after removing empty columns (i.e. columns fully gapped in the p53 clade) for this subset: (A) continuous structural disorder propensities by IUPred [15,62] with a color gradient from blue to white to red mirroring the disorder propensity gradient from low (blue) to high (red), with white being the boundary between order and disorder (remaining alignment gaps are colored in grey). (B) secondary structure predictions by PSIPRED [24,64] displaying 3 states loop (white), alpha helix (purple) and beta strand (yellow), and C) sites predicted to be phosphorylated by NetPhos [25] using a 0.75 cut-off (red). On top of these heat maps, normalized evolutionary rates per site are shown for amino acid sequence (SEQ) in green [26] vs. binary traits [27] of disorder-order transitions (DOT) in orange (upper left), secondary structure elements—loop transitions (SLT) in blue (upper center), and phosphorylation transitions (PT) in pink (upper right). All evolutionary rates were normalized with a mean of zero and standard deviation of 1: negative rates for slow evolving sites and positive rates for fast evolving sites. Grey shaded areas delimitate Pfam domain regions.



Appendix 7. Differential distribution of protein phosphorylations per clade.
Boxplots showing the fractions of serine, threonine and tyrosine residues per protein per clade compared to the fractions of sites predicted to be phosphorylated by NetPhos [25] using a 0.75 cut-off. Significance analysis was carried out using non-parametric tests (Kruskal Wallis test for the comparison of 3 or more samples and Mann-Whitney U test with Bonferroni correction for the pairwise analysis). Differences in means are statistically significant (p-values << 0.05), except for the p53-p63 comparison of predicted phosphorylation factions (p-value = 1).



Appendix 8. Comparison of SEQ, SLT, and PT with DOT rates. Combined profiles of normalized evolutionary rates per aligned site for family and clades (vertebrates set) comparing disorder-order transitions (DOT) with (A) amino acid substitutions (SEQ), (B) secondary structure elements-loop transitions (SLT), and (C) phosphorylation transitions (PT). Grey shaded areas delimitate Pfam domain regions.


A)
p53 family
TAD
P53 DBD
OD
SAM
p53 clade
p63 clade
p73 clade
Normalized Evolutionary Rates
Alignment Position
SEQrate
SLTrate
B)
PTrate

Appendix 9. Comparison of SLT and PT with SEQ rates. Combined profiles of normalized evolutionary rates per aligned site for family and clades (vertebrates set) comparing amino acid substitutions (SEQ) with (A) secondary structure elements-loop transitions (SLT) and (B) phosphorylation transitions (PT). Grey shaded areas delimitate Pfam [66] domain regions.

Appendix 10. PDB files and regions used for mapping DOT and disorder conservation into a structural context

Pfam Domain	PDB code	Protein template	PDB numeration	Sequence positions (human)	Fragment length	Alignment sites
TAD_p53	3dac	p53	17-28	17-28	12	231-242
TAD_p63	3dac	p53	17-28	53-64	12	231-242
TAD_p73	3dac	p53	17-28	13-24	12	231-242
DBD_p53	4hje	p53	94-291	94-291	198	410-660
DBD_p63	4hje	p53	94-291	162-361	198	410-660
DBD_p73	4hje	p53	94-291	112-311	198	410-660
OD_p53	1olg	p53	319-360	319-360	42	717-768
OD_p63	4a9z	p63	358-404	397-443	47	731-802
OD_p73	4a9z	p63	358-404	351-392	47	731-802

Appendix 11. Shared and clade specific predicted phosphorylation patterns. Alignment sites following a 50% majority rule of sequences with phosphorylation predictions based on NetPhos phosphorylation prediction score cut-off=0.75 (gaps included). Information displayed per clade (specific) and per family (shared). Shaded areas correspond to the majority rule (phosphorylation predicted for more than 50% of taxa per clade for the family). Corresponding positions in the canonical human proteins (P53_human NP_000537.3, P63_human NP_003713.3, and P73_human NP_005418.1) are shown.

p53 clade-specific sites	Domain	p53_human	p63_human	p73_human
432	DBD	116	184	134
471	DBD	155	223	173
539	DBD	211	281	231
653	DBD	284	354	304
681	Linker2	303	373	325
706	Linker2	315	385	337
796	Linker3	366	439	392
803	Linker3	367	444	393
809	Linker3	371	448	397
1088	Cter	376	652	596
1091	Cter	378	654	598
1156	Cter	392	679	635
p63 clade-specific sites	Domain	p53_human	p63_human	p73_human
149	Nter	-	25	-
154	Nter	-	30	-
163	Nter	-	34	-
164	Nter	-	35	-
324	Linker1	-	111	65
362	Linker1	-	131	82
386	Linker1	-	142	-
405	Linker1	92	160	110
444	DBD	128	196	146
498	DBD	182	250	200
654	DBD	288	358	308
637	Linker2	297	367	319
712	Linker2	-	389	434
720	OD	322	395	349
744	OD	337	410	364
833	Linker3	-	452	401
839	Linker3	-	458	406
840	Linker3	-	459	407
844	Linker3	-	463	411
862	Linker3	-	477	426

947	Linker3	-	535	479
976	SAM	-	564	508
992	SAM	-	580	524
998	SAM	-	586	530
1058	Cter	-	627	570
1065	Cter	-	630	573
1067	Cter	-	631	575
1081	Cter	-	645	589
p73 clade-specific sites	Domain	p53_human	p63_human	p73_human
253	TAD	30	74	26
255	TAD	32	76	28
332	Linker1	56	119	70
361	Linker1	-	130	81
464	DBD	148	216	166
658	DBD	289	359	309
696	Linker2	311	381	333
735	OD	328	401	355
754	OD	347	420	374
1066	Cter	-	-	574
1130	Cter	380	670	621
1136	Cter	384	674	625
Overlapping sites across paralogs	Domain	p53_human	p63_human	p73_human
255	TAD	15	51	11
364	Linker1	70	132	83
392	Linker1	81	147	97
399	Linker1	87	154	104
415	DBD	99	167	117
419	DBD	103	171	121
434	DBD	118	186	136
437	DBD	121	189	139
533	DBD	205	275	225
543	DBD	215	285	235
638	DBD	269	339	289
689	Linker2	304	374	326
736	OD	329	402	356
770	OD	361	343	388

Appendix 12. Accession numbers for the vertebrate datasets (i) and (ii)

Separate Excel sheet

CHAPTER III

EXPLORING EVOLUTIONARY CONSTRAINTS IN THE PROTEOMES OF ZIKA, DENGUE, AND OTHER FLAVIVIRUSES TO FIND FITNESS-CRITICAL SITES

ABSTRACT

Dengue virus (DENV) challenges vaccine design due to antibody dependent enhancement (ADE) and evidence suggests that Zika virus (ZIKV) experiences ADE with DENV and West Nile virus (WNV) antibodies. Thus, multiple flaviviruses must be considered when developing novel therapies against ZIKV. We analyzed 42 flavivirus polyproteins in their evolutionary context to identify motifs conserved in sequence with low real-time and evolutionary conformational flexibility, thought to be fitness-critical sites. We also analyzed evolutionary rate-shifts between clades for insight on vector specificity. For mosquito-borne flaviviruses, two conserved motifs were identified within the RNA-dependent RNA polymerase (RdRP), critical for flavivirus genome replication. Clade-specific motifs were identified for the ZIKV+DENV and WNV clades, many of which were also in RdRP. Six sites in motifs for WNV experienced significant evolutionary rate-shifts, suggesting their importance for functional divergence. Overall, some of these motifs are prime candidates as broadly neutralizing antiviral drug targets across different mosquito-borne flaviviruses.

INTRODUCTION

The 2015-2016 Zika virus (ZIKV) epidemic in the Americas was caused by a ZIKV outbreak in Brazil in late 2014 and has resulted in over 90 countries and territories being reported as at risk for ZIKV infection (CDC 2018). Although most ZIKV cases are asymptomatic, symptomatic ZIKV cases are associated with increased risk for neurological complications such as Guillain-Barré syndrome (Lessler et al. 2016). In addition, current estimates indicate that 30% of pregnant women symptomatic for ZIKV

will experience ZIKV-associated adverse fetal outcomes, such as microcephaly, compared to only 7% of pregnant women with an asymptomatic ZIKV infection (Lessler et al. 2016).

The presence of other flaviviruses, such as the closely related Dengue virus (DENV), may contribute to the severity and enhancement of ZIKV infection through interference with antibodies against DENV which can lead to both neutralization (Barba-Spaeth et al. 2016) and enhancement (Dejnirattisai et al. 2016; Priyamvada et al. 2016) of the infection. Antibody-dependent enhancement (ADE) between DENV serotypes has long challenged DENV vaccine design (Heinz and Stiasny 2012). ADE has been observed between DENV and ZIKV in instances where DENV is the first infection (Dejnirattisai et al. 2016) and where ZIKV is the first infection (Stettler et al. 2016). Contrasting results have been found regarding ADE activity between ZIKV and West Nile virus (WNV), another member of the flavivirus family. Using a mouse model, one study found that WNV antibodies enhance ZIKV infection (Bardina et al. 2017), while another study found that ZIKV antibodies confer protection against WNV (Vázquez-Calvo et al. 2017). The discrepancy in these results may be related to antibody titer, as it has been observed in DENV infections that antibody titer is correlated to ADE. Low antibody titer does not sufficiently enhance infection while a high antibody tier is able to effectively neutralize infection (Katzelnick et al. 2017). While there are several ZIKV vaccines in clinical trial phases (reviewed by (Makhluf et al. 2018)), vaccine-enhanced DENV disease has also been observed (Hadinegoro et al. 2015) meaning vaccination could actually be counterproductive (Ferguson et al. 2016).

Conformational flexibility in DENV's Envelope protein can alter antibody binding affinity and efficacy among DENV serotypes (Kuhn et al. 2015). Two multifunctional enzymes encoded by the flavivirus polyprotein that are frequent targets for developing antiviral drugs are NS3 and NS5 (Sampath and Padmanabhan 2009; Bollati et al. 2010). NS3 functions as a serine protease, a helicase, and an RTPase (Yon et al. 2005). NS5 functions primarily as a methyltransferase and an RNA-dependent RNA polymerase (Sampath and Padmanabhan 2009). NS5 is also known to inhibit interferon signaling, but although inhibition of interferon signaling is wide-spread across flaviviruses, the mechanism of inhibition is not conserved (Grant et al. 2016). Both NS3 and NS5 are known to interact with many different human proteins, but few NS3 and NS5 proteins from different flaviviruses interact with the same human proteins (Le Breton et al. 2011). NS3 and NS5 have also been found to have conformational flexibility (Assenberg et al. 2009; Bussetta and Choi 2012; Meng et al. 2015; Klema et al. 2016).

Protein conformational flexibility is determined by the properties of the amino acid residues in a protein and correlates strongly with an amino acid's propensity to be disordered (Ruvinsky et al. 2012) as well as its local and global interactions (Zhang et al. 2007). The equilibrium between different conformations can be altered in response to signals in the environment (Smock and Gierasch 2009). Many viral proteins rely on intrinsic disorder for their function. Envelope in DENV is known to undergo functionally important conformational transitions in response to changes in pH (Stiasny et al. 2011). A computational analysis found that Capsid, 2K, NS3, and NS5 in DENV are enriched in disordered regions predominantly predicted to be involved in protein-protein interactions (Meng et al. 2015). Another computational analysis of disorder in the ZIKV proteome

found that the Capsid, NS2B, NS3, NS4A, and NS5 proteins were enriched in intrinsically disordered regions (Giri et al. 2016). Further, a previous study of structural disorder across the individual protein families in the different flaviviruses found rapid evolutionary dynamics of structural disorder in several flavivirus protein families (Ortiz et al. 2013). On the more extreme end, ZIKV was found to be almost completely ordered in the Capsid protein, unlike its close relatives that were 20-30% disordered, but even moderate fluctuations of disorder can be important for functional diversity. Percent disorder in both Membrane and Capsid proteins of flaviviruses has been strongly correlated with virulence (Kian-Meng Goh et al. 2019). The relatively low level of disorder in the Capsid protein of ZIKV compared to other flaviviruses appears to account for its high fetal morbidity rates despite low mortality rates (Kian-Meng Goh et al. 2019). Intrinsic disorder in the Capsid protein of WNV has been implicated in RNA-binding and chaperone activity (Ivanyi-Nagy et al. 2008; Ivanyi-Nagy and Darlix 2010). Overall, the clades for ZIKV, DENV, and WNV have higher disorder content in their Envelope proteins than most other flaviviruses, which enables high conformational flexibility and can have important consequences for antibody binding. Altogether, this suggests that the development of vaccine and antivirals for ZIKV should consider multiple flavivirus taxa, especially the four different DENV serotypes and WNV, in their evolutionary context. It also suggests precaution with targeting conformationally flexible regions.

Both DENV and ZIKV are transmitted by *Aedes* spp. vectors, while WNV is transmitted by a *Culex* spp. vector. Despite that, some studies have shown that ZIKV is able to replicate in *Culex quinquefasciatus* (Guo et al. 2016; Guedes et al. 2017), suggesting that ZIKV may be transmissible by a *Culex* vector. Others assert that the link

between ZIKV and *Culex* needs more support (Roundy et al. 2017; van den Hurk et al. 2017). Further, newer studies have indicated that ZIKV is not able to infect *Culex tarsalis* (Dodson et al. 2018; Main et al. 2018) or *Culex quinquefasciatus* (Lourenço-de-Oliveira et al. 2018; Main et al. 2018). In phylogenetic analyses of individual proteins, ZIKV is sometimes recovered sister to WNV rather than DENV (Ortiz et al. 2013). These results, together with ZIKV's potential to experience ADE with WNV antibodies, raise questions about whether ZIKV could indeed be transmitted by *Culex* mosquitoes today or in the future. It is not unheard of for a virus to evolve to expand its vector association, as Chikungunya virus has been previously found to have expanded its vector specificity through just a single mutation (Tsetsarkin et al. 2007; de Lamballerie et al. 2008). Understanding vector specificity is important because it allows us to determine the potential geographic range for a disease.

Protein sequences diverge with time. The amino acid substitutions at each site in a protein are evolving at a rate that depends on the site's functional (selective) constraint. The presence of significant site-specific rate-shifts between clades in a phylogeny indicates differentiation of functional constraints (Gaucher et al. 2001; Penn et al. 2008). The mosquito-borne flaviviruses originated from a shared ancestor but have since diverged and adapted to different mosquito vectors. To explore if there has been a change in functional constraints for flaviviruses with *Aedes* and *Culex* vectors, respectively, detection of rate-shifting sites can bring further insights. Comparing the amino acid state in ZIKV to the flaviviruses with *Aedes* and *Culex* vector specificity at sites with changing functional constraints can help ascertain which sites may be important for the divergence between these subgroups and inform where ZIKV fits in.

Here, we set out to study evolutionary constraints of the flavivirus proteomes to identify fitness-critical sites that can act as target sites for broadly neutralizing antiviral drugs across flaviviruses. While recent work on intrinsically disordered proteins has shown that structural features conserved across the conformational ensemble of a protein can serve as potential drug targets (Chong et al. 2018), intrinsic disorder has typically posed a challenge for traditional structure-based drug design that relies on a well-defined three-dimensional structure (Cheng et al. 2006; Zhang and Lai 2011; Batool et al. 2019). Thus, we identify regions of the flavivirus proteome under high evolutionary constraint that are conserved in structure and lack conformational flexibility. These fitness-critical target sites can provide us with a faster, cheaper, more successful route towards antiviral drug development against ZIKV, DENV, and other current and emerging flaviviruses. We also identify sites implicated in the determination of vector specificity by analyzing site-specific evolutionary rate-shifts between flaviviruses with an *Aedes* or *Culex* vector.

METHODS

Sequence Retrieval

A dataset of flavivirus polyproteins was constructed by running NCBI BLAST (Altschul et al. 1990) with the Zika virus polyprotein (NCBI reference sequence: YP_002790881.1) against flaviviruses (taxid: 11051) in the RefSeq database (Pruitt et al. 2005). Flaviviruses with a canonical polyprotein representative were selected. For each sequence, domains within the polyprotein were predicted using Pfam v27.0 (Finn et al. 2014). Sequences for which many domains were not predicted by Pfam were removed to ensure conservation of domain composition across the dataset.

Phylogenetic Reconstruction

A multiple sequence alignment was generated using MAFFT v7.123b (Katoh et al. 2002), with the L-INS-I algorithm selected, and for a maximum of 1000 iterations. Following alignment, a phylogenetic tree was estimated using MrBayes v3.2.3 (Ronquist et al. 2012). Bayesian MCMC analysis was performed using a mixed-model for amino acid substitution and gamma distributed rate variation among sites. The program ran for 5,000,000 generations (average standard deviation of split frequencies = 0.000074) with a sampling frequency of 100 generations before building the 50% majority-rule consensus tree with the default burn-in phase to discard the first 25% of trees. The resulting tree was midpoint rooted.

Intrinsic Structural Disorder Prediction

Two predictors were used to infer intrinsic structural disorder based on the full-length polyprotein: IUPred v1.0 (Dosztányi et al. 2005b, a) and DISOPRED2 (Ward et al. 2004). For IUPred, the setting for long disordered regions was selected. Disorder propensity scores from IUPred follow a continuous range from 0 to 1, where 0 indicates a low propensity for structural disorder and 1 indicates a high propensity for structural disorder. While the cut-off value for distinguishing between order and disorder is typically 0.5, a cut-off value of 0.4 is used because it has been shown to have greater accuracy in predicting experimentally verified disorder (Fuxreiter et al. 2007). For DISOPRED2, disorder propensity scores range from 0 to 9, where 0 indicates low propensity for structural disorder and 9 indicates a high propensity for structural disorder. For both predictors, scores below the cut-off are assigned as ordered while scores at or

above the cut-off are assigned as disordered. IUPred and DISOPRED2 predictions were mapped back to their corresponding position on the multiple sequence alignment.

Secondary Structure Prediction

Two predictors were used for secondary structure prediction: PSIPRED v3.4 (Jones 1999; McGuffin et al. 2000) and JPred4 (Drozdetskiy et al. 2015). PSIPRED predictions were generated using default settings against the UniRef90 database (Suzek et al. 2015). PSIPRED predictions were generated using the full-length polyprotein as input, while JPred4 predictions were generated based on the individual proteins due to length restrictions imposed by the program. PSIPRED and JPred4 predictions for alpha helices, beta strands, and loops were mapped back to their corresponding position on the multiple sequence alignment.

Identification of Target Sites

Protein regions with five or more consecutive sites displaying 100% sequence conservation were identified for the full phylogeny or across clades. Identified motifs were then analyzed for 100% conservation in structural order as predicted by IUPred. Motifs conserved in sequence and structural order were further analyzed for 100% conservation in secondary structure element (alpha helix or beta strand) as predicted by PSIPRED (Appendix 1). Protein regions with 100% conservation in sequence, structural order, and secondary structure are henceforth referred to as target sites. Identified target sites were also analyzed for conserved order and conserved secondary structure element as predicted by DISOPRED2 and JPred4, respectively.

To check for solvent accessibility of identified target sites, surface exposed residues were found using the PyMOL script findSurfaceResidues.py (Vertrees 2019)

with a 2.5 Å^2 cut-off. For the sites in Envelope, ZIKV Envelope (PDB ID: 5GZN (Wang et al. 2016)) was used. For the sites in NS3, ZIKV NS3 (PDB ID: 5JWH (Cao et al. 2016)) was used. For the sites in RdRP, the ZIKV RdRP (PDB ID: 5TFR (Upadhay et al. 2017)) was used. Solvent accessibility could not be determined for the sites in NS4 due to a lack of experimentally verified structures.

Analysis of Zika, Dengue, and West Nile Virus Strains

For ZIKV, DENV, and WNV, additional datasets of viral strains were generated. Strain sequence data were retrieved from GenBank by searching for the virus of interest in the NCBI Protein database and filtering by organism (ZIKV, DENV, or WNV), source database (GenBank), and sequence length (length of the polyprotein ± 10 residues). The DENV viral strains were not separated by serotype. When filtering by length, the length of the following RefSeq polyproteins was used: YP_002790881.1 for ZIKV (3419 aa), NP_073286.1 (DENV4) for the DENV lower bound (3387 aa) and NP_059433.1 (DENV1) for the DENV upper bound (3392 aa), and YP_001527877.1 for WNV (3433 aa). After retrieving the viral strain sequences from GenBank, the three datasets were additionally filtered to remove sequences containing X characters. The final datasets were each aligned using Clustal Omega v1.2.1 (Sievers et al. 2011) with default settings before being analyzed for conservation of the identified target sites.

Evolutionary Rate Estimation

Common Core Multiple Sequence Alignment for Mosquito-Borne Flaviviruses

The sub-alignment for the mosquito-borne flaviviruses (MBFVs), including three taxa with no known vector that are sister to the Yellow Fever virus clade, was extracted from the alignment based on the full-length polyprotein. For the resulting sub-alignment

of 27 flaviviruses, all sites with a gap character were removed and we refer to the remaining sites in the alignment (3322 out of originally 3559) as common core sites. The sub-alignment of common core sites for the MBFVs was further divided into three clades of interest: (i) *Aedes*-outgroup clade, (ii) *Aedes* clade, and (iii) *Culex* clade. For each of these common core sub-alignments, site-specific amino acid evolutionary rates were estimated by Rate4Site (Pupko et al. 2002) using empirical Bayesian estimation under the JTT (Jones et al. 1992) model for amino acid substitution. Evolutionary rates estimated by Rate4Site were normalized as Z-scores to have the average rate across all sites be equal to 0 and the standard deviation be equal to 1. Sites with an evolutionary rate < 0 are therefore predicted to be evolving slower than average and those with an evolutionary rate > 0 are predicted to be evolving faster than average. Site rates were drawn from a 16-category gamma distribution, estimated separately for each sequence alignment. Branch lengths were not optimized as the input trees for the clade-specific rates were taken from the full phylogeny. In order to run Rate4Site, the phylogeny was reduced such that it included only the MBFVs. The reduced tree was then re-rooted on the branch that denotes the split between the majority of the *Aedes* clade and the *Culex* clade. This allows for all three previously mentioned clades of interest to be recovered, with each clade being used as the input tree for their respective rates.

The clade-specific common core sub-alignments were further divided into individual proteins based on the protein boundaries in the Zika virus polyprotein. The 2K protein was included with the NS4B protein due to its short length. Amino acid evolutionary rates per site were estimated by Rate4Site as previously described for each of the individual protein alignments corresponding to the clades of interest.

Rate-Shift Calculations

Site-specific rate-shifts were calculated for each pairwise comparison between the *Aedes*-outgroup, *Aedes*, and *Culex* clades. The clade-specific sequence rates based both on the full-length polyprotein and on each individual protein for the common core sites were used in the analysis. Rate4Site yields a normalized estimate of the conservation score at each site in the sub-alignment for the clade on which they are based, represented by a mean and standard deviation. By assuming that the distributions of these estimates are roughly Gaussian, they can be compared to estimates at homologous sites in the sub-alignments for other clades using a simple *t* test. The sample sizes (*N*) used to estimate the conservation scores are a function of both the number of sequences with residues (non-gap characters) at a site and a measure of sequence divergence based on their branch lengths in the phylogenetic tree, preventing us from estimating the true sample size. Thus, the analysis was performed on all sites with > 1 non-gap character (i.e., residue), and the number of non-gap characters in a clade was used as an estimate of sample size for that site. To calculate a *p* value based on the above *t*-statistic, we used the Satterthwaite approximation of degrees of freedom (Satterthwaite 1946) and the same estimate of *N* as before.

In this way, the statistical significance of rate-shifts between all homologous site pairs among all clade-specific sub-alignments may be evaluated. Importantly, because we are making $\binom{A}{2}$ pairwise comparisons across *n* alignment positions, the operational confidence limit (α) must be corrected accordingly: corrected $\alpha = \frac{\alpha}{\binom{A}{2}n}$. Here, A is the 3 clades of interest and n is 3322, the length of the common core alignment. Thus, based on an alpha of 0.05, the corrected alpha used here is 5.017×10^{-6}. To achieve higher

specificity, significant rate-shifting sites were filtered such that the absolute mean difference as a site is greater than the sum of the standard deviations at that site.

Site-specific rate-shifts were also calculated using DIVERGE 3.0 (Gu and Vander Velden 2002; Gu et al. 2013). A tree was provided of the MBFVs rooted as described above and the three clades of interest were selected. The alignment provided was for the common core sites of the full-length polyprotein. The Gu99 function (Gu 1999) for Type-I functional divergence was used and sites with a posterior probability greater than or equal to 0.5 were identified as experiencing a significant rate-shift.

Structural Alignment

A structural alignment was generated for the Zika virus RNA-dependent RNA polymerase domain (PDB id: 5U0C (Zhao et al. 2017)) of the NS5 protein with the Hepatitis C virus NS5B protein (functions as an RNA-dependent RNA polymerase) bound to the nucleoside analog inhibitor sofosbuvir (PDB id: 4WTG (Appleby et al. 2015)). Each of the 8 entities of 5U0C was aligned with the single entity of 4WTG with CATH-SSAP v0.16.2 (Taylor and Orengo 1989; Orengo and Taylor 1996). The SSAP alignment for entity 1 of 5U0C and 4WTG was used to superpose the structures using CATH-superpose v0.16.2 (Taylor and Orengo 1989).

Visualization

The phylogeny, multiple sequence alignment, prediction heatmaps, and evolutionary rates were visualized using the Python packages ETE3 (Huerta-Cepas et al. 2016) and Matplotlib (Hunter 2007) as previously implemented (Rahaman and Siltberg-Liberles 2016). 3D protein structures were visualized using PyMOL (Schrödinger 2014).

RESULTS

Polyprotein and Phylogeny

A phylogeny was reconstructed for 42 flaviviruses based on their full-length polyprotein. The polyprotein can be divided into 3 structural and 7 non-structural proteins (Fig. 1, Table 1). The structural proteins include Capsid (C), pre-membrane (prM), and Envelope (E). The non-structural proteins include NS1, NS2A, NS2B, NS3, NS4A, NS4B, and NS5. An additional peptide, 2K, is located between NS4A and NS4B (not pictured). Many of the proteins are composed of multiple domains. The prM protein has the propeptide (pr) and glycoprotein M (M) domains. Envelope has the glycoprotein central (Domain I, DI) and dimerization (Domain II, DII) domains (shown together), as well as an immunoglobulin-like (Domain III, DIII) domain, NS3 contains a protease (NS3Pro) domain, followed by the DEAD domain, and ends with a helicase (NS3Hel) domain. Last, NS5 is composed of a methyltransferase domain (MTase) and an RNA-dependent RNA polymerase domain (RdRP).

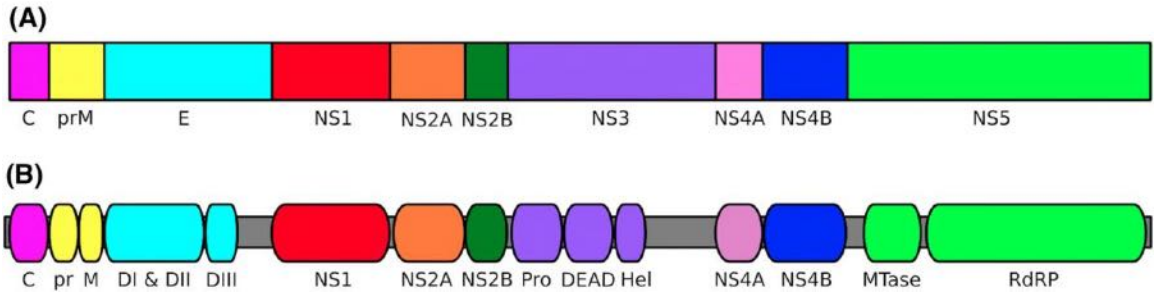


Figure 1. Schematic of the flavivirus polyprotein illustrating **a** the proteins that make up the polyprotein and **b** the domains that make up the proteins.

Table 1. Flavivirus protein function

Protein	Function
Capsid	Housing of viral genome, interaction with host proteins, and determination of viral infectivity (Oliveira et al. 2017)
Premembrane	prM acts as a chaperone for envelope folding and prevents premature fusion of envelope proteins. prM is cleaved to pr and M proteins for mature virion formation (Hsieh et al. 2014)
Envelope	Facilitates viral infection through receptor binding and membrane fusion. Triggers neutralizing antibody response (Zhang et al. 2017)
NS1	Formation of viral replication complex (dimer), and eliciting and evading immune response (hexamer) (Rastogi et al. 2016)
NS2A	Formation of viral replication complex and eliciting host immune response (Xie et al. 2013)
NS2B	Necessary cofactor for serine protease activity of NS3. Also has roles in viral replication and virion formation (Li et al. 2016)
NS3	Serine protease domain performs autocleavage and cleaves many sites in viral polyprotein. Helicase domain unwinds RNA secondary structure to assist in RNA replication (Bollati et al. 2010)
NS4A	Formation of viral replication complex. Regulates ATPase activity of NS3 helicase (Shiryaev et al. 2009)
NS4B	Formation of viral replication complex (Kaufusi et al. 2014)
NS5	N-terminus has methyltransferase domain for RNA cap methylation. C-terminus has RNA-dependent RNA polymerase domain for viral genome replication (Bollati et al. 2010)

At the polyprotein level, many of the phylogenetic relationships observed among the flaviviruses show a strong correlation with their vector association (Fig. 2). Arthropod-borne flaviviruses can be divided into two groups: tick-borne flaviviruses (TBFVs) and mosquito-borne flaviviruses (MBFVs). MBFVs can be further categorized

as being *Aedes* spp. specific or *Culex* spp. specific. The two arthropod-borne flavivirus groups are recovered in separate clades, resulting from the earliest divergence event. A third group of flaviviruses with no known vector (NKV) can be found within each of the two major clades of the phylogeny.

The upper major clade of the phylogeny branches off into two subclades. The upper subclade is composed of five NKV flaviviruses: Apoi virus (APOIV), Rio Bravo virus (RBV), Montana myotis leukoencephalitis virus (MMLV), Jutiapa virus (JUTV), and Modoc virus (MODV). The lower subclade is composed of 10 TBFVs: Tyuleniy virus (TYUV), Kama virus (KAMV), Karshi virus (KSIV), Powassan virus (POWV), Alkhurma hemorrhagic fever virus (AHFV), Langat virus (LGTV), Omsk hemorrhagic fever virus (OHFV), Tick-borne encephalitis virus (TBEV), Spanish goat encephalitis virus (SGEV), and Louping ill virus (LIV).

The lower major clade is composed of an outgroup consisting of the three MBFVs belonging to the Yellow fever virus (YFV) clade, which includes YFV, Sepik virus (SEPV), and Wesselsbron virus (WESSV), as well as its sister clade of 3 NKV flaviviruses, which includes Yokose virus (YOKV), Sokoluk virus (SOKV), and Entebbe bat virus (ENTV). *In vitro*, these three NKV viruses have shown the ability to replicate within *Aedes* spp. cells (Kuno 2007). Two more *Aedes*-associated MBFVs, Chaoyang virus (CHAOV) and Donggang virus (DONV), branch off at the next node junction. These two viruses, along with the YFV clade and the sister NKV clade, create the *Aedes*-outgroup clade that is referred to later when discussing significant rate-shifts between groups of MBFVs (Fig. 2). Following, the next branching event results in two clades, one composed of 7 *Aedes* spp. flaviviruses and the other of 12 *Culex* spp. flaviviruses. The

Aedes group includes Kedougou virus (KEDV), Zika virus (ZIKV), Spondweni virus (SPOV), and the four Dengue virus serotypes (DENV1-4). The *Culex* group includes Kokobera virus (KOKV) Aroa virus (AROAV), Ilheus virus (ILHV), Tembusu virus (TMUV), Ntaya virus (NTAV), Bagaza virus (BAGV), St. Louis encephalitis virus (SLEV), Cacipacore virus (CPCV), West Nile virus (WNV), Murray Valley encephalitis virus (MVEV), Japanese encephalitis virus (JEV), and Usutu virus (USUV).

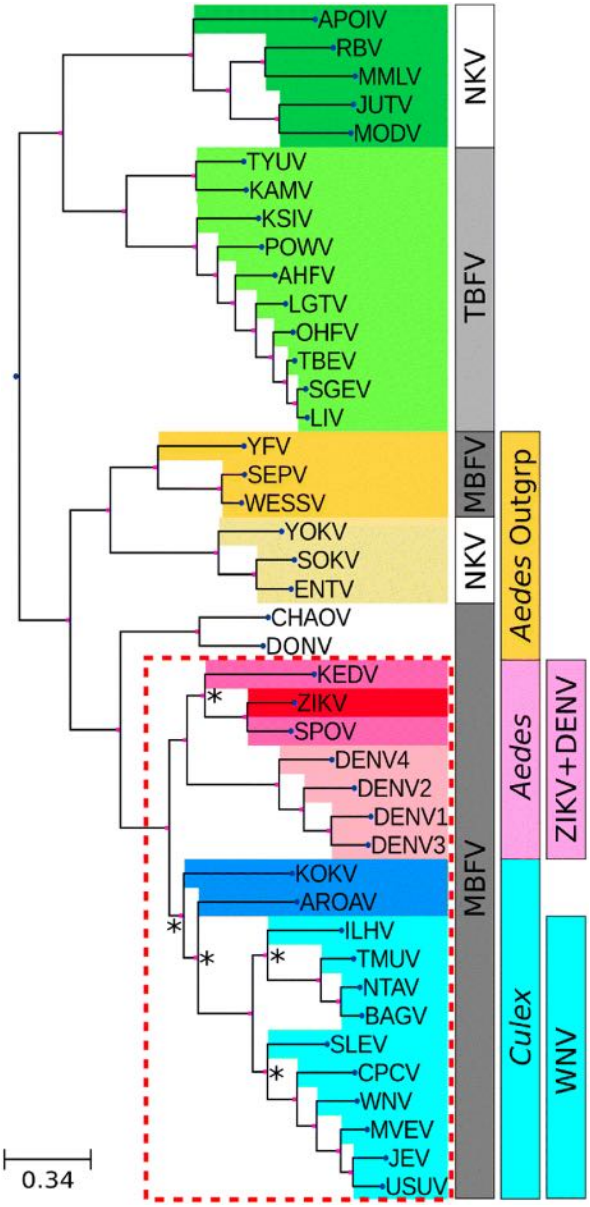


Figure 2. Phylogenetic reconstruction of 42 flavivirus polyproteins. Taxa often associate by vector association: no known vector (NKV), tick-borne flaviviruses (TBFV), and mosquito-borne flaviviruses (MBFV). A second NKV group is found within the MBFVs. These viruses have been found to replicate *in vitro* within *Aedes* spp. cells (Kuno 2007). Nodes indicated with an asterisk have a posterior probability greater than 0.9 but less than 1, all other nodes have a posterior probability of 1.

Identification of Target Sites

When discussing clade-specific target sites, we will refer to the clades by their taxa composition (e.g., ZIKV+DENV clade). Protein regions with five or more consecutive sites displaying 100% sequence conservation were identified for the full phylogeny (42 taxa), as well as for the MBFV clade (27 taxa), a subgroup of the MBFVs (19 taxa, Fig. 2 – boxed clade), the ZIKV+DENV clade (7 taxa, Fig. 2), the WNV clade with AROAV and KOKV (12 taxa, Fig. 2) and the WNV clade alone (10 taxa, Fig. 2). Identified motifs were then analyzed for 100% conservation in structural order as predicted by IUPred. Motifs conserved in sequence and structural order were further analyzed for 100% conservation in secondary structure element (alpha helix or beta strand) as predicted by PSIPRED (Appendix 1). Protein regions with 100% conservation in sequence, structural order, and secondary structure are considered fitness-critical and henceforth referred to as target sites.

When considering the alignment and structural predictions for the full phylogeny, there were no identifiable target sites. While 11 regions conserved in amino acid sequence were identified, only three of these regions were conserved in structural order, and none of those regions were conserved in secondary structure. Examining structural order and secondary structure individually, there were 82 and 90 regions, respectively, with full conservation for that feature. Given the relatively small number of regions

conserved in amino acid sequence, conservation at the sequence level is the limiting factor.

When considering only the 27 MBFVs (Fig. 2), we identified 16 regions conserved in amino acid sequence, 7 of which were conserved in structural order, but only 1 was also conserved in secondary structure and classified as a target site (Table 2, Fig. 3). The identified site, LEFEA, is in the RNA-dependent RNA polymerase (RdRP) domain of the NS5 protein. Further limiting the taxa under consideration to a subset of 19 MBFVs (Fig. 2, boxed clade) allowed for the identification of an additional target site, RRDLR, which is also located in RdRP (Table 2, Fig. 3).

Table 2. Summary of target sites

Clade	Protein—Domain	Conserved Sites in the MSA[a,b]	Surface Exposed Residues[c]
MBFV Subgroup	NS5—RdRP	3133-**LEFEA**-3137	L**E*
	NS5—RdRP	3422-RR**DLR**-3426	*R**R
ZIKV+DENV	Envelope—DIII	630-**GHLKC**-634	GH*K*
	NS3—DEAD	1872-HA**TFT**-1876	*AT**
	NS5—RdRP	3133-**LEFEAL**-3138	L**E**
	NS5—RdRP	3411-**YAQMW**-3415	Y*QM*
	NS5—RdRP	3421-HR**RDLRL**-3427	**R**RL
WNV	NS3—DEAD	1901-PA**SIAARGY**I-1910	PA********
	NS4B—NS4B	2516-**WQAEA**-2520	N/A[d]
	NS4B—NS4B	2527-**RTAAG**-2531	N/A[d]
	NS5—RdRP	3133-**LEFEA**-3137	L**E*
	NS5—RdRP	3331-**LHFLN**-3335	LH*LN
	NS5—RdRP	3422-RR**DLR**-3426	*R**R
	NS5—RdRP	3428-**MANAIC**-3433	**N**C
	NS5—RdRP	3513-**TWAEN**-3517	TWAEN

[a]Bold means also conserved with JPred4 and DISOPRED2

[b]Underline means also conserved in ZIKV

[c]Asterisk means residue is not exposed on the surface

[d]No experimentally determined structure available

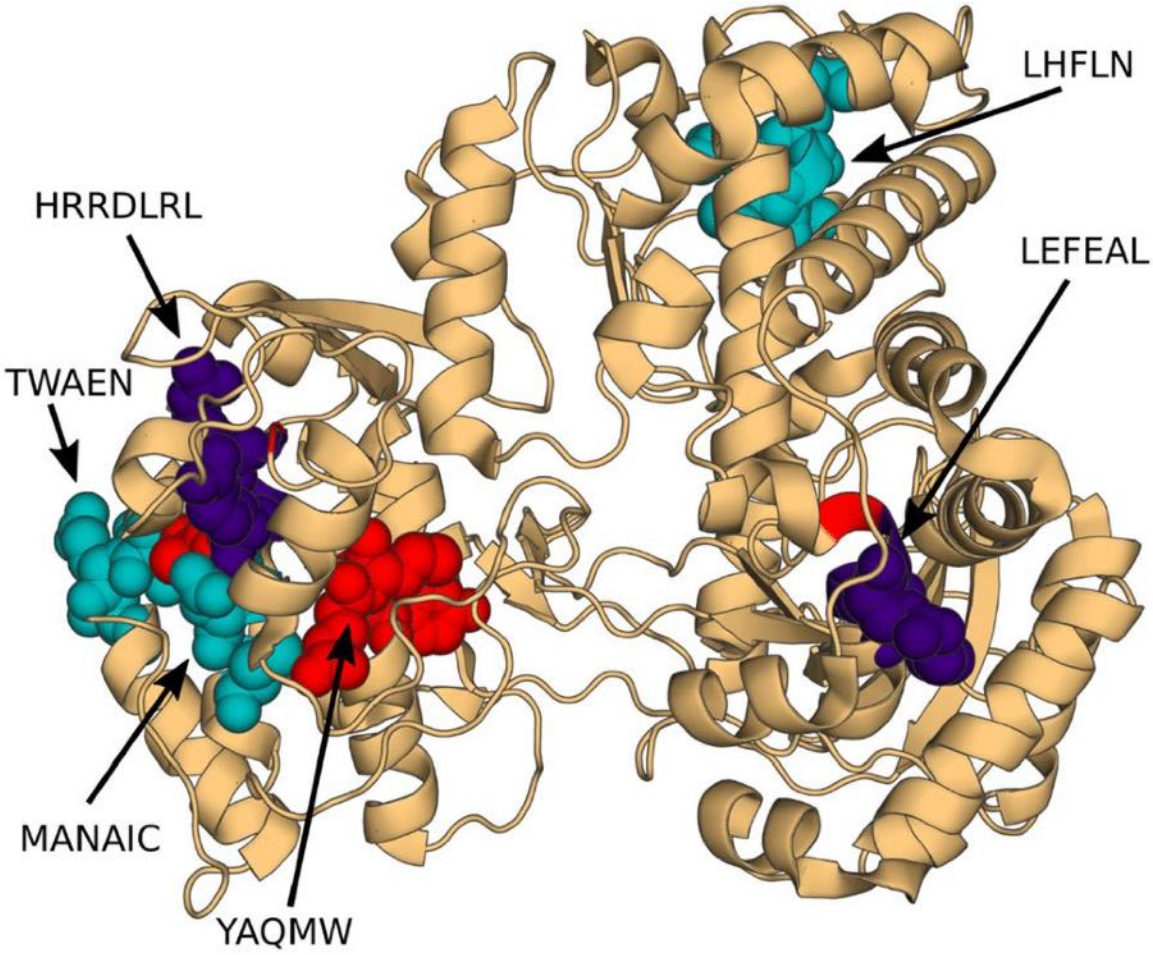


Figure 3. Target sites mapped to the RNA-dependent RNA polymerase structure for ZIKV (PDB id: 5TFR (Upadhyay et al. 2017)). Sites shown in purple are shared among 19 MBFVs. Sites in red are ZIKV+DENV clade specific. Sites in blue are WNV clade specific. Residues shown as spheres are exposed to the surface as determined by the PyMOL script findSurfaceResidues.py (Vertrees 2019) using a 2.5 Å^2 cut-off.

