Investigating the role of mutational interdependencies on viral protein function and the evolution of drug resistance

A Dissertation Presented

By

Neha Shirish Samant

Submitted to the Faculty of the

Morningside Graduate School of Biomedical Sciences at UMass Chan Medical School

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

January 24th, 2023

BIOCHEMISTRY AND MOLECULAR BIOTECHNOLOGY
Investigating the role of mutational interdependencies on viral protein function and the evolution of drug resistance

A Dissertation Presented

By

Neha Shirish Samant

This work was undertaken in the Morningside Graduate School of Biomedical Sciences

Interdisciplinary Program

Under the mentorship of

Daniel N. Bolon, Ph.D., Thesis Advisor

Celia A. Schiffer, Ph.D., Member of Committee

Jeremy Luban, MD., Member of Committee

Carol Schrader, Ph.D., Member of Committee

Harmit Malik, Ph.D., External Member of Committee

William E. Royer, Ph.D., Chair of Committee

Mary Ellen Lane, Ph.D.,

Dean of the Graduate School of Biomedical Sciences

January 24, 2023
Dedication

To, आई and बाबा
बाबा (01/12/1955-01/06/2017), you left us too soon, life has never been the same. I hope I have made you proud.
Acknowledgement

Many people have contributed directly and indirectly to this journey. Firstly, I would like to thank Dan for the opportunity to work in such a wonderful lab. I had very little background on protein biochemistry when I joined for rotation. Dan, you provided the support and direction for me to get accustomed to what was a relatively new field to me. Thank you Dan, for being patient and believing in me. You have been a very strong support, in my professional life and outside of work. I owe one of the major parts of my fulfilling PhD to your kindness and support. Thank you.

To my co-workers, past and current members of the Bolon lab for being the most wonderful colleagues I have ever worked with. I want to give special thanks to Dr. Julia and Dr. Gily, for always being there, I looked forward to coming to the lab everyday and seeing your smiling faces. You not only had great technical and experimental input but most importantly you showed kindness and love and still do. You both have helped me grow as a scientist, but also as a person, and have been a constant source of support and in my most difficult times. I will cherish the times that we had. To Aneth, Anne and Pam, for your friendship, you both continue to inspire me to not only to become a better scientist but also to efficiently juggle work and life balance, thank you for all the life conversations. To Tsepal, Mohan, Parul, David for your help over the years.
I would like to thank Dr. Jennifer Wang for letting me perform experiments in your lab and for your guidance. To my TRAC members, Dr. Celia Schiffer, Dr. William Royer, Dr. Jeremy Luban, thank you for your time and experimental advice on my projects. To Dean Mary Ellen Lane, for continuous support and thoughtful advice through all these years.

I want to thank my friends, Tejaswi, Sneha, Digvijay, Rahul, Mary, Jennifer, and Asmita for their continued support over the years.

I want to thank my incredible family, Tanu and Pillu for being the loving cousins I could ever ask for. To Aji, Mama, Mami, Sini and Vallu and Gauri tai for continued support and fun memories. I am beyond lucky to have you.

I am deeply grateful to my mom for her endless support and patience, for the endless days and nights you have worked hard to provide for our family. You teach me daily to have faith and a smiling face even through hardships. And to my dad, for teaching me that being a good person is what matters the most. I am extremely grateful to my husband, Gaurav, for his unwavering support, encouragement and patience. Thank you for taking care of Sia so that I could finish my graduate work. I could not have done this without you by my side, and to my daughter Sia, you are the sunshine of my life, thank you for filling our life with pure joy and love!

ॐ सर्वं भक्तं सुखिनः सर्वं सन्तं निरामयः।
सर्वं भद्राणि पश्चि-ित शमामभवेत्।
Interaction of mutations is ubiquitous in understanding protein fitness landscapes. Fitness landscapes are critical in understanding protein evolution and drug resistance. I aim to elucidate functional consequences of mutations in viral proteins. Retroviral proteases cleave highly diverse substrates, for example, HIV-1 protease (PR) cleaves dramatically different cleavage sites, making it a challenging and interesting system to investigate epistasis. Epistasis also plays an important role in shaping the emergence and evolution of drug resistance, for example in Oseltamivir resistance in Influenza A virus (IAV). To systematically investigate interaction of mutations in important proteins of RNA viruses, we used a mutational scanning approach, EMPIRIC, to investigate the fitness landscape of cleavage sites of HIV-1 PR. We observed that the cleavage sites had higher preferences for hydrophobic and aromatic amino acids. We also observed that negatively charged amino acids are preferred at positions distal to the scissile bond, where these positions are not involved in binding in the PR active site. Studying the fitness landscapes revealed that biophysical features and context-dependencies both mediate cutting of the cleavage sites. However, in-depth analysis of long-range and short-range contextuality would provide further insights on functional determinants of PR cleavage. I also explored the interaction of mutations in the neuraminidase (NA) of influenza A virus in response to inhibitor oseltamivir and identified positive epistasis between drug resistant mutation and a permissive mutation. Our data revealed the potential of epistasis in the
evolution of drug resistance in circulating viruses. In summary, these studies provide a framework to examine evolutionary constraints and biochemical mechanisms of viral proteins that can contribute to the evolution of drug resistance.
Table of Contents