Shorter evolutionary time scales were evaluated: the ZIKV+DENV clade, the WNV clade plus KOKV and AROAV, and the WNV clade, respectively. For the ZIKV+DENV clade, five target sites were identified across three proteins: Envelope,

NS3, and NS5. For the WNV clade with KOKV and AROAV, four target sites were identified across two proteins: NS4B and NS5. By excluding KOKV and AROAV and reducing the taxa to only the WNV clade, the number of target sites identified increased so that nine target sites were identified within three proteins: NS3, NS4B, and NS5 (Table 2). For both the ZIKV+DENV and WNV clades, the majority of the identified target sites were located within RdRP (Fig. 3). Given ZIKV's association with the WNV clade in phylogenies based on individual proteins (Ortiz et al. 2013), we also investigated any target sites shared by the WNV clade and ZIKV, shown underlined (Table 2).

Predictions for structural order and secondary structure by DISOPRED2 and JPred4, respectively, were also taken into consideration. Sites where both structural disorder predictors and both secondary structure predictors agree are indicated in bold (Table 2).

The ability of a drug to bind to a protein requires that the protein binding site is exposed to solvent. For all target sites found in experimentally determined structures, solvent accessibility was determined by checking for surface exposed residues based on a 2.5 $Å^2$ cut-off. The WNV clade target sites in NS4B could not be assessed for surface accessibility due to the lack of an experimentally structure for that protein. For all other target sites across clades and proteins, only one is fully exposed on the surface: TWAEN in RdRP for the WNV clade. However, the remaining sites have at least two residues exposed to the surface, and none of the target sites are fully buried (Table 2).

Polyprotein sequences for ZIKV, DENV, and WNV strains were analyzed to ensure that the target sites identified were conserved in amino acid identity across strains. For each of these viruses, all target sites were conserved in 99-100% of the strains

analyzed. For 567 ZIKV strains analyzed, all target sites were conserved across all strains except for GHLKC which lacked conservation for one of the strains. For 5078 DENV strains (DENV1-4 combined), no target site was found to be 100% conserved across strains. Regardless, all target sites can be considered highly conserved in DENV given that conservation levels varied between 99.8 and 99.9%, with only between 3 and 10 strains lacking conservation, depending on the site. For 2125 WNV strains, the target sites LEFEA and RRDLR were conserved across all strains, while the remaining target sites showed 99.8-99.9% conservation, with between 1 and 4 strains lacking conservation, depending on the site (Table 3).

Table 3. Conservation of target sites in ZIKV, DENV, and WNV strains

Virus (strain count)	Target site	Percent conservation (%)
ZIKV (567)	GHLKC	99.82
	HATFT	100
	LEFEAL	100
	YAQMW	100
	HRRDLR	100
DENV (5078)	GHLKC	99.92
	HATFT	99.80
	LEFEAL	99.94
	YAQMW	99.86
	HRRDLR	99.86
WNV (2124)	PASIAARGYI	99.95
	WQAEA	99.95
	RTAAG	99.95
	LEFEA	100
	LHFLN	99.81
	RRDLR	100

MANAIC	99.86
TWAEN	99.86

Evolutionary Rate-Shifts – Functional Determinants for Mosquito-Borne Flaviviruses

Understanding vector specificity is important as it allows for the appropriate management of vector populations to mitigate the spread of disease. Additionally, the presence of a vector in certain geographic areas can indicate the potential for a previously absent disease to spread to that area. As a proxy for sites that play a role as functional determinants of vector specificity, we determined significant site-specific evolutionary rate-shifts between three clades of interest: *Aedes*-outgroup, *Aedes*, and *Culex* (Fig. 2). To this end, when discussing the clades used for the identification of site-specific evolutionary rate-shifts, we will refer to them by their vector association (e.g., *Aedes* clade). Clade-specific evolutionary rates were determined for the common core sites (ungapped sites) between these clades based on both individual proteins and the full-length polyprotein. A modified *t* test was used to determine which sites presented a significant shift in evolutionary rate (for further details, please see Methods – Evolutionary rate estimation and Rate-shift calculations). For both sets of rates, rate-shifts are analyzed in a per-protein context. Evolutionary rate-shifts between the three clades of interest were also estimated using DIVERGE (Gu and Vander Velden 2002) and the results were analyzed in a per-protein context.

The results from DIVERGE yielded fewer sites predicted to have significant evolutionary rate-shifts. This method identified 23, 0, and 71 rate-shifting sites for *Aedes* vs *Culex*, *Aedes* vs *Aedes*-outgroup, and *Culex* vs *Aedes*-outgroup, respectively.

Generally, for each individual protein there are more sites with significant rate-shifts identified based on the polyprotein rates than there are based on the per-protein rates (Fig. 4). The primary exception to this trend is the NS5 protein, where there are more sites with significant rate-shifts identified for all clade comparisons for the individual protein rates versus the polyprotein rates. Additionally, the NS1 and NS2A proteins for the *Aedes* vs *Culex* comparison had more sites identified for the per-protein rates, as did the prM protein for the *Culex* vs *Aedes*-outgroup comparison.

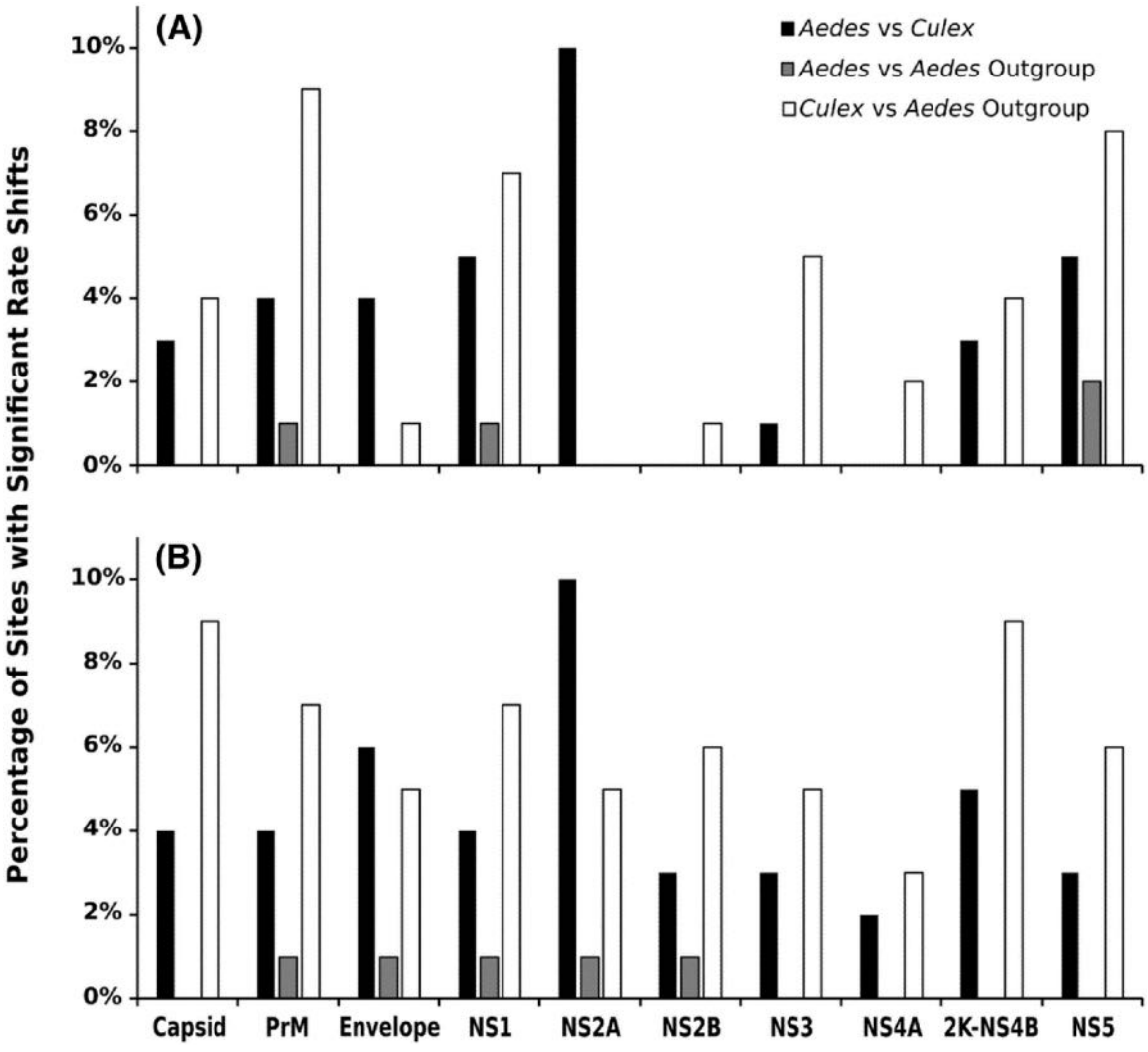


Figure 4. Percent of sites exhibiting significant evolutionary rate-shifts based on **a** evolutionary rates for individual proteins and **b** evolutionary rates for the full-length polyprotein. Percent significant rate-shifts are shown in a per-protein context.

Regardless of method, *Aedes* vs *Aedes*-outgroup consistently had less sites with significant rate-shifts identified, with DIVERGE being unable to identify any. Similarly, *Aedes* vs *Culex* had the second highest number of sites identified, while *Culex* vs *Aedes*-outgroup had the highest number of sites identified. As a method, DIVERGE identified the fewest sites when compared to the *t* test based on individual protein rates and the *t* test based on the polyprotein rates, which were relatively comparable. Significant rate-shifts identified by the polyprotein rates generally recovered all sites also identified by DIVERGE, with there being a few exceptions of some sites that were only identified by DIVERGE for the *Culex* vs *Aedes*-outgroup comparison (six total).

Of greatest interest, some significant rate-shifts occur in sites identified as target sites for the WNV clade (Table 4, Fig. 5). For these sites, we only focus on comparisons between *Aedes* vs *Culex* and *Culex* vs *Aedes*-outgroup. These significant rate-shifts mostly occur in rates based on the polyprotein for *Aedes* vs *Culex* and *Culex* vs *Aedes*-outgroup comparisons, with some also having a significant rate-shift across comparisons when based on the individual protein rates. Three sites are also identified by DIVERGE for the *Culex* vs *Aedes*-outgroup comparison. One site is identified to have a significant rate-shift across both comparisons featuring *Culex* for all methods of analysis.

Table 4. Target sites of the WNV clade with sites identified as having significant evolutionary rate-shifts across clades

	Aedes vs *Culex*		*Culex* vs *Aedes*-outgroup		
	Individual	Polyprotein	DIVERGE	Individual	Polyprotein
2517					X
3332	X	X	X	X	X
3430	X	X			
3433		X	X		X

3516	X	X		X
3517	X			X

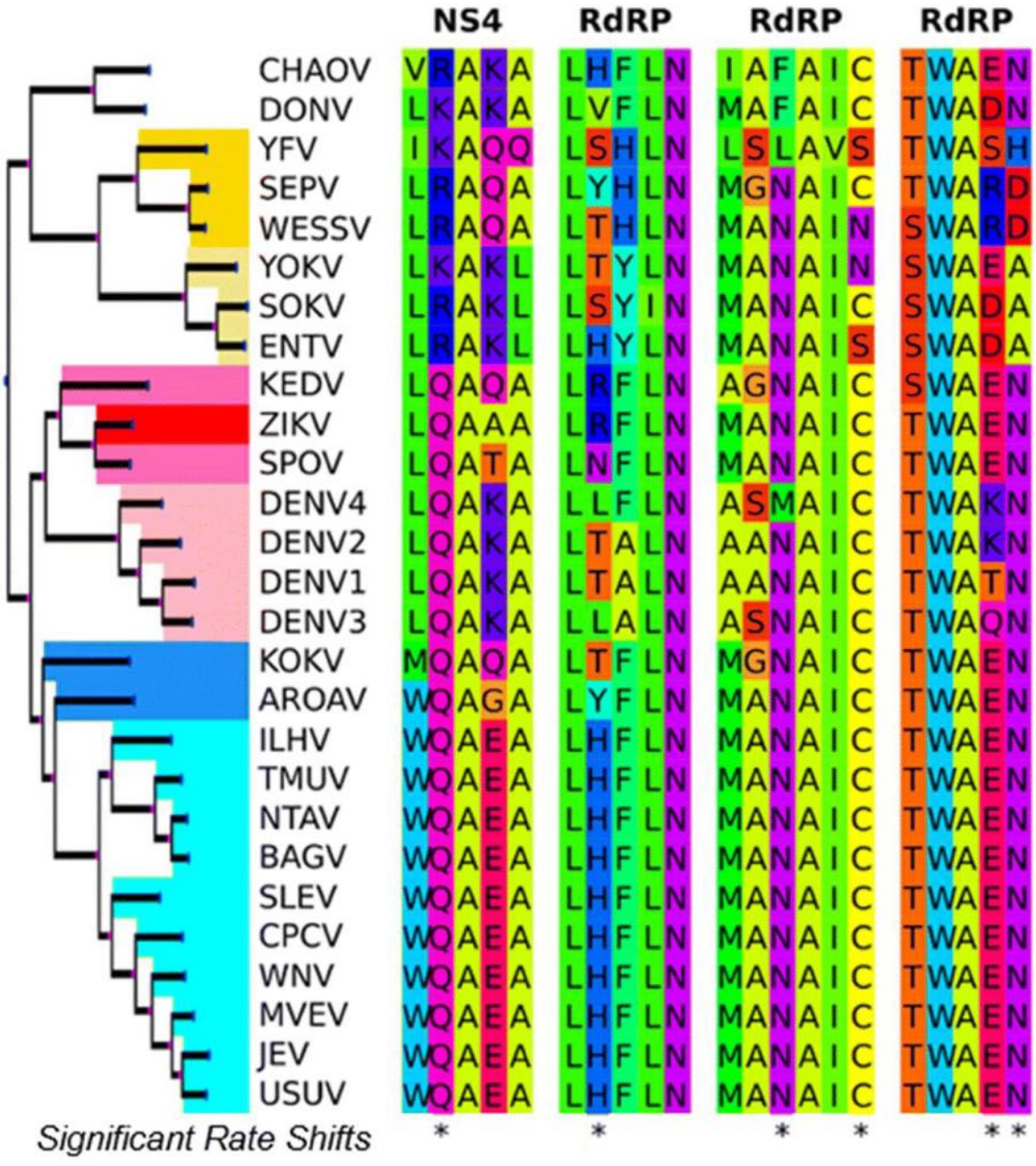


Figure 5. WNV target sites with significant evolutionary rate-shifts shown in a phylogenetic context. Phylogeny on the left is shown re-rooted to recover the three clades used in the rate-shift analysis. Taxa are colored as in Fig. 2. Target sites for the WNV clade are shown on the right in the context of the multiple sequence alignment for the

MBFVs. The last row of the alignment, labeled Significant Rate Shifts, has stars at the sites that were identified to be experiencing significant evolutionary rate-shifts.

DISCUSSION

We have performed a comparative study of the flavivirus proteome to identify its evolutionary constraints on sequence and structural properties. The structural properties discussed here are predicted intrinsic disorder and secondary structure. We chose to base the structural properties on predictions to treat all sites equally instead of only including sites with experimental structural data. Predictions are not in perfect agreement with experimental structural data which also depend on experimental conditions (See Ahrens et al. 2018 for further discussion). We sought to identify sites that are critical for viral fitness with potential as broadly neutralizing antiviral target sites. We define fitness-critical target sites as 5 or more consecutive residues that are conserved in sequence, order (not intrinsically disordered), and secondary structure over a specific clade in the flavivirus polyprotein phylogeny.

Our phylogeny included tick-borne and mosquito-borne flaviviruses, as well as flaviviruses with no known vector (Fig. 2). For the mosquito-borne flaviviruses, three main clades were identified: the outgroup clade that contains YFV and that uses *Aedes* as vector (*Aedes*-outgroup), the clade that contains ZIKV and DENV and uses *Aedes* as vector (ZIKV+DENV clade/*Aedes* clade), and the clade that contains WNV and uses *Culex* as a vector (WNV clade/*Culex* clade). Two target sites within RdRP were identified for the 19 mosquito-borne flaviviruses in the ZIKV+DENV and WNV clades. Additional target sites were found for the ZIKV+DENV and WNV clades separately. Five and nine target sites were found in the ZIKV+DENV clade and the WNV clade, respectively. At the sequence level, the identified target sites were either fully or nearly

fully (>99%) conserved across viral strains for ZIKV, DENV, and WNV, strengthening the position that these are evolutionarily constrained fitness-critical sites that may be taken advantage of for the development of broadly neutralizing antiviral drugs.

Next, when we compared rate-shifts in amino acid sequence across these clades, we found that significant rate-shifts have occurred between all clades. However, rate-shifts between the *Aedes* clade and *Aedes*-outgroup clade were sparse (Fig. 4), suggesting that the rate-shifting sites between the *Culex* clade and the *Aedes* and *Aedes*-outgroup clades could be important for vector specificity. Six of the rate-shifting sites for the *Culex* and *Aedes*-outgroup comparison fall within four target sites for the WNV clade. Two of these motifs (MANAIC and TWAEN) are located close to each other in the 3D structure of RdRP (Fig. 3). These two motifs may be ancestral as MANAIC occurs in SOKV and TWAEN in CHAOV, both from the *Aedes*-outgroup clade. Both are also conserved in the *Aedes*-associated ZIKV and SPOV (Table 2, Fig. 5). As such, the presence of both motifs may contribute to a distinct function found in ZIKV, SPOV, and the WNV clade with potential implications for *Aedes* and *Culex* vector specificity.

One of the ZIKV+DENV target sites is found within the Envelope (E) protein. The E protein of flaviviruses is a conformationally flexible protein (Kuhn et al. 2015) responsible for viral entry into cells (Modis et al. 2004). The E protein is recognized by potentially neutralizing antibodies and thus plays a large role in ADE (Kuhn et al. 2015). A pocket, called the βOG pocket, in the flexible hinge region of the DENV2 E protein involved in the conformational changes necessary for viral infection has been found to bind to a small detergent, β-octylglucoside (Modis et al. 2003), suggesting that blocking this pocket may inhibit viral entry. The βOG pocket has been previously found to bind to

small molecules that block this region in DENV (Clark et al. 2016). More recent work found that several small compounds targeted at the βOG pocket were able to inhibit DENV, ZIKV, WNV, and JEV activity, though with variable viral specificity (de Wispelaere et al. 2018). Interestingly, one of our identified target sites for the ZIKV+DENV clade (GHLKC) is at the top of this pocket, highlighting its potential as a target site for broadly neutralizing antiviral drugs. While the motif GHLKC was only found conserved in sequence, structural order, and secondary structure for the ZIKV+DENV clade, the ability to target the βOG pocket in JEV and WNV indicates that the clade-specific target sites we identify may serve as broader targets than anticipated.

For the ZIKV+DENV clade and for the WNV clade, one target was found for each clade in the DEAD domain of NS3. This domain is critical for the helicase activity of NS3. While the sites for WNV clade are buried, the two accessible sites from the ZIKV+DENV clade hold promise as a potential antiviral target site. These two sites participate in coordinating ssRNA (ZIKV, PDB id: 5GJB, (Tian et al. 2016)) and are found in a deep pocket when ssRNA is not bound (ZIKV, PDB id: 5JPS) (Fig. 6). Flavivirus helicases are considered important drug targets (Luo et al. 2015).

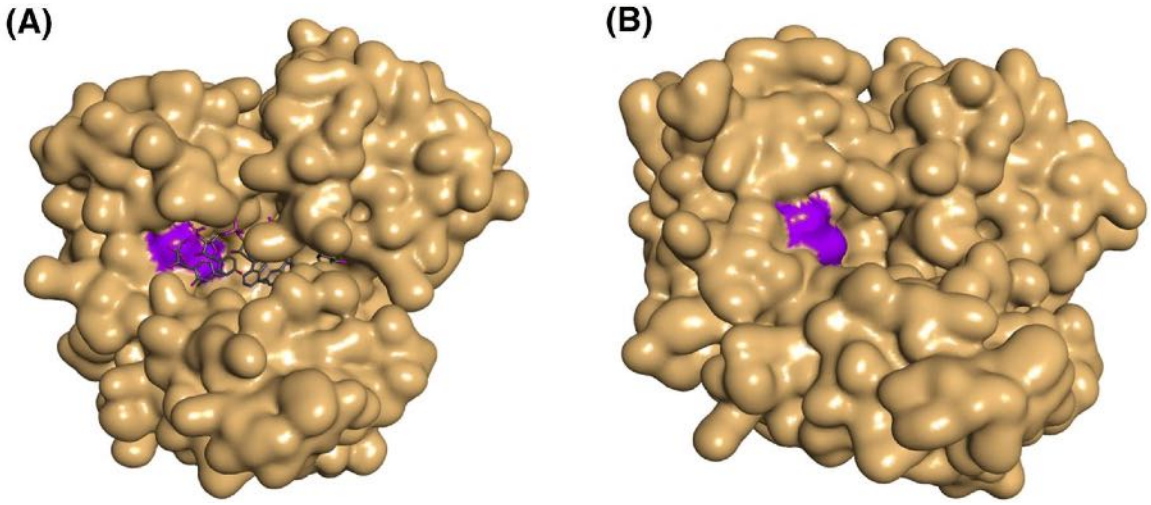


Figure 6. The target sites in the NS3 helicase (DEAD domain) for the ZIKV+DENV clade has the motif HATFT shown in purple. Only two of the five sites (positions 2 and 3, A and T) are surface accessible. However, in ZIKV NS3, these two sites **a** participate in coordinating ssRNA (PDB id: 5GJB (Tian et al. 2016)) and **b** are found in a deep pocket when ssRNA is not bound (PDB id: 5JPS, not published).

Most of the target sites identified are located in RdRP, which is already a frequently considered drug target (Malet et al. 2008; Sampath and Padmanabhan 2009; Bollati et al. 2010). We find that RdRP has potential target sites that are conserved across a subset of 19 MBFVs and even more if we consider only the WNV clade and the ZIKV+DENV clade. One target site for the MBFVs, RRDLR, forms an arginine patch in RdRP. This patch binds to a 3'-UTR of the flaviviral genome called the 3' stem-loop top loop (3'-SL-TL). The 3'-SL-TL contains a highly conserved motif (ACAG) that functions as a recognition site for RdRP. Using DENV2 as a model, it was found that site-directed mutagenesis of two arginines in R**R**DL**R** (bolded) resulted in loss of interaction with the 3'-SL-TL and in reduced viral replication as a result (Hodge et al. 2016). These two arginines are also found to be solvent accessible based on the ZIKV RdRP structure (PDB id: 5U0C (Upadhyay et al. 2017)). The apparent functional significance and relative accessibility of RRDLR in addition to its conservation across a multitude of flaviviruses further indicate its potential as a broadly neutralizing antiviral target site.

The flavivirus RdRP is distantly related to the RdRP in Hepatitis C virus (HCV) (Potisopon et al. 2014). RdRP from HCV is inhibited by the nucleoside analog sofosbuvir, a proven therapy against HCV (Bhatia et al. 2014). Various studies have supported the use of sofosbuvir in treating ZIKV using both cell-line studies (Mumtaz et al. 2017; Sacramento et al. 2017) and mouse models (Ferreira et al. 2017). Retallack and

co-workers found that sofosbuvir could reduce ZIKV viral load (Retallack et al. 2016). A structural alignment between the ZIKV RdRP and the HCV RdRP illustrates the structural similarities of these proteins. Sequence similarity, however, is low and conservation between the flavivirus target sites and HCV is hardly observed (Fig. 7). While sofosbuvir binding to the flaviviruses may not be specific or long-term, it shows that RdRP is a potential target against these viruses. Compounds that bind to the target sites in RdRP can be used as broadly neutralizing antivirals across several flaviviruses.

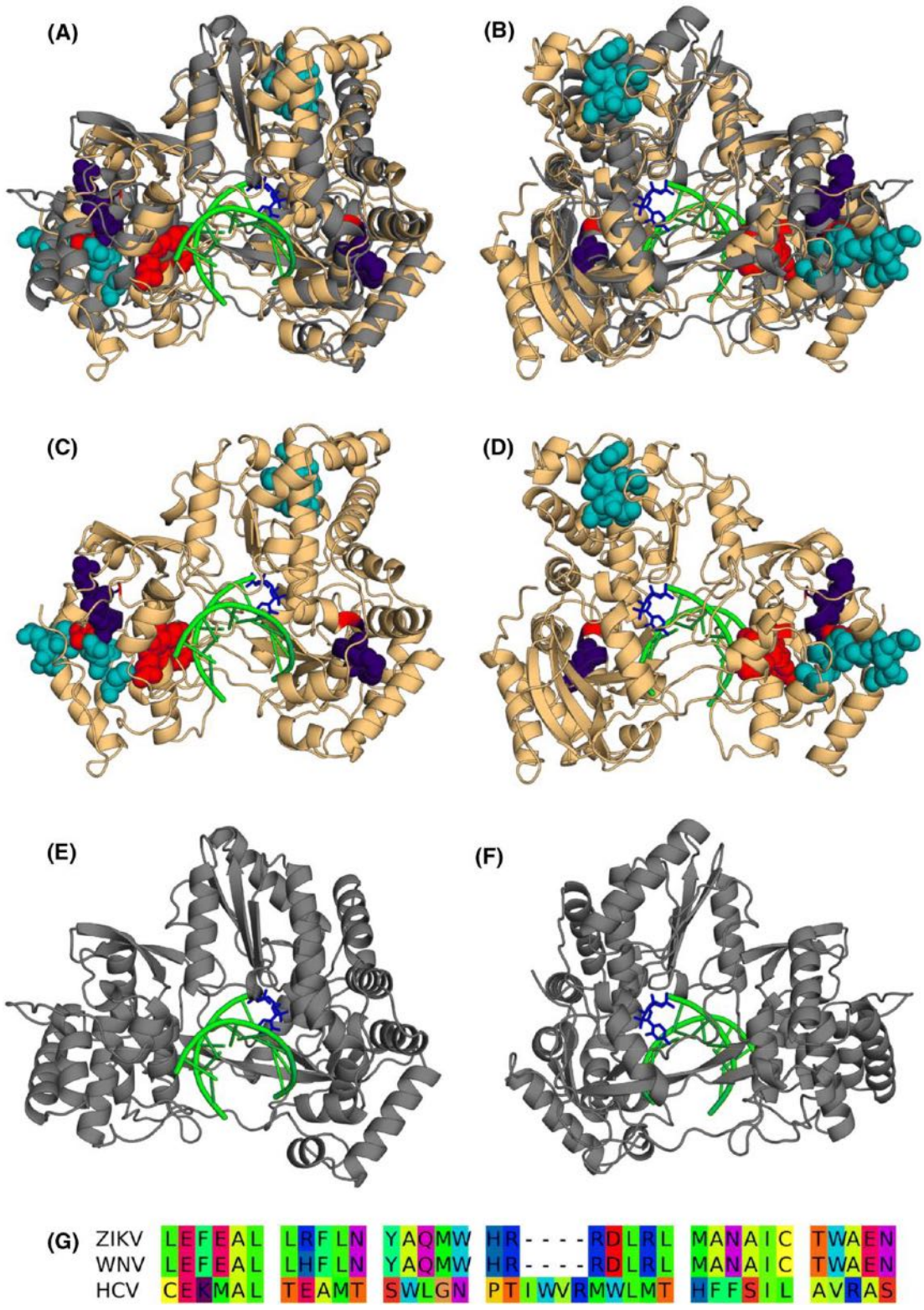


Figure 7. Structural alignment for ZIKV RdRP (PDB id: 5U0C (Zhao et al. 2017)) and HCV RdRP (PDB id: 4WTG (Appleby et al. 2015)). Each of the 8 entities of 5U0C was aligned with the single entity of 4WTG with CATH-SSAP v0.16.2 (Taylor and Orengo

1989; Orengo and Taylor 1996). Entity 1 of 5U0C had the lowest RMSD (4.57 Å) and the highest SSAP score (74.17) with 4WTG. The SSAP alignment for these two entities was used to superpose the structures using CATH-superpose v0.16.2 (Taylor and Orengo 1989). ZIKV is shown in beige, while HCV is shown in gray. Target sites for the MBFVs, ZIKV+DENV clade, and WNV clade are shown mapped onto the ZIKV RdRP and are colored as in Fig. 3. RNA is shown in green. Sofosbuvir is shown in blue. Views of the structural alignment are shown from **a** the front and **b** the back. Additionally, ZIKV and HCV RdRPs are shown from **c**, **e** the front and **d**, **f** the back. **g** A sequence alignment for the target sites in the ZIKV+DENV clade, the WNV clade, and HCV. ZIKV and WNV are shown as the representative sequences of their clades.

Altogether, our results identify evolutionarily constrained protein regions, both in sequence and structure, that can serve as promising target sites for the development of broadly neutralizing antivirals against flaviviruses while aiming to avoid complications caused by ADE. Identifying the majority of these target sites in proteins already often used as drug targets, such as NS5 (RdRP), further supports the plausibility of these sites as candidate targets for antiviral drug development. We find significant evolutionary rate-shifts between *Culex* and *Aedes,* or WNV and ZIKV+DENV, in some of these target sites. Rate-shifts between clades indicate functional divergence. In this case, the rate-shifts may be implicated in vector specificity (*Culex* for WNV and *Aedes* for ZIKV+DENV) and can provide ways to address the spread of these viruses by disrupting vector-virus interactions. Notably, ZIKV shares some target sites with the WNV clade, two of which contain sites experiencing significant evolutionary rate-shifts. The implication of functional divergence made by rate-shifting sites suggests that ZIKV may not only share functional determinants with DENV but also with WNV. This, together with ZIKV's association with WNV in certain phylogenies, raises concerns as to whether ZIKV can easily evolve to expand its vector association or if it already has under certain conditions.

ACKNOWLEDGMENTS

We thank Carlos Urbina for assistance in the lab and for helpful discussions. The authors would also like to acknowledge the Instructional & Research Computing Center (IRCC) at Florida International University for providing HPC computing resources that have contributed to the research results reported within this article, web: https://ircc.fiu.edu.

LITERATURE CITED

Ahrens J, Rahaman J, Siltberg-Liberles J (2018) Large-scale analyses of site-specific evolutionary rates across eukaryote proteomes reveal confounding interactions between intrinsic disorder, secondary structure, and functional domains. Genes 9:553. https://doi.org/10.3390/genes9110553

Altschul SF, Gish W, Miller W, et al (1990) Basic local alignment search tool. J Mol Biol 245:403–410

Appleby TC, Perry JK, Murakami E, et al (2015) Viral replication. Structural basis for RNA replication by the hepatitis C virus polymerase. Science 347:771–775. https://doi.org/10.1126/science.1259210

Assenberg R, Mastrangelo E, Walter TS, et al (2009) Crystal structure of a novel conformational state of the flavivirus NS3 protein: implications for polyprotein processing and viral replication. J Virol 83:12895–12906. https://doi.org/10.1128/JVI.00942-09

Barba-Spaeth G, Dejnirattisai W, Rouvinski A, et al (2016) Structural basis of potent Zika–dengue virus antibody cross-neutralization. Nature 536:48–53. https://doi.org/10.1038/nature18938

Bardina SV, Bunduc P, Tripathi S, et al (2017) Enhancement of Zika virus pathogenesis by preexisting antiflavivirus immunity. Science 356:175–180. https://doi.org/10.1126/science.aal4365

Batool M, Ahmad B, Choi S (2019) A structure-based drug discovery paradigm. Int J Mol Sci. https://doi.org/10.3390/ijms20112783

Bhatia HK, Singh H, Grewal N, Natt NK (2014) Sofosbuvir: A novel treatment option for chronic hepatitis C infection. J Pharmacol Pharmacother 5:278–284. https://doi.org/10.4103/0976-500X.142464

Bollati M, Alvarez K, Assenberg R, et al (2010) Structure and functionality in flavivirus NS-proteins: Perspectives for drug design. Antiviral Res 87:125–148. https://doi.org/10.1016/j.antiviral.2009.11.009

Bussetta C, Choi KH (2012) Dengue virus nonstructural protein 5 adopts multiple conformations in solution. Biochemistry 51:5921–5931. https://doi.org/10.1021/bi300406n

Cao X, Li Y, Jin X, et al (2016) Molecular mechanism of divalent-metal-induced activation of NS3 helicase and insights into Zika virus inhibitor design. Nucleic Acids Res 44:gkw941. https://doi.org/10.1093/nar/gkw941

CDC (2018) World Map of Areas with Risk of Zika. https://wwwnc.cdc.gov/travel/page/world-map-areas-with-zika

Cheng Y, LeGall T, Oldfield CJ, et al (2006) Rational drug design via intrinsically disordered protein. Trends Biotechnol 24:435–442. https://doi.org/10.1016/J.TIBTECH.2006.07.005

Chong B, Li M, Li T, et al (2018) Conservation of Potentially Druggable Cavities in Intrinsically Disordered Proteins. ACS omega 3:15643–15652. https://doi.org/10.1021/acsomega.8b02092

Clark MJ, Miduturu C, Schmidt AG, et al (2016) GNF-2 Inhibits Dengue Virus by Targeting Abl Kinases and the Viral E Protein. Cell Chem Biol 23:443–452. https://doi.org/10.1016/J.CHEMBIOL.2016.03.010

de Lamballerie X, Leroy E, Charrel RN, et al (2008) Chikungunya virus adapts to tiger mosquito via evolutionary convergence: a sign of things to come? Virol J 5:33. https://doi.org/10.1186/1743-422X-5-33

de Wispelaere M, Lian W, Potisopon S, et al (2018) Inhibition of Flaviviruses by Targeting a Conserved Pocket on the Viral Envelope Protein. Cell Chem Biol 25:1006-1016.e8. https://doi.org/10.1016/J.CHEMBIOL.2018.05.011

Dejnirattisai W, Supasa P, Wongwiwat W, et al (2016) Dengue virus sero-cross-reactivity drives antibody-dependent enhancement of infection with zika virus. Nat Immunol 17:1102–1108. https://doi.org/10.1038/ni.3515

Dodson BL, Pujhari S, Rasgon JL (2018) Vector competence of selected North American *Anopheles* and *Culex* mosquitoes for Zika virus. PeerJ 6:e4324. https://doi.org/10.7717/peerj.4324

Dosztányi Z, Csizmok V, Tompa P, Simon I (2005a) IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated

energy content. Bioinformatics 21:3433–3434. https://doi.org/10.1093/bioinformatics/bti541

Dosztányi Z, Csizmók V, Tompa P, Simon I (2005b) The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. J Mol Biol 347:827–839. https://doi.org/10.1016/j.jmb.2005.01.071

Drozdetskiy A, Cole C, Procter J, Barton GJ (2015) JPred4: a protein secondary structure prediction server. Nucleic Acids Res. https://doi.org/10.1093/nar/gkv332

Ferguson NM, Rodríguez-Barraquer I, Dorigatti I, et al (2016) Benefits and risks of the Sanofi-Pasteur Dengue vaccine: Modeling optimal deployment. Science 353:1033–1036. https://doi.org/10.1126/science.aaf9590

Ferreira AC, Zaverucha-do-Valle C, Reis PA, et al (2017) Sofosbuvir protects Zika virus-infected mice from mortality, preventing short- and long-term sequelae. Sci Rep 7:9409. https://doi.org/10.1038/s41598-017-09797-8

Finn RD, Bateman A, Clements J, et al (2014) Pfam: the protein families database. Nucleic Acids Res 42:D222–D230. https://doi.org/10.1093/nar/gkt1223

Fuxreiter M, Tompa P, Simon I (2007) Local structural disorder imparts plasticity on linear motifs. Bioinformatics 23:950–956. https://doi.org/10.1093/bioinformatics/btm035

Gaucher EA, Miyamoto MM, Benner SA (2001) Function-structure analysis of proteins using covarion-based evolutionary approaches: Elongation factors. Proc Natl Acad Sci U S A 98:548–552. https://doi.org/10.1073/pnas.98.2.548

Giri R, Kumar D, Sharma N, Uversky VN (2016) Intrinsically disordered side of the Zika virus proteome. Front Cell Infect Microbiol. https://doi.org/10.3389/fcimb.2016.00144

Grant A, Ponia SS, Tripathi S, et al (2016) Zika Virus Targets Human STAT2 to Inhibit Type I Interferon Signaling. Cell Host Microbe 19:882–890. https://doi.org/10.1016/j.chom.2016.05.009

Gu X (1999) Statistical methods for testing functional divergence after gene duplication. Mol Biol Evol 16:1664–1674. https://doi.org/10.1093/oxfordjournals.molbev.a026080

Gu X, Vander Velden K (2002) DIVERGE: phylogeny-based analysis for functional-structural divergence of a protein family. Bioinformatics 18:500–501. https://doi.org/10.1093/bioinformatics/18.3.500