Title Page II
Dedication III
Acknowledgement IV
Abstract VII
Table of Contents VIII
List of Tables XIII
List of Figures XIV
List of Abbreviations XVI
List of copyrighted materials produced by the author XVIII
Third party copyrighted material XIX
Chapter I: Introduction 1

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pandemics of the past century</td>
<td>1</td>
</tr>
<tr>
<td>HIV - Pandemic, treatment and drug resistance</td>
<td>1</td>
</tr>
<tr>
<td>HIV Life cycle</td>
<td>3</td>
</tr>
<tr>
<td>HIV Structure</td>
<td>3</td>
</tr>
<tr>
<td>HIV infection</td>
<td>4</td>
</tr>
<tr>
<td>Function and significance of Protease and cleavage sites</td>
<td>5</td>
</tr>
<tr>
<td>Protease inhibitors and drug resistance</td>
<td>11</td>
</tr>
<tr>
<td>Treatment of HIV-1 – Enzyme inhibitors</td>
<td>11</td>
</tr>
<tr>
<td>Mutations in the substrate cleavage sites affect protease function</td>
<td>12</td>
</tr>
</tbody>
</table>
Analysis of protein function using systematic mutagenesis 14
Understanding biophysical and biochemical functions in high-throughput by harnessing display technologies 19
Influenza A Virus Life Cycle 22
IAV Structure 22
IAV Infection 23
Function and significance of Surface protein – NA 24
NA inhibitors and drug resistance 27
Scope of this thesis 29

Chapter II: Sequence dependencies and biophysical features both govern cleavage of diverse cut-sites by HIV protease 31

Abstract 31
Introduction 33
Results 38

Experimental approach to investigate features underlying cleavage efficiency 38
Comparing fitness landscape for the three cut-sites 43
Analysis of common physical underpinnings of fitness effects 47
Distribution of functional effects (DFE) 50
Variants with increased cutting function 54
Comparison of yeast display findings with natural variation in HIV-1 56
Conclusion 59

Methods 64

Engineered libraries of cut-site variants 64
Yeast surface display 65
Protease treatment, labeling, and FACS of yeast display cells 66
Preparation of DNA and next-generation sequencing 67
Estimation of functional effects from sequencing results 68
Gel-based analyses of the cutting of full-length MA/CA variants 69
Analysis of cut-site variation in sequenced isolates 70

Acknowledgement 71

Chapter III: Identification of a Permissive Secondary Mutation That Restores the Enzymatic Activity of Oseltamivir Resistance Mutation H275Y 72

Abstract 73

Importance 74

Introduction 74

Results 87

Y276F rescues the fitness defect of H275Y 87
Y276F restores the enzymatic activity and surface expression of H275Y NA 87
Y276F has little effect on oseltamivir resistance 90
Y276F rescues H275Y defects in a modern strain of IAV  90
Discussion  93
Materials and Methods  99
Generation of single and double mutations by site-directed
mutagenesis  99
  Cell culture  99
  Fitness analyses  100
  Generation of recombinant influenza viruses  101
  NA enzymatic activity in vitro  102
  Surface enzymatic activity and expression of NA on 293T cells  102
Acknowledgements  103

Chapter IV: Discussion  104

Summary  104
  Biophysical features and sequence contextuality mediate
  HIV-1 PR cleavage  104
  Permissive mutation hold potential for evolution of
  oseltamivir resistance  108
  Advantages and shortcomings of using EMPIRIC combined
  with yeast display approach to probe viral protein function  110
Future directions  111
Broad Impact

Appendix I: Investigating co-evolution of HIV-1 cleavage sites and drug resistance proteases

Introduction

Result and discussion

FACS analysis with high resolution with increased cutting

DFE across all drug resistance proteases show a few hyperactive mutations

Analysis of hyperactive mutations

Correlation of relative functional scores reveal that moderate sequence dependencies mediate cleavage

Methods and Materials

Appendix II: Yeast-display based Flow cytometry assay

Bibliography
List of Tables

Table 3.1 Relative plaque size measurements for A/WSN/33 viruses with different NA variants on MDCK cells 81

Table 3.2 IC50s of substitutions with Y276 (WT) or F276 93

Table 5.1. Panel of drug resistant mutant proteases 118

Table 5.2. Common and unique hyperactive mutations 126

Table 5.3 Mutant proteases catalytic efficiency 136

Table 6.1. Plasmid PCR amplification parameters 139
List of Figures

Figure 1.1 A Schematic Outline of the Sequential Proteolytic Processing of HIV-1 Gag Polyprotein 7

Figure 1.2 A schematic of HIV-1 protease polyproteins, amino acid sequences of cleavage sites and HIV-1 protease bound to MA/CA polyprotein 8

Figure 2.1 Yeast display approach to examine physical underpinnings of cut-sites for HIV-1 protease 35

Figure 2.2 Impacts of amino acid changes in different cleavage sites 45

Figure 2.3 Correlation of physical properties with relative function of positions in different cut-sites 48

Figure 2.4 Distribution of mutant effects in cut-sites 52

Figure 2.5 Examining variants that were cut fast by yeast display assay 54

Figure 2.6 Comparison of functional effects with sequence variation in circulating isolates 58

Figure 2.7 Model of potential distinctions between functional impacts of mutations in engineered cut-sites compared to their native context 61

Figure 3.1 Plaque size assessment of the experimental fitness of oseltamivir-responsive mutations with or without Y276F 78

Figure 3.2 Viral expansion assessment of the experimental fitness effects of H275Y and Y276F in MDCK and A549 cells 85

Figure 3.3 Enzyme activity of NA with oseltamivir resistance mutations with or without Y276F 88
Figure 3.4 Inhibition of NA variants by oseltamivir

Figure 3.5 Impacts of H275Y and Y276F in pH1N1 A/California/04/2009 strain

Figure 3.6 Position of Y276 relative to drug resistance mutations in NA

Figure 5.1 Resolution on faster cutting mutations depend on how much WT is cut

Figure 5.2 DFE of NC/p1 variants across drug resistance proteases

Figure 5.3 Identifying variants that were hyperactive in drug resistance proteases

Figure 5.4 Correlation of RF scores in different cleavage sites across P6-P6’

Figure 6.1 Flow cytometry controls

Figure 6.2 Codon distribution in the library
List of Abbreviations Used in This Work

AIDS: HIV Acquired immunodeficiency syndrome

CA: Capsid

DMS: Deep mutational scanning

DR: Drug resistance

DRM: Drug resistance mutations

Env: Envelope

HAART: Highly active antiretroviral therapy

HIV: Human immune deficiency

HTLV-1: Human T cell Lymphotropic Virus Type I

IAV: Influenza A Virus

IBV: Influenza B

INT: Integrase

MA: Matrix

NNRTIs: Non-Nucleoside Reverse Transcriptase Inhibitors
NRTIs: Nucleoside Reverse Transcriptase Inhibitors

PIs: Protease Inhibitors

PR: Protease

RT: Reverse Transcriptase

SARS: Severe acute respiratory syndrome

SIV: Simian immunodeficiency

SP1 or p2: Spacer peptide 1

SP2 or p1: Spacer peptide 2

TFP: TransFrame Protein
List of copyrighted materials produced by the author

Chapter II represents work previously published and is presented in accordance with copyright law.

Neha Samant, Gily Nachum, Tenzin Tsepal, Daniel N.A. Bolon. “Sequence dependencies and biophysical features both govern cleavage of diverse cut-sites by HIV protease” Protein Science. 2022;31(7):e4366. doi.org/10.1101/2022.04.18.488666

Chapter III represents work previously published and is presented in accordance with copyright law.

### Third Party Copyrighted Material

The following figure was reproduced from journal: No permission required

<table>
<thead>
<tr>
<th>Figure</th>
<th>Publisher</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1.1</td>
<td>Plos Pathogens</td>
</tr>
<tr>
<td>Figure 1.2</td>
<td>BioChem and Viruses</td>
</tr>
</tbody>
</table>
Chapter I: Introduction

Pandemics of the past century

The past century has seen several pandemics resulting from viral zoonotic infections (Cleaveland et al., 2001). Zoonotic infections are caused by cross-species transmission, including transmission from avian and mammalian species. For example, human immune deficiency (HIV) originated from simian immunodeficiency (SIV) in chimpanzees, Severe acute respiratory syndrome (SARS) from bats, Measles, Mumps and Smallpox from livestock, and Influenza A Virus (IAV) from aquatic wildfowl. Some of these pathogens stably adapt to humans leading to sustained human-to-human transmission. HIV, IAV and SARS have caused severe pandemics in the past and the world is currently in an active pandemic caused by SARS CoV-2.