Gu X, Zou Y, Su Z, et al (2013) An update of DIVERGE software for functional divergence analysis of protein family. Mol Biol Evol. https://doi.org/10.1093/molbev/mst069

Guedes DR, Paiva MH, Donato MM, et al (2017) Zika virus replication in the mosquito *Culex quinquefasciatus* in Brazil. Emerg Microbes Infect 6:e69. https://doi.org/10.1038/emi.2017.59

Guo X, Li C, Deng Y, et al (2016) *Culex pipiens quinquefasciatus*: a potential vector to transmit Zika virus. Emerg Microbes Infect 5:1–5. https://doi.org/10.1038/emi.2016.102

Hadinegoro SR, Arredondo-García JL, Capeding MR, et al (2015) Efficacy and long-term safety of a Dengue vaccine in regions of endemic disease. N Engl J Med 373:1195–1206. https://doi.org/10.1056/NEJMoa1506223

Heinz FX, Stiasny K (2012) Flaviviruses and flavivirus vaccines. Vaccine 30:4301–4306. https://doi.org/10.1016/j.vaccine.2011.09.114

Hodge K, Tunghirun C, Kamkaew M, et al (2016) Identification of a conserved RNA-dependent RNA polymerase (RdRp)-RNA interface required for flaviviral replication. J Biol Chem 291:17437–17449. https://doi.org/10.1074/jbc.M116.724013

Hsieh S-C, Wu Y-C, Zou G, et al (2014) Highly conserved residues in the helical domain of Dengue virus type 1 precursor membrane protein are involved in assembly, precursor membrane (prM) protein cleavage, and entry. J Biol Chem 289:33149–33160. https://doi.org/10.1074/jbc.M114.610428

Huerta-Cepas J, Serra F, Bork P (2016) ETE 3: Reconstruction, analysis and visualization of phylogenomic data. Mol Biol Evol. https://doi.org/10.1093/molbev/msw046

Hunter JD (2007) Matplotlib: A 2D graphics environment. Comput Sci Eng 9:90–95

Ivanyi-Nagy R, Darlix J-L (2010) Intrinsic disorder in the core proteins of flaviviruses. Protein Pept Lett 17:1019–1025. https://doi.org/10.2174/092986610791498911

Ivanyi-Nagy R, Lavergne J-P, Gabus C, et al (2008) RNA chaperoning and intrinsic disorder in the core proteins of Flaviviridae. Nucleic Acids Res 36:712–725. https://doi.org/10.1093/nar/gkm1051

Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 292:195–202. https://doi.org/10.1006/jmbi.1999.3091

Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 8:275–282

Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 30:3059–3066

Katzelnick LC, Gresh L, Halloran ME, et al (2017) Antibody-dependent enhancement of severe dengue disease in humans. Science 358:929–932. https://doi.org/10.1126/science.aan6836

Kaufusi PH, Kelley JF, Yanagihara R, Nerurkar VR (2014) Induction of endoplasmic reticulum-derived replication-competent membrane structures by West Nile virus non-structural protein 4B. PLoS One 9:e84040. https://doi.org/10.1371/journal.pone.0084040

Kian-Meng Goh G, Dunker AK, Foster JA, Uversky VN (2019) Zika and flavivirus shell disorder: Virulence and fetal morbidity. Biomolecules. https://doi.org/10.3390/biom9110710

Klema VJ, Ye M, Hindupur A, et al (2016) Dengue virus nonstructural protein 5 (NS5) assembles into a dimer with a unique methyltransferase and polymerase interface. PLoS Pathog 12:e1005451. https://doi.org/10.1371/journal.ppat.1005451

Kuhn RJ, Dowd KA, Beth Post C, Pierson TC (2015) Shake, rattle, and roll: impact of the dynamics of flavivirus particles on their interactions with the host. Virology 479–480:508–517. https://doi.org/10.1016/j.virol.2015.03.025

Kuno G (2007) Host range specificity of flaviviruses: correlation with in vitro replication. J Med Entomol 44:93–101

Le Breton M, Meyniel-Schicklin L, Deloire A, et al (2011) Flavivirus NS3 and NS5 proteins interaction network: a high-throughput yeast two-hybrid screen. BMC Microbiol 11:234. https://doi.org/10.1186/1471-2180-11-234

Lessler J, Chaisson LH, Kucirka LM, et al (2016) Assessing the global threat from Zika virus. Science 353:aaf8160. https://doi.org/10.1126/science.aaf8160

Li X-D, Deng C-L, Ye H-Q, et al (2016) Transmembrane domains of NS2B contribute to both viral RNA replication and particle formation in Japanese Encephalitis virus. J Virol 90:5735–5749. https://doi.org/10.1128/JVI.00340-16

Lourenço-de-Oliveira R, Marques JT, Sreenu VB, et al (2018) *Culex quinquefasciatus* mosquitoes do not support replication of Zika virus. J Gen Virol 99:258–264. https://doi.org/10.1099/jgv.0.000949

Luo D, Vasudevan SG, Lescar J (2015) The flavivirus NS2B-NS3 protease-helicase as a target for antiviral drug development. Antiviral Res 118:148–158. https://doi.org/10.1016/j.antiviral.2015.03.014

Main BJ, Nicholson J, Winokur OC, et al (2018) Vector competence of *Aedes aegypti*, *Culex tarsalis*, and *Culex quinquefasciatus* from California for Zika virus. PLoS Negl Trop Dis 12:e0006524. https://doi.org/10.1371/journal.pntd.0006524

Makhluf H, Shresta S, Makhluf H, Shresta S (2018) Development of Zika Virus Vaccines. Vaccines 6:7. https://doi.org/10.3390/vaccines6010007

Malet H, Massé N, Selisko B, et al (2008) The flavivirus polymerase as a target for drug discovery. Antiviral Res 80:23–35. https://doi.org/10.1016/j.antiviral.2008.06.007

McGuffin LJ, Bryson K, Jones DT (2000) The PSIPRED protein structure prediction server. Bioinformatics 16:404–405

Meng F, Badierah RA, Almehdar HA, et al (2015) Unstructural biology of the dengue virus proteins. FEBS J 282:3368–3394. https://doi.org/10.1111/febs.13349

Modis Y, Ogata S, Clements D, Harrison SC (2004) Structure of the Dengue virus envelope protein after membrane fusion. Nature 427:313–319. https://doi.org/10.1038/nature02165

Modis Y, Ogata S, Clements D, Harrison SC (2003) A ligand-binding pocket in the dengue virus envelope glycoprotein. Proc Natl Acad Sci U S A 100:6986–6991. https://doi.org/10.1073/pnas.0832193100

Mumtaz N, Jimmerson LC, Bushman LR, et al (2017) Cell-line dependent antiviral activity of sofosbuvir against Zika virus. Antiviral Res 146:161–163. https://doi.org/10.1016/J.ANTIVIRAL.2017.09.004

Oliveira ERA, Mohana-Borges R, de Alencastro RB, Horta BAC (2017) The flavivirus capsid protein: Structure, function and perspectives towards drug design. Virus Res. 227:115–123

Orengo CA, Taylor WR (1996) SSAP: sequential structure alignment program for protein structure comparison. Methods Enzymol 266:617–635

Ortiz JF, MacDonald ML, Masterson P, et al (2013) Rapid evolutionary dynamics of structural disorder as a potential driving force for biological divergence in flaviviruses. Genome Biol Evol 5:504–513. https://doi.org/10.1093/gbe/evt026

Penn O, Stern A, Rubinstein ND, et al (2008) Evolutionary modeling of rate shifts reveals specificity determinants in HIV-1 subtypes. PLoS Comput Biol 4:e1000214. https://doi.org/10.1371/journal.pcbi.1000214

Potisopon S, Priet S, Selisko B, Canard B (2014) Comparison of dengue virus and HCV: from impact on global health to their RNA-dependent RNA polymerases. Future Virol 9:53–67. https://doi.org/10.2217/fvl.13.121

Priyamvada L, Quicke KM, Hudson WH, et al (2016) Human antibody responses after Dengue virus infection are highly cross-reactive to Zika virus. Proc Natl Acad Sci U S A 113:7852–7857. https://doi.org/10.1073/pnas.1607931113

Pruitt KD, Tatusova T, Maglott DR (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 33:D501–D504. https://doi.org/10.1093/nar/gki025

Pupko T, Bell RE, Mayrose I, et al (2002) Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. Bioinformatics 18:S71–S77. https://doi.org/10.1093/bioinformatics/18.suppl_1.S71

Rahaman J, Siltberg-Liberles J (2016) Avoiding regions symptomatic of conformational and functional flexibility to identify antiviral targets in current and future coronaviruses. Genome Biol Evol 8:3471–3484. https://doi.org/10.1093/gbe/evw246

Rastogi M, Sharma N, Singh SK (2016) Flavivirus NS1: a multifaceted enigmatic viral protein. Virol J 13:131. https://doi.org/10.1186/s12985-016-0590-7

Retallack H, Di Lullo E, Arias C, et al (2016) Zika virus cell tropism in the developing human brain and inhibition by azithromycin. Proc Natl Acad Sci U S A 113:14408–14413. https://doi.org/10.1073/pnas.1618029113

Ronquist F, Teslenko M, van der Mark P, et al (2012) MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Softw Syst Evol 61:539–542

Roundy CM, Azar SR, Brault AC, et al (2017) Lack of evidence for Zika virus transmission by *Culex* mosquitoes. Emerg Microbes Infect 6:e90. https://doi.org/10.1038/emi.2017.85

Ruvinsky AM, Kirys T, Tuzikov AV, Vakser IA (2012) Structure fluctuations and conformational changes in protein binding. J Bioinform Comput Biol 10:1241002. https://doi.org/10.1142/S0219720012410028

Sacramento CQ, de Melo GR, de Freitas CS, et al (2017) The clinically approved antiviral drug sofosbuvir inhibits Zika virus replication. Sci Rep 7:40920. https://doi.org/10.1038/srep40920

Sampath A, Padmanabhan R (2009) Molecular targets for flavivirus drug discovery. Antiviral Res 81:6–15. https://doi.org/10.1016/j.antiviral.2008.08.004

Satterthwaite FE (1946) An Approximate Distribution of Estimates of Variance Components. Biometrics Bull 2:110. https://doi.org/10.2307/3002019

Schrödinger L (2014) The PyMOL molecular graphics system, Version 1.7.2

Shiryaev SA, Chernov AV, Aleshin AE, et al (2009) NS4A regulates the ATPase activity of the NS3 helicase: a novel cofactor role of the non-structural protein NS4A from West Nile virus. J Gen Virol 90:2081–2085. https://doi.org/10.1099/vir.0.012864-0

Sievers F, Wilm A, Dineen D, et al (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 7:539. https://doi.org/10.1038/msb.2011.75

Smock RG, Gierasch LM (2009) Sending signals dynamically. Science 324:198–203. https://doi.org/10.1126/science.1169377

Stettler K, Beltramello M, Espinosa DA, et al (2016) Specificity, cross-reactivity, and function of antibodies elicited by Zika virus infection. Science 353:823–826. https://doi.org/10.1126/science.aaf8505

Stiasny K, Fritz R, Pangerl K, Heinz FX (2011) Molecular mechanisms of flavivirus membrane fusion. Amino Acids 41:1159–1163. https://doi.org/10.1007/s00726-009-0370-4

Suzek BE, Wang Y, Huang H, et al (2015) UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics 31:926–932. https://doi.org/10.1093/bioinformatics/btu739

Taylor WR, Orengo CA (1989) Protein structure alignment. J Mol Biol 208:1–22

Tian H, Ji X, Yang X, et al (2016) Structural basis of Zika virus helicase in recognizing its substrates. Protein Cell 7:562–570. https://doi.org/10.1007/s13238-016-0293-2

Tsetsarkin KA, Vanlandingham DL, McGee CE, Higgs S (2007) A single mutation in Chikungunya virus affects vector specificity and epidemic potential. PLoS Pathog 3:e201. https://doi.org/10.1371/journal.ppat.0030201

Upadhyay AK, Cyr M, Longenecker K, et al (2017) Crystal structure of full-length Zika virus NS5 protein reveals a conformation similar to Japanese encephalitis virus NS5. Acta Crystallogr Sect F Struct Biol Commun 73:116–122. https://doi.org/10.1107/S2053230X17001601

van den Hurk AF, Hall-Mendelin S, Jansen CC, Higgs S (2017) Zika virus and *Culex quinquefasciatus* mosquitoes: a tenuous link. Lancet Infect Dis 17:1014–1016. https://doi.org/10.1016/S1473-3099(17)30518-2

Vázquez-Calvo Á, Blázquez A-B, Escribano-Romero E, et al (2017) Zika virus infection confers protection against West Nile virus challenge in mice. Emerg Microbes Infect 69(6):e81. https://doi.org/10.1038/emi.2017.68

Vertrees J (2019) FindSurfaceResidues - PyMOLWiki. https://pymolwiki.org/index.php/FindSurfaceResidues

Wang Q, Yang H, Liu X, et al (2016) Molecular determinants of human neutralizing antibodies isolated from a patient infected with Zika virus. Sci Transl Med 8:369ra179. https://doi.org/10.1126/scitranslmed.aai8336

Ward JJ, McGuffin LJ, Bryson K, et al (2004) The DISOPRED server for the prediction of protein disorder. Bioinformatics 20:2138–2139. https://doi.org/10.1093/bioinformatics/bth195

Xie X, Gayen S, Kang C, et al (2013) Membrane topology and function of Dengue virus NS2A protein. J Virol 87:4609–4622. https://doi.org/10.1128/JVI.02424-12

Yon C, Teramoto T, Mueller N, et al (2005) Modulation of the nucleoside triphosphatase/RNA helicase and 5'-RNA triphosphatase activities of Dengue virus type 2 nonstructural protein 3 (NS3) by interaction with NS5, the RNA-dependent RNA polymerase. J Biol Chem 280:27412–27419. https://doi.org/10.1074/jbc.M501393200

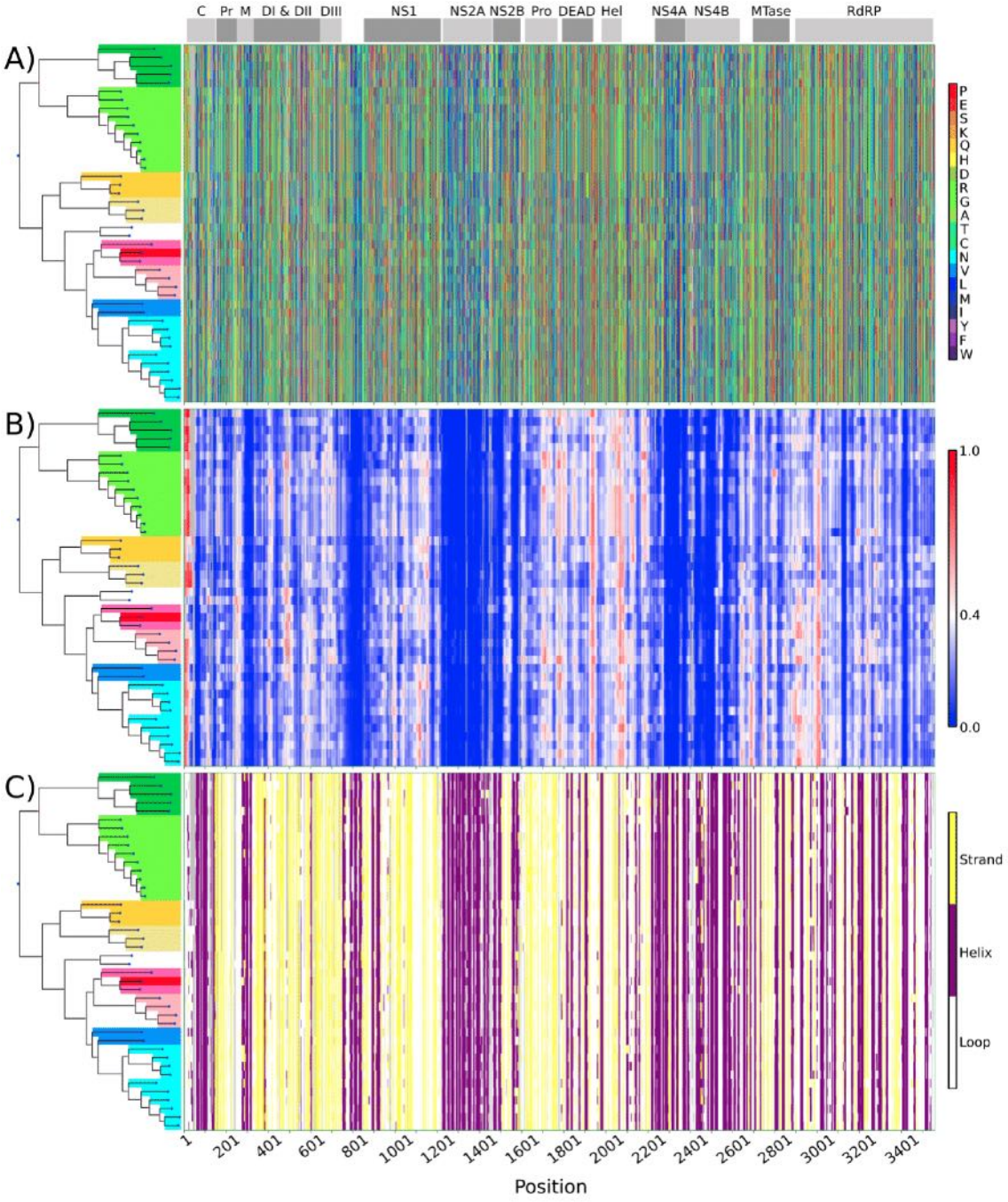
Zhang C, Lai L (2011) Towards structure-based protein drug design. Biochem Soc Trans 39:1382–1386. https://doi.org/10.1042/BST0391382

Zhang X, Jia R, Shen H, et al (2017) Structures and functions of the envelope glycoprotein in flavivirus infections. Viruses 9(11):338.

Zhang Y, Stec B, Godzik A (2007) Between order and disorder in protein structures: analysis of "dual personality" fragments in proteins. Structure 15:1141–1147. https://doi.org/10.1016/j.str.2007.07.012

Zhao B, Yi G, Du F, et al (2017) Structure and function of the Zika virus full-length NS5 protein. Nat Commun 8:14762. https://doi.org/10.1038/ncomms14762

Appendices



Appendix 1. Sequence-based predictions. Heatmaps for the three properties of interest mapped onto the polyprotein multiple sequence alignment context, with gaps colored gray. Phylogeny shown to the left is colored as shown in Fig. 2. Gray shaded blocks above the heatmaps illustrate domain boundaries as predicted by Pfam (Finn et al. 2016). **a** Multiple sequence alignment. Amino acids are colored based on TOP-IDP scale for measuring intrinsic disorder propensity (red is highest, purple is lowest; (Campen et al.

2008)). **b** Intrinsic disorder propensity as predicted for the full-length polyprotein by IUPred (Dosztányi et al. 2005). Blue-to-white-to-red illustrates low propensity towards disorder (blue, 0) to high propensity towards disorder (red, 1), with 0.4 (white) acting as the cut-off. **c** Secondary structure as predicted for the full-length polyprotein by PSIPRED (Jones 1999) showing beta strands (yellow), alpha helices (purple), and coils (white).

CHAPTER IV

EPITOPEDIA: IDENTIFYING MOLECULAR MIMICRY BETWEEN PATHOGENS AND KNOWN IMMUNE EPITOPES

ABSTRACT

Upon infection, foreign antigenic proteins stimulate the host's immune system to produce antibodies targeting the pathogen. These antibodies bind to regions on the antigen called epitopes. Structural similarity (molecular mimicry) of epitopes between an infecting pathogen and host proteins or other pathogenic proteins the host has previously encountered can impact the host immune response to the pathogen and may lead to cross-reactive antibodies. The ability to identify potential regions of molecular mimicry in a pathogen can illuminate immune effects which are especially important to pathogen treatment and vaccine design. Here we present Epitopedia, a software pipeline that facilitates the identification of regions that may exhibit potential three-dimensional molecular mimicry between an antigenic pathogen protein and known immune epitopes as catalogued by the Immune Epitope Database (IEDB). Epitopedia is open-source software released under the MIT license and is freely available on GitHub, including a Docker container with all other software dependencies preinstalled. We performed an analysis describing how various secondary structure states, identity between pentapeptide pairs, and identity between the parent sequences of pentapeptide pairs affects RMSD. We found that pentapeptides pairs in a helical conformation had considerably lower RMSD values than those in extended or coil conformations. We also found that RMSD is significantly increased when pentapeptide pairs are from non-homologous sequences.

INTRODUCTION

Pathogens present antigenic molecules that can elicit a host immune response. For proteins, an epitope is the portion of the antigen that is recognized and bound by an

antibody. Occasionally, pathogen epitopes may share similar chemical and structural properties to unrelated host epitopes, leading to unexpected interactions between the pathogen's epitope and host proteins (Getts et al., 2013). Molecular mimicry can also potentially lead to autoimmune disorders where infection with a pathogen can trigger the production of antibodies that mistakenly cross-react with an epitope in a host protein, potentially resulting in autoimmune complications involving both B-cell and T-cell response (Cusick et al., 2012). Alternatively, molecular mimicry between two pathogens may lead to heterologous immunity where infection with one pathogen can provide protection against other pathogens that exhibit molecularly similar antigenic proteins (Agrawal, 2019).

Epitopes can be linear or conformational. Linear epitopes consist of short local sequence stretches while conformational epitopes consist of sequence stretches across the protein sequence that come together in the 3D structure. Prediction of molecular mimicry for conformational epitopes presents a challenge, while the prediction of molecular mimicry at linear epitopes using a sequence-based approach followed by structural comparison is more straightforward. To the best of our knowledge there are currently no computational programs or pipelines readily available for the prediction of molecular mimicry of known epitopes, although programs exist to map peptides (mimotopes) onto the antigenic protein structure to identify a native epitope (Chen et al., 2012; Huang et al., 2008; Mayrose et al., 2007; Negi & Braun, 2009), to identify molecular mimicry in remote homologs (Armijos-Jaramillo et al., 2021), and to identify molecular mimicry in antibody-binding interfaces (Stebliankin et al., 2022).

We present Epitopedia, a computational pipeline for the prediction of molecular mimicry. Epitopedia identifies sequence and structural similarity between an antigenic protein of interest and experimentally verified linear epitopes found in the Immune Epitope Database (IEDB) (Vita et al., 2019). Given the structural similarity between these epitopes and the pathogenic protein, it follows that binding of the same antibody may be possible.

EPITOPEDIA IMPLEMENTATION

Internal Database Generation

Epitopedia utilizes data from the Immune Epitope Database (IEDB) (Vita et al., 2019), the Protein Data Bank (PDB) (Berman et al., 2000), and, optionally, the AlphaFold Protein Structure Database (Varadi et al., 2022) for the human proteome (Tunyasuvunakool et al., 2021). The data are organized into four internal tables (IEDB-FILT, mmcif-seqs, EPI-3D, and 3D-DSSP) stored in a SQLite3 database. IEDB-FILT is derived from a reduced IEDB that only includes the necessary data (epitope sequence, epitope identifier, antigen source sequence, range, accession, organism, etc.) for epitope mimicry search, including the full-length antigen source sequences from all assays available for T Cell, B Cell, and MHC Ligand available in IEDB. Based on the epitopes with positive assays from IEDB-FILT, a database for BLASTP (referred to as EPI-SEQ) of linear epitope sequences (mean length of 13 residues) and associated taxonomic origin of the epitopes is generated. Sequences from all PDB structures and human AlphaFold models were extracted and stored in mmcif-seqs. To find structural representatives for the antigen source sequences from IEDB, a sensitive (s=7.5) MMseqs2 (Steinegger & Söding, 2017) many-against-many

search of antigen source sequences against mmcif-seqs is performed and the results are stored in EPI-3D. For a structural representative to be included in EPI-3D, the MMseqs2 pairwise alignment between the antigen source sequence and the structure sequence must have at least 90% identity and 20% query coverage. Lastly, DSSP (Kabsch & Sander, 1983) is used to determine secondary structure and compute the accessible surface area (ASA) for every residue in each chain in EPI-3D and the results are stored in 3D-DSSP.

Searching for 1-Dimensional Molecular Mimics

The Epitopedia pipeline is executed with one or more PDB IDs as input. The protein sequence (seqres) is extracted from the input structure and used in a BLASTP search against EPI-SEQ. The BLASTP parameters evalue and max_target_seq are both set to 2,000,000 to avoid discarding hits due to large evalues or reaching the match limit, respectively. The BLAST hits are filtered to only include hits with regions containing 5 or more consecutive, identical amino acids between the query (input protein based on the PDB ID input) and subject (epitope). If a hit meets this requirement in more than one region, the regions are split into subalignments so that one epitope may have >1 region.

Further, to be considered molecular mimics, the regions must have at least 3 consecutive accessible amino acids with a relative accessible surface area (RASA) > 20%. Based on ASA from 3D-DSSP and the maximum allowed solvent accessibility (MaxASA) values per amino acid as defined in Wilke (Tien et al., 2013), RASA is calculated according to the equation $RASA = ASA/MaxASA$. Regions meeting these qualifications are considered one dimensional mimics (1D-mimics). Regions that do not meet the aforementioned criteria to be considered a 1D-mimic are discarded.

Identifying 3-Dimensional Molecular Mimics

For 1D-mimics where the antigen source protein containing the epitope hit is represented in EPI-3D, the structural regions of the input structure corresponding to the 1D-mimic regions are evaluated to ensure that all residues are solved. To avoid missing potential mimics due to regions of missing electron density in an input structure, several structures can simultaneously be used as an input. Further, providing multiple PDB IDs for the same protein as input allows for a conformational ensemble approach to search for structural mimics. The structural fragments of 1D-mimics represented in EPI-3D and the corresponding hit fragment from the input structure are extracted. To compliment structural representation of human antigen source proteins in PDB, structural fragments can also be extracted from AlphaFold2 models for the human proteome (Tunyasuvunakool et al., 2021). Although AlphaFold2 models are used, we refer to them as AlphaFold models from here on after.

TM-align (Zhang & Skolnick, 2005) is used to evaluate the structural similarity based on the RMSD for each extracted peptide structure pair based on its BLAST hit pairwise alignment. To ensure that the structural superposition step is in agreement with the peptide pair sequence alignment, the pairwise alignment of the 100% identical 1D-mimic peptide pair is provided to TM-align. Pairs with an RMSD $\leq$ 1Å are considered three dimensional mimics (3D-mimics).

Handling Redundancy and Quantifying Results

Given the nature of epitopes and IEDB, it is common to have several overlapping epitopes where both the epitope mimic region and the antigen source sequence are identical. Internal accession numbers for all antigen source sequences in IEDB-FILT

were assigned to ensure that any two or more identical sequences will have the same internal accession number to allow for filtering of redundancy at the output stage of the pipeline.

Epitopedia outputs results in CSV, JSON, and a simple web interface. The web interface is built using Flask, Bootstrap, and NGL Viewer (Rose et al., 2018) and provides an interactive visualization of the 3D-mimic region in both the input and epitope-containing source protein. The distribution of RMSD values for the 3D-mimics is plotted as a histogram, with grey lines denoting the points of -1, 0, 1 standard deviations, respectively, and a red line denoting the hit's RMSD value amongst the distribution. he Z-score for the hit is also computed, allowing for a comparative assessment of the hit quality against other hits for a particular run. An additional score termed EpiScore is calculated by dividing the mimic length by the RMSD (length of alignment/RMSD) to emphasize the significance of longer mimics. For example, given several mimics of varying length with the same RMSD, a longer mimic would have a higher EpiScore than a shorter mimic. Further, the EpiScore can reflect a more notable hit for a longer mimic with a higher RMSD than a shorter mimic with a lower RMSD. Thus, a higher EpiScore represents a more remarkable hit.

User Customization

For each provided input structure, the following main steps allow for customization of the run. For the BLASTP search in *Step 1* (Figure 1), the user can specify a taxonomy filter for a focused search. With the taxonomy filter, epitopes from the specified taxonomic id will be excluded from the search.

For extracting potential epitope hits based on the input structure in *Step 2*, the minimum span length of an identical hit and the minimum accessibility of the hit in the input structure can be specified, with default values set to 5 and 3 residues, respectively. The user determines the cutoff for RASA, with the default set to 0.2. The sequence motifs from the epitope hits that meet span length and accessibility cutoffs are considered 1D-mimics, because although they are valid epitope hits based on the input structure, the structure of the epitope hit fragment is yet unknown. The structural fragments corresponding to the motif of each 1D-mimic are excised from the input structure.

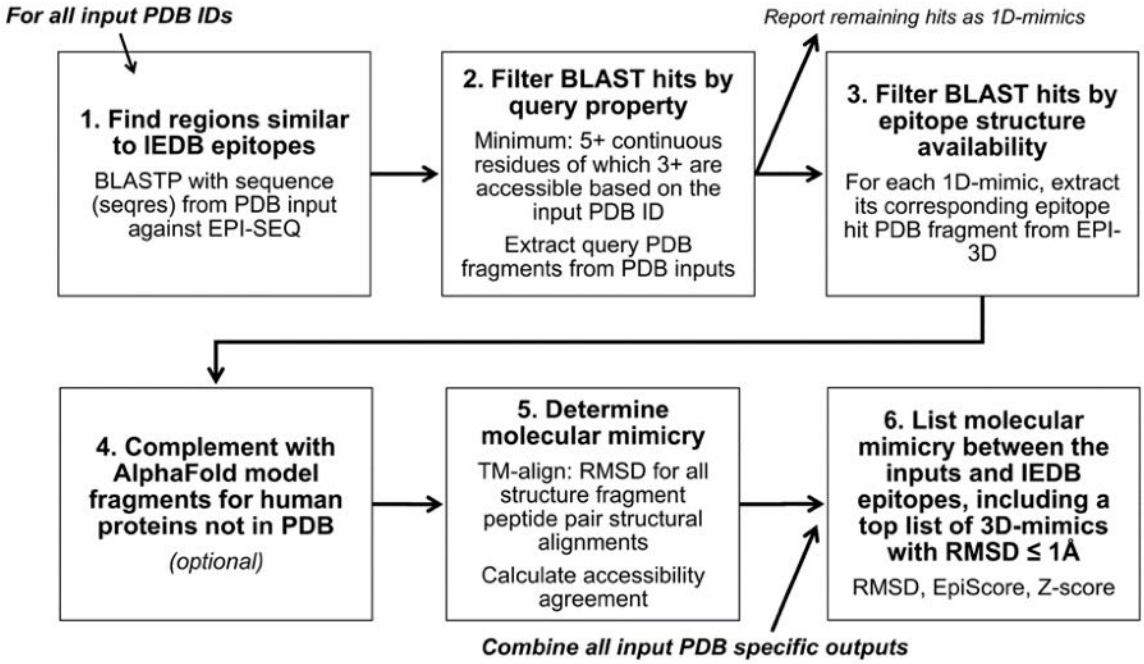


Figure 1. Overview of Epitopedia. Epitopedia is initiated with one or more PDB structures as input. In *Step 1*, a BLASTP search against linear epitope sequences in EPI-SEQ is performed with the corresponding sequence (seqres) from each PDB input as query. In *Step 2*, BLASTP hits that include sequence fragments from the query that do not contain at least 5 consecutive identical amino acids and where less than 3 amino acids are surface accessible based on the input structure are discarded. For the remaining hits, the PDB fragment is extracted from the input structure. These are considered 1D-mimics. In *Step 3*, structural fragments from the hits from EPI-SEQ that correspond to the 1D-mimics are extracted from PDB structural representatives of the source antigens. In *Step 4* (optional), for hits against epitopes in human source antigens that are not represented in

PDB, structural fragments are extracted from AlphaFold models for regions with a certain confidence level (specified by the user). In *Step 5*, TM-align is used to calculate the RMSD of the structural alignment of the BLAST hit fragment or peptide pairs. In *Step 6*, RMSD results for all fragment pairs for all inputs for the run are combined. EpiScore (length of alignment/RMSD) and RMSD histograms are generated, and Z-scores are calculated based on the whole run. A top list of fragment pairs with RMSD ≤ 1Å is created. These fragment pairs are referred to as 3D-mimics.

In *Step 3*, for epitope hits corresponding to 1D-mimics from *Step 2*, the PDB structure of their source antigen protein is extracted from EPI-3D, if such a structure exists. Fragments matching the motifs of the 1D-mimics are excised for later comparison to the corresponding motif of each 1D-mimic from the input structure. Further, accessibility of the residues in the motifs is extracted from 3D-DSSP based on the whole protein structure.

Similarly, the user can choose to extract representative structures from an AlphaFold model of the human proteome (Tunyasuvunakool et al., 2021) based on EPI-3D in *Step 4*. The user can specify the confidence level of the AlphaFold models to consider using a motif (local) and a protein (global) confidence score. Both scores are based on pLDDT, which is the primary confidence score reported for AlphaFold models (Jumper et al., 2021). For the motif confidence score (m-pLDDT), no residue within the 1D-mimic motif can be below the cutoff. For the protein confidence score (p-pLDDT), the average of pLDDT for the entire model cannot be below the cutoff. The defaults are set to 0.9 and 0.7 for m-pLDDT and p-pLDDT, respectively. Structural fragments matching the motifs of the 1D-mimics are excised for later comparison to the corresponding motif of each 1D-mimic from the input structure. Further, accessibility of the residues in the motifs is extracted from 3D-DSSP based on the whole AlphaFold model.

In *Step 5*, structural comparisons of each motif fragment from the input structure to the corresponding fragments from *Step 3* or *Step 4* are performed using TM-align for the exact pairwise sequence alignment (Zhang & Skolnick, 2005). TM-score and RMSD are reported. However, because only short structural fragments are compared, the TM-score is not meaningful, while the RMSD of the structural alignment and agreement in RASA (based on the whole structural context) are meaningful. The user can set an RMSD cutoff for hits to be reported but the default is no RMSD cutoff.

In *Step 6*, all results for all input structures are compiled into a list. The EpiScore and Z-scores are computed. Hits with RMSD of at most 1Å are considered 3D-mimics. For the 3D-mimics, a web interface output is generated. The web interface includes the settings used to execute Epitopedia and basic information about the motif in the input structure, the epitope it mimics, and the source antigen in addition to RMSD, accessibility, EpiScore, Z-scores, and a link to a visualization of the results (Figure 2). For motifs with a 3D-mimic, the best hit is shown but the other hits are included under a dropdown menu. Structural visualization of 3D-mimics highlights the location of each mimic in the input structure and in the antigen source structural representative (Figure 3).

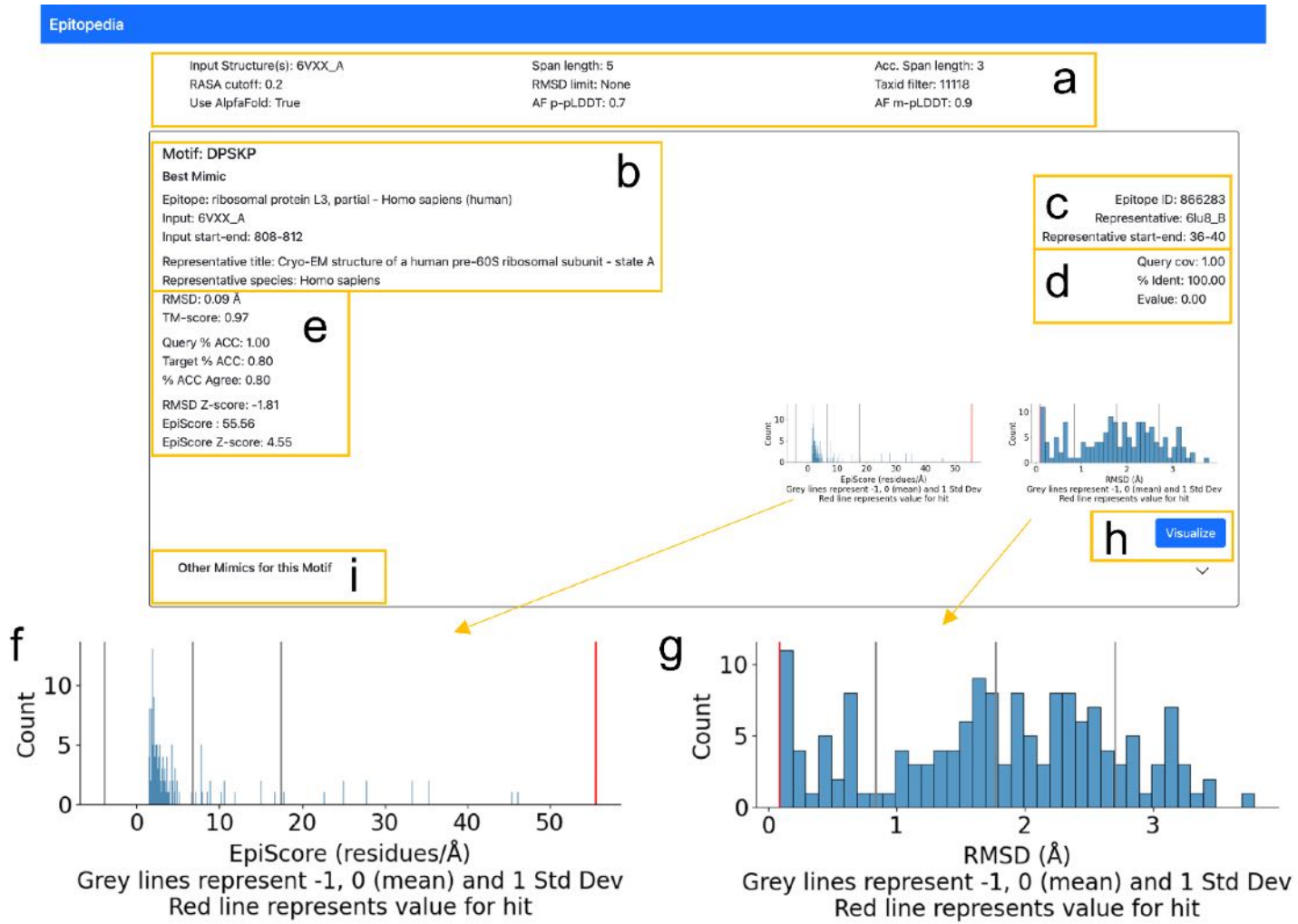


Figure 2. Overview of the Epitopedia web interface for 3D-mimics. For each run, (a) information about the run; (b) the mimic and protein in which the mimic was identified; (c) the epitope and its structural representative; (d) identification of the structural

representative with MMseqs2; (e) structural comparison of the mimics including EpiScore, EpiScore Z-Score, and RMSD Z-Score; (f) EpiScore distribution for all structurally represented mimics (blue) during the given run including the EpiScore Z-score (grey), with the current mimic in red; (g) RMSD distribution for all structurally represented mimics (blue) during the given run including the RMSD Z-score (grey), with the location of the current mimic in red; (h) link to 3D visualization of the mimic; (i) and while the Best Mimic is shown from the start, additional mimics for the same motif from the same or different proteins but with higher RMSD are included in a dropdown menu.