HIV - Pandemic, treatment and drug resistance

Health impact of HIV, vaccines and antivirals

In 1983, soon after the discovery of the first human retrovirus, the Human T cell Lymphotropic Virus Type I (HTLV-1) (Poiesz et al., 1980), Montagnier and his colleagues isolated another retrovirus, from the lymph nodes of immune-deficient patients (Barré-Sinoussi et al., 1983). The virus was named Human Immunodeficiency Virus (HIV). Since the first outbreak of HIV Acquired immunodeficiency syndrome (AIDS) in June of 1981, this ongoing worldwide public threat has killed approximately 36.3 million people as of
2021, and approximately 37.7 million people are currently infected with HIV globally.

The peak of the viral load post-exposure is between two and four weeks, which leads to rapid depletion in CD4+ T-cells. A severe case of immune system collapse is when AIDS is diagnosed, this often leads to opportunistic infections and cancers. If not treated, most patients succumb to the disease within approximately 15 months (Alcami et al., 2002; Cohen et al., 2008). Due to the propensity of HIV to mutate rapidly, traditional vaccines have not worked, while the protein-based or inactivated virus fail to generate effective antibodies. However, a modest success rate of a RV144 vaccine trial in Thailand, that combined trimeric Env and monomeric gp120, has led to continued vaccine research. The current treatment of HIV is anti-retroviral therapy, a combinatorial therapy utilizing different classes of anti-retroviral drugs, which has converted the disease from being a death sentence to a chronic infection that requires continuous therapy for survival (Deeks et al., 2013). There are 30 FDA approved enzyme inhibitors targeting different stages of viral replication cycle including, Nucleoside Reverse Transcriptase Inhibitors (NRTIs), Non-Nucleoside Reverse Transcriptase Inhibitors (NNRTIs), Protease Inhibitors (PIs), CCR5 Antagonists, Fusion Inhibitors, and Integrase Strand Transfer Inhibitor (INSTIs).
**HIV Life cycle**

**Structure**

HIV is a single-stranded positive sense RNA virus. Each viral particle has two copies of the RNA genome, which is about 9 kb in length and encodes for 15 different proteins (Frankel & Young, 1998). The two main polyproteins, Gag and Gag-Pol are proteolytically processed by HIV protease to produce structural and functional units of the virus. Gag encodes for Matrix (MA) - anchors to the inner membrane of the viral particle via a myristilated domain and forms an envelope associated with the outer protein. Capsid (CA) - forms a core structure that encompasses Nucleocapsid (NC) in complex with the viral RNA, p6 – involved in recruiting ESCRT machinery and release of the immature virus, and two spacer peptides SP1 (p2) and SP2 (p1) (Sundquist and Krausslich 2012; Berkowitz et al. 1996; Martin-Serrano et al. 2003; Kräusslich and Welker 1996). Gag-Pol encodes for three enzymes, (Reverse Transcriptase-RT) – transcribe viral RNA into cDNA, Integrase (INT) – integrates the viral cDNA into the host genome, and Protease (PR) – essential for proteolytic cleavage of the Gag and Gag-Pol polypeptides for proper viral maturation. The Envelope (Env) is also transcribed as a polyprotein, and is proteolytically processed by cellular proteases into a transmembrane subunit (gp41) and a surface protein (gp120). The other proteins encoded by the genome are accessory proteins Vif, Vpr, Vpu and Nef, which are required for proper replication and regulatory proteins, Tat, which recruits host
transcription factors and aids in viral transcription, and Rev which is required for viral RNA nuclear export.

**HIV Infection**

HIV infects host cells via binding its Env surface protein to the CD4 receptor on the host cells (Kwong et al., 2002). Upon binding to CD4 receptors, the Env protein additionally binds to the chemokine receptors CCR5 or CXCR4 (Dalgleish et al., 1984; Dragic et al., 1996). These receptors are necessary for viral entry (Picard et al., 1997). Following the entry, the virus undergoes uncoating to release the viral genome and proteins into host cytoplasm, followed by the transport of the RNA into the nucleus. The RNA is then reverse transcribed into cDNA and integrated into the host genome by viral RT and INT proteins respectively. RT and INT enzymes enter the nucleus along with the viral RNA as a component of the reverse transcription complex (Arhel, 2010). The integrated viral RNA is then transcribed and the full-length transcripts along with two copies of viral RNA are exported to the cytoplasm (Emerman et al., 1989; Swanson et al., 2004). The polyproteins and the viral RNA then assemble at the plasma membrane before budding of the immature viral particles (Fuller et al., 1997). Gag interacts with the plasma membrane via insertion of myristic acid into the plasma membrane (Paillart & Göttlinger, 1999). In fact, the formation of the HIV-1 virus particle is mainly driven by Gag, although non-infectious particles are produced in the absence of other viral proteins and processing events. (Gheysen et al., 1989). Upon release, HIV-1 Protease (PR) plays a major role in transforming the immature viral
particles into a mature virus, by proteolytically cleaving Gag and Gag-pol into their functional subunits (Wiegers et al., 1998). The Gag is the predominant translation product while the Gag-pol is produced with a frequency of about 5-10% due to a -1 ribosomal frameshift event. (Jacks et al., 1988). The Gag-pol polyprotein is required to dimerize in order for PR to self-cleave itself out of the Gal-pol polypeptide (Frankel & Young, 1998). The self-cleavage of PR then begins a cascade of substrate cleavage events, leading to an ordered cleavage of Gag and Gag-Pol at 12 different cleavage sites. (Wiegers et al., 1998)

Function and significance of Protease and cleavage sites

HIV-1 Protease and substrate cleavage

Similar to other retroviral PRs, HIV-1 PR belongs to the aspartyl proteinase family (Navia et al., 1989). HIV-1 PR is a homodimer of 99 amino acids, and contains aspartic acid residues (Asp25) at position 25 in each monomer central to its active site. After the Gag and Gag-pol polyproteins in conjunction with dimers of genomic HIV-1 RNA assemble at the plasma membrane, the multimerization then initiates the budding. Concurrently or immediately after budding the dimerization of Gag-Pol and PR self-cleavage occurs by an intramolecular mechanism, where cleavage sites SP1/NC, an internal TransFrame Protein (TFP) site, and the N-terminus of PR, TFP/PR of Gag-Pol are cleaved (Frankel & Young, 1998; Kaplan et al., 1994; Lindhofer et al., 1995; Pettit et al., 2004; Wan et al., 1996; Zybarth & Carter, 1995).
Upon the formation of a functional dimer, PR intermolecularly cleaves the rest of the cleavage sites, in a specific order. The initial cleavage in the Gag polyprotein occurs at the SP1/NC site, releasing MA-CA-SP1 and NC-SP2-p6 polyprotein, followed by cleavage at NC-SP2/p6 site releasing NC-SP2 and p6 proteins, and at MA/CA releasing MA and CA. Lastly, NC/SP2 and CA/SP1 are cleaved (Erickson-Viitanen et al., 1989; Pettit et al., 2005; Wiegers et al., 1998).
Figure 1.1

Figure 1.1. A Schematic Outline of the Sequential Proteolytic Processing of HIV-1 Gag Polyprotein.

Arrowheads represent the cleavage sites. The order of cleavage is shown as per the rates of cleavage observed in vitro. Figure from (de Marco et al., 2010).
Figure 1.2

A) Schematic representation of A) HIV-1 Gag and Gag-Pro-Pol polyprotein. B) Cleavage site amino acid sequences within Gag polyprotein recognized by HIV-1 protease for residues P6-P6’. C) HIV-1 protease bound to the MA/CA cleavage site within the MA/CA region of the polyprotein. Figures adapted from (Laco, 2021; Majerová & Novotný, 2021).

<table>
<thead>
<tr>
<th>Cleavage site</th>
<th>Sequence P6 P5 P4 P3 P2 P1' P1' P2' P3' P4' P5' P6'</th>
</tr>
</thead>
<tbody>
<tr>
<td>MA/CA</td>
<td>QVSQNY/PIVQNL</td>
</tr>
<tr>
<td>CA/p2</td>
<td>HKARVL/AEAMSQ</td>
</tr>
<tr>
<td>p2/NC</td>
<td>NPATIM/IQKGNF</td>
</tr>
<tr>
<td>NC/p1</td>
<td>TERQAN/FLGKIW</td>
</tr>
<tr>
<td>p1/p6</td>
<td>GRPGNF/LQSRPE</td>
</tr>
</tbody>
</table>