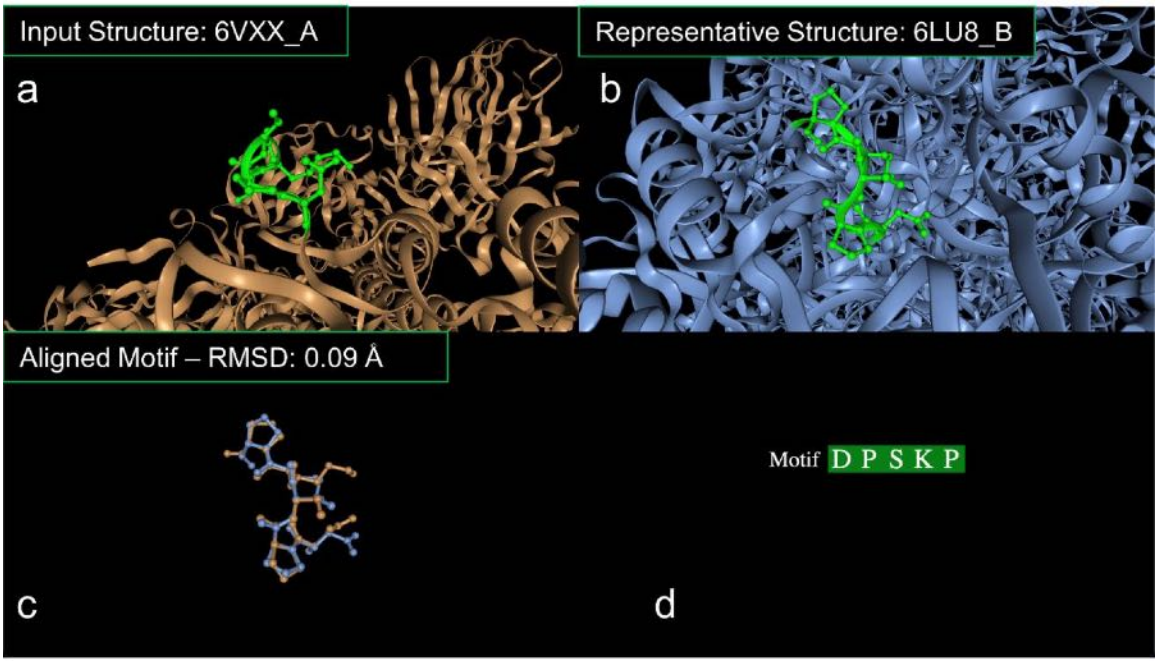


Figure 3. Visualization of the mimic pair in 3D. (a) The motif (green) shown in input protein (brown) (b) and in the structural representative protein (blue). (c) The TM-align structural superimposition for the motif in the input protein (brown) and the structural representative (blue). Panels a-c are interactive. (d) The mimic motif is interactive, hovering over a residue in the motif will highlight it in panels a-c.

EPITOPEDIA DEMONSTRATION

To demonstrate an Epitopedia run, we provide an example using an electron microscopy structure of the SARS-CoV-2 Spike protein (PDB ID: 6VXX, chain A (Walls et al., 2020)) as input (Figure 4). The taxid-filter flag with a taxid of 11118 was utilized to ensure neither the input protein nor other Coronavirus proteins were included as mimics (since these are homologous proteins and not mimics). The search for mimic

representatives was performed against both PDB and the Human AlphaFold Protein Structure Database with default settings.

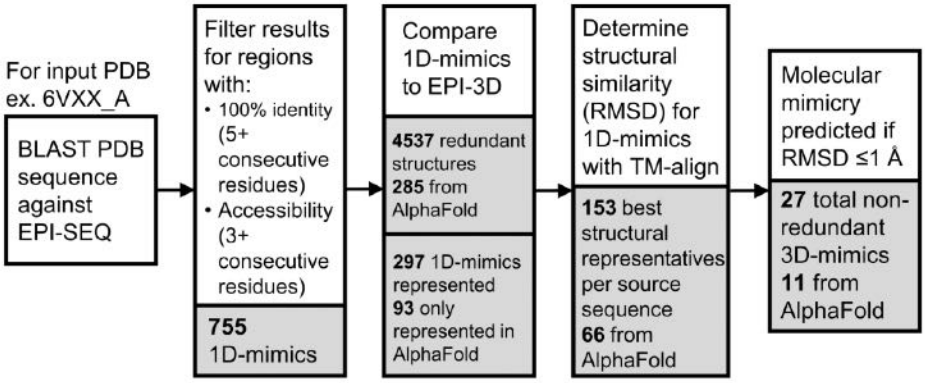


Figure 4. Epitopedia output overview using PDB 6VXX, chain A as input. For detailed output see example_output folder on the GitHub repository.

The run resulted in 755 1D-mimics, where 297 1D-mimics are structurally represented, of which 93 are only represented in the Human AlphaFold Protein Structure Database. After ensuring that only the best mimic per source sequence progresses, there were 153 mimics, with 66 of them mimicked with an AlphaFold structure. Finally, after filtering the results so that only 3D-mimics with an RMSD ≤ 1Å remain and removing redundant hits, there were 27 mimics, of which 11 are mimicked with an AlphaFold structure. Of the 16 3D-mimics from PDB, 13 are from human (such as integrin beta-1), and one each are from *Mycobacterium tuberculosis*, *Bacillus anthracis*, and Timothy grass (Appendix 1, Appendices 4-9). The remaining 11 3D-mimics are from the Human AlphaFold Protein Structure Database (Tunyasuvunakool et al., 2021; Varadi et al., 2022) and thus are all from human epitopes (Appendix 2). The mimic with the lowest RMSD (0.09 Å) is shown in Figures 2 and 3.

We also applied Epitopedia to a different SARS-CoV-2 Spike structure (PDB ID: 6XR8) and identified additional molecular mimicry with potential implications for COVID-19 (Nunez-Castilla et al., 2022).

PENTAPEPTIDE STRUCTURAL SPACE ANALYSIS

To provide guidance on how to interpret structural mimicry based on RMSD for the 3D-mimic pentapeptide fragment pairs identified by Epitopedia, we performed an investigation of RMSD for random pentapeptide pairs for the three main secondary structure states helix, extended, and coil from any sequence pair regardless of sequence similarity and for sequence pairs with low sequence similarity representing non-homologous proteins.

Methods

To understand how secondary structure state and sequence identity affect the distribution of RMSD values for pentapeptide pairs, an analysis of RMSD distributions of pentapeptide pairs across various secondary structure states and pentapeptide sequence identity levels was performed.

All possible pentapeptides based on PDB structures were generated and annotated with a DSSP secondary structure state reduction based on 3D-DSSP. The DSSP state reduction was performed such that if all residues in a pentapeptide were classified as turn (T), bend (S) or none (-), the pentapeptide was labeled coil, if all residues were strand (E) or beta-bridge (B) the pentapeptide was labeled extended, and if all residues were alpha helix (H), 3-10 helix (G), or pi-helix (I) the pentapeptide was labeled helix. Any pentapeptides that did not fit into one of these 3 categories were discarded.

Around 1,000 pentapeptide pairs (Appendix 3) were generated for each secondary structure state per identity level (0%, 20%, 40%, 60%, 80%, and 100%) from the labeled pentapeptide database described above. The number of pentapeptide pairs per category is not exactly the same across all categories because matches of a pentapeptide against itself (same PDB ID) are discarded. The pentapeptide regions were extracted from the parent structures using GEMMI (*GitHub - Project-Gemmi/Gemmi: Macromolecular Crystallography Library and Utilities*, n.d.) and superposed using TM-align (Zhang & Skolnick, 2005), with a fixed alignment as described for the Epitopedia implementation above.

To reduce the influence that parent sequence homology may have on the above analysis, we performed a similar analysis starting with 2,000 pentapeptides for each secondary structure state per identity level. Here, an added filtering step was performed to ensure that the parent sequences of the pentapeptide pairs were no more than 30% identical according to a local pairwise Smith-Waterman alignment of the parent sequences generated with EMBOSS Water (Madeira et al., 2019). Pentapeptide matches where the identity filter could not be enforced were discarded, thus, the number of pentapeptide pairs per category is not exactly the same across categories. For instance, if a query pentapeptide had been paired with over 100 other pentapeptides to generate a pentapeptide pair, yet a pentapeptide pair with a parent sequence identity of less than 30% was not found, the query pentapeptide was discarded. This scenario disproportionately affected pentapeptide pairs with higher pentapeptide identity, as there is a lower chance of parent sequences having less than 30% identity as the pentapeptide

pair identity increases. In total, all pentapeptide identity and secondary structure combinations have greater than 900 pentapeptide pairs (Appendix 3).

Statistical comparisons were performed with Mann Whitney U using SciPy (Virtanen et al., 2020). Alpha values were corrected for multiple comparisons using simple Bonferroni correction. For a confidence level of 99%:

$$corrected\ alpha = \frac{0.01}{N\ pairwise\ comparisons}$$

Results

An analysis was performed to better understand how the RMSD distribution for pentapeptides pairs varies with differing pentapeptide pair sequence identity, parent sequence identity and secondary structure state. For the first analysis that did not consider the percent identity of the parent sequences for a pentapeptide pair, a decrease in the median RMSD is observed at the 100% identity levels (Figure 5, Table 1). For helix pentapeptide pairs, the median RMSD for the 0% to 80% pentapeptide identity levels is 0.20-0.22Å, while at the 100% identity level the median is 0.13Å, which is a significant decrease when compared to all other identity levels for the helical state (Table 2). For extended pentapeptide pairs, the median RMSD for the 0% to 80% pentapeptide identity levels is 0.69-0.84Å, while at the 100% identity level the median is 0.14Å. This is a significant decrease when compared to all other identity levels for the extended state (Table 2). Lastly, for coil pentapeptide pairs, the median RMSD for the 0% to 80% pentapeptide identity levels is 1.79-1.95Å, while at the 100% identity level the median is 0.31Å. This large decrease of ~1.5Å is significant when compared to all other identity levels for coil (Table 2).

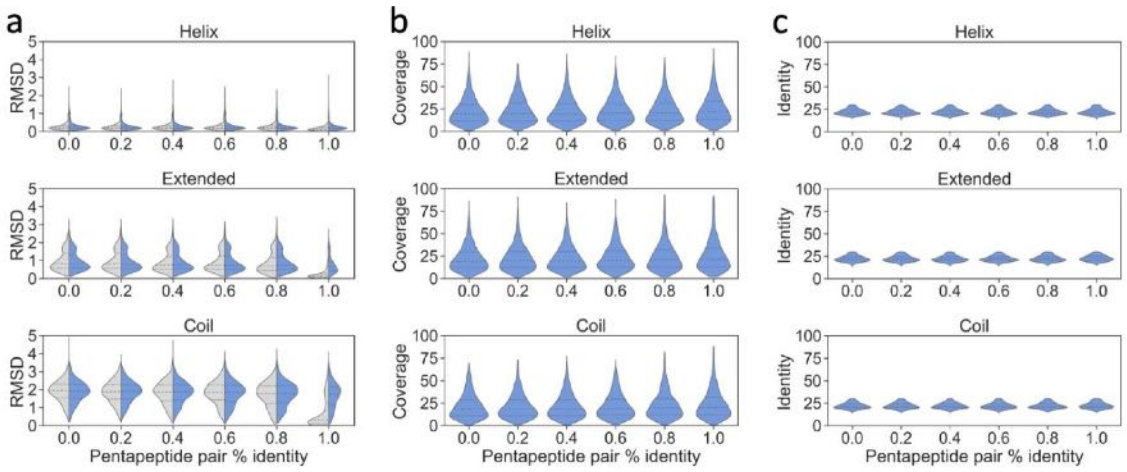


Figure 5. (a) Violin plots of the resulting RMSD distribution from pentapeptide structure analysis. The distributions for the analysis without the 30% parent sequence identity filter are shown in grey while the corresponding distributions for the pentapeptides from the 30% parent sequence identity set are shown in blue**.** (b) Violin plots showing the distribution of query coverage between the parent sequences for pentapeptide pairs at various identity levels and secondary structure categories. (c) Violin plots showing the distribution of pairwise identity between the parent sequences for pentapeptide pairs at various identity levels and secondary structure categories.

Table 1. Median RMSD values resulting from RMSD distribution for structural space analysis of pentapeptide pairs of various identity levels and secondary structural categories shown in Figure 5.

Pentapeptide % Identity	Median RMSD (Å)					
	No Parent Identity Filter			30% Parent Identity Filter		
	Helix	Extended	Coil	Helix	Extended	Coil
0	0.22	0.84	1.95	0.22	0.83	1.91
20	0.21	0.82	1.88	0.21	0.80	1.90
40	0.21	0.75	1.87	0.21	0.77	1.88
60	0.20	0.74	1.85	0.21	0.74	1.85
80	0.21	0.69	1.79	0.21	0.76	1.79
100	0.13	0.14	0.31	0.20	0.63	1.74

Table 2. Comparisons across pentapeptide identity levels for the same structural class for the no filter and 30% filter datasets, respectively. Significantly different comparisons are shown in blue.[1]

Pentapeptide % Identity		No Parent Identity Filter			30% Parent Identity Filter		
		Helix	Extended	Coil	Helix	Extended	Coil
0	20	0.197694	0.578304	0.070875	0.084507	0.107303	0.464566
0	40	0.377122	0.000931	0.001582	0.212098	0.002897	0.013426
0	60	0.076141	0.000368	0.000528	0.15162	1.64E-06*	7.41E-06*
0	80	0.313838	3.52E-12*	5.1E-09*	0.0284	9.37E-08*	1.97E-10*
0	100	8.4E-59*	1.3E-212*	5.5E-156*	5.63E-06*	2.77E-31*	3.33E-17*
20	40	0.688416	0.006825	0.145049	0.636335	0.170195	0.083497
20	60	0.634811	0.00257	0.092602	0.779538	0.001278	0.000172
20	80	0.791704	1.43E-10*	2.84E-05*	0.639867	0.000168	1.84E-08*
20	100	3.63E-52*	2.8E-210*	2.3E-149*	0.002963	4.31E-25*	6.02E-15*
40	60	0.362381	0.785088	0.797398	0.838454	0.06416	0.038889
40	80	0.891271	7.14E-05*	0.006063	0.35334	0.016001	8.83E-05*
40	100	1.27E-54*	4.5E-203*	1.9E-141*	0.00072	2.59E-20*	2.54E-10*
60	80	0.457191	0.000238	0.011388	0.457478	0.571466	0.067424
60	100	1.06E-50*	7.3E-200*	2.5E-140*	0.001357	1.51E-14*	6.59E-06*
80	100	3.08E-53*	2.2E-174*	1.5E-124*	0.01095	5.00E-13*	0.003258

[1] Based on simplified Bonferroni correction at 99% confidence level, corrected alpha = 0.000111.

For the follow-up analysis we enforced a 30% parent sequence identity filter to better resemble molecular mimics from unrelated protein pairs. To ensure that the pentapeptide pairs were from proteins that are not closely related, we performed local alignments and extracted the percent sequence identity and query cover for each parent sequence pair. By design, no parent sequence pair has a pairwise sequence identity above 30%, with a median around 20% (Figure 5). The query cover for the parent sequence pair alignments is low, with a median of around 20% (Figure 5). For these pairwise sequence alignments with 20% sequence identity and query cover, we can assume that these are primarily non-homologous parent sequence pairs although some remote homologs may be included in this dataset.

For the pentapeptide pairs from these non-homologous sequence pairs, the sharp decrease in the median RMSD at the 100% pentapeptide identity level has faded for extended and coil conformations (Figure 5). For helix pentapeptide pairs, the median RMSD at the 0% to 100% pentapeptide identity level is 0.20-0.22Å. Only the 100% vs 0% identity level comparison yields significant difference for the helix pentapeptide pairs (Table 2). For extended pentapeptide pairs, the median RMSD at the 0% to 100% pentapeptide identity level is 0.63-0.83Å. For coil pentapeptide pairs, the median RMSD at the 0% to 100% pentapeptide identity level is 1.74-1.91Å. For extended and coil pentapeptide pairs, the 100% identity level is significantly different when compared against every other identity level except for one comparison, 80% vs 100% in the coil state (Table 2).

When comparing the same identity level for the pentapeptide pairs across the set with no parent sequence identity filter and the set with the 30% sequence identity filter,

we found that the pairwise parent sequence identity has an impact on the RMSD for identical pentapeptide pairs in the helical state, but not for the less identical peptide pairs (Table 3). This pattern is shared for the coil state, but for the extended state, the pairwise parent sequence identity seems to impact RMSD for identical and 80% identical pentapeptides (Table 3).

Table 3. Comparisons between pentapeptide identity levels for the same structural class for the 30% filter vs the no filter dataset. Significantly different comparisons are shown in blue.[1]

Pentapeptide % Identity		30% vs no Parent Identity Filter		
		Helix	Extended	Coil
0	0	0.438284984	0.56859737	0.1896511
20	20	0.381648164	0.255207316	0.7826223
40	40	0.445432037	0.339928559	0.5969043
60	60	0.091388265	0.844626827	0.4058797
80	80	0.856800634	0.000269319*	0.5128634
100	100	3.41E-55*	4.83E-171*	1.93E-131*

[1]Based on simplified Bonferroni correction at 99% confidence level, corrected alpha = 0.000555556.

Altogether, this analysis shows that for pentapeptides, the secondary structure state is important to consider when identifying molecular mimics using RMSD for random proteins. We used TM-align to calculate RMSD and this method, like many others, calculates RMSD based on the spatial coordinates for C-alpha in each amino acid residue. Our observation that pentapeptide pairs in a helical state have lower RMSD is not surprising given the regular geometry of the α-helix. For identical pentapeptide pairs in extended and coil conformations, the median RMSD for the non-homologous parent sequences are 0.63Å and 1.74Å, respectively, compared to 0.20Å for helix (Table 1).

Guidelines

Our interpretation, as far as molecular mimicry goes, is that mimics with identical sequences in α-helices are likely to appear very similar if they are oriented the same way in their parent proteins. As such, they are likely to be able to participate in similar interactions with, for example, an antibody. Mimics with identical sequences with low RMSDs, approaching the median RMSD of the unfiltered set (Table 1), are likely to present a similar interaction interface, if oriented similarly. A pentapeptide in a helix, given its winding structure, is relatively short while a pentapeptide in the extended or coil conformation may present a larger accessible area.

Pathogen proteins that mimic known epitopes in antigenic proteins may stimulate the production of cross-reactive antibodies that can interact with the pathogen protein as well as the human antigen. Pathogen proteins that mimic known epitopes in other pathogens may trigger an immune memory that could lead to protective immunity or complex immune effects such as anti-body dependent enhancement.

CONSLUSION

Here, we have developed Epitopedia, a pipeline for the discovery of potential molecular mimics of immune epitopes found in IEDB. Importantly, Epitopedia is designed to only predict molecular mimicry for linear epitopes, that are continuous in sequence, as opposed to conformational epitopes, that are discontinuous in sequence and come together in three-dimensional space. As such, molecular mimics found in conformational epitopes cannot be identified using our approach. Additionally, Epitopedia is reliant on publicly available data found in both IEDB and PDB and cannot

predict instances of molecular mimicry *de novo*. Molecular mimics that are not yet found in IEDB or PDB will not be identified by Epitopedia. Furthermore, relying on public databases can lead to biased results because proteins with greater perceived relevance (e.g. those involved in more common human diseases) are more likely to be well-studied and thus have functional and structural information deposited in these databases, while other proteins remain underrepresented. PDB is also biased towards proteins that lack intrinsic disorder and the more stable conformation of a dynamic protein. Therefore, Epitopedia may not predict molecular mimics in conformationally flexible regions. Importantly, results produced by Epitopedia are only predictions, subject to both false positives and negatives. It is critical to further investigate this output with both literature searches and experimental validation.

Epitopedia can facilitate our understanding of how pathogens may interfere with the known epitopes from the human proteome and also known epitopes from other species. Epitopes shared between pathogens can impact immune responses for secondary infections and identification of mimics of epitopes can provide insights to the mechanism behind the widely differing clinical manifestations and complications of infection with certain pathogens, such as SARS-CoV-2. Identification of molecular mimicry between known epitopes from the human proteome and a human pathogen protein can provide clues to the autoimmune potential of an infection caused by the pathogen. Further, by pinpointing regions in the pathogen's proteome that may cause an autoimmune response if a cross-reactive antibody is created against it, these regions can be avoided in future vaccine design. Lastly, by highlighting which human proteins may be at risk for autoimmune targeting in response to a pathogen infection, therapeutics to counteract

autoimmune effects can be used (or developed). Epitopedia provides a starting point for generating a better understanding of the autoimmune potential of pathogens and can benefit large-scale data mining and experimental *in-vitro* and *in-vivo* design to solve autoimmune conundrums in infectious disease.

ACKNOWLEDGMENTS

We thank Teresa Liberatore for discussions. This work was partially supported by the National Science Foundation under Grant No. 2037374.

DATA AND CODE AVAILABILITY

Epitopedia is primarily written in Python and relies on established software and databases. Epitopedia is available at https://github.com/cbalbin-bio/Epitopedia under the opensource MIT license and also as a docker container at https://hub.docker.com/r/cbalbin/epitopedia.

LITERATURE CITED

Agrawal, B. (2019). Heterologous Immunity: Role in Natural and Vaccine-Induced Resistance to Infections. *Frontiers in Immunology*, *10*, 2631. https://doi.org/10.3389/FIMMU.2019.02631

Armijos-Jaramillo, V., Espinosa, N., Vizcaıno, K., & Santander-Gordon, D. (2021). A Novel In Silico Method for Molecular Mimicry Detection Finds a Formin with the Potential to Manipulate the Maize Cell Cytoskeleton. *Molecular Plant-Microbe Interactions: MPMI*, *34*(7), 815–825. https://doi.org/10.1094/MPMI-11-20-0332-R

Chen, W., Guo, W. W., Huang, Y., & Ma, Z. (2012). Pepmapper: A collaborative web tool for mapping epitopes from affinity-selected peptides. *PLoS ONE*, *7*(5), 37869. https://doi.org/10.1371/journal.pone.0037869

Cusick, M. F., Libbey, J. E., & Fujinami, R. S. (2012). Molecular mimicry as a mechanism of autoimmune disease. *Clinical Reviews in Allergy and Immunology*, *42*(1), 102–111. https://doi.org/10.1007/s12016-011-8294-7

GitHub - project-gemmi/gemmi: macromolecular crystallography library and utilities. (n.d.). Retrieved February 16, 2022, from https://github.com/project-gemmi/gemmi

Huang, Y. X., Bao, Y. L., Guo, S. Y., Wang, Y., Zhou, C. G., & Li, Y. X. (2008). Pep-3D-Search: A method for B-cell epitope prediction based on mimotope analysis. *BMC Bioinformatics*, *9*(538). https://doi.org/10.1186/1471-2105-9-538

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*, 583–589. https://doi.org/10.1038/s41586-021-03819-2

Kabsch, W., & Sander, C. (1983). Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, *22*(12), 2577–2637. https://doi.org/10.1002/bip.360221211

Madeira, F., Park, Y. M., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., Basutkar, P., Tivey, A. R. N., Potter, S. C., Finn, R. D., & Lopez, R. (2019). The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Research*, *47*(W1), W636–W641. https://doi.org/10.1093/nar/gkz268

Mayrose, I., Penn, O., Erez, E., Rubinstein, N. D., Shlomi, T., Freund, N. T., Bublil, E. M., Ruppin, E., Sharan, R., Gershoni, J. M., Martz, E., & Pupko, T. (2007). Pepitope: Epitope mapping from affinity-selected peptides. *Bioinformatics*, *23*(23), 3244–3246. https://doi.org/10.1093/bioinformatics/btm493

Negi, S. S., & Braun, W. (2009). Automated detection of conformational epitopes using phage display peptide sequences. *Bioinformatics and Biology Insights*, *2009*(3), 71–81. https://doi.org/10.4137/bbi.s2745

Nunez-Castilla, J., Stebliankin, V., Baral, P., Balbin, C. A., Sobhan, M., Cickovski, T., Mondal, A. M., Narasimhan, G., Chapagain, P., Mathee, K., & Siltberg-Liberles, J. (2022). Potential Autoimmunity Resulting from Molecular Mimicry between SARS-CoV-2 Spike and Human Proteins. *Viruses*, *14*(7), 1415. https://doi.org/10.3390/V14071415

Rose, A. S., Bradley, A. R., Valasatava, Y., Duarte, J. M., Prlic, A., & Rose, P. W. (2018). NGL viewer: web-based molecular graphics for large complexes.

Bioinformatics, *34*(21), 3755–3758. https://doi.org/10.1093/BIOINFORMATICS/BTY419

Stebliankin, V., Baral, P., Balbin, C., Nunez-Castilla, J., Sobhan, M., Cickovski, T., Mohan Mondal, A., Siltberg-Liberles, J., Chapagain, P., Mathee, K., & Narasimhan, G. (2022). EMoMiS: A Pipeline for Epitope-based Molecular Mimicry Search in Protein Structures with Applications to SARS-CoV-2. *BioRxiv*. https://doi.org/10.1101/2022.02.05.479274

Steinegger, M., & Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, *35*(11), 1026–1028. https://doi.org/10.1038/nbt.3988

Tien, M. Z., Meyer, A. G., Sydykova, D. K., Spielman, S. J., & Wilke, C. O. (2013). Maximum allowed solvent accessibilites of residues in proteins. *PLoS ONE*, *8*(11), e80635. https://doi.org/10.1371/journal.pone.0080635

Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Žídek, A., Bridgland, A., Cowie, A., Meyer, C., Laydon, A., Velankar, S., Kleywegt, G. J., Bateman, A., Evans, R., Pritzel, A., Figurnov, M., Ronneberger, O., Bates, R., Kohl, S. A. A., … Hassabis, D. (2021). Highly accurate protein structure prediction for the human proteome. *Nature*, *596*, 590–596. https://doi.org/10.1038/s41586-021-03828-1

Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., Žídek, A., Green, T., Tunyasuvunakool, K., Petersen, S., Jumper, J., Clancy, E., Green, R., Vora, A., Lutfi, M., … Velankar, S. (2022). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, *50*(D1), D439–D444. https://doi.org/10.1093/NAR/GKAB1061

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., … Vázquez-Baeza, Y. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2

Vita, R., Mahajan, S., Overton, J. A., Dhanda, S. K., Martini, S., Cantrell, J. R., Wheeler, D. K., Sette, A., & Peters, B. (2019). The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Research*, *47*(D1), D339–D343. https://doi.org/10.1093/nar/gky1006

Walls, A. C., Park, Y.-J., Tortorici, M. A., Wall, A., McGuire, A. T., & Veesler, D. (2020). Structure, Function, and Antigenicity of the SARS-CoV-2 Spike

Glycoprotein. *Cell*, *181*(2), 281-292.e6. https://doi.org/10.1016/J.CELL.2020.02.058

Zhang, Y., & Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research*, *33*(7), 2302–2309. https://doi.org/10.1093/NAR/GKI524

Appendices

Appendix 1. 3D-mimics found for SARS-CoV-2 Spike (PDB id 6VXX_A)

Motif	Protein	Species	RMSD	RMSD Z-Score	EpiScore	PDB_chain
DPSKP	60S ribosomal protein L3	Human	0.09 Å	-1.81	55.56	6LU8_B
DPSKP	Alanine and proline-rich secreted protein apa precursor	*Mycobacterium tuberculosis*	0.22 Å	-1.67	22.73	5ZX9_A
LPDPS	BRCA1-A complex subunit BRE	Human	0.18 Å	-1.71	27.78	6GVW_C
EHVNN	Casein kinase 2 alpha isoform	Human	0.30 Å	-1.58	16.67	2ZJW_A
NLLLQ	DNA polymerase subunit gamma 1	Human	0.42 Å	-1.45	11.90	5C51_A
LLQYG	Ankyrin 1	Human	0.49 Å	-1.38	10.20	1N11_A
GEVFN	Integrin beta 1	Human	0.56 Å	-1.30	8.93	7NWL_B
QEVFA	Lethal factor precursor	Anthrax	0.59 Å	-1.27	8.47	6ZXL_H
DPFLG	NAD-dependent deacetylase sirtuin-2	Human	0.64 Å	-1.22	7.81	5D7P_B
KIADY	Nucleoporin NUP188	Human	0.64 Å	-1.22	7.81	5IJO_J
IDGYF	lanosterol 14-alpha demethylase	Human	0.64 Å	-1.22	7.81	4UHI_A
PFLGV	CTP synthase 1	Human	0.64 Å	-1.22	7.81	7MH0_B
FTVEKG	Pollen allergen Phl p 2	Timothy grass	0.67 Å	-1.18	8.96	1WHP_A
HAPAT	Activator of 90 kDa heat shock protein ATPase homolog 1	Human	0.76 Å	-1.09	6.58	1X53_A

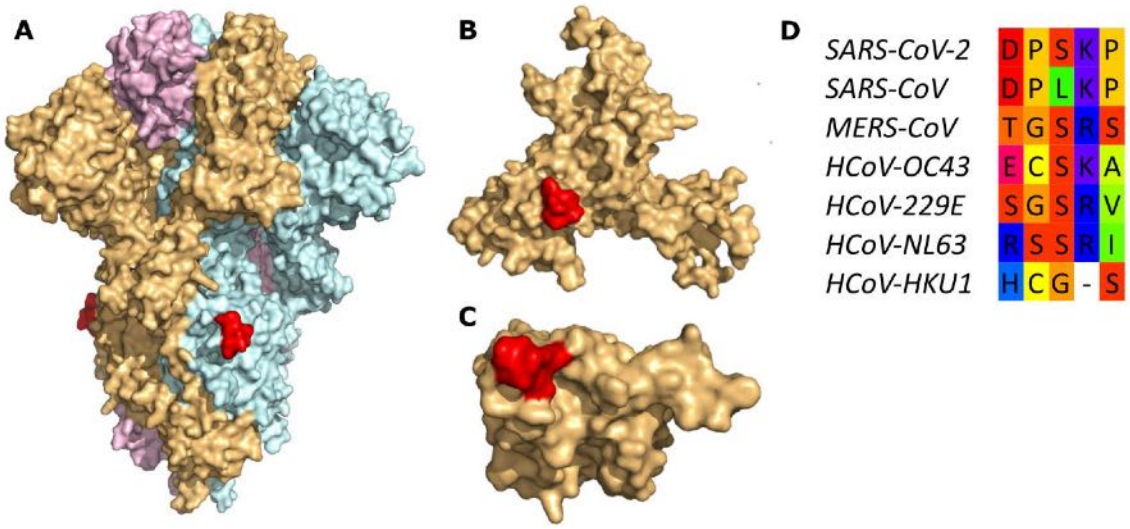
STASAL	40S ribosomal protein S13	Human	0.84 Å	-1.00	7.14	6G5I_N
PPEAE	Integrin alpha-5	Human	0.96 Å	-0.87	5.21	7NXD_A

Appendix 2. Human AF-3D-mimics for SARS-CoV-2 Spike

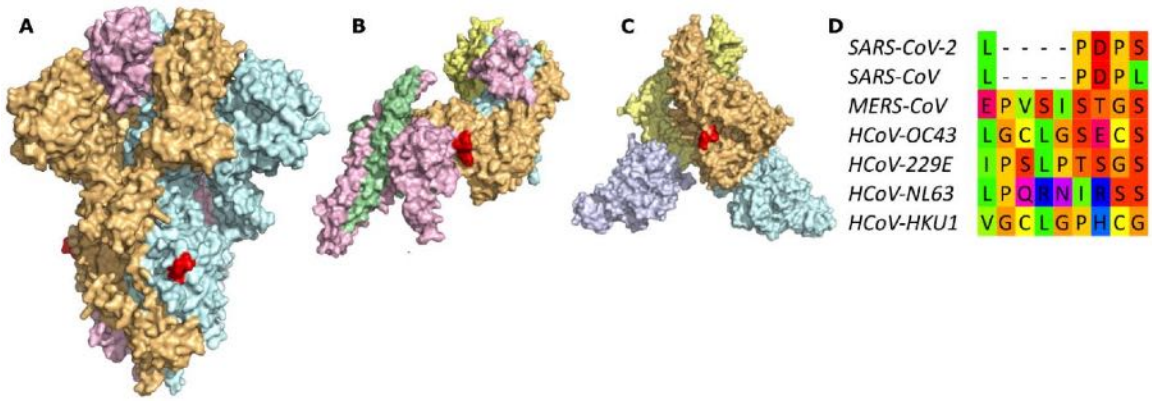
Motif	Protein	RMSD	RMSD Z-Score	EpiScore	AlphaFold2 ID
TGIAV	Phosphofructokinase	0.09 Å	-1.81	55.56	AF-P17858-F1-model_v1_A
TGIAV	Low affinity immunoglobulin gamma Fc region receptor II-b	0.11 Å	-1.78	45.45	AF-P31995-F1-model_v1_A
KIQDSL	Phosphorylase b kinase regulatory subunit beta	0.13 Å	-1.76	46.15	AF-Q93100-F1-model_v1_A
KIQDSL	Long-chain-fatty-acid-CoA ligase 5	0.40 Å	-1.47	15.00	AF-Q9ULC5-F1-model_v1_A
VYDPL	Actin-binding protein IPP	0.15 Å	-1.74	33.33	AF-Q9Y573-F1-model_v1_A
SAIGKI	Ran-GTP binding protein	0.17 Å	-1.72	35.29	AF-O60518-F1-model_v1_A
LPDPS	Semaphorin 7a	0.63 Å	-1.23	7.94	AF-O75326-F1-model_v1_A
VLYNS	U2 snRNP-associated SURP motif-containing protein	0.20 Å	-1.69	25.00	AF-O15042-F1-model_v1_A
KLPDD	F-box only protein 3	0.28 Å	-1.60	17.86	AF-Q9UK99-F1-model_v1_A
NLLLQ	Ankyrin 3	0.47 Å	-1.40	10.64	AF-Q12955-F1-model_v1_A
DNTFV	N-acetylgalactosamine-6-sulfatase	0.47 Å	-1.40	10.64	AF-P34059-F1-model_v1_A

Appendix 3. Number of pentapeptide pairs per pentapeptide identity / secondary structure category for both analyses (with and without parent sequence identity filter).

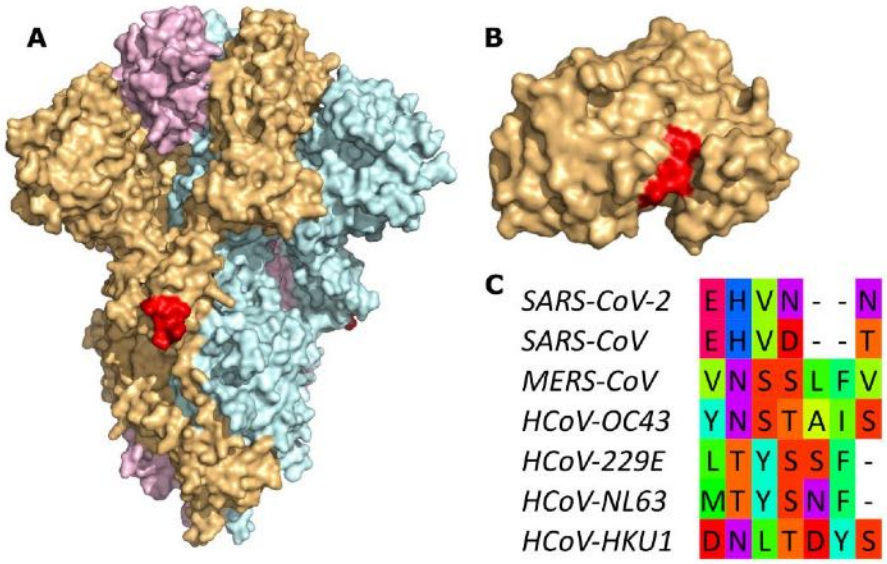
Pentapeptide % Identity	Number of Pentapeptide Pairs					
	No Parent Identity Filter			30% Parent Identity Filter		
	Helix	Extended	Coil	Helix	Extended	Coil
0	962	962	954	1999	2000	1999
20	964	967	955	1999	2000	1999
40	967	966	954	1999	1999	1999
60	965	969	953	1999	2000	1999
80	966	966	953	1999	1992	1998
100	944	925	903	1508	921	1236



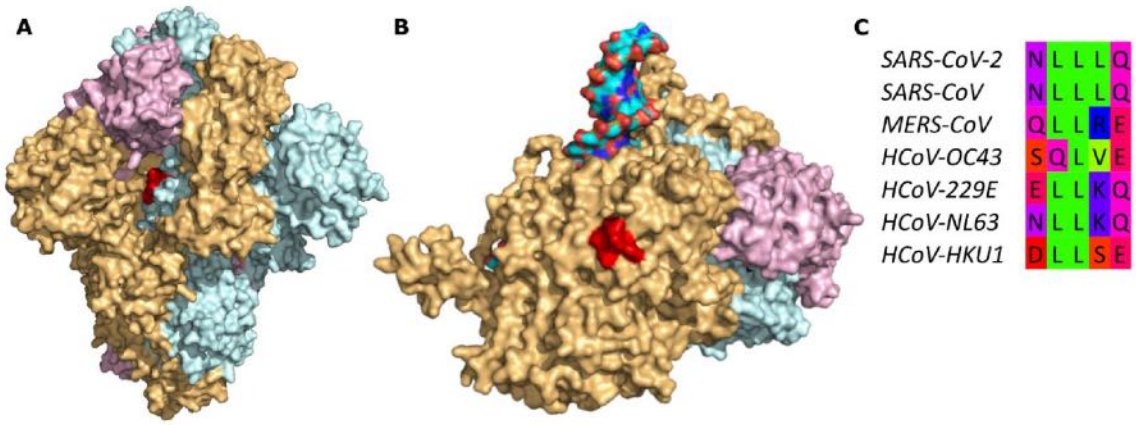
Appendix 4. The molecular mimicry motif DPSKP (red) from Spike (A, colored by chain) matches ribosomal protein L3 (B, beige) from *Homo sapiens* with an RMSD of 0.09 Å and alanine and proline-rich secreted protein apa precursor (C, beige) from *Mycobacterium tuberculosis* with an RMSD of 0.22 Å. The motif is not conserved in human betacoronaviruses (D). Protein structures visualized can be found in Appendix 1. Sequences for human betacoronavirus Spike proteins were aligned using MAFFT. The molecular mimicry motif region was extracted from the alignment according to Appendix 1. Accessions for the sequences in order of appearance are: YP_009724390, YP_009825051, YP_009047204, YP_009555241, NP_073551, YP_003767, YP_173238.



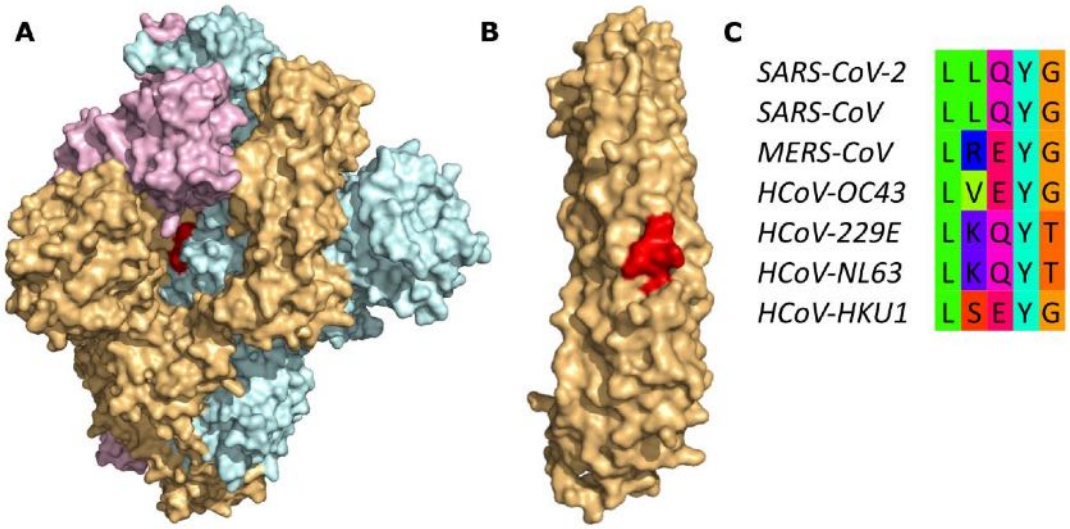
Appendix 5. The molecular mimicry motif LPDPS (red) from Spike (A, colored by chain) matches BRCA1-A complex subunit BRE (B, colored by chain) from *Homo sapiens* with an RMSD of 0.18 Å and semaphorin-7A (C, colored by chain) from *Homo sapiens* with an RMSD of 0.66 Å. The motif is not conserved in human betacoronaviruses (D). For details, see legend of Appendix 4.



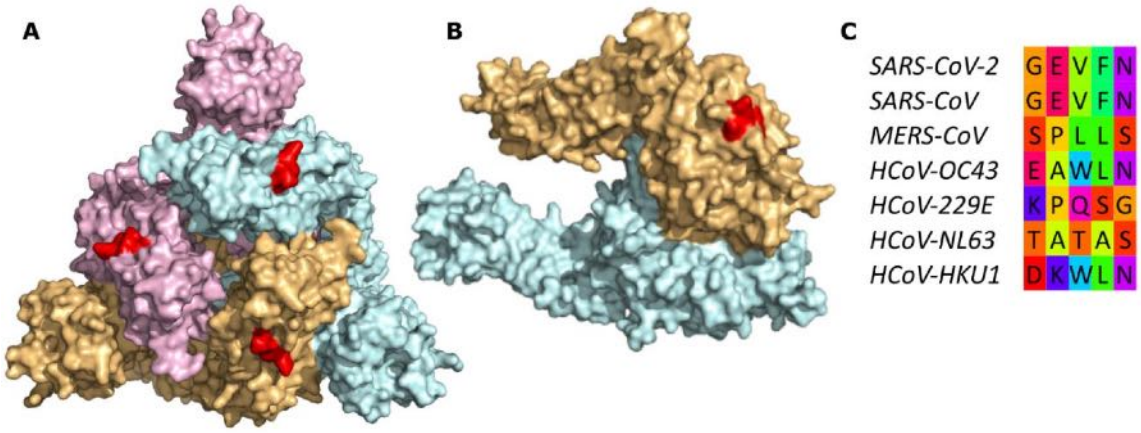
Appendix 6. The molecular mimicry motif EHVNN (red) from Spike (A, colored by chain) matches casein kinase 2 alpha isoform (B, beige) from Homo sapiens with an RMSD of 0.30 Å. The motif is not conserved in human betacoronaviruses (C). For details, see legend of Appendix 4.



Appendix 7. The molecular mimicry motif NLLLQ (red) from Spike (A, colored by chain) matches DNA polymerase subunit gamma-1 (B, colored by chain, with DNA colored by element) from *Homo sapiens* with an RMSD of 0.42 Å. The motif is semi-conserved in human betacoronaviruses (C). For details, see legend of Appendix 4.



Appendix 8. The molecular mimicry motif LLQYG (red) from Spike (A, colored by chain) matches ankyrin-1 (B, beige) from *Homo sapiens* with an RMSD of 0.49 Å. The motif is semi-conserved in human betacoronaviruses (C). For details, see legend of Appendix 4.



Appendix 9. The molecular mimicry motif GEVFN (red) from Spike (A, colored by chain) matches integrin beta-1 (B, colored by chain) from *Homo sapiens* with an RMSD of 0.67 Å. The motif is not conserved in human betacoronaviruses (C). For details, see legend of Appendix 4.

CHAPTER V

POTENTIAL AUTOIMMUNITY RESULTING FROM MOLECULAR MIMICRY BETWEEN SARS-COV-2 SPIKE AND HUMAN PROTEINS

ABSTRACT

Molecular mimicry between viral antigens and host proteins can produce cross-reacting antibodies leading to autoimmunity. The coronavirus SARS-CoV-2 causes COVID-19, a disease curiously resulting in varied symptoms and outcomes, ranging from asymptomatic to fatal. Autoimmunity due to cross-reacting antibodies resulting from molecular mimicry between viral antigens and host proteins may provide an explanation. Thus, we computationally investigated molecular mimicry between SARS-CoV-2 Spike and known epitopes. We discovered molecular mimicry hotspots in Spike and highlight two examples with tentative high autoimmune potential and implications for understanding COVID-19 complications. We show that a TQLPP motif in Spike and thrombopoietin shares similar antibody binding properties. Antibodies cross-reacting with thrombopoietin may induce thrombocytopenia, a condition observed in COVID-19 patients. Another motif, ELDKY, is shared in multiple human proteins, such as PRKG1 involved in platelet activation and calcium regulation, and tropomyosin, which is linked to cardiac disease. Antibodies cross-reacting with PRKG1 and tropomyosin may cause known COVID-19 complications such as blood-clotting disorders and cardiac disease, respectively. Our findings illuminate COVID-19 pathogenesis and highlight the importance of considering autoimmune potential when developing therapeutic interventions to reduce adverse reactions.

INTRODUCTION

The coronavirus SARS-CoV-2 is the causative agent of the COVID-19 pandemic. COVID-19 is an infectious disease whose typical symptoms include fever, cough,

shortness of breath [1,2], and loss of taste or smell [3]. Curiously, despite over half a billion confirmed cases worldwide [4], roughly one-third are estimated to be asymptomatic [5]. Yet, other SARS-CoV-2 infected individuals may also experience a variety of disease-related complications including liver injury [6], kidney injury [7], and cardiovascular complications including myocarditis, heart failure, thrombosis [8], and thrombocytopenia [9]. COVID-19 can trigger a range of antibody response levels [10] and an enrichment in autoantibodies that react with human proteins has been found for patients with severe disease [11]. While the reason for the variety of disease severity affecting people with COVID-19 is not well understood, molecular mimicry may provide an avenue for explanations.

Molecular mimicry occurs when unrelated proteins share regions of high molecular similarity, such that they can perform similar and unexpected interactions with other proteins. When molecular mimicry involves antigens to which antibodies are produced, cross-reactive antibodies can result. Molecular mimicry between pathogen antigens and human proteins can cause an autoimmune response, where antibodies against the pathogen erroneously interact with human proteins, sometimes leading to transient or chronic autoimmune disorders [12]. Alternatively, molecular mimicry could be viewed through the lens of heterologous immunity, where previous exposure to one pathogen antigen can result in short- or long-term complete or partial immunity to another pathogen with a similar antigen [13]. Moreover, antigen-driven molecular mimicry can also lead to cross-reactive antibody immunity which has been reported against SARS-CoV-2 for uninfected individuals [14].

The SARS-CoV-2 Spike protein is responsible for enabling SARS-CoV-2 entry into host cells [15]. Spike protrudes from the virus surface and is one of the main antigenic proteins for this virus [16]. Additionally, Spike is the primary component in the vaccines against SARS-CoV-2. Consequently, mimicry between Spike and human proteins or Spike and other human pathogens can result in cross-reactive antibodies in response to SARS-CoV-2 infection or vaccination. Cross-reactive antibodies may yield complex outcomes such as diverse symptoms with varying severity across populations and developmental stages as observed for COVID-19. It must be noted that there are a variety of genetic and environmental factors that contribute to an individual's likelihood to develop an autoimmune response [17]. Still, identifying autoimmune potential and heterologous immunity through instances of molecular mimicry between Spike and proteins from humans or human pathogens can serve to better understand disease pathogenesis, improve therapeutic treatments, and inform vaccine design as they relate to SARS-CoV-2 infection. Previous studies have predicted molecular mimicry between SARS-CoV-2 Spike and human proteins using sequence similarity [18] to known epitopes in the Immune Epitope Database (IEBD) [19] and sequence and structural similarity in general [20,21]. We combine these approaches and investigate molecular mimicry between Spike and human proteins by considering both sequence and structural similarity and searching against known epitopes from IEDB [19]. We define molecular mimicry as a match of at least five identical consecutive amino acids that appear in both Spike and in a known epitope, where at least three amino acids are surface accessible on Spike and the match from the epitope has high structural similarity to the corresponding sequence from Spike. In light of our findings, we discuss autoimmune potential and

heterologous immunity with implications for vaccine design and the side-effects of SARS-CoV-2 infection.

METHODS

Identifying Epitopes with Molecular Mimicry

To identify known epitopes with positive assays from IEBD, we used Epitopedia [22] with a full-length Cryo-EM structure of Spike from SARS-CoV-2 (PDB id: 6XR8, chain A, RBD: 0up3down (solved residues:14-69, 77-244, 254-618, 633-676, 689-1162) [23]) as input. Hits containing 5 or more consecutive residues with 100% sequence identity where at least 3 of the input residues are surface accessible are considered sequence-based molecular mimics (termed as "1D-mimics"). For all 1D-mimics with corresponding structural representation from either the Protein Data Bank (PDB) [24] or AlphaFold2 [25] 3D models of human proteins, TM-align [26] was used to generate a structural alignment and Root Mean Square Deviation (RMSD) for all input-hit (1D-mimic) alignment pairs using only the structural regions corresponding to the hit for the source antigenic protein containing the epitope and the input. Epitopes with an RMSD $\leq$ 1 Å to Spike were considered structure-based molecular mimics (termed as "3D-mimics").

Conformational Ensemble of TQLPP Structural Mimicry

To gather all structures of the TQLPP motif in Spike, an NCBI BLASTP search against PDB was performed with the SARS-CoV-2 Spike reference sequence as the query and a SARS-CoV-2 taxa filter. Of 75, close to full-length, hits (>88% query cover),

20 included a solved structure for the TQLPP motif. The TQLPP region of the PDB structure was extracted for all chains in the 20 structures (all were trimers, as in Spike's biological state) resulting in a TQLPP Spike ensemble of 60 different chains from SARS-CoV-2. Each sequence in the TQLPP Spike ensemble was superimposed with chain X of the two PDB structures of human thrombopoietin (hTPO, PDB ids: 1V7M and 1V7N) to generate an RMSD value distribution for Spike's conformational ensemble vs hTPO for the structural mimicry region (Appendix 7).

Modeling Spike-Antibody Complexes

We constructed a composite model of the Spike-TN1 complex using the hTPO-TN1 complex (PDB id: 1V7M) as a template. For this, we first aligned the TQLPP segment of hTPO in the hTPO-TN1 complex with the TQLPP segment of the fully glycosylated model of Spike (PDB id: 6VSB [27]) retrieved from the CHARMM-GUI Archive [28]. We then removed hTPO, leaving TN1 complexed with Spike. For the Spike-TN1 simulations, only the TN1 interacting N-terminal domain of Spike (residues 1-272) was considered. Similarly, a composite model of the Spike-S2P6 complex was modeled by using the stem helix-S2P6 complex with ELDKY in the stem helix segment of Spike (PDB id: 6XR8) retrieved from the Protein Data Bank. We then removed the stem helix segment from the stem helix-S2P6 complex, leaving S2P6 complexed with Spike. For the Spike-S2P6 simulations, only the S2P6 interacting stem helix segment of Spike (residues 1146-1159) was considered. Geometrical alignments, as well as visualization, were performed with PyMOL version 2.5.0 [30] and Visual Molecular Dynamics (VMD 1.9.3 [31]).

To confirm that the modeled Spike TQLPP region is in agreement with the TQLPP region of solved Spike structures, these regions were extracted. TM-align was used to superimpose the TQLPP regions from the different structures, including the modeled TQLPP region from the Spike-TN1 complex, and to calculate the respective RMSD values. Three states of the model were included (before and after equilibration, and after molecular dynamics (described in the following paragraph)) together with the 60 experimentally determined Spike structures in Appendix 7) and compared in an all-against-all manner (Appendix 1, Appendix 8). A Mann-Whitney U test was used to compare the TQLPP region from 60 experimentally determined Spike structures based on RBD state: (1) both down, (2) 1 down and 1 up, (3) both up (Appendix 2). Further, TM-align was used to calculate RMSD between wild-type TQLPP (PDB id: 6XR8, chain A) and the corresponding region in known variants of concern with available structures (Appendix 9).

Molecular Dynamics Simulation

A simulation system for the modeled Spike-antibody systems was prepared using CHARMM-GUI [32,33,34]. The complexes were solvated using a TIP3P water model and 0.15 M concentration of KCl and equilibrated for 1 ns at 303 K. All-atom simulations were performed with NAMD2.14 [35] using CHARMM36m force-field. The production runs were performed under constant pressure of 1 atm, controlled by a Nose−Hoover Langevin piston [36] with a piston period of 50 fs and a decay of 25 fs to control the pressure. The temperature was set to 303 K and controlled by Langevin temperature coupling with a damping coefficient of 1/ps. The Particle Mesh Ewald method (PME) [37] was used for long-range electrostatic interactions with periodic boundary conditions

and all covalent bonds with hydrogen atoms were constrained by Shake [38]. The contact area of the interface was calculated as $(S_1+S_2-S_{12})/2$, where S_1 and S_2 represent the solvent accessible surface areas of the antigen and antibody and S_{12} represents that for the complex (Appendix 3). We performed MD simulations of the hTPO-TN1 complexes (PDB ids: 1V7M and 1V7N) as well as the Spike-TN1 complexes modeled from PDB ids: 1V7M and 1V7N to generate interaction matrices of protein-antibody hydrogen bonds during the last 50 ns of 200 ns MD simulation for each run.

Binding Affinity

The PRODIGY webserver [39] was used to calculate the binding affinity and intermolecular contacts for Spike-TN1 (described above) and hTPO-TN1 complexes (PDB ids: 1V7M and 1V7N) at the TQLPP region. We retrieved five intermediate structures from 200 ns MD simulations of each of these complexes at an interval of 40 ns. Similarly, PRODIGY was used to calculate the binding affinity and intermolecular contacts for the modeled Spike-S2P6 complex (from PDB id 7RNJ [29]) at the ELDKY region. We retrieved five intermediate structures from a 50 ns MD simulation at an interval of 10 ns. The average binding affinity for each complex is reported (Appendix 10).

Antibody Interface Complementarity

We used the MaSIF-search geometric deep learning tool designed to uncover and learn from complementary patterns on the surfaces of interacting proteins [40]. The surface properties of proteins are captured using radial patches. A radial patch is a fixed-sized geodesic around a potential contact point on a solvent-excluded protein surface [41]. In MaSIF-search, the properties include both geometric and physicochemical

properties characterizing the protein surface [40]. This tool exploits that a pair of patches from the surfaces of interacting proteins exhibit interface complementarity in terms of their geometric shape (e.g., convex regions would match with concave surfaces) and their physicochemical properties. The data structure of the patch is a grid of 80 bins with 5 angular and 16 radial coordinates and ensures that its description is rotation invariant. Each bin is associated with 5 geometric and chemical features: shape index, distance-dependent curvature, electrostatics, hydropathy, and propensity for hydrogen bonding. The model converts patches into 80-dimensional descriptor vectors, such that the Euclidian distance between interacting patches is minimized. Here, we define the binding confidence score as a measure of distance between the descriptor vectors of the two patches. Thus, lower "MaSIF binding confidence scores" represent better complementarity and therefore better matches. The pre-trained MaSIF-search model "sc05" with a patch radius of 12 Å was used.

Using the MaSIF protocol, we evaluated complexes of the TN1 antibody bound to Spike in the TQLPP region. The antibody-antigen patch pairs were extracted using scripts from the molecular mimicry search pipeline EMoMiS [42]. To accommodate multiple Spike configurations, we extracted patches from 40 SARS-CoV-2 Spike-antibody complexes from the SabDab structural antibody database [43]. Patches centered at Q23 from Spike and W33 from TN1 were selected as representative pairs for the Spike-TN1 interaction type because this potential contact point has the most hydrogen bonds in the modeled Spike-TN1 TQLPP region. Binding confidence scores of randomly formed complexes (Random), complexes between Spike and its native antibodies (Spike-Ab), and complexes between hTPO and TN1 (hTPO-TN1) were extracted and tabulated

(Appendix 11). The distribution of binding confidence scores from randomly formed complexes was obtained by pairing patches from random locations on Spike with patches from its antibodies. For native antibody-antigen Spike-Ab and hTPO-TN1 complexes, we obtained patch pairs from known interface regions using the MaSIF-search strategy for the selection of interacting patches [40]. Columns “Contact AB” and “Contact AG” in Appendix 12 indicate the residue used as the center of the patch from the antibody and the corresponding antigen.

Evaluating Further Cross-Reactivity

All 3D-mimics and AlphaFold-3D-mimics (termed as “AF-3D-mimics”) were split into pentapeptides (if mimicry motif exceeded 5 residues) which were used as queries for NCBI BLASTP searches against the RefSeq Select [44] set of proteins from the human proteome. Results for the BLAST searches can be found in Appendix 13.

For the TQLPP sequence motif, 10 representative isoforms in proteins containing the complete motif were found, including hTPO. The other 9 proteins lacked a solved structure containing TQLPP. However, AlphaFold2 3D models were available for all 10 of these RefSeq Select sequences [25,45], allowing us to extract the region corresponding to TQLPP in these hits and structurally superimpose this region with Spike TQLPP (from PDB id 6XR8) with TM-align as described above.

TN1-protein complexes were generated for three of the remaining 9 proteins (Fc receptor-like protein 4 (residues 190-282), serine/threonine-protein kinase NEK10 (residues 1029-1146), ALG12 (Mannosyltransferase ALG12 homolog (residues 1-488)). The TQLPP segment in hTPO was structurally aligned with each of the TQLPP segments of the modeled proteins, after which, hTPO was removed resulting in the complex of

TN1 with the modeled proteins following the methods mentioned for Spike above. The equilibrated structures of these complexes show that TN1 stays firmly with these proteins without any structural clash. Further, to evaluate the shape complementarity of these three proteins and TN1, MaSIF was used to calculate binding confidence scores as described above (Appendix 14).

It should also be noted that two additional human genes (GeneIDs 8028 and 57110) also have one TQLPP motif, but not in the RefSeq Select isoforms. Since no structure or structural prediction was available for these proteins, they were excluded from further analysis.

For the ELDKY sequence motif, 6 additional representative isoforms containing the complete motif were found, in addition to the human proteins identified by Epitopedia to contain 3D-mimics of the motif. Solved structures of the ELDKY motif were available for 3 of the proteins, while the others had AlphaFold2 3D models available. In all instances, the region corresponding to the ELDKY motif was extracted and structurally superimposed with Spike ELDKY (from PDB id 6XR8) with TM-align as previously described.

Statistical Analysis

Distributions were visualized as violin plots with ggpubr (Version 0.40) and ggplot2 (Version 3.3.6) from R (Version 4.2.1). Following Shapiro-Wilk normality testing, statistical analysis comparing the different distributions was performed using Mann-Whitney U with *SciPy* [46], followed by a simplified Bonferroni correction (alpha/n comparisons) when appropriate.

RESULTS AND DISCUSSION

We used Epitopedia [22] to predict molecular mimicry for the structure of the SARS-CoV-2 Spike protein (PDB id: 6XR8, chain A [23]) against all linear epitopes in IEDB, excluding those from Coronaviruses. Epitopedia returned 789 sequence-based molecular mimics (termed as "1D-mimics"). 1D-mimics are protein regions from epitopes that share at least five consecutive amino acids with 100% sequence identity to a pentapeptide in SARS-CoV-2 Spike, where at least three of the amino acids are surface accessible on Spike. Most 1D-mimics (627 epitopes) were found in human. Additionally, 1D-mimics were found in non-human vertebrates (65 epitopes, 7 species), viruses (58 epitopes, 17 species), bacteria (18 epitopes, 7 species), parasitic protists (12 epitopes, 2 species), plants (5 epitopes, 2 species), and invertebrates (4 epitopes, 2 species). Seemingly redundant 1D-mimics from the same protein may represent different epitopes and, thus, all 789 1D-mimics were kept at this step.

Structural representatives from the Protein Data Bank (PDB) were identified for 284 1D-mimics based on their source sequence using the minimum cutoffs of 90% for sequence identity and 20% for query cover. The 284 1D-mimics are represented by 7992 redundant structures from 1514 unique PDB chains. From these, structure-based molecular mimics (termed as "3D-mimics") were identified. 3D-mimics are 1D-mimics that share structural similarity with SARS-CoV-2 Spike as determined by an RMSD of at most 1 Å. We found 20 3D-mimics for Spike. Unsurprisingly, as with the 1D-mimics, most 3D-mimics were found for human proteins. Additionally, one 3D-mimic was found for *Mus musculus* (mouse), *Mycobacterium tuberculosis*, *Phleum pratense* (Timothy grass), and respiratory syncytial virus, respectively (Table 1). For each 3D-mimic,

Epitopedia computes a Z-score based on the distribution of RMSD values for all resulting hits for the input structure. This allows for a comparative assessment of the quality of a hit, with respect to RMSD, to other hits for a given run. Epitopedia also computes an EpiScore for each hit. EpiScore, calculated as (*motif length/RMSD*), favors longer motifs over shorter ones with the same RMSD values.

Table 1. 3D-mimics found for SARS-CoV-2 Spike

Motif	Protein	Species	RMSD	Z-Score	Epi Score	PDB_chain
TQLPP	Thrombopoietin	Human	0.46 Å	-1.34	10.87	1V7N_X
QLPPA	SMYD3 protein	Human	0.38 Å	-1.42	13.16	5CCL_A
KNLRE	Toll-like receptor 8	Human	0.87 Å	-0.92	5.75	6WML_D
FTVEKG	Pollen allergen Phl p2	*Phleum pratense*	0.76 Å	-1.03	7.89	1WHP_A
GEVFN	Integrin beta 1	Human	0.63 Å	-1.16	7.94	7NWL_B
HAPAT	Activator of 90 kDa heat shock protein ATPase homolog 1	Human	0.74 Å	-1.05	6.76	7DME_A
YSTGS	Argininosuccinate lyase	Human	0.48 Å	-1.31	10.42	1K62_B
EHVNN	Casein kinase 2 alpha isoform	Human	0.29 Å	-1.51	17.24	2ZJW_A
NLLLQ	DNA polymerase subunit gamma 1	Human	0.57 Å	-1.22	8.77	5C51_A
LLQYG	Ankyrin 1	Human	0.20 Å	-1.60	25.00	1N11_A
LPDPS	BRCA1-A complex subunit BRE	Human	0.32 Å	-1.48	15.62	6GVW_C
LPDPS	Semaphorin 7a	Human	0.84 Å	-0.91	5.95	3NVQ_A
DPSKP	60S ribosomal protein L3	Human	0.10 Å	-1.70	50.00	6LU8_B
DPSKP	Alanine and proline-rich secreted protein apa precursor	*Mycobacterium tuberculosis*	0.21 Å	-1.59	23.81	5ZXA_A

IAARD	Talin	*Mus musculus*	0.74 Å	-1.05	6.76	6R9T_A
GNCDV	Tryptophan-tRNA ligase	Human	0.91 Å	-0.88	5.49	1O5T_A
SFKEE	Small subunit processome component 20 homolog	Human	0.32 Å	-1.48	15.62	7MQA_SP
EELDK	Kynureninase	Human	0.22 Å	-1.58	22.73	2HZP_A
ELDKY	Fusion glycoprotein F0	Respiratory syncytial virus	0.12 Å	-1.68	41.67	6EAE_F
DKYFK	Cytoplasmic FMR1-interacting protein 1	Human	0.14 Å	-1.66	35.71	4N78_A

For the 402 human 1D-mimics where no PDB structural representative could be identified for their source sequence, AlphaFold2 3D models were used. 3D model representatives were found for 102 human 1D-mimics. Of these, 10 are predicted to be AlphaFold-3D-mimics (termed as “AF-3D-mimics”) based on the RMSD (Table 2).

Table 2. Human AF-3D-mimics for SARS-CoV-2 Spike

Motif	Protein	RMSD	Z-Score	EpiScore	AlphaFold2 ID
NLLLQ	Ankyrin 3	0.61 Å	-1.18	8.20	AF-Q12955-F1-model_v1_A
LLQYG	Olfactory receptor 10Q1	0.66 Å	-1.13	7.58	AF-Q8NGQ4-F1-model_v1_A
TGIAV	Phosphofructokinase	0.17 Å	-1.63	29.41	AF-P17858-F1-model_v1_A
TGIAV	Low affinity immunoglobulin gamma Fc region receptor II-b	0.17 Å	-1.63	29.41	AF-P31995-F1-model_v1_A
KIQDSL	Phosphorylase b kinase regulatory subunit beta	0.19 Å	-1.61	31.58	AF-Q93100-F1-model_v1_A
KIQDSL	Long-chain-fatty-acid-CoA ligase 5	0.37 Å	-1.43	16.22	AF-Q9ULC5-F1-model_v1_A

VYDPL	Actin-binding protein IPP	0.17 Å	-1.63	29.41	AF-Q9Y573-F1-model_v1_A
EELDK	Tight junction-associated protein 1	0.20 Å	-1.60	25.00	AF-Q5JTD0-F1-model_v1_A
EELDKY	Keratin, type I cytoskeletal 18	0.22 Å	-1.58	27.27	AF-P05783-F1-model_v1_A
ELDKY	Tropomyosin alpha-3 chain	0.18 Å	-1.62	27.78	AF-P06753-F1-model_v1_A

The 3D- and AF-3D-mimics (hereinafter referred to as "molecular mimics") mapped to a few clusters on Spike. Ten molecular mimics were singletons, six overlapping molecular mimics were found as pairs in three small clusters, and the remaining 14 were found in three larger clusters with at least four overlapping molecular mimics (Figure 1a). The largest cluster, with six molecular mimics, was also adjacent to three additional molecular mimics. All mimics are displayed on the surface of Spike's functional trimer, but the large cluster centered around LLLQY is in a deep pocket and is an unlikely antibody binding epitope in this conformation (Figure 1b). Only one molecular mimic is predicted for the RBD, despite RBD being an immunodominant region in Spike to which many antibodies naturally bind [47]. This molecular mimic (HAPAT) corresponds to the activator of 90 kDa heat shock protein ATPase homolog 1 (AHA1). Two molecular mimics are predicted near the S1/S2 boundary that is a site for proteolytic cleavage [48]. The first is YSTGS from argininosuccinate lyase. The second is EHVNN from casein kinase 2 alpha (CK2). CK2 has been found to play an important role in SARS-CoV-2 infection [49]. Activation of CK2 is promoted by SARS-CoV-2 infection [50] and inhibiting CK2 has been suggested as a therapeutic strategy against both SARS-CoV and SARS-CoV-2 [49]. If a cross-reactive antibody intended for SARS-CoV-2 can interact with CK2, it may impact its activity and perhaps the antibody can

stabilize conformations that make CK2 more active, but these are speculations and more work along these lines is needed.

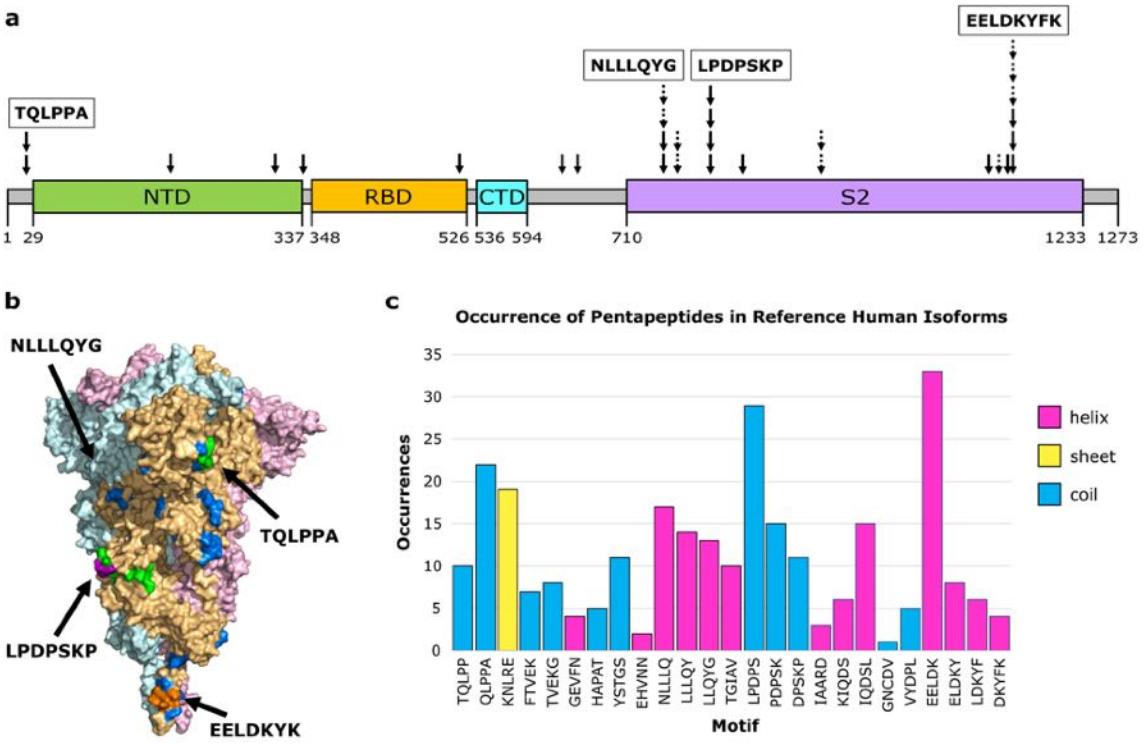


Figure 1. Molecular mimicry with autoimmune potential across SARS-CoV-2 Spike. (**a**) Overview of molecular mimics (solid arrow: 3D-mimic, dashed arrow: AF-3D-mimic) for Spike in the linear sequence showing Spike domains (NTD: N-terminus domain of S1 subunit (green), RDB: receptor binding domain of S1 subunit (orange), CTD: C-terminus domain of S1 subunit (cyan), S2: S2 domain (purple)) as predicted by Pfam [51] based on the NCBI reference sequence (YP:009724390.1). The boundary between the S1 and S2 subunits is indicated by S1/S2. (**b**) Surface representation of Spike (PDB id: 6XR8 [23]) colored by subunit (pink, beige, light blue) with residues colored by number of occurrences in a molecular mimic (blue: 1, green: 2, purple: 3, orange: 4 or more). Structural visualization generated with PyMOL 2.5.0 [30]. (**c**) The number of occurrences of the sequence motif in human RefSeq Select isoforms arranged in order from the N-terminus to the C-terminus and colored by predominant secondary structure element based on Spike PDB id 6XR8 chain A.

To further evaluate the autoimmune potential of the human mimics, we identified all occurrences of the motifs in the human RefSeq Select proteome [44]. The pentapeptides from the molecular mimicry regions are found from four to 33 times in

human proteins (Figure 1c, Appendix 13). The human protein thrombopoietin that includes the 3D-mimic TQLPP (Table 1) also has an occurrence of the sequence mimic LPDPS (Appendix 13). Further, another protein family that occurs twice for the same pentapeptide is tropomyosin. Tropomyosin alpha-3 is an AF-3D-mimic (Table 2), and tropomyosin alpha-1 has one occurrence of the same pentapeptide (ELDKY). The same motif, ELDKY, is a 3D-mimic in the fusion F0 glycoprotein of respiratory syncytial virus (Table 1). Altogether, based on the known epitopes in IEDB, heterologous immunity is rare with Spike while regions of autoimmune potential form hotspots.

To further evaluate molecular mimicry and, indirectly, autoimmune potential, we performed a deeper investigation of two motifs, TQLPP and ELDKY, that mapped to positions 22-26 (small cluster) and 1151-1155 (largest cluster) in Spike, respectively. For TQLPP, a 3D-mimic with human thrombopoietin was identified. The only structure in our dataset where a 3D-mimic was located at an antibody interface was for human thrombopoietin (hTPO). Thrombopoietin is a cytokine that regulates platelet production [52] (Figure 2). Interestingly, COVID-19 patients often suffer from thrombocytopenia, characterized by low platelet levels [53], which correlates with an almost 5-fold increase in mortality [54]. Thrombocytopenia in COVID-19 patients resembles immune thrombocytopenia (ITP), where hTPO and/or its receptor are mistakenly targeted by autoantibodies leading to reduced platelet count [55]. Treatments with hTPO Receptor Agonists improve thrombocytopenia in both general [56] and COVID-19 [57] patients, suggesting the mistaken targeting occurs before hTPO activates the hTPO receptor. ITP is a heterogenous disease caused by numerous mechanisms. In ITP patients, about half have antibodies against the major platelet glycoproteins while 28.1% have autoantibodies

against hTPO, 28.1% against the hTPO receptor, and 6.8% against the hTPO-hTPO receptor complex. While autoantibodies often seem to play a role in ITP, other mechanisms are possible [58]. For ELDKY, we identified one 3D-mimic in the fusion F0 glycoprotein of respiratory syncytial virus (Table 1) and two AF-3D-mimics from keratin type I cytoskeletal 18 and tropomyosin alpha-3 (Table 2). Additional 3D-mimics partially overlapping with ELDKY were identified. The ELDKY motif in Spike is part of a highly reactive epitope [59] found in an α-helix located towards the C-terminus. This motif is conserved across beta-coronaviruses and can bind an antibody effective against all human-infecting beta-coronaviruses [29]. Altogether, the numerous molecular mimics of the ELDKY motif suggests a potential for both protective and autoimmune cross-reactivity.

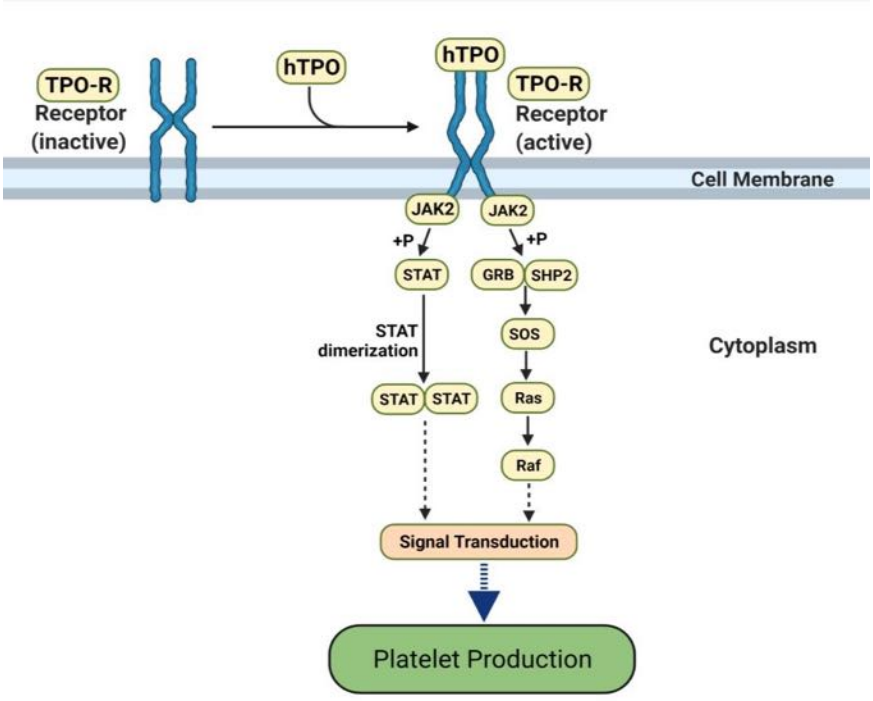


Figure 2. The hTPO pathway to induce platelet production. Simplified JAK-STAT signaling pathway in megakaryocytes where hTPO activates the TPO receptor and triggers signaling cascades that stimulate platelet production [60,61]. Created with BioRender.com (accessed on 12 August 2021).

Molecular Mimicry between Spike and Thrombopoietin Mediated through TQLPP

The shared five-amino acid motif, TQLPP (Figure 3a), is located on the surface of Spike's N-terminal Domain (NTD) (Figure 3b, c), whereas it is found at the interface with a neutralizing antibody in hTPO [62] (Figure 3d). The TQLPP motifs from the two proteins are found in coil conformations but exhibit high structural similarity (Figure 3e, f). On Spike, the motif is adjacent to the NTD supersite that is known to be targeted by neutralizing antibodies [63]. We hypothesized that COVID-19 may trigger the production of TQLPP-specific antibodies against this epitope that can cross-react with hTPO. In the absence of Spike TQLPP antibodies, we used molecular modeling and machine learning to investigate the binding of the neutralizing mouse Fab antibody (TN1) from the hTPO structure [64] to the Spike TQLPP epitope.

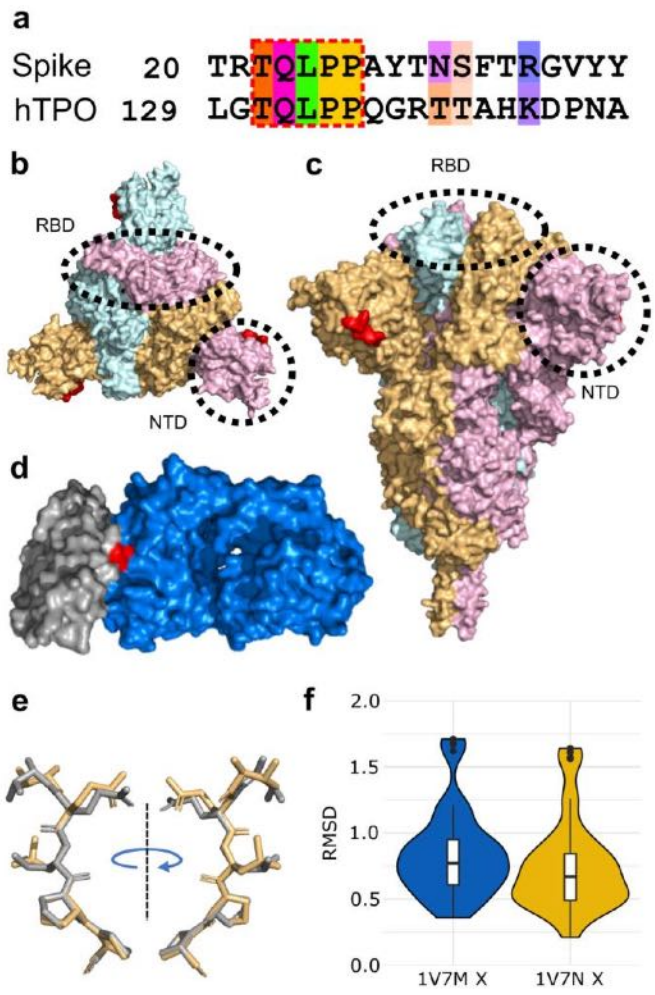


Figure 3. Structural mimicry between a TQLPP motif in SARS-CoV-2 Spike and an antibody binding epitope in thrombopoietin. (**a**) Pairwise sequence alignment for the TQLPP motif in the epitope for human thrombopoietin (hTPO, IEDB Epitope ID: 920946) and Spike, amino acids colored by Taylor [65] for sites with ≥ 50% conservation in the amino acid property [66]. The region of molecular mimicry is highlighted in the red dashed box. Surface representation of Spike from (**b**) the top and (**c**) the side, with Spike trimer (PDB id: 6XR8 [23]) colored by subunit (pink, beige, light blue) and red indicating the location of the TQLPP epitope fragment, illustrating the surface accessibility of TQLPP and highlighting the location of RBD (dashed oval) and NTD (dashed circle). (**d**) Surface representation shown for hTPO (gray, PDB id: 1V7M [62]) and its TN1 antibody (blue) with the TQLPP motif (red) at the interface. (**e**) TM-align generated structural alignment for TQLPP in Spike (beige) and hTPO (gray), with RMSD = 0.61 Å. (**f**) Violin plots of RMSD values resulting from the comparison of the TQLPP region in 20 Spike structures (60 chains) vs TQLPP in two hTPO structures (PDB ids: 1V7M and 1V7N, chain X for both [62]). Statistical analysis with Mann-Whitney U reveals no statistical significance between the sets. Box plots, bounded by the 1st and 3rd quartiles, show median value (horizontal solid bold line), vertical lines (whiskers)

represent 1.5 × IQR, while outliers are marked as black points. For further details, see methods. Alignment representations were generated with Jalview 2.11.2.2 [66] and structural visualizations were generated with PyMOL 2.5.0 [30].