**Figure 1.2.** A Schematic representation of A) HIV-1 Gag and Gag-Pro-Pol polyprotein. B) Cleavage site amino acid sequences within Gag polyprotein recognized by HIV-1 protease for residues P6-P6’. C) HIV-1 protease bound to the MA/CA cleavage site within the MA/CA region of the polyprotein. Figures adapted from (Laco, 2021; Majerová & Novotný, 2021).
The cleavage events are initiated upon binding of the substrate in the active site pocket, the PR flaps enclose nearly eight residues of the substrate cleavage site. The general nomenclature in defining the substrate positions are P1/P1’ where, / denotes the scissile bond and P1 (N-Terminal), and P1’(C-Terminal) denote position 1 on the either side of the scissile bond, as described by Schechter and Berger (Schechter & Berger, 1967). Some studies have implicated that the PR does not recognize a specific sequence in a cleavage site but recognizes a conserved shape and volume that forms the basis for substrate recognition called the “Substrate envelope”. (Ozen et al., 2011; Prabu-Jeyabalan et al., 2002). Thus, one of the key determinants of the order of cleavage of each site is its ability to fill the space in the substrate envelope. Most of the research aimed at understanding the specificity and cleavability of a substrate is focused on using short peptide substrates targeting P4-P3’ positions that occupy the active site pocket (Schneider and Kent 1988; Kotler et al. 1988; Darke et al. 1988; Moore et al. 1989; Ermolieff et al. 1997; Schock et al. 1996; Kräusslich et al. 1989; Billich et al. 1988). A few studies have also investigated cleavage specificity using full length viral Gag and Gag-Pro-Pol polyproteins (Erickson-Viitanen et al., 1989; Lee et al., 2012; Pettit et al., 1991, 1994, 2002, 2005; Tritch et al., 1991; Wiegers et al., 1998). These two approaches provide differing results with respect to the order of substrate cleavage. For example in the peptide assays, the efficiency of cleavage of the CA/p2 site is approximately similar to that of the MA/CA site (Tözsér et al., 1991). However, in the context of the native substrate, the full length Gag polyprotein, CA/p2 is cleaved less efficiently (Pettit et al., 1994).
Different computational approaches have also been applied to predict the cleavability of substrates (Chou et al., 1996; Poorman et al., 1991; Yang & Chou, 2004). Although these methods address substrate cleavage with some success, all of them were based on short peptides or full length Gag studying only a handful of mutations. Dauber and coworkers compared the substrate specificity at four positions of the NC/p1 cleavage site of WT HIV-1 PR and clinically relevant mutant PRs. Using Positional-scanning and synthetic combinatorial libraries, they found that the context is important in defining substrate specificity and providing information on the preferences of amino acids at a few positions. Although changing one residue to 20 other amino acids allowed simultaneous analysis of thousands of substrates, the experiment does not provide a complete analysis of all residues at every position in addition to the lack of information about the non-prime side of the substrate cleavage site (Dauber et al., 2002). For example, according to their study, hydrophobic amino acids (Beck et al., 2000), as well as Glu are preferred at P2. Position P3 is more context interdependent and has a preference for larger hydrophobic residues (Billich and Winkler 1991; Tozser et al. 1992).

A few recent studies have shown that amino acids further away from the scissile bond, P4’ and P5’, may also affect substrate binding and processing; however, detailed studies investigating the impact of distal positions on cleavage are not fully understood (Dam et al., 2009; Maguire et al., 2002; Nijhuis et al., 2007; Shibata et al., 2011; van Maarseveen et al., 2012).
Protease inhibitors and drug resistance

Treatment of HIV-1 – Enzyme inhibitors.

Highly active antiretroviral therapy (HAART) is a regimen that consists of enzyme inhibitors, targeting all three enzymes encoded by the Gag-Pol polyprotein, RT, INT and PR. AIDS related morbidities and mortalities have reduced tremendously since the introduction of HAART. Current regimens usually consist of combination therapy with RT inhibitors which can be nucleoside reverse transcriptase inhibitors (NRTIs) or non-nucleoside reverse transcriptase inhibitors (NNRTIs), PR and/or INT inhibitors (Palella et al., 1998). RT inhibitors were the first enzyme inhibitors approved for treating HIV-1, followed by Protease inhibitors (PIs) (Mitsuya, 1999). The success of HAART led to further research and development of extremely potent PIs. Currently there are nine FDA approved PIs (Lv et al. 2015).

Despite the potency of these inhibitors and success of HAART, antiviral drug resistance is a major concern, mainly due to the ability of the virus to evolve rapidly under selection pressure of the inhibitors, along with low bioavailability of the first generation PIs (Cameron et al., 1999). PR accumulates mutations mainly in the active site in response to the PIs (A. M. J. Wensing et al., 2010), that affect inhibitor binding as well as substrate processing, resulting in decreased viral fitness (Baldwin et al., 1995; Mahalingam et al., 1999; Nijhuis et al., 1999). Prolonged inhibitor therapy also leads to additional mutations in the PR outside of the active site as well as in other HIV-1 regions such as in the substrates Gag and Gag-Pol (Doyon et al.).
Some of these secondary mutations act as compensatory mutations to restore viral fitness (Clavel & Mammano, 2010).

**Mutations in the substrate cleavage sites affect protease function**

Most of these compensatory mutations in the PR substrate are located in the cleavage sites of Gag polyprotein, mainly in the C-terminal region of the Gag, that encompasses the NC/p1 and p1/p6 cleavage sites (Mammano et al., 1998). NC/p1 and p1/p6 are cleaved at a slower rate than other sites, with the exception of P2/NC which is cleaved the slowest (Doyon et al. 1996; Maquire et al. 2002; Côté et al. 2001; Sheng and Erickson-Viitanen 1994; Tritch et al. 1991; Wondrak et al. 1993; Zhang et al. 1997).

The NC/p1 site contains the frameshift sequence around codon 432 (P1’) (Jacks et al., 1988). Ribosomal −1 frameshifting in HIV-1 requires a consensus slippery sequence U UUU UUA and a RNA stem-loop structure downstream for a ribosomal pause (Gesteland & Atkins, 1996). This controlled event ensures the correct Gag to Gag-Pol ratio (Park & Morrow, 1991). Mutations in NC/p1 are frequently observed in in vivo studies and are associated with primary drug resistance mutations (DRMs) in PR. P1’ and P5’ mutations in NC/p1 are found to compensate for the fitness defect caused by primary DRMs in PR. These improve proteolytic processing as seen in both peptide and viral studies (Doyon et al., 1996). Interestingly, these are also capable of providing a processing advantage and may serve as a primary
DRM in the absence of any DRMs in the PR. (Dam et al., 2009; Fehér et al., 2002; Nijhuis et al., 2007). In vivo studies have also observed compensatory mutations in PI-resistant viruses. For example, viruses containing PR mutations M46L, I54V, and V82A have been associated with cleavage site mutations in NC/p1 and p1/p6, and are shown to increase the rate of viral replication (Bally et al. 2000; Koch et al. 2001; Zhang et al. 1997).

Many retroviruses contain a L-domain (Late domain), PTAP, PPXY, or YPDL type motifs in the Gag polyprotein that bind host cellular factors necessary for viral budding (Göttlinger et al., 1991; Huang et al., 1995; Xiang et al., 1996). In HIV-1, the N-terminus of the p6 protein contains the L-domain as PTAPP motif, which facilitates the binding of ESCRT complexes necessary for viral budding, located at positions P7’-P10’. Single mutations in this domain resulted in non-productive viral particles (Garrus et al., 2001; Huang et al., 1995). Mutation of only a few residues upstream of this motif, PP5’R is observed to cause reduced virus growth in the WT-PR background (Dettenhofer Markus & Yu Xiao-Fang, 1999). PP5’R shows improved viral replicative capacity and PR processing in the presence of a PR mutation which is observed in patients under Amprenavir, APV therapy. Another p1/p6 cleavage site mutation LP1’F is also associated with PR I50V and V82A mutations and is observed to improve PR processing and in turn the viral Replicative Capacity (RC). (Maguire et al. 2002; Fehér et al. 2002; Kolli et al. 2006; Zhang et al. 1997) . The LP1’F mutation in p1/p6 is a C-T base transition in the leucine codon, making the surrounding sequence very similar to AAU UUU UUU when transcribed to RNA. A study proposes that this
secondary slippery sequence could promote frameshifting under selection pressure. These studies and others have found that cleavage sites co-evolve with DRMs in PR, harboring mutations that compensate for fitness defects. Given the importance of cleavage sites in viral production, infectivity and drug resistance, they serve as an important model for studying cut-site recognition and cleavability.

While there have been plenty of studies investigating the structural and molecular basis of substrate recognition, the biophysical and sequence determinants of substrate recognition remain unclear (Prabu-Jeyabalan et al., 2002; Ridky & Leis, 1995; Tözsér et al., 1991; Tozser et al., 1992). Some studies have shown that the cleavage is context dependent, while some studies have observed context-independent cleavage (Lee et al., 2012). In addition, the kinetic studies have revealed different results in peptide vs natural substrate context further emphasizing the importance of doing a comprehensive analysis to determine the sequence and biophysical determinants of the substrate cleavage (Pettit et al., 1994; Tritch et al., 1991).