To construct a composite model of Spike and TN1 Fab, a full-length glycosylated model of the Spike trimer, based on PDB id 6VSB [27] with the first 26 residues (including the TQLPP motif) reconstructed [67], was coupled to three copies of TN1 Fab from the structure of hTPO complexed with TN1 Fab [62]. The Spike-TN1 complex was energy minimized and equilibrated with molecular dynamics (MD) simulation. The final model of the Spike trimer complexed with three TN1 Fab antibodies (Figure 4a, b) shows that the TQLPP epitope is accessible to the antibody and the adjacent glycan chains do not shield the antibody-binding site (Figure 4c, Appendix 4). To confirm the conformation of TQLPP, we calculated the RMSD for TQLPP regions from 20 Spike trimer structures (60 chains) from PDB, plus the modeled states (before and after equilibration, and upon 200 ns MD simulation) in an all-vs-all manner (Appendix 1, Appendix 8), paying particular attention to the orientation (up or down) of the RBD. For 1953 pairwise comparisons, 1306 had an RMSD ≤ 1 Å and 32 had an RMSD ≥ 2 Å. Three groups were compared using a Mann-Whitney U test based on RBD state: (1) both down (N = 666, mean = 0.78 Å , median = 0.66 Å), (2) 1 down 1 up (N = 962, mean = 0.81 Å, median = 0.73 Å), and (3) both up (N = 325, mean = 0.85 Å, median = 0.78 Å). Here, comparisons between groups 1 and 2 (*p*-value = 0.030) and 1 and 3 (*p*-value = 0.003) were significantly different, while that between groups 2 and 3 (*p*-value = 0.055) was not (Appendix 2) The reconstructed TQLPP region falls within the conformational ensemble from PDB, suggesting that the modeled representation of TQLPP is valid. Furthermore, the Spike-TN1 complexes (with TN1 from PDB ids 1V7M and 1V7N) and

hTPO-TN1 complexes (PDB ids 1V7M and 1V7N) are all stable and have comparable binding affinities, with averages ranging from -9.2 to -9.56 kcal/mol (Appendix 10). The predominant intermolecular contacts for these four complexes are between polar-apolar and apolar-apolar residues (Appendix 10).

To evaluate the molecular mimicry between the antibody interface areas, we performed MD simulations of hTPO and Spike NTD with TQLPP complexed with the TN1 antibody. The hydrogen bonds were calculated between the TN1 antibody with hTPO and Spike, respectively, from the last 50 ns of both trajectories (Appendix 3). Both the Spike-TN1 and the hTPO-TN1 complexes showed similar contact areas (Appendix 3). Notably, critical hydrogen bonds were observed for residues Q and L in the TQLPP motif with TN1 for both Spike and hTPO (Figure 4d, e and Appendix 3).

To further support our findings, we evaluated the antibody-antigen interface complementarity with MaSIF-search, a recent tool that uses deep learning techniques [40], on a pair of circular surface regions (patches) from an antibody-antigen complex. MaSIF-search produces a score associated with the confidence of binding when forming a stable complex. We refer to this score here as the binding confidence score, where lower scores indicate a higher probability of protein-protein binding. The results show that Spike-TN1 complexes have a better (lower) binding confidence score than random complexes and that complexes including Spike from PDB id 7LQV [63] have three of the four best binding confidence scores (0.86, 1.05, 1.42) and may bind to TN1 as well as, or better than, hTPO (Figure 4h, Appendices 11-12). Notably, in 7LQV, COVID-19 antibodies bind to Spike at the NTD supersite [63]. These results strongly argue for the

possibility of cross-reactivity between Spike and hTPO driven by the molecular mimicry of TQLPP (Figure 4).

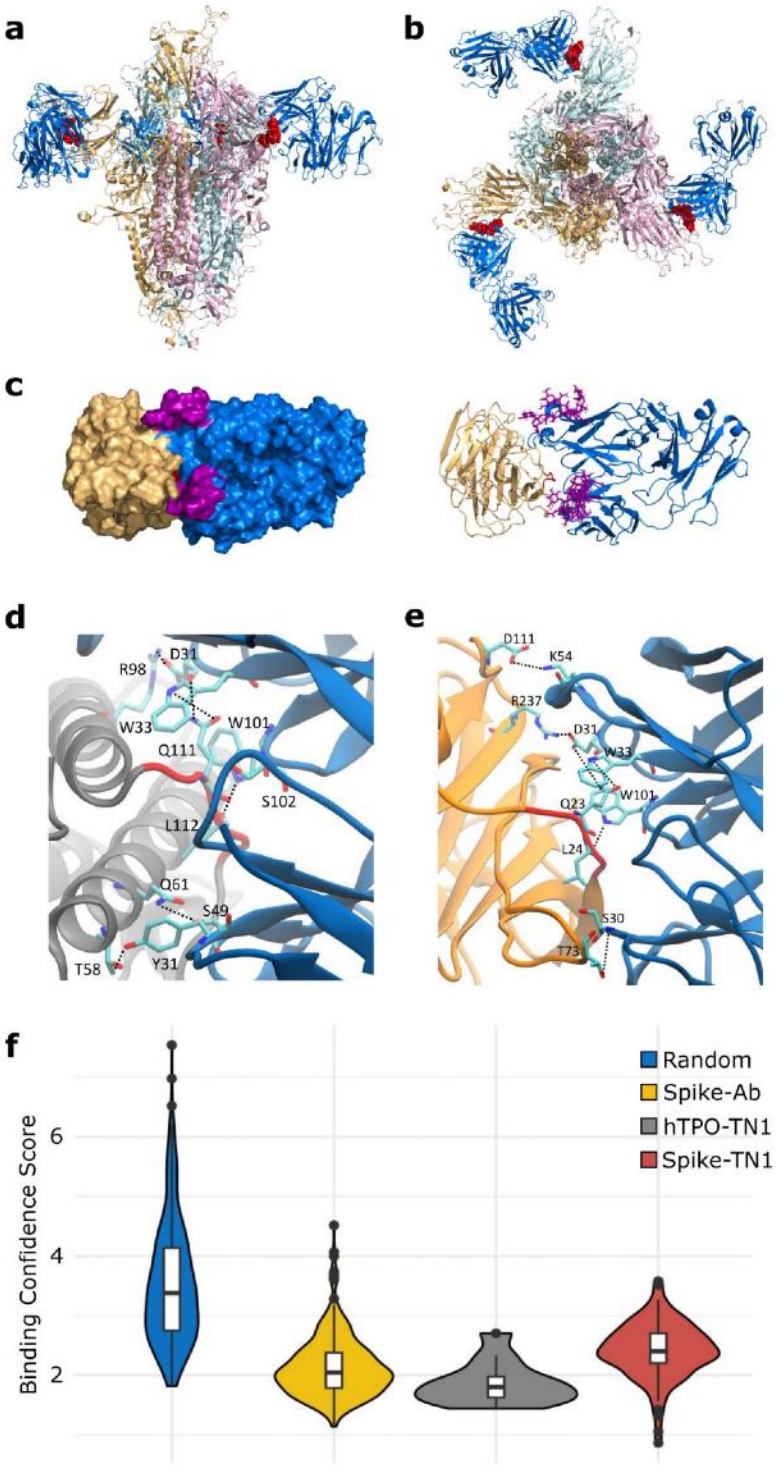


Figure 4. Binding of SARS-CoV-2 Spike to TN1 Fab antibody. Equilibrated structure (1 ns) of the modeled TN1 Fab antibody (blue, PDB id: 1V7M) complexed with Spike trimer model (pink, beige, light blue) shown from (**a**) the side and (**b**) the top, with TQLPP shown as red spheres. (**c**) The Spike NTD (beige) and TN1 Fab complex used for MD simulations (200 ns), with adjacent glycans at N17 and N74 highlighted in purple. The representative amino acids contributing to hydrogen bonds (dashed lines) during the last 50 ns of simulations for the (**d**) hTPO-TN1 and (**e**) Spike-TN1 complexes are highlighted as cyan sticks. (**f**) Violin plot showing the distribution of the MaSIF binding confidence scores for randomly selected patch pairs (blue), the interacting region of Spike-antibody (yellow) and hTPO-TN1 (gray) complexes, and for modeled Spike-TN1 complexes across 40 Spike configurations (red). Statistical analysis with Mann-Whitney U shows that all pairwise comparisons except for Spike-Ab and hTPO-TN1 are significantly different after Bonferroni correction (Appendix 12). Box plots, bounded by the 1st and 3rd quartiles, show median value (horizontal solid bold line), vertical lines (whiskers) represent 1.5 × IQR, while outliers are marked as black points. For further details, see methods. Structural visualizations were generated with PyMOL 2.5.0 [30] and VMD 1.9.3 [31].

The human proteome contains nine additional occurrences of the TQLPP motif. Two of these motifs, found in Hermansky-Pudlak syndrome 4 protein and ALG12 (Mannosyltransferase ALG12 homolog), have been associated with thrombosis and hemostasis disorder [68]. To evaluate structural mimicry between Spike-TQLPP and all human-TQLPP motifs, we utilized AlphaFold2 3D models [25,45] (Appendix 5). The closest structural mimicry region is in hTPO (RMSD = 0.39 Å), followed by coiled-coil domain-containing protein 185, Fc receptor-like protein 4 (FCRL4), and far upstream element-binding protein 1 (Appendix 5). These results indicate that TQLPP motifs have similar conformations (Appendix 1), strengthening the notion of structural mimicry. We investigated the potential cross-reactivity of an antibody targeting TQLPP in these proteins, after discarding six that display the TQLPP motif in low confidence or unstructured regions. The remaining three proteins, NEK10 (ciliated cell-specific kinase), FCRL4, and ALG12 were complexed with TN1 (Figure 5). The binding confidence score for NEK10-TN1 (1.44) is comparable to the hTPO-TN1 complex (Figure 5). NEK10

regulates motile ciliary function responsible for expelling pathogens from the respiratory tract [69]. Dysfunction of NEK10 can impact mucociliary clearance and lead to respiratory disorders such as bronchiectasis [69]. Based on our results, it is plausible that the function of NEK10 and thus mucociliary clearance can be affected by cross-reactive Spike antibodies targeting TQLPP.

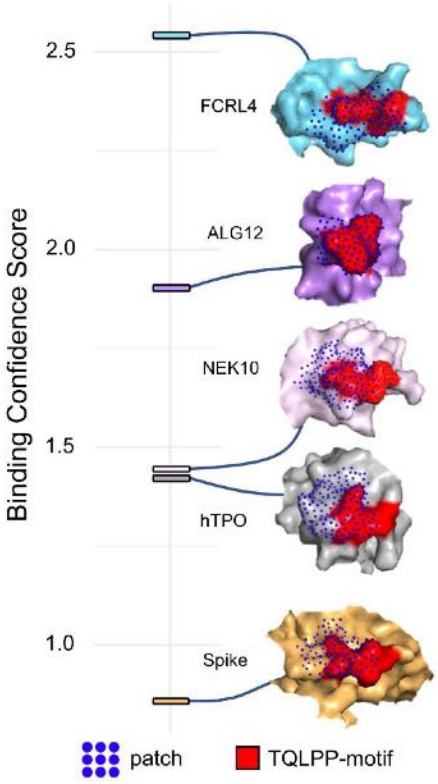


Figure 5. Predicted interaction patches between TN1 Fab antibody (PDB id: 1V7N) and the TQLPP motif. The best (lowest) binding confidence score is shown for Spike (PDB id: 7LQV, chain A, beige), hTPO (PDB id: 1V7N, chain X, gray), NEK10 (Uniprot: Q6ZWH5, pink), ALG12 (Uniprot: Q9BV10, purple), and FCRL4 (Uniprot: Q96PJ5, light blue). For all, red indicates the TQLPP motif and dark blue dots represent the surface points included in the predicted MaSIF patches.

Molecular Mimicry between Spike, RSV, and Many Human Proteins Mediated through ELDKY

Another motif, ELDKY, is in a region with several partially overlapping pentamer motifs including three 3D-mimics and three AF-3D-mimics (Figure 6a). For the 3D-

mimics, two are from the human proteins kynureninase (hKYNU; motif: EELDK) and cytoplasmic FMR1-interacting protein 1 (hCYFIP1; motif: DKYFK), while the last is found in the fusion F0 glycoprotein of respiratory syncytial virus (RSV; motif: ELDKY). For the AF-3D-mimics, the motif is found in human tight junction-associated protein 1 (hTJAP1; motif: EELDK), keratin type I cytoskeletal 18 (hKRT18; motif: EELDKY), and tropomyosin alpha-3 (hTPM3; motif: ELDKY). In Spike, the ELDKY motif is in a stem helix region near the C-terminus. This motif is well-conserved across beta-coronaviruses and is found in a highly reactive epitope [59] that has been shown to bind to a broadly neutralizing antibody (S2P6) effective against all human-infecting beta-coronaviruses [29]. The S2P6 antibody (from PDB id 7RNJ [29]) forms a stable complex with the Spike helix, with an average binding affinity of -9.52 ± 0.26 kcal/mol (Appendix 10). Here, the predominant intermolecular contacts are formed between charged-apolar, polar-apolar, and apolar-apolar residues (Appendix 10). In COVID-19, stronger antibody responses to the epitope containing the ELDKY motif have been recorded for severe (requiring hospitalization) vs moderate cases, while fatal cases had a weaker response than surviving cases [16]. A synthetic epitope containing the ELDKY motif has also been shown to elicit antibody production following COVID-19 immunization [70]. Together with the 3D-mimics identified here, these results suggest interesting possibilities for the ELDKY motif from the perspective of both protective immunity and an autoimmune response. First, while not an example of molecular mimicry but evolutionary conservation across beta-coronaviruses, prior exposure to an endemic cold-causing coronavirus (ex. HCoV-OC43) could result in the production of a broadly neutralizing antibody against an epitope containing the ELDKY motif that would be effective against

SARS-CoV-2 infection, which could result in milder or asymptomatic infection. Further, a protective effect due to molecular mimicry is suggested by the 3D-mimic identified for the fusion F0 glycoprotein of RSV, a common virus that infects most children in the United States by the time they are 2 years old [71], where antibodies against the ELDKY-containing epitope in RSV may be effective in combatting SARS-CoV-2 infection. In contrast, the potential for an autoimmune response against this motif is suggested by its presence in both two human 3D- and AF-3D-mimics (Figure 6).

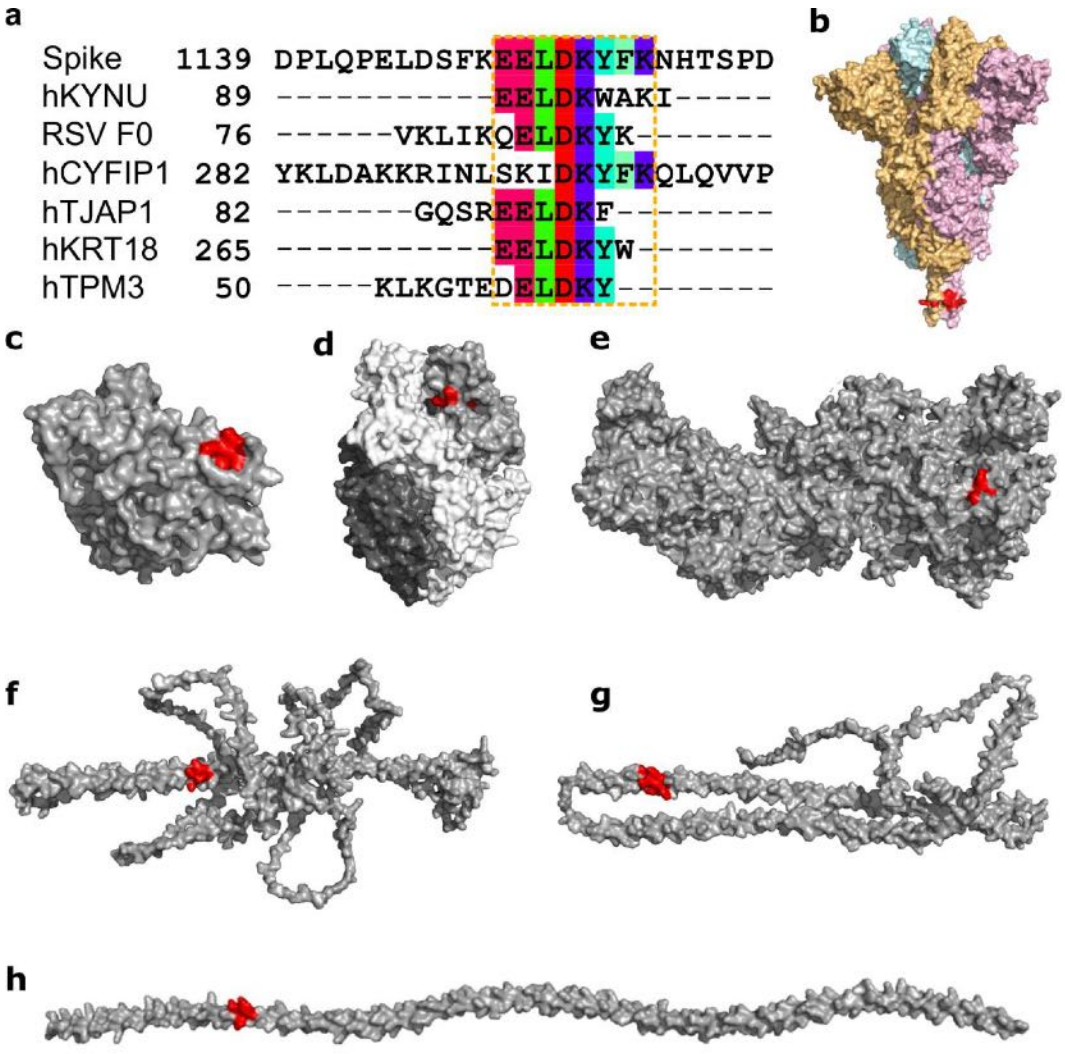


Figure 6. Structural mimicry between an ELDKY motif in SARS-CoV-2 Spike and epitopes in six other proteins. (**a**) Sequence alignment between SARS-CoV-2 Spike and

the epitopes containing the 3D-mimicry motif for human kynureninase (hKYNU, IEDB Epitope ID: 1007556), respiratory syncytial virus fusion F0 glycoprotein (RSV F0, IEDB Epitope ID: 1087776), human cytoplasmic FMR1-interacting protein 1 (hCYFIP1, IEDB Epitope ID: 1346528), human tight junction-associated protein 1 (hTJAP1, IEDB Epitope ID: 1016424), human keratin type I cytoskeletal 18 (hKRT18, IEDB Epitope ID: 1331545), and human tropomyosin alpha-3 (hTPM3, IEDB Epitope ID: 938472). Residues in the molecular mimicry motifs are colored by Taylor [65]. The extended molecular mimicry region is highlighted by the orange dashed box. (**b**) Surface representation of Spike (PDB id: 6XR8) colored by subunit (beige, pink, light blue) with ELDKY motif indicated in red. Surface representation of proteins (gray) with full or partial 3D-mimics of the ELDKY motif (red): (**c**) hKYNU (PDB id: 2HZP), (**d**) RSV F0 (PDB id: 6EAE), (**e**) hCYFIP1 (PDB id: 4N78), (**f**) hTJAP1 (Uniprot: Q5JTD0), (**g**) hKRT18 (Uniprot: P05783), (**h**) hTPM3 (Uniprot: P06753). Alignment representations were generated with Jalview 2.11.2.2 [66] and structural visualizations were generated with PyMOL 2.5.0 [30].

There are six additional occurrences of the ELDKY motif in the human proteome (Appendix 6). Structural similarity between Spike-ELDKY and human-ELDKY was assessed based on experimentally determined structures (if available) or AlphaFold2 3D models. RMSDs for the ELDKY motif ranged from 0.12-0.20 Å for 5 of the structures, with one hit being an outlier at an RMSD = 0.46 Å. In all instances, the ELDKY motif is found in an α-helix, resulting in the high degree of structural similarity found for this motif across proteins and bolstering the possibility for molecular mimicry. The ELDKY occurrence with the largest RMSD (0.46 Å) is found in the leucine-zipper dimerization domain of cGMP-dependent protein kinase 1 (PRKG1) (Appendix 6) whose phosphorylation targets have roles in the regulation of platelet activation and adhesion [72], smooth muscle contraction [73], and cardiac function [74]. Additionally, PRKG1 regulates intracellular calcium levels via a multitude of signaling pathways [75]. The ELDKY motif is also found in tropomyosin alpha-1 (TPM1), a homolog of the AF-3D-mimic tropomyosin alpha-3 (TPM3). Tropomyosins (TPMs) are involved in regulation of the calcium-dependent contraction of striated muscle [76]. TPM1 is a 1D-mimic but due

to a discrepancy in IEDB it was not identified as a 3D-mimic, although there is high structural similarity between ELDKY in Spike and ELDKY in TPM1 (Appendix 6). A previous study identified a longer match with 53% sequence identity between Spike and TPM1 that included the ELDKY motif [77]. However, in a separate search for structural similarity, Marrama and colleagues were unable to identify structural mimicry at the ELDKY motif due to using a structure for Spike lacking the motif, leading to a conclusion against molecular mimicry contributing to myocarditis in COVID-19 [77], in contrast to our work. These results illustrate the importance of structural representative selection when performing structural comparisons and in taking both sequence and structural similarity together into account when performing molecular mimicry searches, as we have done. For PRKG1, cross-reactive Spike antibodies targeting ELDKY may react with the motif, affecting PRKG1's role in the regulation of platelet activation and adhesion and thus providing another avenue for thrombocytopenia or other blood clotting disorders. Antibodies that cross-react with PRKG1 may also alter calcium levels, thus affecting TPM function. For TPM1, cross-reactive Spike antibodies targeting the ELDKY motif may result in coronary heart disease, as low-level autoantibodies against this protein have been associated with increased risk for this condition [78] and TPM1 and TPM3 are cardiac disease-linked antigens [77]. Cardiac disease, including myocardial injury and arrhythmia, can be induced by SARS-CoV-2 infection [79] and myocarditis has been found to develop in some individuals following vaccination against SARS-CoV-2 [80]. Furthermore, COVID-19 has been found to increase risk and long-term burden of several cardiovascular diseases, with COVID-19 severity being proportionate to increased risk and incidence [81].

CONCLUSION

We find that molecular mimics with high autoimmune potential are often found in clusters within Spike. Some clusters have several molecular mimics whose motifs are also found multiple times in the human proteome. Molecular mimics located in α-helices tend to have high structural similarity as can be expected based on the regular conformation of the helix, but also some molecular mimics in coil regions are remarkably similar. Our results on the TQLPP motif, located in a coil region, suggest a worrisome potential for cross-reactivity due to molecular mimicry between Spike and hTPO involving the TQLPP epitope that may affect platelet production and lead to thrombocytopenia. Further, cross-reactivity with other TQLPP-containing proteins such as NEK10 cannot be dismissed based on our *in-silico* results, but *in-vivo* validation is required. The presence of neutralizing antibodies against peptides with TQLPP in COVID-19 patients' convalescent plasma [82], particularly in severe and fatal cases [16] adds credence to our result. It is also interesting to note that antibodies against a TQLPP-containing peptide were found in the serum of pre-pandemic, unexposed individuals [83]. Prior infection with a different human coronavirus cannot explain the cross-reactivity observed in the unexposed group because TQLPP is situated in a region with low amino acid conservation [83]. Rather, this suggests the presence of an antibody for an unknown epitope with affinity for the TQLPP region in Spike. The COVID-19 vaccines designed to deliver the Spike protein from SARS-CoV-2, like COVID-19 infection itself, can cause thrombocytopenia [53,84,85,86] and it is plausible that cross-reactivity can titrate the serum concentration of free hTPO. The TQLPP motif is changing in the SARS-CoV-2 variants and evolutionary trends in the motif suggest it may not remain in Spike. RMSD

values between wild-type TQLPP and TQLPP in five variants of concern range from 0.21-1.78 Å (Appendix 9). In the Gamma variant, a P26S mutation has changed TQLPP to TQLPS and two additional mutations are located just before the motif at L18F and T20N in the NTD supersite, while the Delta variant is mutated at T19R [87]. The first Omicron variant (21K or BA.1), however, has no amino acid substitutions near the TQLPP motif, while a closely related Omicron variant (21L or BA.2) contains a 9 nucleotide deletion that results in the loss of 60% of the TQLPP motif (L24-, P25-, P26-) [87]. Neutralizing antibodies targeting the NTD supersite may rapidly lose efficacy against the evolving SARS-CoV-2. While the current COVID-19 vaccines remain safe and efficacious, we postulate that protein engineering of the TQLPP motif and possibly the NTD supersite for future COVID-19 vaccines may reduce the risk for thrombocytopenia and improve long-term vaccine protection against evolving variants.

We illuminated the cross-reactivity mediated through the ELDKY motif between Spike and PRKG1, TPM1, and TPM3. While PRKG1 provides a connection between blood clotting disorders and cardiac complications, it has a larger RMSD than other ELDKY motifs. ELDKY motifs in α-helices have high similarity and make good candidates for molecular mimicry. We find ELDKY in the homologous proteins TPM1 and TPM3 suggesting a conserved importance for structure and function. In contrast to TQLPP, the ELDKY motif is highly conserved among beta-coronaviruses [29] and there are presently no SARS-CoV-2 variants with mutations in this region [87]. Further, while the existence of a broadly neutralizing antibody against an epitope containing ELDKY [29] illustrates the potential of this motif as a pan-coronavirus vaccine target, the viability may be diminished by the possibility for autoimmune cross-reactivity due to this motif.

We present an extended application of Epitopedia [22] to identify molecular mimicry between Spike and known epitopes. We do not attempt to discover all possible epitopes for Spike. Epitopedia is only capable of predicting molecular mimicry for linear epitopes with positive assays that have been deposited in IEDB [19] and cannot predict molecular mimicry *de novo*. By design, Epitopedia does not predict molecular mimicry for conformational epitopes. Epitopedia relies primarily on structures available in PDB [24] when assessing structural similarity between 1D-mimics and the corresponding region on SARS-CoV-2 Spike. This can result in the nonidentification of potentially genuine molecular mimics if they are only present as 1D-mimics but have yet to have their structure experimentally determined. Moreover, the composition of the PDB is biased towards proteins that crystallize well, thus a molecular mimic can additionally go nonidentified if the 1D-mimic is found in an intrinsically disordered protein region. Proteins are dynamic molecules and the structures present in PDB may only represent a fraction of a protein's full conformational ensemble [88]. Further, IEDB and PDB both have a biased data composition in that more well-studied proteins are likely to be the ones whose functions and structures are published while other proteins are underrepresented. Lastly, it is important to be mindful that Epitopedia output it strictly a prediction and can have false positives. It is therefore of utmost importance to follow up on the results with both literature searches and experimental validation.

We highlight two epitopes of particular interest in our investigation of molecular mimicry in SARS-CoV-2. For one epitope, we find the TQLPP motif and an interacting antibody with which we perform a computational investigation into antibody binding properties of the tentative molecular mimic. The results show that the same antibody may

be able to bind TQLPP-containing epitopes in different proteins and that the TQLPP motif tends to be found in similar conformations despite being in a loop or coil. For the other epitope, we find the ELDKY motif with potential for protective immunity and with high structural similarity. High structural similarity can be expected for α-helical structures, and, if the sequence is identical, molecular mimicry results. Altogether, these are examples of molecular mimicry that may play a role in autoimmune or cross-reactive potential of antibodies generated by the immune system against SARS-CoV-2 Spike, but it must be noted that these results have not been experimentally verified. Still, computational investigations into the autoimmune potential of pathogens like SARS-CoV-2 are important for therapeutic intervention and when designing vaccines to avoid potential predictable autoimmune interference.

ACKNOWLEDGMENTS

We thank Dr. Sathibalan Ponniah (Immune Analytics LLC, Columbia, MD), Dr. Charles Dimitroff (Florida International University), Dr. Sixto Leal (University of Alabama – Birmingham), and Kevin Bennett (Florida International University) for discussions. This work was partially supported by the National Science Foundation under Grant No. 2037374. The authors would also like to acknowledge the Instructional & Research Computing Center (IRCC) at Florida International University for providing HPC computing resources that have contributed to the research results reported within this article.

LITERATURE CITED

1. Guan, W.; Ni, Z.; Hu, Y.; Liang, W.; Ou, C.; He, J.; Liu, L.; Shan, H.; Lei, C.; Hui, D.S.C.; et al. Clinical Characteristics of Coronavirus Disease 2019 in China. *N. Engl. J. Med.* **2020**, *382*, 1708–1720.

2. Wang, D.; Hu., B.; Hu, C.; Zhu, F.; Liu, X.; Zhang, J.; Wang, B.; Xiang, H.; Cheng, Z.; Xiong, Y.; et al. Clinical Characteristics of 138 Hospitalized Patients with 2019 Novel Coronavirus-Infected Pneumonia in Wuhan, China. *JAMA* **2020**, *323*, 1061–1069.

3. Dawson, P.; Rabold, E.M.; Laws, R.L.; Conners, E.E.; Gharpure, R.; Yin, S.; Buono, S.A.; Dasu, T.; Bhattacharyya, S.; Westergaard, R.P.; et al. Loss of Taste and Smell as Distinguishing Symptoms of Coronavirus Disease. *Clin. Infect. Dis.* **2021**, *72*, 682–685.

4. Dong, E.; Du, H.; Gardner, L. An Interactive Web-Based Dashboard to Track COVID-19 in Real Time. *Lancet Infect, Dis.* **2020**, *20*, 533–534.

5. Sah, P.; Fitzpatrick, M.C.; Zimmer, C.F.; Abdollahi, E.; Juden-Kelly, L.; Moghadas, S.M.; Singer, B.H.; Galvani, A.P.; Asymptomatic SARS-CoV-2 Infection: A Systematic Review and Meta-Analysis. *Proc. Natl. Acad. Sci. USA* **2021**, *118*, e2109229118.

6. Saviano, A.; Wrensch, F.; Ghany, M.G.; Baumert, T.F. Liver Disease and Coronavirus Disease 2019: From Pathogenesis to Clinical Care. *Hepatology* **2021**, *74*, 1088–1100.

7. Han, X.; Ye, Q. Kidney Involvement in COVID-19 and its Treatments. *J. Med. Virol*. **2021**, *93*, 1387–1395.

8. Long, B.; Brady, W.J.; Koyfman, A.; Gottlieb, M. Cardiovascular Complications in COVID-19. *Am. J. Emerg. Med*. **2020**, *38*, 1504–1507.

9. Mei, H.; Luo, L.; Hu, Y. Thrombocytopenia and Thrombosis in Hospitalized Patients with COVID-19. *J. Hematol. Oncol.* **2020**, *13*, 161.

10. Wei, J.; Matthews, P.C.; Stoesser, N.; Maddox, T.; Lorenzi, L.; Studley, R.; Bell, J.I.; Newton, J.N.; Farrar, J.; Diamond, I.; et al. Anti-Spike Antibody Response to Natural SARS-CoV-2 Infection in the General Population. *Nat. Commun.* **2021**, *12*, 6250.

11. Wang, E.Y.; Mao, T.; Klein, J.; Dai, Y.; Huck, J.D.; Jaycox, J.R.; Liu, F.; Zhou, T.; Israelow, B.; Wong, P.; et al. Diverse Functional Autoantibodies in Patients with COVID-19. *Nature* **2021**, *595*, 283–288.

12. Getts, D.R., Chastain, E.M.; Terry, R.L.; Miller, S.D. Virus Infection, Antiviral Immunity, and Autoimmunity. *Front. Immunol*. **2013**, *255*, 197–209.

13. Agrawal, B. Heterologous Immunity: Role in Natural and Vaccine-Induced Resistance to Infections. *Front. Immunol.* **2019**, *10*, 2631.

14. Fraley, E.; LeMaster, C.; Banerjee, D.; Khanal, S.; Selvarangan, R.; Bradley, T. Cross-Reactive Antibody Immunity against SARS-CoV-2 in Children and Adults. *Cell. Mol. Immunol.* **2021**, *18*, 1826–1828.

15. Shang, J.; Wan, Y.; Luo, C.; Ye, G.; Geng, Q.; Auerbach, A.; Li, F. Cell Entry Mechanisms of SARS-CoV-2. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 11727–11734.

16. Voss, C.; Esmail, S.; Liu, X.; Knauer, M.J., Ackloo, S.; Kaneko, T.; Lowes, L.; Stogios, P.; Seitova, A.; Hutchinson, A.; et al. Epitope-Specific Antibody Responses Differentiate COVID-19 Outcomes and Variants of Concern. *JCI Insight* **2021**, *6*, e148855.

17. Segal, Y.; Shoenfeld, Y. Vaccine-Induced Autoimmunity: The Role of Molecular Mimicry and Immune Crossreaction. *Cell. Mol. Immunol.* **2018**, *15*, 586–594.

18. Kanduc, D. Anti-SARS-CoV-2 Immune Responses to COVID-19 via Molecular Mimicry. *Antibodies* **2020**, *9*, 33.

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

20. O'donoghue, S.I.; Schafferhans, A.; Sikta, N.; Stolte, C.; Kaur, S.; Ho, B.K.; Anderson, S.; Procter, J.B.; Dallago, C.; Bordin, N.; et al. SARS-CoV-2 Structural Coverage Map Reveals Viral Protein Assembly, Mimicry, and Hijacking Mechanisms. *Mol. Syst. Biol.* **2021**, *17*, e10079.

21. Khavinson, V.; Terekhov, A.; Kormilets, D.; Maryanovich, A. Homology between SARS-CoV-2 and Human Proteins. *Sci. Rep.* **2021**, *11*, 17199.

22. Balbin, C.A.; Nunez-Castilla, J.; Stebliankin, V.; Baral, P.; Sobhan, M.; Cickovski, T.; Mondal, A.M.; Narasimhan, G.; Chapagain, P.; Mathee, K.; et al. Epitopedia: Identifying Molecular Mimicry of Known Immune Epitopes. *BioRxiv* **2021**.

23. Cai, Y.; Zhang, J.; Xiao, T.; Peng, H.; Sterling, S.M.; Walsh, R.M.; Rawson, S.; Rits-Volloch, S.; Chen, B. Distinct Conformational States of SARS-CoV-2 Spike Protein. *Science* **2020**, *369*, 1586–1592.

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

25. Tunyasuvunakool, K.; Adler, J.; Wu, Z.; Green, T.; Zielinski, M.; Žídek, A.; Bridgland, A.; Cowie, A.; Meyer, C.; Laydon, A.; et al. Highly Accurate Protein Structure Prediction for the Human Proteome. *Nature* **2021**, *596*, 590–596.

26. Zhang, Y.; Skolnick, J. TM-Align: A Protein Structure Alignment Algorithm Based on the TM-Score. *Nucleic Acids Res.* **2005**, *33*, 2302–2309.

27. Wrapp, D.; Wang, N.; Corbett, K.S.; Goldsmith, J.A.; Hsieh, C.L.; Abiona, O.; Graham, B.S.; McLellan, J.S. Cryo-EM Structure of the 2019-NCoV Spike in the Prefusion Conformation. *Science* **2020**, *367*, 1255–1260.

28. Choi, Y.K.; Cao, Y.; Frank, M.; Woo, H.; Park, S.J.; Yeom, M.S.; Croll, T.I.; Seok, C.; Im, W. Structure, Dynamics, Receptor Binding, and Antibody Binding of the Fully Glycosylated Full-Length SARS-CoV-2 Spike Protein in a Viral Membrane. *J. Chem. Theory Comput.* **2021**, *17*, 2479–2487.

29. Pinto, D.; Sauer, M.M.; Czudnochowski, N.; Low, J.S.; Tortorici, M.A.; Housley, M.P.; Noack, J.; Walls, A.C.; Bowen, J.E.; Guarino, B.; et al. Broad Betacoronavirus Neutralization by a Stem Helix–Specific Human Antibody. *Science* **2021**, *373*, 1109–1116.

30. Schrödinger. *The PyMOL Molecular Graphics System*; Version 2.5.0; Schrödinger: New York, NY, USA, 2015.

31. Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. *J. Mol. Graph.* **1996**, *14*, 33–38.

32. Jo, S.; Kim, T.; Iyer, V.G.; Im, W. CHARMM-GUI: A Web-Based Graphical User Interface for CHARMM. *J. Comput. Chem.* **2008**, *29*, 1859–1865.

33. Brooks, B.R.; Brooks, C.L.; Mackerell, A.D.; Nilsson, L.; Petrella, R.J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; et al. CHARMM: The Biomolecular Simulation Program. *J. Comput. Chem.* **2009**, *30*, 1545–1614.

34. Lee, J.; Cheng, X.; Swails, J.M.; Yeom, M.S.; Eastman, P.K.; Lemkul, J.A.; Wei, S.; Buckner, J.; Jeong, J.C.; Qi, Y.; et al. CHARMM-GUI Input Generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM Simulations Using the CHARMM36 Additive Force Field. *J. Chem. Theory Comput.* **2016**, *12*, 405–413.

35. Phillips, J.C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R.D.; Kalé, L.; Schulten, K. Scalable Molecular Dynamics with NAMD. *J. Comput. Chem.* **2005**, *26*, 1781–1802.

36. Nosé, S.; Klein, M.L. Constant Pressure Molecular Dynamics for Molecular Systems. *Mol. Phys.* **1983**, *50*, 1055–1076.

37. Essmann, U.; Perera, L.; Berkowitz, M.L.; Darden, T.; Lee, H.; Pedersen, L.G. A Smooth Particle Mesh Ewald Method. *J. Chem. Phys.* **1995**, *103*, 8577–8593.

38. Ryckaert, J.P.; Ciccotti, G.; Berendsen, H.J.C. Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of n-Alkanes. *J. Comput. Phys.* **1977**, *23*, 327–341.

39. Xue, L.C.; Rodrigues, J.P.; Kastritis, P.L.; Bonvin, A.M.; Vangone, A. PRODIGY: A Web Server for Predicting the Binding Affinity of Protein-Protein Complexes. *Bioinformatics* **2016**, *32*, 3676–3678.

40. Gainza, P.; Sverrisson, F.; Monti, F.; Rodolà, E.; Boscaini, D.; Bronstein, M.M.; Correia, B.E. Deciphering Interaction Fingerprints from Protein Molecular Surfaces Using Geometric Deep Learning. *Nat. Methods* **2019**, *17*, 184–192.

41. Sanner, M.F.; Olson, A.J.; Spehner, J.C. Reduced Surface: An Efficient Way to Compute Molecular Surfaces. *Biopolymers* **1996**, *38*, 305–320.

42. Stebliankin, V.; Baral, P.; Balbin, C.; Nunez-Castilla, J.; Sobhan, M.; Cickovski, T.; Mohan Mondal, A.; Siltberg-Liberles, J.; Chapagain, P.; Mathee, K.; et al. EMoMiS: A Pipeline for Epitope-Based Molecular Mimicry Search in Protein Structures with Applications to SARS-CoV-2. *BioRxiv* **2022**.

43. Dunbar, J.; Krawczyk, K.; Leem, J.; Baker, T.; Fuchs, A.; Georges, G.; Shi, J.; Deane, C.M. SAbDab: The Structural Antibody Database. *Nucleic Acids Res.* **2014**, *42*, D1140–D1146.

44. NCBI RefSeq Select. Available online: https://www.ncbi.nlm.nih.gov/refseq/refseq_select/ (accessed on 4 August 2021).

45. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596*, 583–589.

46. Virtanen, P.; Gommers, R.; Oliphant, T.E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; et al. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nat. Methods* **2020**, *173*, 261–272.

47. Premkumar, L.; Segovia-Chumbez, B.; Jadi, R.; Martinez, D.R.; Raut, R.; Markmann, A.J.; Cornaby, C.; Bartelt, L.; Weiss, S.; Park, Y.; et al. The Receptor-Binding Domain of the Viral Spike Protein Is an Immunodominant and Highly Specific Target of Antibodies in SARS-CoV-2 Patients. *Sci. Immunol.* **2020**, *5*, 8413.

48. Takeda, M. Proteolytic Activation of SARS-CoV-2 Spike Protein. *Microbiol. Immunol.* **2022**, *66*, 15–23.

49. Borgo, C.; D'Amore, C.; Sarno, S.; Salvi, M.; Ruzzene, M. Protein Kinase CK2: A Potential Therapeutic Target for Diverse Human Diseases. *Signal Transduct. Target. Ther.* **2021**, *6*, 183.

50. Bouhaddou, M.; Memon, D.; Meyer, B.; White, K.M.; Rezelj, V.V.; Correa Marrero, M.; Polacco, B.J.; Melnyk, J.E.; Ulferts, S.; Kaake, R.M.; et al. The Global Phosphorylation Landscape of SARS-CoV-2 Infection. *Cell* **2020**, *182*, 685–712.e19.

51. Finn, R.D.; Bateman, A.; Clements, J.; Coggill, P.; Eberhardt, R.Y.; Eddy, S.R.; Heger, A.; Hetherington, K.; Holm, L.; Mistry, J.; et al. Pfam: The Protein Families Database. *Nucleic Acids Res.* **2014**, *42*, D222–D230.

52. Varghese, L.N.; Defour, J.-P.; Pecquet, C.; Constantinescu, S.N. The Thrombopoietin Receptor: Structural Basis of Traffic and Activation by Ligand, Mutations, Agonists, and Mutated Calreticulin. *Front. Endocrinol.* **2017**, *8*, 59.

53. Yang, X.; Yang, Q.; Wang, Y.; Wu, Y.; Xu, J.; Yu, Y.; Shang, Y. Thrombocytopenia and Its Association with Mortality in Patients with COVID-19. *J. Thromb. Haemost.* **2020**, *18*, 1469–1472.

54. Shi, C.; Wang, L.; Ye, J.; Gu, Z.; Wang, S.; Xia, J.; Xie, Y.; Li, Q.; Xu, R.; Lin, N. Predictors of Mortality in Patients with Coronavirus Disease 2019: A Systematic Review and Meta-Analysis. *BMC Infect. Dis.* **2021**, *21*, 663.

55. Nazy, I.; Kelton, J.G.; Moore, J.C.; Clare, R.; Horsewood, P.; Smith, J.W.; Ivetic, N.; D'Souza, V.; Li, N.; Arnold, D.M. Autoantibodies to Thrombopoietin and the Thrombopoietin Receptor in Patients with Immune Thrombocytopenia. *Br. J. Haematol.* **2018**, *181*, 234–241.

56. Audia, S.; Bonnotte, B. Emerging Therapies in Immune Thrombocytopenia. *J. Clin. Med.* **2021**, *10*, 1004.

57. Watts, A.; Raj, K.; Gogia, P.; Gahona, C.C.T.; Porcelli, M. Secondary Immune Thrombocytopenic Purpura Triggered by COVID-19. *Cureus* **2021**, *13*, e14501.

58. Frankel, A.E.; Wylie, D.; Peters, B.; Marrama, D.; Ahn, C. Bioinformatic Analysis Underpinning the Frequent Occurence of Immune Thrombocytopenic Purpura in COVID-19 Patients. *Isr. Med. Assoc. J.* **2022**, *24*, 320–326.

59. Mishra, N.; Huang, X.; Joshi, S.; Guo, C.; Ng, J.; Thakkar, R.; Wu, Y.; Dong, X.; Li, Q.; Pinapati, R.S.; et al. Immunoreactive Peptide Maps of SARS-CoV-2. *Commun. Biol.* **2021**, *4*, 225.

60. Kanehisa, M.; Furumichi, M.; Sato, Y.; Ishiguro-Watanabe, M.; Tanabe, M. KEGG: Integrating Viruses and Cellular Organisms. *Nucleic Acids Res.* **2021**, *49*, D545–D551.

61. Kuter, D.J. The Biology of Thrombopoietin and Thrombopoietin Receptor Agonists. *Int. J. Hematol.* **2013**, *98*, 10–23.

62. Feese, M.D.; Tamada, T.; Kato, Y.; Maeda, Y.; Hirose, M.; Matsukura, Y.; Shigematsu, H.; Muto, T.; Matsumoto, A.; Watarai, H.; et al. Structure of the Receptor-Binding Domain of Human Thrombopoietin Determined by Complexation with a Neutralizing Antibody Fragment. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 1816–1821.

63. Cerutti, G.; Guo, Y.; Zhou, T.; Gorman, J.; Lee, M.; Rapp, M.; Reddem, E.R.; Yu, J.; Bahna, F.; Bimela, J.; et al. Potent SARS-CoV-2 Neutralizing Antibodies Directed against Spike N-Terminal Domain Target a Single Supersite. *Cell Host Microbe* **2021**, *29*, 819–833.e7.

64. Tahara, T.; Kuwaki, T.; Matsumoto, A.; Morita, H.; Watarai, H.; Inagaki, Y.; Ohashi, H.; Ogami, K.; Miyazaki, H.; Kato, T. Neutralization of Biological Activity and Inhibition of Receptor Binding by Antibodies against Human Thrombopoietin. *Stem Cells* **1998**, *16*, 54–60.

65. Taylor, W.R. Residual Colours: A Proposal for Aminochromography. *Protein Eng.* **1997**, *10*, 743–746.

66. Waterhouse, A.M.; Procter, J.B.; Martin, D.M.A.; Clamp, M.; Barton, G.J. Jalview Version 2—A Multiple Sequence Alignment Editor and Analysis Workbench. *Bioinformatics* **2009**, *25*, 1189–1191.

67. Woo, H.; Park, S.J.; Choi, Y.K.; Park, T.; Tanveer, M.; Cao, Y.; Kern, N.R.; Lee, J.; Yeom, M.S.; Croll, T.I.; et al. Developing a Fully Glycosylated Full-Length SARS-CoV-2 Spike Protein Model in a Viral Membrane. *J. Phys. Chem. B* **2020**, *124*, 7128–7137.

68. Kanduc, D. Thromboses and Hemostasis Disorders Associated with Coronavirus Disease 2019: The Possible Causal Role of Cross-Reactivity and Immunological Imprinting. *Glob. Med. Genet.* **2021**, *8*, 162–170.

69. Chivukula, R.R.; Montoro, D.T.; Leung, H.M.; Yang, J.; Shamseldin, H.E.; Taylor, M.S.; Dougherty, G.W.; Zariwala, M.A.; Carson, J.; Daniels, L.A.; et al. A Human Ciliopathy Reveals Essential Functions for NEK10 in Airway Mucociliary Clearance. *Nat. Med.* **2020**, *26*, 244.

70. Andries, J.; Viranaicken, W.; Cordonin, C.; Herrscher, C.; Planesse, C.; Roquebert, B.; Lagrange-Xelot, M.; El-Kalamouni, C.; Meilhac, O.; Mavingui, P.; et al. The SARS-CoV-2 Spike Residues 616/644 and 1138/1169 Delineate Two Antibody Epitopes in COVID-19 MRNA COMINARTY Vaccine (Pfizer/BioNTech). *Sci. Rep.* **2022**, *12*, 5999.

71. Respiratory Syncytial Virus (RSV)|NIH: National institute of Allergy and Infectious Diseases. Available online: https://www.niaid.nih.gov/diseases-conditions/respiratory-syncytial-virus-rsv (accessed on 8 January 2022).

72. Li, Z.; Xi, X.; Gu, M.; Feil, R.; Ye, R.D.; Eigenthaler, M.; Hofmann, F.; Du, X. A Stimulatory Role for CGMP-Dependent Protein Kinase in Platelet Activation. *Cell* **2003**, *112*, 77–86.

73. Sauzeau, V.; Le Jeune, H.; Cario-Toumaniantz, C.; Smolenski, A.; Lohmann, S.M.; Bertoglio, J.; Chardin, P.; Pacaud, P.; Loirand, G. Cyclic GMP-Dependent Protein Kinase Signaling Pathway Inhibits RhoA-Induced Ca^{2+} Sensitization of Contraction in Vascular Smooth Muscle. *J. Biol. Chem.* **2000**, *275*, 21722–21729.

74. Francis, S.H. The Role of CGMP-Dependent Protein Kinase in Controlling Cardiomyocyte CGMP. *Circ. Res.* **2010**, *107*, 1164.

75. Francis, S.H.; Busch, J.L.; Corbin, J.D. CGMP-Dependent Protein Kinases and CGMP Phosphodiesterases in Nitric Oxide and CGMP Action. *Pharmacol. Rev.* **2010**, *62*, 525.

76. Szent-Györgyi, A.G. Calcium Regulation of Muscle Contraction. *Biophys. J.* **1975**, *15*, 707–723.

77. Marrama, D.; Mahita, J.; Sette, A.; Peters, B. Lack of Evidence of Significant Homology of SARS-CoV-2 Spike Sequences to Myocarditis-Associated Antigens. *EBioMedicine* **2022**, *75*, 103807.

78. Zhang, Y.; Zhao, H.; Liu, B.; Li, L.; Zhang, L.; Bao, M.; Ji, X.; He, X.; Yi, J.; Chen, P.; et al. Low Level Antibodies Against Alpha-Tropomyosin Are Associated With Increased Risk of Coronary Heart Disease. *Front. Pharmacol.* **2020**, *11*, 195.

79. Nishiga, M.; Wang, D.W.; Han, Y.; Lewis, D.B.; Wu, J.C. COVID-19 and Cardiovascular Disease: From Basic Mechanisms to Clinical Perspectives. *Nat. Rev. Cardiol.* **2020**, *17*, 543–558.

80. Patone, M.; Mei, X.W.; Handunnetthi, L.; Dixon, S.; Zaccardi, F.; Shankar-Hari, M.; Watkinson, P.; Khunti, K.; Harnden, A.; Coupland, C.A.C.; et al. Risks of Myocarditis, Pericarditis, and Cardiac Arrhythmias Associated with COVID-19 Vaccination or SARS-CoV-2 Infection. *Nat. Med.* **2021**, *28*, 410–422.

81. Xie, Y.; Xu, E.; Bowe, B.; Al-Aly, Z. Long-Term Cardiovascular Outcomes of COVID-19. *Nat. Med.* **2022**, *28*, 583–590.

82. Li, Y.; Lai, D.; Zhang, H.; Jiang, H.; Tian, X.; Ma, M.; Qi, H.; Meng, Q.; Guo, S.; Wu, Y.; et al. Linear Epitopes of SARS-CoV-2 Spike Protein Elicit Neutralizing Antibodies in COVID-19 Patients. *Cell. Mol. Immunol.* **2020**, *17*, 1095–1097.

83. Stoddard, C.I.; Galloway, J.; Chu, H.Y.; Shipley, M.M.; Sung, K.; Itell, H.L.; Wolf, C.R.; Logue, J.K.; Magedson, A.; Garrett, M.E.; et al. Epitope Profiling Reveals Binding Signatures of SARS-CoV-2 Immune Response in Natural Infection and Cross-Reactivity with Endemic Human CoVs. *Cell Rep.* **2021**, *35*, 109164

84. Helms, J.M.; Ansteatt, K.T.; Roberts, J.C.; Kamatam, S.; Foong, K.S.; Labayog, J.M.S.; Tarantino, M.D. Severe, Refractory Immune Thrombocytopenia Occurring after SARS-CoV-2 Vaccine. *J. Blood Med.* **2021**, *12*, 221–224.

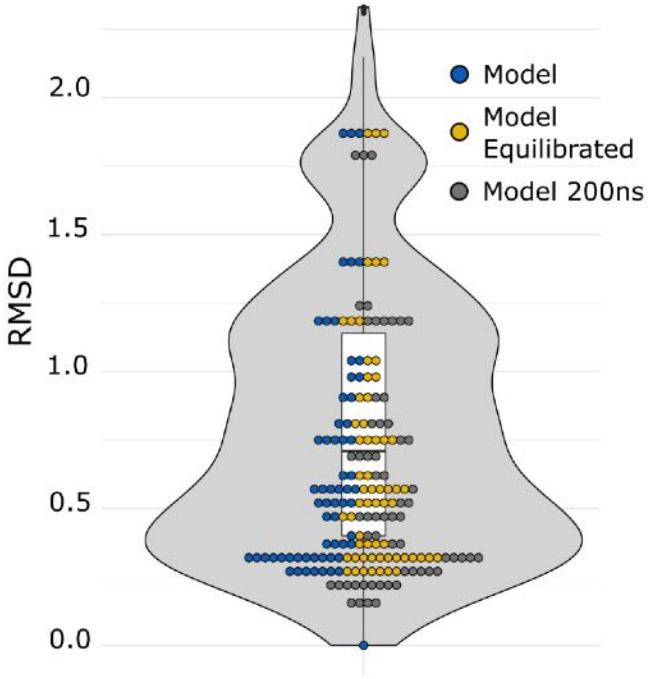
85. Schultz, N.H.; Sørvoll, I.H.; Michelsen, A.E.; Munthe, L.A.; Lund-Johansen, F.; Ahlen, M.T.; Wiedmann, M.; Aamodt, A.-H.; Skattør, T.H.; Tjønnfjord, G.E.; et al. Thrombosis and Thrombocytopenia after ChAdOx1 NCoV-19 Vaccination. *N. Engl. J. Med.* **2021**, *384*, 2124–2130.

86. Greinacher, A.; Thiele, T.; Warkentin, T.E.; Weisser, K.; Kyrle, P.A.; Eichinger, S. Thrombotic Thrombocytopenia after ChAdOx1 NCov-19 Vaccination. *N. Engl. J. Med.* **2021**, *384*, 2092–2101.

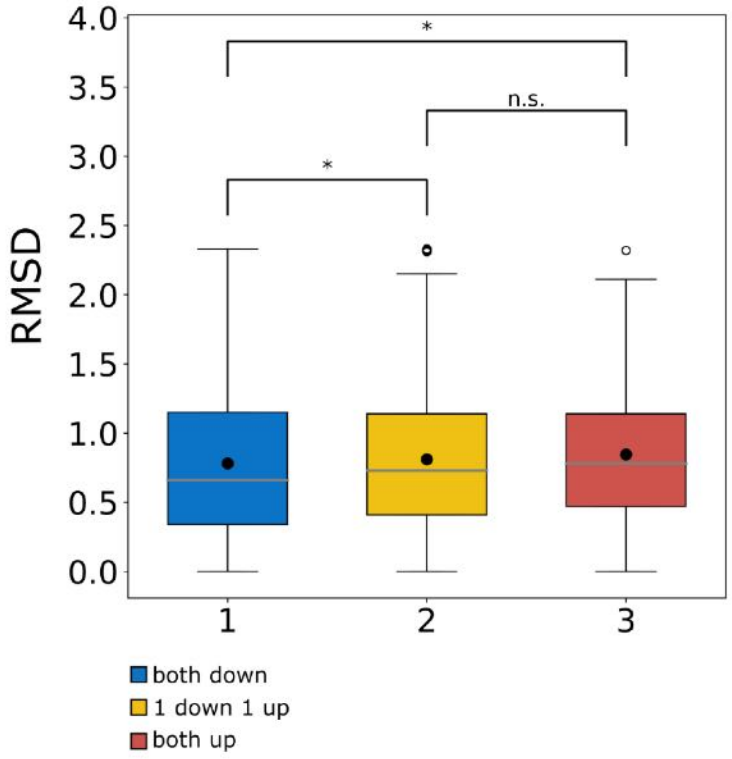
87. Hodcroft, E.B. CoVariants: SARS-CoV-2 Mutations and Variants of Interest. **2022**. Available online: https://covariants.org/shared-mutations (accessed on 8 January 2022).

88. Slabinski, L.; Jaroszewski, L.; Rodrigues, A.P.C.; Rychlewski, L.; Wilson, I.A.; Lesley, S.A.; Godzik, A. The Challenge of Protein Structure Determination—Lessons from Structural Genomics. *Protein Sci.* **2007**, *16*, 2472–2482.

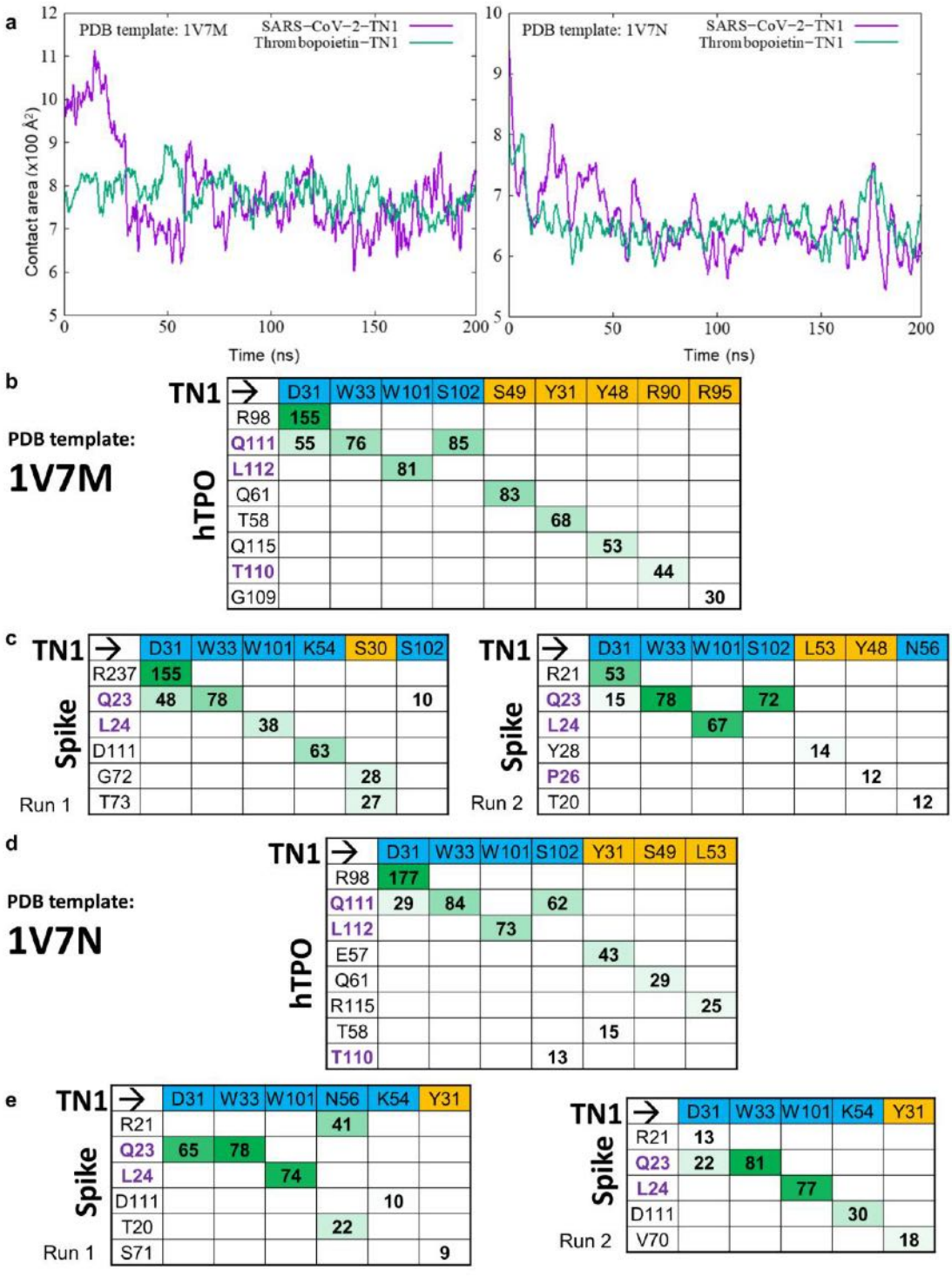
Appendices



Appendix 1. RMSD value distribution for solved and modeled Spike TQLPP regions. RMSD values resulting from an all-against-all comparison of the Spike TQLPP region of 63 structures, including the model in 3 states (shown as dots).

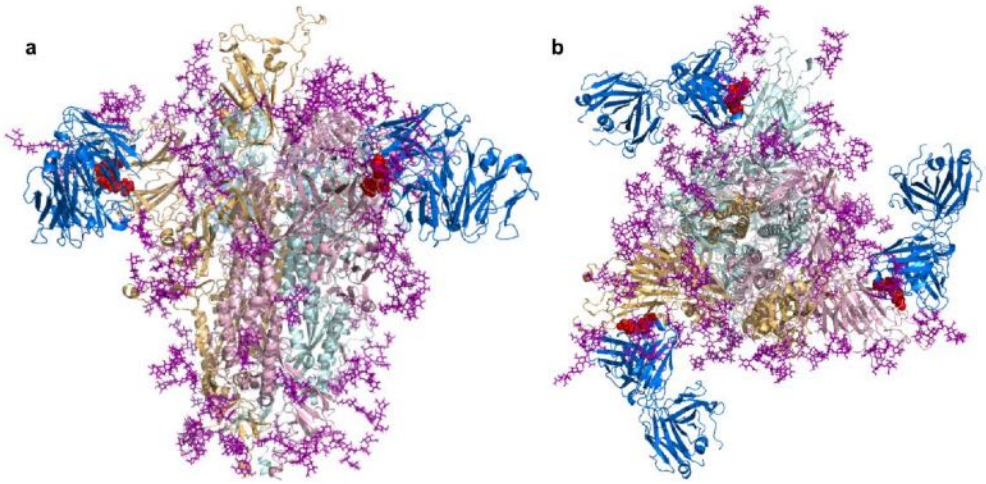


Appendix 2. Comparison of RMSD values for TQLPP region from 20 Spike trimer structures based on RBD state. Box plots show distribution of RMSD values for Spike TQLPP where RBDs are: (1) both down (blue, N = 666, mean = 0.78 Å, median = 0.66 Å), (2) 1 down and 1 up (yellow, N = 962, mean = 0.81 Å, median = 0.73 Å), (3) both up (red, N = 325, mean = 0.85 Å, median = 0.78 Å). Statistical testing was performed using the Mann-Whitney U test. Brackets marked with an asterisk (*) denote statistically significant comparisons while those marked "n.s." denote non-significant comparisons. Groups 1 and 2 (p-value = 0.30) and 1 and 3 (p-value = 0.003) are significantly different but groups 2 and 3 (p-value = 0.055) are not. Box plots, bounded by the 1st and 3rd quartiles, show mean (black dot) and median values (horizontal solid gray line), vertical lines (whiskers) represent 1.5 × IQR, while outliers are marked as open circles.

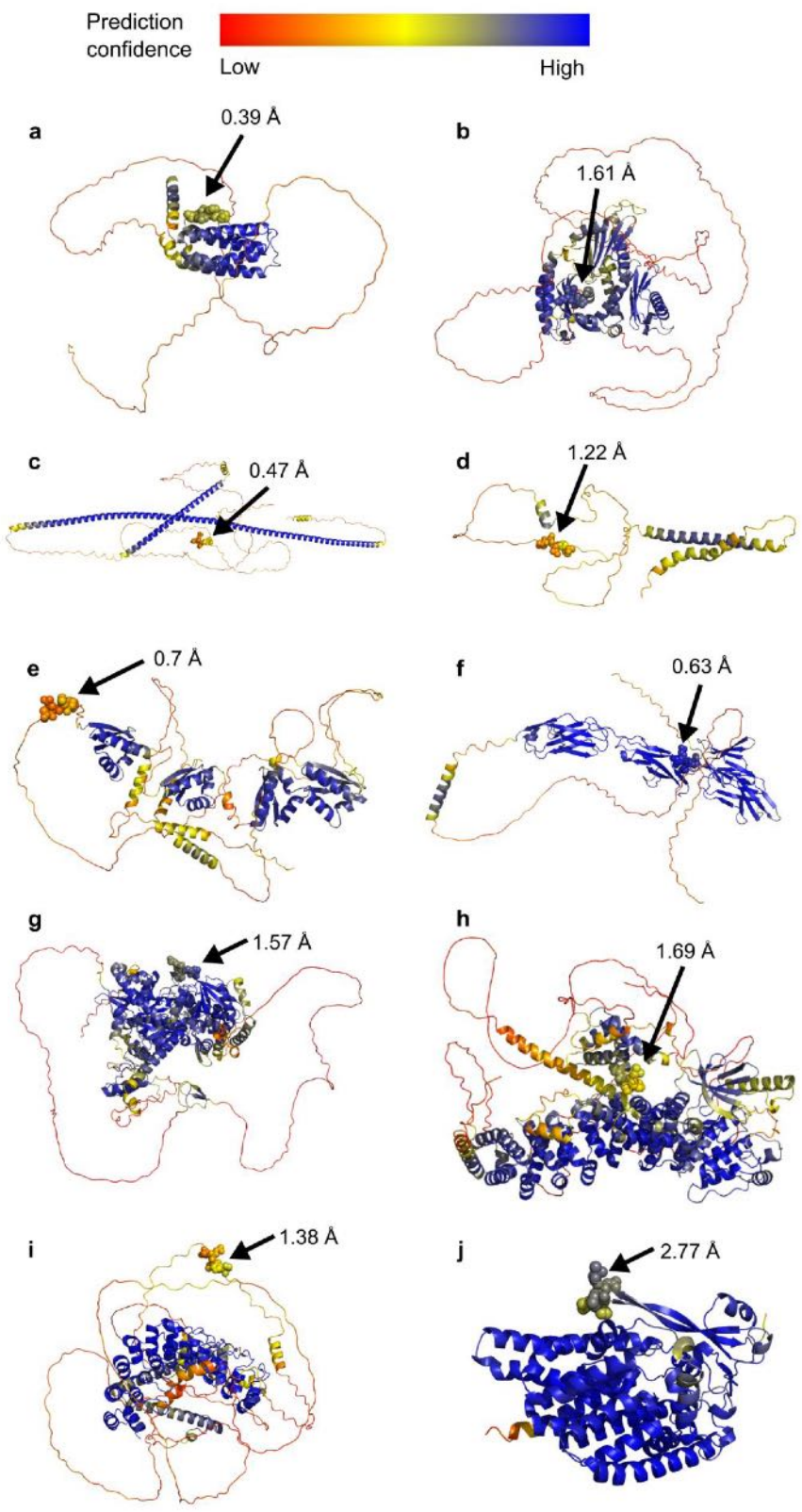


Appendix 3. Molecular dynamics simulations overview. (**a**) Time evolution of the protein-antibody binding interface contact areas (100x Å2) for Spike-TN1 (purple) and thrombopoietin-TN1 (green) in the molecular dynamics trajectories for PDB templates 1V7M (left) and 1V7N (right). Interaction matrices showing hydrogen bond contribution during the last 50 ns of 200 ns simulations between amino acid residue pairs ordered

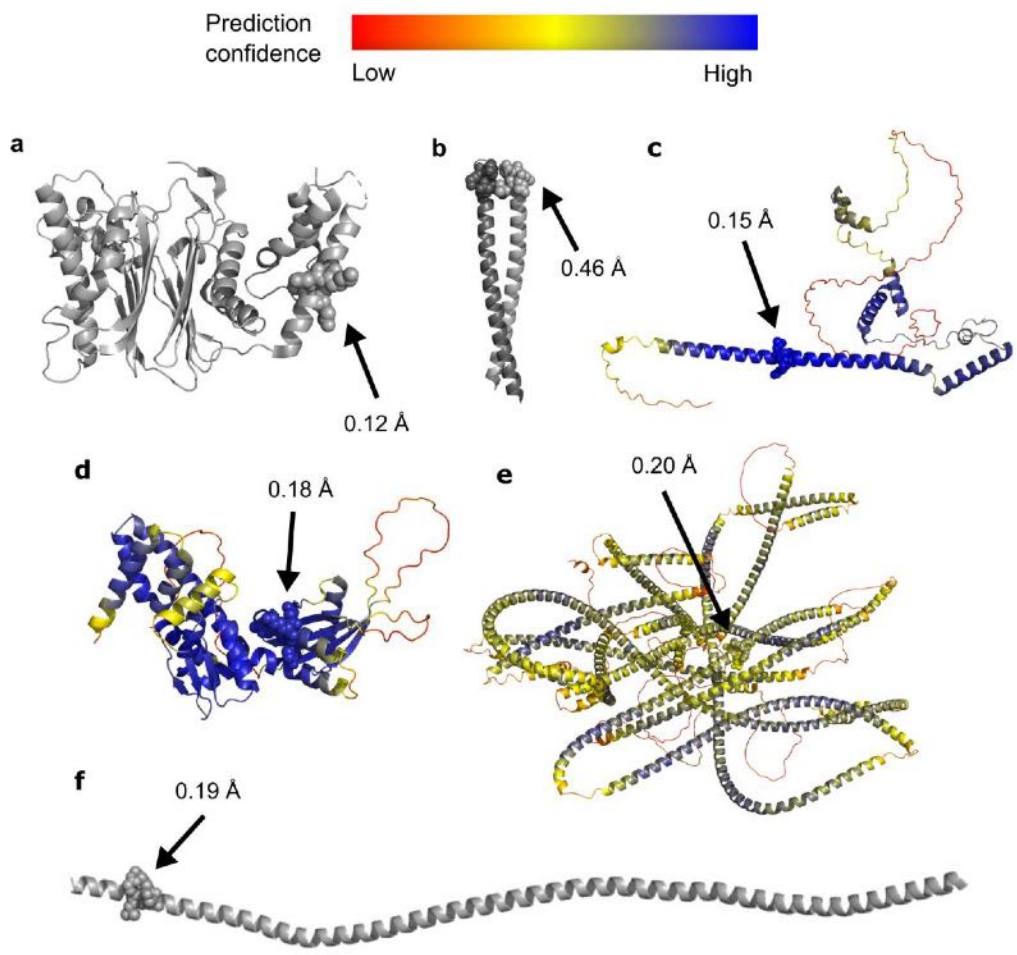
according to their hydrogen-bond occupancies for the (**b**, **d**) hTPO-TN1 and (**c**, **e**) Spike-TN1 complexes for PDB template 1V7M and 1V7N, respectively. Residues belonging to TQLPP are colored in purple and positions for hTPO are based on the PDB template. TN1 Fab residues from heavy and light chains are shaded blue and yellow, respectively.



Appendix 4. SARS-CoV-2 Spike bound to TN1 Fab antibody. SARS-CoV-2 Spike shown in the trimeric state (PDB id: 6VSB) bound to TN1 Fab antibody (blue, PDB id: 1V7M) as viewed from (**a**) the side and (**b**) the top. The TQLPP motifs are shown as red spheres and glycans are shown in purple. Structural visualization generated with PyMOL [30].


Prediction confidence
Low
High
a
0.39 Å
b
1.61 Å
c
0.47 Å
d
1.22 Å
e
0.7 Å
f
0.63 Å
g
1.57 Å
h
1.69 Å
i
1.38 Å
j
2.77 Å

Appendix 5. TQLPP motif for 10 human proteins modeled by AlphaFold2. Protein structure models are colored by AlphaFold confidence estimate according to the color bar where red = 25 (low confidence) and blue = 100 (high confidence). TQLPP motif is shown as spheres. RMSD for human TQLPP in the 10 proteins compared to SARS-CoV-2 Spike (PDB id: 6XR8, chain A) is shown. The proteins are (**a**) thrombopoietin (Uniprot: P40225), (**b**) Hermansky-Pudlak syndrome 4 protein (Uniprot: Q9NQG7), (**c**) coiled-coil domain containing protein 85 (Uniprot: Q8N715), (**d**) transmembrane protein 52 precursor (Uniprot: Q8NDY8), (**e**) far upstream element-binding protein 1 (Uniprot: Q96AE4), (**f**) Fc receptor-like protein 4 (Uniprot: Q96PJ5), (**g**) DNA annealing helicase and endonuclease ZRANB3 (Uniprot: Q5FWF4), (**h**) serine/threonine-protein kinase NEK10 (Uniprot: Q6ZWH5), (**i**) espin (Uniprot: B1AK53), and (**j**) ALG12 (Mannosyltransferase ALG12 homolog, Uniprot: Q9BV10). Structural visualization generated with PyMOL [30].



Appendix 6. Structure of ELDKY motif for 5 human proteins. Protein structures from PDB are colored gray while AlphaFold2 3D models are colored by AlphaFold confidence estimate according to the color bar where red = 25 (low confidence) and blue = 100 (high confidence). ELDKY motif is shown as spheres. RMSD for human ELDKY in the 5 proteins compared to SARS-CoV-2 Spike (PDB id: 6XR8, chain A) is shown. The proteins are (**a**) protein phosphatase 1A (PDB id: 3FXJ), (**b**) leucine zipper domain of cGMP-dependent protein kinase 1 (PDB id: 3NMD), (**c**) protein FAM228B (Uniprot: P0C875), (**d**) protein Njmu-R1 (Uniprot: Q9HAS0), (**e**) thyroid receptor interacting protein 11 (Uniprot: Q15643), and (f) tropomyosin alpha-1 (PDB id: 6X5Z). Structural visualization generated with PyMOL [30].

Appendix 7. RMSD values resulting from the alignment of the TQLPP region of 1V7M chain X and 1V7N chain X against the TQLPP region of 60 Spike structures. Sorted by RMSD.

1V7M X		1V7N X	
Spike Structure	**RMSD**	**Spike Structure**	**RMSD**
6ZGG A	0.36	7DCC E	0.21
6ZGG B	0.40	7DCC I	0.27
7BNN B	0.43	7DCC K	0.28
6ZGG C	0.44	7BNN B	0.42
7DCC E	0.44	7BNM C	0.44
7DCC I	0.48	7BNM B	0.44
7DCC K	0.48	7BNM A	0.44
7BNM A	0.52	7A25 C	0.46
7BNM B	0.52	6XR8 A	0.46
7BNM C	0.52	6ZGE C	0.47
7A25 C	0.59	6ZGE A	0.47
6ZGE A	0.60	6ZGE B	0.48
6ZGE B	0.60	6ZGG A	0.49
6ZGE C	0.60	7KMK B	0.49
7BNN A	0.60	7LRT B	0.49
6XR8 A	0.61	6ZGG C	0.49
7A25 A	0.63	7A25 A	0.50
7LRT B	0.66	6XR8 B	0.51
7LRT C	0.66	7LRT C	0.51
6XR8 B	0.68	6ZGG B	0.53
6XR8 C	0.71	6XR8 C	0.55
7A25 B	0.71	7A25 B	0.57
6ZP2 A	0.72	7LRT A	0.58
6ZP2 B	0.72	7N1U A	0.59
6ZP2 C	0.72	7KRQ A	0.61
7KMK B	0.72	7BNN A	0.61
7LRT A	0.73	7KRQ B	0.62
7KRQ A	0.76	7E8C A	0.64
7KRQ B	0.76	7BNN C	0.64
7N1U A	0.76	7KRQ C	0.66
7BNN C	0.78	7KMK C	0.68
7KRQ C	0.79	7E8C C	0.71
7E8C A	0.82	7E8C B	0.72
7LQV A	0.85	7N1U C	0.73

7KMK C	0.87	7JJI A	0.75
7N1U C	0.87	7JJI B	0.75
7E8C C	0.88	7JJI C	0.75
7LQV C	0.88	7N1U B	0.76
7E8C B	0.90	7KMK A	0.78
7N1U B	0.90	7LQV A	0.78
7LQV B	0.91	7LQV C	0.80
7JJI A	0.92	7LQV B	0.82
7JJI B	0.92	6ZP2 A	0.84
7JJI C	0.92	6ZP2 C	0.84
7CWL B	0.95	6ZP2 B	0.84
7CWS R	0.95	7MJG B	0.85
7KMK A	0.96	7MJG C	0.93
7MJG B	1.02	7MJG A	0.93
7MJG A	1.09	7CWL B	0.94
7MJG C	1.10	7CWS R	0.94
7CWL A	1.17	7CWS O	1.03
7CWS O	1.17	7CWL A	1.03
7C2L A	1.21	7C2L B	1.26
7C2L B	1.21	7C2L C	1.26
7C2L C	1.21	7C2L A	1.26
7N1Q B	1.62	7N1Q B	1.56
7CWL C	1.67	7CWS Q	1.57
7CWS Q	1.67	7CWL C	1.57
7N1Q A	1.68	7N1Q A	1.61
7N1Q C	1.71	7N1Q C	1.64

Appendix 8. RMSD values resulting from the alignment of the TQLPP region from 60 Spike structures and three modeled states, representing a conformational ensemble of TQLPP in Spike, sorted by RMSD.

Separate Excel sheet

Appendix 9. RMSD values for SARS-CoV-2 Spike wild-type TQLPP compared to corresponding region in known variants of concern.

VARIANT OF CONCERN	SPIKE PDB_CHAIN	RMSD (Å)	NOTES
Alpha	7N1U_A	0.21	
Beta	7N1Q_A	1.78	
Gamma	7SBS_A	1.17	P26S turns TQLPP to TQLPS
Delta	7SBK_A	0.69	
Omicron BA.1	7WE7_D	0.33	
Omicron BA.2	7UB0_A	N/A	Contains deletion of LPP
Omicron BA.2.12.1	Not available	N/A	
Omicron BA.4	Not available	N/A	
Omicron BA.5	Not available	N/A	

Appendix 10. PRODIGY binding affinities for antigen-antibody complexes

Complex	Frame	Binding Affinity (kcal/mol)	Intermolecular Contacts					
			Charged-Charged	Charged-Polar	Charged-Apolar	Polar-Polar	Polar-Apolar	Apolar-Apolar
Spike-TN1 Fab (1V7M template)	1	-11.5	3	18	11	20	35	19
	2	-8.6	0	5	2	7	16	20
	3	-9.2	2	8	8	6	15	16
	4	-8.8	3	6	3	7	15	14
	5	-9.7	3	7	8	10	20	22
	Mean	-9.56	2.2	8.8	6.4	10	20.2	18.2
	Std Dev	1.16	1.30	5.26	3.78	5.79	8.53	3.19
Spike-TN1 Fab (1V7N template)	1	-8.7	4	5	3	6	13	13
	2	-9.6	1	6	4	5	18	16
	3	-9.1	0	6	7	3	13	17
	4	-10.1	0	5	6	5	19	17
	5	-8.5	0	6	4	4	12	14
	Mean	-9.2	1	5.6	4.8	4.6	15	15.4
	Std Dev	0.66	1.73	0.55	1.64	1.14	3.24	1.82
hTPO-TN1 Fab (1V7M)	1	-9.2	1	11	9	10	20	16
	2	-10.3	1	9	10	7	21	20
	3	-9.2	1	9	9	7	17	16
	4	-9.5	2	10	8	9	20	15
	5	-9.3	3	8	6	9	20	18
	Mean	-9.5	1.6	9.4	8.4	8.4	19.6	17
	Std Dev	0.46	0.89	1.14	1.52	1.34	1.52	2.00
	1	-9.3	1	8	5	4	16	11
	2	-9.2	1	9	8	3	14	13

hTPO-TN1 Fab (1V7N)	**3**	-9.1	1	6	6	4	15	11
	4	-9.7	2	7	6	5	14	9
	5	-9.9	2	7	5	5	20	13
	Mean	-9.24	1.4	7.4	6	4.2	15.8	11.4
	Std Dev	0.43	0.55	1.14	1.22	0.84	2.49	1.67
Spike-S2P6 Fab (7RNJ template)	**1**	-9.5	2	5	10	0	10	13
	2	-9.2	2	6	8	0	10	16
	3	-9.6	1	5	11	0	11	16
	4	-9.9	1	6	11	0	11	17
	5	-9.4	2	5	10	0	11	16
	Mean	-9.52	1.6	5.4	10	0	10.6	15.6
	Std Dev	0.26	0.55	0.55	1.22	0.00	0.55	1.52

Appendix 11. Distribution of MaSIF binding confidence scores.

Separate Excel sheet

Appendix 12. Statistical comparison of MaSIF binding confidence scores for antibody complexes.

MaSIF binding confidence scores			
Comparison		**p-value**	**Significant[1]**
Random	Spike-Ab	4.83E-75	Yes
Random	hTPO-TN1	3.00E-08	Yes
Random	Spike-TN1	5.24E-26	Yes
Spike-Ab	hTPO-TN1	3.37E-02	No
Spike-Ab	Spike-TN1	7.68E-09	Yes
hTPO-TN1	Spike-TN1	1.92E-04	Yes

[1] Compared to Bonferroni corrected p-value (<8.33E-03) for alpha = 0.05

Appendix 13. RefSeq Select human isoforms that contain pentapeptides found in the 3D-mimics and AF-3D-mimics for SARS-CoV-2 Spike.

Separate Excel sheet

Appendix 14. MASIF binding confidence scores of other human proteins in complex with TN1.

UNIPROT ACCESSION[1]	PROTEIN NAME	MASIF BINDING SCORE	CONTACT PROTEIN[2]	CONTACT TN1
Q6ZWH5	NEK10	1.44195044	1047 GLN	102 SER
Q9BV10	ALG12	1.897805691	466 GLN	102 SER
Q96PJ5	FCRL4	2.539466143	215 GLN	102 SER

[1] Reference for AlphaFold2 prediction

[2] Corresponds to Q in TQLPP

CHAPTER VI

CONCLUSIONS AND FUTURE DIRECTIONS

Functional Diversification After Gene Duplication: Paralog Specific Regions of Structural Disorder and Phosphorylation in p53, p63, and p73

I evaluated roughly 300 protein sequences in the vertebrate p53 family and close to an additional 50 protein sequences from the p53 DNA-binding domain (p53 DBD) in invertebrates. In doing so, I illustrated how this protein family is functionally diverging based on sequence, structural, and regulatory properties. Broadly, it appears that vertebrate p63 is more constrained from diverging at the sequence level than either of its two paralogs, p53 and p73. This is further reflected in the high conservation seen in the p63 clade for intrinsic disorder, secondary structure, and phosphorylation. In contrast, vertebrate p53 appears to be the least constrained of the three paralogs as reflected in the large proportion of rapid evolutionary rates in this clade across sequence, structural, and regulatory properties.

In a phylogeny reconstructed based on vertebrate and invertebrate p53 DBDs, clades appear to form primarily based on the domain composition of the full-length protein. Arranging the species from this phylogeny according to their taxonomy revealed that the precursor of the metazoan p53 protein must have contained three of the four domains (p53 DBD, oligomerization domain [OD], and sterile alpha motif domain [SAM]) found in the vertebrate p53 family. Proteins with all four domains (the aforementioned three, plus the transactivation domain [TAD]) are found in gastropods, hemichordates, early chordates, suggesting the presence of a four-domain protein in the bilaterian ancestor. TAD and other non-p53 DBD domains are frequently lost in Ecdysozoa, whether due to loss of the sequence segment or depletion of the domain signature within the sequence. Overall, this indicates that many ecdysozoan p53 family

proteins, which have often lost most of their domains and consist only of a p53 DBD, are more divergent than p53 family proteins in early metazoans.

TAD exhibits high evolutionary rates for sequence and disorder-to-order transitions in the p53 clade, and in the p63 and 73 clades TAD has diverged beyond recognition by Pfam. MDM2 is a critical regulator of p53 (Lane et al., 2010) that, upon binding to TAD, ubiquinates p53 and marks it for proteasomal degradation (Chao, 2015). Binding of MDM2 to TAD is likely an ancestral function, as remnants of MDM2 binding sites have been found in p53 from early chordates (Lane et al., 2010). For p73, studies have shown that binding of MDM2 does not always result in ubiquitination (Bálint et al., 1999), and even after successful ubiquitination does not lead to degradation (Wu & Leng, 2015). The interaction between p63 and MDM2 is even weaker (Zdzalik et al., 2010). Divergent functional dependence on MDM2 is supported by the differential patterns of disorder seen in the MDM2 binding region for this protein family.

Clade-specific patterns of phosphorylation between the three paralogs are further indicative of functional divergence in this protein family. Changes in phosphorylation pattern can lead to functional diversification following gene duplication because the phosphorylation can be performed by different kinases in response diverse signals. As previously mentioned, p63 is more constrained in sequence divergence than its two paralogs. It follows, then, that the p63 clade would have more clade-specific predicted phosphorylation sites with over 50% conservation than either the p53 or p73 clades, as was observed to be the case. This would suggest that phosphorylation sites have been lost in p53 and p73. In addition, human p53 presents a different experimentally verified posttranslational modification at two of the clade-specific phosphorylation sites for p63,

further supporting functional divergence between the paralogs through different regulatory mechanisms.

As a tumor suppressor protein, p53 is responsible for preventing cancer. Despite that, mutated p53 is found in roughly 50% of human cancers (Soussi & Béroud, 2001). p53 is also often found mutated in non-cancerous cells (Martincorena et al., 2015). p63 and p73 likely represent most of the ancestral function for this family, given their higher conservation in sequence and structural properties, although divergent regions within these proteins are suggestive of ongoing functional divergence. Knockout studies reveal that p63- and p73-null mice experience high mortality while p53-null mice survive to adulthood (Stiewe, 2007), indicating that p63 and p73 are more vital than p53. Lineage-specific changes in p53 and functional redundancy between p53 and its two paralogs may allow p53 to functionally diversify in a near-neutral manner. Differing phosphorylation patterns between the paralogous clades hint at diverging signaling and interaction networks for these proteins. Further, the p53 DBD of the p53 clade has rapid disorder-to order transitions, while disorder is more conserved in the p63 and p73 clades. Increased order for some species in the p53 DNA binding region suggests functional divergence that may result in changes to DNA binding regulation. Non-conserved disorder may allow for lineage-specific modulation of fine-tuned signaling and allow for gain or loss of function(s). Ultimately, functional divergence is ongoing in the p53 family and is particularly pronounced for the p53 clade. As p53 appears to still be exploring its function, referring to it as the Guardian of the Genome seems like a misnomer. It may be more aptly referred to as the Gambler of the Genome.

Exploring Functional Constraints in the Proteomes of Zika, Dengue, and Other Flaviviruses to Identify Fitness-Critical Sites

I investigated fitness-critical sites in the flavivirus proteome experiencing evolutionary constraints in sequence and structural properties. Fitness-critical sites are considered regions of 5 or more amino acids conserved in sequence, order (lacking intrinsic disorder), and secondary structure element. These fitness-critical sites are also referred to as target sites, with the intention that they may be appropriate targets for broadly neutralizing antiviral drugs.

While flaviviruses are sufficiently divergent that I was unable to find any target sites for the full phylogeny, focusing on shorter evolutionary timescales allowed for the identification of multiple clade-specific target sites for the mosquito-borne flaviviruses (MBFVs) and the ZIKV+DENV and WNV clades. Here, two target sites were found for 19 MBFVs, while five and nine target sites were found for the ZIKV+DENV and WNV clades, respectively. Furthermore, all target sites were found to have >99% sequence conservation among ZIKV, DENV, and WNV strains, bolstering the possibility for these sites to be used as targets for broadly neutralizing antivirals.

One target site, GHLKC, identified for the ZIKV+DENV clade is found within the conformationally flexible (Kuhn et al., 2015) Envelope protein that enables viral entry (Modis et al., 2004). GHLKC is located above a flexible hinge region, called the βOG pocket, whose conformational changes promote viral infection. Targeting the βOG pocket with small molecules has inhibited viral activity for viruses in both the ZIKV+DENV and WNV clades (de Wispelaere et al., 2018). While GHLKC was only identified as a target site for the ZIKV+DENV clade, the ability of the βOG pocket to be targeted in viruses

corresponding to the WNV clade suggests that the clade-specific target sites I identified may serve as target sites for a broader group of viruses than anticipated.

I also identified two target sites for the DEAD domain of the NS3 protein, one each for the ZIKV+DENV and WNV clades. While the target site for the WNV clade is not viable given its lack of solvent accessibility, the target site for the ZIKV+DENV clade, HATFT, contains two residues that coordinate with ssRNA (Tian et al., 2016) and are found in a deep pocket when not bound to ssRNA. NS3 performs helicase functions in flaviviruses (Bollati et al., 2010) and flavivirus helicases are popular drug targets (Luo et al., 2015).

Most target sites I identified are found in RdRP, another highly popular drug target (Bollati et al., 2010; Malet et al., 2008; Sampath & Padmanabhan, 2009). For 19 MBFVs, I found two target sites in RdRP. Narrowing my search to the ZIKV+DENV and WNV clades, I was able to identify one and four additional target sites in RdRP, respectively. One MBFV target site, RRDLR, is found in an arginine patch within RdRP that interacts with the flavivirus genome. Mutations at this motif disrupt this interaction and result in reduced viral replication (Hodge et al., 2016). The functional relevance of this motif, together with its solvent accessibility and conservation across multiple flaviviruses, highlight its potential to serve as a target site for broadly neutralizing antiviral drugs.

In the rate-shift analysis, I observe significantly rate-shifting sites between all three clades (*Aedes*, *Aedes*-outgroup, and *Culex*). Unsurprisingly, the fewest rate-shifting sites are found between the *Aedes* and *Aedes*-outgroup clades. Rate-shifting sites between the *Culex* clade and both the *Aedes* and *Aedes*-outgroup clades are suspected to be

important for vector specificity. Six of the rate-shifting sites I found for the *Culex* and *Aedes*-outgroup comparison are located within four target sites pertaining to the WNV clade. Here, two of the target site motifs may be ancestral as one each is conserved in two different viruses from the *Aedes*-outgroup clade. Both motifs are also conserved in ZIKV and SPOV from the *Aedes* clade. Thus, these two motifs may play distinct roles in ZIKV, SPOV, and the viruses in the WNV clade and may be involved in determination of vector specificity. While support for transmissibility of ZIKV via a *Culex* vector remains inconclusive (Viveiros-Rosa et al., 2020), results herein suggest that the possibility should continue to be monitored in future.

The target sites I identified remain to be validated *in silico* and *in vitro*. For a similar study identifying target sites in the coronavirus family prior to the COVID-19 pandemic (Rahaman & Siltberg-Liberles, 2016), a follow-up study was conducted wherein a virtual screening of FDA-approved drugs was performed for the SARS-CoV-2 RdRP (Pokhrel et al., 2020). While this is speculative, I believe that if the 2016 study by Rahaman and Siltberg-Liberles had been paid heed to, perhaps we could have been better prepared to respond to the COVID-19 pandemic. Thus, a similar follow-up study could, and should, be performed for the flavivirus targets sites I identified. The highest-ranking targets should then have their efficacy tested in *in vitro* experiments. Identifying an effective broadly neutralizing antiviral against flaviviruses now would be of incredible value should a new or existing member of this family become a significant threat in the future.

Epitopedia: Identifying Molecular Mimicry Between Pathogens and Known Immune Epitopes

I describe Epitopedia (Balbin et al., 2021), a novel computational pipeline for the prediction of molecular mimicry based on known immune epitopes from the Immune Epitope Database (IEDB) (Vita et al., 2019). I emphasize the importance of considering the secondary structure element in which a molecular mimic is found when interpreting Epitopedia results. Analysis of pentapeptide pairs with varying pairwise sequence identities (0%, 20%, 40%, 60%, 80%, and 100%) from parent sequences with no identity filter revealed a significant decrease in RMSD for 100% identity compared to all other identity levels across the three secondary structure states (helix, extended, coil). The pentapeptide analysis was repeated while enforcing a 30% pairwise sequence identity filter on the parent sequences to better represent non-homologous sequence pairs. In doing so, the steep decrease in RMSD for pentapeptide pairs with 100% identity is no longer observed and is particularly noticeable for the extended and coil states. Still, for pentapeptide pairs with 100% identity, there is a significant difference in RMSD between pairs from parent sequences with no identity filter in place and those from parent sequences with a 30% identity filter. Altogether, molecular mimics found in helices will tend to have low RMSDs, which is not surprising given the regular geometry of that secondary structure element. Low RMSDs observed for molecular mimics found in extended or coiled states are expected to increase confidence in the predicted molecular mimic's validity, especially if the parent sequences for the molecular mimic share low pairwise sequence identity.

Despite the contributions made by Epitopedia in its current form, it's important to recognize its limitations. One limitation is that Epitopedia is only capable of predicting molecular mimicry for epitopes that have been deposited in IEDB and cannot predict molecular mimicry *de novo*. Similarly, Epitopedia relies primarily on structures deposited in the Protein Data Bank (PDB) (Berman et al., 2000) when assessing structural similarity between the query and hit. Therefore, Epitopedia cannot detect a molecular mimic (even if it is real) if the epitope is either not in IEDB or if it has no structural representation in PDB. Additionally, due to feasibility of implementation, Epitopedia only focuses on linear epitopes. However, not all epitopes are linear. In fact, many epitopes are conformational, meaning that the relevant amino acids are brought together in 3-dimensional space upon protein folding. Information on conformational epitopes can be found in IEDB. In future, it would be interesting to see Epitopedia (or any other program) expanded to include the ability to predict molecular mimicry for conformational epitopes in addition to linear ones.

To my knowledge, Epitopedia is the first program of its kind, although there exist programs for the prediction of molecular mimicry in remote homologs (Armijos-Jaramillo et al., 2021) and programs are under development for the prediction of molecular mimicry based on antibody-binding interfaces (Stebliankin et al., 2022). Additionally , there exist several programs that focus on mimotopes, which are macromolecules (often peptides and obtained by phage display) recognized by an antibody primed for a different epitope. It is assumed that the mimotopes and the native epitope for the antibody to which they bind share similar components (Geysen et al., 1986). These programs map mimotopes onto the antigenic protein structure to identify

the native epitope (Chen et al., 2012; Mayrose et al., 2007; Negi & Braun, 2009). The means to predict molecular mimicry is of great value in understanding disease complications, developing therapeutic interventions, and informing vaccine design to prevent autoimmune outcomes. Predictions generated by Epitopedia can initiate hypotheses on potential pathogenic and autoimmune cross-reactivity that can be tested *in vitro* and *in vivo*.

Potential Autoimmunity Resulting from Molecular Mimicry Between SARS-CoV-2 Spike and Human Proteins

In an effort to explain the variety of disease severity seen in COVID-19, I use Epitopedia to predict molecular mimicry in the SARS-CoV-2 Spike protein. 1D-mimics are epitopes with at least five consecutive amino acids with 100% sequence identity to a corresponding protein region in Spike. 3D-mimics are 1D-mimics that have at least three amino acids surface accessible on Spike and for which the RMSD between the epitope and Spike fragment is at most 1 Å. I found 789 1D-mimics, of which 284 had structural representation in the Protein Data Bank (PDB) (Berman et al., 2000), and 20 had a sufficiently low RMSD to be considered a 3D-mimic. Of the 402 human 1D-mimics lacking a structural representative in PDB, Epitopedia identified AlphaFold2 models for 102, of which 10 had a sufficiently low RMSD to be considered AF-3D-mimics. 3D-mimics and AF-3D-mimics are collectively referred to as molecular mimics. Most predicted molecular mimics for SARS-CoV-2 Spike are found in human, while few are found in other pathogens. Altogether, this suggests that Spike has autoimmune potential,

and that heterologous immunity is rare or unlikely. Furthermore, many of the molecular mimics can be found clustered near one another when mapped to the Spike sequence.

One of the molecular mimics I found, TQLPP, is located at an antibody-binding interface in human thrombopoietin, a protein responsible for regulation of platelet production (Varghese et al., 2017). On Spike, the TQLPP motif is located near the N-terminus domain supersite, a known antibody-binding site (Cerutti et al., 2021). Thrombocytopenia, a condition characterized by low platelet levels, has been observed in COVID-19 patients (Yang et al., 2020) and in individuals vaccinated against SARS-CoV-2 (Greinacher et al., 2021). I therefore hypothesized that SARS-CoV-2 Spike may trigger the production of TQLPP-specific antibodies that may cross-react with human thrombopoietin and result in thrombocytopenia. This was investigated using molecular dynamics (MD) simulations and machine learning to assess the binding of SARS-CoV-2 Spike to the TN1 Fab antibody from the human thrombopoietin structure. MD simulations revealed that TQLPP is accessible to the antibody and that glycans on Spike's surface do not hinder binding. Additionally, residues Q and L of TQLPP were found to be the largest hydrogen bond contributors between Spike and TN1 Fab. The machine learning tool MaSIF-search (Gainza et al., 2019) assessed the antibody-antigen interface complementarity for different complexes to provide a binding confidence score. Spike-TN1 complexes had better binding confidence scores than random complexes, further supporting the possibility of molecular mimicry between Spike and thrombopoietin. Fortunately for humankind, it seems the TQLPP motif may not remain in Spike for long as Gamma and Omicron SARS-CoV-2 variants have already experienced

mutations in this region (Hodcroft, 2021). Protein engineering of the TQLPP motif for the next generation of SARS-CoV-2 vaccines may reduce the risk of thrombocytopenia.

I found another motif, ELDKY, in many molecular mimicry candidates. On SARS-CoV-2 Spike, ELDKY is in a C-terminal stem helix region that has been found to bind a broadly neutralizing antibody effective against all human-infecting beta-coronaviruses (Pinto et al., 2021). One of the molecular mimics containing ELDKY was human tropomyosin alpha-3 (TPM3). Moreover, while these are not molecular mimics identified by Epitopedia, ELDKY is found in human tropomyosin alpha-1 (TPM1) and human cGMP-dependent protein kinase 1 (PRKG1), both of which had an RMSD less than 1 Å for this motif compared to Spike. PRKG1 plays a role in regulation of intracellular calcium levels (Francis, 2010) and its phosphorylation targets have roles in platelet activation and adhesion (Li et al., 2003), smooth muscle contraction (Sauzeau et al., 2000), and cardiac function (Francis, 2010). TPMs are involved in the calcium-dependent contraction of striated muscle (Szent-Györgyi, 1975) and are known cardiac disease-linked antigens (Marrama et al., 2022). Altogether, this suggests that cross-reactive antibodies against ELDKY in PRKG1 may provide an alternate avenue for the development of thrombocytopenia in COVID-19 patients. Additionally, cross-reactive antibodies targeting ELDKY in PRKG1, TPM1, and TPM3 may result in cardiac disease. Cardiac diseases have been observed following both SARS-CoV-2 infection (Nishiga et al., 2020) and vaccination (Patone et al., 2021). The ELDKY motif is highly conserved across beta-coronaviruses and is not mutated in any current SARS-CoV-2 variants, hinting at its potential as a viable pan-coronavirus vaccine target. However, the potential for autoimmune cross-reactivity at this motif may reduce its suitability.

Importantly, the findings I present in this chapter are the results of computational predictions and must still be validated experimentally. Still, computational analyses such as these can help to inform experimental design when investigating phenomena such as molecular mimicry in the SARS-CoV-2 Spike protein.

LITERATURE CITED

Armijos-Jaramillo, V., Espinosa, N., Vizcaıno, K., & Santander-Gordon, D. (2021). A Novel In Silico Method for Molecular Mimicry Detection Finds a Formin with the Potential to Manipulate the Maize Cell Cytoskeleton. *Molecular Plant-Microbe Interactions : MPMI*, *34*(7). https://doi.org/10.1094/MPMI-11-20-0332-R

Balbin, C.A., Nunez-Castilla, J., Stebliankin, V., Baral, P., Sobhan, M., Cickovski, T., Mondal, A.M., Narasimhan, G., Chapagain, P., Mathee, K. and Siltberg-Liberles, J. (2021). Epitopedia: identifying molecular mimicry of known immune epitopes. *BioRxiv*. https://doi.org/https://doi.org/10.1101/2021.08.26.457577

Bálint, E., Bates, S., & Vousden, K. (1999). Mdm2 binds p73 alpha without targeting degradation. *Oncogene*, *18*(27), 3923–3929. https://doi.org/10.1038/SJ.ONC.1202781

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research*, *28*(1), 235–242.

Bollati, M., Alvarez, K., Assenberg, R., Baronti, C., Canard, B., Cook, S., Coutard, B., Decroly, E., de Lamballerie, X., Gould, E. A., Grard, G., Grimes, J. M., Hilgenfeld, R., Jansson, A. M., Malet, H., Mancini, E. J., Mastrangelo, E., Mattevi, A., Milani, M., … Bolognesi, M. (2010). Structure and functionality in flavivirus NS-proteins: Perspectives for drug design. *Antiviral Research*, *87*(2), 125–148. https://doi.org/10.1016/j.antiviral.2009.11.009

Cerutti, G., Guo, Y., Zhou, T., Gorman, J., Lee, M., Rapp, M., Reddem, E. R., Yu, J., Bahna, F., Bimela, J., Huang, Y., Katsamba, P. S., Liu, L., Nair, M. S., Rawi, R., Olia, A. S., Wang, P., Zhang, B., Chuang, G. Y., … Shapiro, L. (2021). Potent SARS-CoV-2 neutralizing antibodies directed against spike N-terminal domain target a single supersite. *Cell Host & Microbe*, *29*(5), 819-833.e7. https://doi.org/10.1016/J.CHOM.2021.03.005

Chao, C. (2015). Mechanisms of p53 degradation. *Clinica Chimica Acta; International Journal of Clinical Chemistry*, *438*, 139–147. https://doi.org/10.1016/J.CCA.2014.08.015

Chen, W., Guo, W. W., Huang, Y., & Ma, Z. (2012). Pepmapper: A collaborative web tool for mapping epitopes from affinity-selected peptides. *PLoS ONE*, *7*(5), 37869. https://doi.org/10.1371/journal.pone.0037869

Chen, W. H., Sun, P. P., Lu, Y., Guo, W. W., Huang, Y. X., & Ma, Z. Q. (2011). MimoPro: A more efficient Web-based tool for epitope prediction using phage display libraries. *BMC Bioinformatics*, *12*. https://doi.org/10.1186/1471-2105-12-199

de Wispelaere, M., Lian, W., Potisopon, S., Li, P.-C., Jang, J., Ficarro, S. B., Clark, M. J., Zhu, X., Kaplan, J. B., Pitts, J. D., Wales, T. E., Wang, J., Engen, J. R., Marto, J. A., Gray, N. S., & Yang, P. L. (2018). Inhibition of Flaviviruses by Targeting a Conserved Pocket on the Viral Envelope Protein. *Cell Chemical Biology*, *25*(8), 1006-1016.e8. https://doi.org/10.1016/J.CHEMBIOL.2018.05.011

Francis, S. H. (2010). The Role of cGMP-dependent Protein Kinase in Controlling Cardiomyocyte cGMP. *Circulation Research*, *107*(10), 1164. https://doi.org/10.1161/CIRCRESAHA.110.233239

Gainza, P., Sverrisson, F., Monti, F., Rodolà, E., Boscaini, D., Bronstein, M. M., & Correia, B. E. (2019). Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nature Methods*, *17*(2), 184–192. https://doi.org/10.1038/s41592-019-0666-6

Geysen, H. M., Rodda, S. J., & Mason, T. J. (1986). A priori delineation of a peptide which mimics a discontinuous antigenic determinant. *Molecular Immunology*, *23*(7), 709–715. https://doi.org/10.1016/0161-5890(86)90081-7

Greinacher, A., Thiele, T., Warkentin, T. E., Weisser, K., Kyrle, P. A., & Eichinger, S. (2021). Thrombotic Thrombocytopenia after ChAdOx1 nCov-19 Vaccination. *New England Journal of Medicine*, *384*(22), 2092–2101. https://doi.org/10.1056/nejmoa2104840

Hodcroft, E. B. (2021). *CoVariants: SARS-CoV-2 Mutations and Variants of Interest*.

Hodge, K., Tunghirun, C., Kamkaew, M., Limjindaporn, T., Yenchitsomanus, P.-T., & Chimnaronk, S. (2016). Identification of a Conserved RNA-dependent RNA Polymerase (RdRp)-RNA Interface Required for Flaviviral Replication. *The Journal of Biological Chemistry*, *291*(33), 17437–17449. https://doi.org/10.1074/jbc.M116.724013

Huang, J., Gutteridge, A., Honda, W., & Kanehisa, M. (2006). MIMOX: A web tool for phage display based epitope mapping. *BMC Bioinformatics*, *7*(1), 1–10. https://doi.org/10.1186/1471-2105-7-451

Huang, Y. X., Bao, Y. L., Guo, S. Y., Wang, Y., Zhou, C. G., & Li, Y. X. (2008). Pep-3D-Search: A method for B-cell epitope prediction based on mimotope analysis. *BMC Bioinformatics*, 9. https://doi.org/10.1186/1471-2105-9-538

Kuhn, R. J., Dowd, K. A., Beth Post, C., & Pierson, T. C. (2015). Shake, rattle, and roll: Impact of the dynamics of flavivirus particles on their interactions with the host. *Virology*, *479–480*, 508–517. https://doi.org/10.1016/j.virol.2015.03.025

Lane, D. P., Cheok, C. F., Brown, C., Madhumalar, A., Ghadessy, F. J., & Verma, C. (2010). Mdm2 and p53 are highly conserved from placozoans to man. *Cell Cycle*, *9*(3), 540–547. https://doi.org/10.4161/CC.9.3.10516

Li, Z., Xi, X., Gu, M., Feil, R., Ye, R. D., Eigenthaler, M., Hofmann, F., & Du, X. (2003). A Stimulatory Role for cGMP-Dependent Protein Kinase in Platelet Activation. *Cell*, *112*(1), 77–86. https://doi.org/10.1016/S0092-8674(02)01254-0

Luo, D., Vasudevan, S. G., & Lescar, J. (2015). The flavivirus NS2B-NS3 protease-helicase as a target for antiviral drug development. *Antiviral Research*, *118*, 148–158. https://doi.org/10.1016/j.antiviral.2015.03.014

Malet, H., Massé, N., Selisko, B., Romette, J.-L., Alvarez, K., Guillemot, J. C., Tolou, H., Yap, T. L., Vasudevan, S. G., Lescar, J., & Canard, B. (2008). The flavivirus polymerase as a target for drug discovery. *Antiviral Research*, *80*, 23–35. https://doi.org/10.1016/j.antiviral.2008.06.007

Marrama, D., Mahita, J., Sette, A., & Peters, B. (2022). Lack of evidence of significant homology of SARS-CoV-2 spike sequences to myocarditis-associated antigens. *EBioMedicine*, *75*, 103807. https://doi.org/10.1016/J.EBIOM.2021.103807

Martincorena, I., Roshan, A., Gerstung, M., Ellis, P., Van Loo, P., McLaren, S., Wedge, D. C., Fullam, A., Alexandrov, L. B., Tubio, J. M., Stebbings, L., Menzies, A., Widaa, S., Stratton, M. R., Jones, P. H., & Campbell, P. J. (2015). High burden and pervasive positive selection of somatic mutations in normal human skin. *Science*, *348*(6237), 880–886. https://doi.org/10.1126/science.aaa6806

Mayrose, I., Penn, O., Erez, E., Rubinstein, N. D., Shlomi, T., Freund, N. T., Bublil, E. M., Ruppin, E., Sharan, R., Gershoni, J. M., Martz, E., & Pupko, T. (2007). Pepitope: Epitope mapping from affinity-selected peptides. *Bioinformatics*, *23*(23), 3244–3246. https://doi.org/10.1093/bioinformatics/btm493

Modis, Y., Ogata, S., Clements, D., & Harrison, S. C. (2004). Structure of the dengue virus envelope protein after membrane fusion. *Nature*, *427*(6972), 313–319. https://doi.org/10.1038/nature02165

Negi, S. S., & Braun, W. (2009). Automated detection of conformational epitopes using phage display peptide sequences. *Bioinformatics and Biology Insights*, *3*, 71–81. https://doi.org/10.4137/bbi.s2745

Nishiga, M., Wang, D. W., Han, Y., Lewis, D. B., & Wu, J. C. (2020). COVID-19 and cardiovascular disease: from basic mechanisms to clinical perspectives. *Nature Reviews Cardiology*, *17*(9), 543–558. https://doi.org/10.1038/s41569-020-0413-9

Patone, M., Mei, X. W., Handunnetthi, L., Dixon, S., Zaccardi, F., Shankar-Hari, M., Watkinson, P., Khunti, K., Harnden, A., Coupland, C. A. C., Channon, K. M., Mills, N. L., Sheikh, A., & Hippisley-Cox, J. (2021). Risks of myocarditis, pericarditis, and cardiac arrhythmias associated with COVID-19 vaccination or SARS-CoV-2 infection. *Nature Medicine*, *28*, 1–13. https://doi.org/10.1038/s41591-021-01630-0

Pinto, D., Sauer, M. M., Czudnochowski, N., Low, J. S., Tortorici, M. A., Housley, M. P., Noack, J., Walls, A. C., Bowen, J. E., Guarino, B., Rosen, L. E., Iulio, J. di, Jerak, J., Kaiser, H., Islam, S., Jaconi, S., Sprugasci, N., Culap, K., Abdelnabi, R., … Veesler, D. (2021). Broad betacoronavirus neutralization by a stem helix–specific human antibody. *Science*, *373*(6559), 1109–1116. https://doi.org/10.1126/SCIENCE.ABJ3321

Pokhrel, R., Chapagain, P., & Siltberg-Liberles, J. (2020). Potential RNA-dependent RNA polymerase inhibitors as prospective therapeutics against SARS-CoV-2. *Journal of Medical Microbiology*, *69*(6), 864–873. https://doi.org/10.1099/JMM.0.001203/CITE/REFWORKS

Rahaman, J., & Siltberg-Liberles, J. (2016). Avoiding Regions Symptomatic of Conformational and Functional Flexibility to Identify Antiviral Targets in Current and Future Coronaviruses. *Genome Biology and Evolution*, *8*(11), 3471–3484. https://doi.org/10.1093/gbe/evw246

Sampath, A., & Padmanabhan, R. (2009). Molecular targets for flavivirus drug discovery. *Antiviral Research*, 81, 6–15. https://doi.org/10.1016/j.antiviral.2008.08.004

Sauzeau, V., Le Jeune, H., Cario-Toumaniantz, C., Smolenski, A., Lohmann, S. M., Bertoglio, J., Chardin, P., Pacaud, P., & Loirand, G. (2000). Cyclic GMP-dependent Protein Kinase Signaling Pathway Inhibits RhoA-induced Ca2+ Sensitization of Contraction in Vascular Smooth Muscle. *Journal of Biological Chemistry*, *275*(28), 21722–21729. https://doi.org/10.1074/JBC.M000753200

Soussi, T., & Béroud, C. (2001). Assessing TP53 status in human tumours to evaluate clinical outcome. *Nature Reviews Cancer*, *1*(3), 233–240. https://doi.org/10.1038/35106009

Stebliankin, V., Baral, P., Balbin, C., Nunez-Castilla, J., Sobhan, M., Cickovski, T., Mohan Mondal, A., Siltberg-Liberles, J., Chapagain, P., Mathee, K., & Narasimhan, G. (2022). EMoMiS: A Pipeline for Epitope-based Molecular Mimicry Search in Protein Structures with Applications to SARS-CoV-2. *BioRxiv*. https://doi.org/10.1101/2022.02.05.479274

Stiewe, T. (2007). The p53 family in differentiation and tumorigenesis. *Nature Reviews Cancer*, *7*(3), 165–168. https://doi.org/10.1038/nrc2072

Szent-Györgyi, A. G. (1975). Calcium regulation of muscle contraction. *Biophysical Journal*, *15*(7), 707–723. https://doi.org/10.1016/S0006-3495(75)85849-8

Tian, H., Ji, X., Yang, X., Zhang, Z., Lu, Z., Yang, K., Chen, C., Zhao, Q., Chi, H., Mu, Z., Xie, W., Wang, Z., Lou, H., Yang, H., & Rao, Z. (2016). Structural basis of Zika virus helicase in recognizing its substrates. *Protein & Cell*, *7*(8), 562–570. https://doi.org/10.1007/s13238-016-0293-2

Varghese, L. N., Defour, J.-P., Pecquet, C., & Constantinescu, S. N. (2017). The Thrombopoietin Receptor: Structural Basis of Traffic and Activation by Ligand, Mutations, Agonists, and Mutated Calreticulin. *Frontiers in Endocrinology*, *8*(59), 1–13. https://doi.org/10.3389/FENDO.2017.00059

Vita, R., Mahajan, S., Overton, J. A., Dhanda, S. K., Martini, S., Cantrell, J. R., Wheeler, D. K., Sette, A., & Peters, B. (2019). The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Research*, *47*(D1), D339–D343. https://doi.org/10.1093/nar/gky1006

Viveiros-Rosa, S. G., Regis, E. G., & Santos, W. C. (2020). Vector competence of Culex mosquitoes (Diptera: Culicidae) in Zika virus transmission: an integrative review. *Panamerican Journal of Public Health*, *44*(e7), 1–9. https://doi.org/10.26633/RPSP.2020.7

Wu, H., & Leng, R. (2015). MDM2 mediates p73 ubiquitination: a new molecular mechanism for suppression of p73 function. *Oncotarget*, *6*(25), 21479–21492. https://doi.org/10.18632/ONCOTARGET.4086

Yang, X., Yang, Q., Wang, Y., Wu, Y., Xu, J., Yu, Y., Shang, Y., & Lillicrap, D. (2020). Thrombocytopenia and its association with mortality in patients with COVID-19. *J Thromb Haemost*, *18*, 1469–1472. https://doi.org/10.1111/jth.14848

Zdzalik, M., Pustelny, K., Kedracka-Krok, S., Huben, K., Pecak, A., Wladyka, B., Jankowski, S., Dubin, A., Potempa, J., & Dubin, G. (2010). Interaction of

regulators Mdm2 and Mdmx with transcription factors p53, p63 and p73. *Cell Cycle*, *9*(22), 4584–4591. https://doi.org/10.4161/CC.9.22.13871

VITA

JANELLE NUNEZ-CASTILLA

Born, Miami, Florida

2010-2013	B.S., Biology Florida International University Miami, Florida
2016-2017	NIGMS RISE Fellow Florida International University Miami, Florida
2016-2017	Vice-President, Biology Graduate Student Association Florida International University Miami, Florida
2019-2022	Doctoral Candidate Florida International University Miami, Florida
2022	3MT People's Choice, University Graduate School Florida International University Miami, Florida

PUBLICATIONS AND PRESENTATIONS

Nunez-Castilla, J., Stebliankin, V., Baral, P., Balbin, C. A., Sobhan, M., Cickovski, T., Mondal, A. M., Narasimhan, G., Chapagain, P., Mathee, K., and Siltberg-Liberles, J. (2022) Potential autoimmunity resulting from molecular mimicry between SARS-CoV-2 Spike and human proteins. *Viruses*, *14*(7): 1415. https://doi.org/10.3390/v14071415

Vater, A., Mayoral, J., Nunez-Castilla, J., Labonte, J. W., Briggs, L. A., Gray, J. J., Makarevitch, I., Rumjahn, S. M., and Siegel, J. B. (2021) Development of a broadly accessible, computationally guided biochemistry course-based undergraduate research experience. *Journal of Chemical Education*, *98*(2):400-409. https://doi.org/10.1021/acs.jchemed.0c01073

Nunez-Castilla, J. and Siltberg-Liberles, J. (2020) An easy protocol for evolutionary analysis of intrinsically disordered proteins. In: Kragelund, B., Skriver, K. (eds) Intrinsically Disordered Proteins. Methods in Molecular Biology, vol 2141. Humana, New York, NY. https://doi.org/10.1007/978-1-0716-0524-0_7

Nunez-Castilla, J., Rahaman, J., Ahrens, J. B., Balbin, C. A., and Siltberg-Liberles, J. (2020) Exploring evolutionary constraints in the proteomes of Zika, Dengue and other flaviviruses to find fitness-critical sites. *Journal of Molecular Evolution, 88*:399-414. https://doi.org/10.1007/s00239-020-09941-5

Ahrens, J. B., Nunez-Castilla, J., and Siltberg-Liberles, J. (2017) Evolution of intrinsic disorder in eukaryotic proteins. *Cellular and Molecular Life Sciences, 74*(17):3163-3174. https://doi.org/10.1007/s00018-017-2559-0

Dos Santos, H.G., Nunez-Castilla, J., and Siltberg-Liberles, J. (2016) Functional diversification after gene duplication: paralog specific regions of structural disorder and phosphorylation in p53, p63, and p73. *PLoS ONE, 11*(3):e0151961. https://doi.org/10.1371/journal.pone.0151961

Nunez-Castilla, J., and Siltberg-Liberles, J. (2021) Identification of potential molecular mimicry in the SARS-CoV-2 Spike protein. American Society for Virology's 40th Annual Meeting, 19-23 Jul, Virtual Symposium (poster presentation).

Nunez-Castilla, J., and Siltberg-Liberles, J. (2021) Coronavirus CURE (no, not that kind) – Biology students see increased bioinformatics self-efficacy, knowledge, and skill while researching SARS-CoV-2. Biology Research Symposium, Florida International University, 5 Feb, Miami, Fl, Virtual Symposium (oral presentation).

Nunez-Castilla, J. and Siltberg-Liberles, J. (2019) Gotta catch 'em all – Using Pokemon to introduce students to phylogenetics and bioinformatics. Network for Integrating Bioinformatics in Life Science Education, 9-11 Oct, Omaha, Ne (poster presentation).

Nunez-Castilla, J., Rahaman, J., Balbin, C. A., Ahrens, J. B, and Siltberg-Liberles, J. (2017) Avoiding real-time and evolutionary conformational flexibility to find broadly neutralizing antiviral targets towards Zika and other flaviviruses. Society for Molecular Biology & Evolution, 2-6 Jul, Austin, Tx (poster presentation).

Siltberg-Liberles, J., Ahrens, J. B., and Nunez-Castilla, J. (2016) Computational molecular biology: deciphering the past from the current to prepare for the future. Conference for Undergraduate Research at FIU, Florida International University, 31 Mar, Miami, Fl (oral presentation).

Nunez-Castilla, J., Gomes dos Santos, H, and Siltberg-Liberles, J. (2016) Guardian or gambler of the genome? Intelligent Systems for Molecular Biology, International Society for Computation Biology, 8-12 Jul, Orlando, Fl (poster presentation).