**Analysis of protein function using systematic mutagenesis**

How molecular phenotypes are affected by amino acid sequence is a complex but fundamental question. Very small changes in sequence can have a very large impact on the function of a protein. Mutations distal to the enzyme’s active site can as well have a strong influence on the function of the protein (Shimotohno et al., 2001). Therefore, predicting the effects of sequence change on function remains elusive. Computational tools that predict the
impact of mutations with programs like phyloP (Hubisz et al., 2011), GERP (Davydov et al., 2010), and SIFT (Ng & Henikoff, 2003), use evolutionary conservation as a singular measurement or in combination with biophysical constraints of amino acids. However, these methods lack accuracy.

Directed evolution on the other hand, is very effective in engineering new proteins. It’s an iterative process of mutations and screening to engineer proteins with new functions or properties. Numerous examples of directed evolution prove that proteins have an extraordinary ability to adapt to new challenges (Aharoni et al., 2005). For example, directed evolution on fluorescent proteins has transformed biological imaging by introducing many new features of fluorescent proteins, like emission and excitation properties (Tsien, 2009). One major caveat of directed evolution is that it does not reflect a protein’s natural evolution. The proteins in natural evolution undergo different selections like cellular environment, non specific interactions, regulation, etc exploring its biological relevance.

Targeted mutagenesis approaches use amino acid substitutions to screen for a desired function, typically this requires prior knowledge of structure and is based on mutating mechanistically important residues.

Mutational scanning on the other hand is an unbiased approach in which a systematic substitution of amino acids at each position in a protein provides insight into residues critical for protein function. For example, alanine scanning is a technique that uses systematic screening of alanines at every position in a protein to identify residues critical for protein function and stability (Cunningham & Wells, 1989). It has been extensively used to identify
functional sites in HIV-1 surface glycoprotein and CD4 binding sites (Ashkenazi et al., 1990; Dragic et al., 1998; Jiang Jiyang & Aiken Christopher, 2007; Lu et al., 2001; Sen et al., 2010). Other site-directed mutagenesis approaches using random mutagenesis or error-prone PCR test single or combinations of mutations and screen them using activity or growth assays, for example, to determine stability (Ibarra-Molero et al., 2004), phosphorylation, etc. (Oh et al., 2012).

Despite the success of these approaches, low throughput and limited ability to screen for position specific biophysical properties of amino acids like steric hindrance, electrostatic effects, polarity etc. in the case of Alanine scanning, demands further development of mutational scanning approaches.

Site-saturation mutational scanning has also been used to determine specific enzyme functions like binding and specificity, for example, to determine the enzyme activity and residues that contribute towards substrate specificity of a kinase, APH(3 )II, the Melnikov and coworkers used a distinct site saturation mutagenesis technique called Mutagenesis by Integrated TiEs (MITE) (Melnikov et al., 2014), but this technique is also limited to maximum 200 bases and a high error rate in generating the single-substitution library. Another study generated a comprehensive map of substrate specificity of thrombin and ADAMTS1 (Kretz et al., 2018). The mechanisms that govern the substrate recognition are different for different proteins leading to wide substrate specificity for some compared to other proteins. For example, DMS studies of an amidase from Pseudomonas aeruginosa found the residues that contribute towards specificity are globally distributed throughout the protein,
while DMS studies on a metallo-β-lactamases, VIM-2, found that residues that contribute towards specificity are localized to the active site (Chen et al., 2020; Wrenbeck et al., 2017). Large scale mutational data has been used to infer sequence determinants of protein properties.

Two independent high throughput saturation mutagenesis methodologies have been able to efficiently delineate effects of hundreds or thousands of mutations on protein structure and function in a single experiment – Deep mutational scanning (DMS) and Exceedingly methodical and parallel investigation of randomized individual codons (EMPIRIC). Both these techniques have been used to probe protein function in high-throughput. The pioneering work by Fowler and coworkers on the human WW domain provided a high resolution fitness landscape of mutational preferences in a WW domain. They used chemical gene synthesis to create a mutant library on a phage display and next-generation sequencing (NGS) to parallelly screen for mutant frequency before and after selection to measure binding affinities to the peptide ligand (Fowler et al., 2010). DMS has since then been modified and applied to dissect interaction of mutations in the same gene of WW domain (Araya et al., 2012), to construct a local fitness map and epistasis network of avGFP (Sarkisyan et al., 2016). More recently, DMS was used to map the fitness landscape of ZIKV envelope protein to determine host tropism (Setoh et al., 2019), and was used with yeast-display to determine impact of all amino acid mutations on SARS-CoV-2 RBD expression and its binding affinity for ACE2 (Starr et al., 2020).
Heitpas and coworkers developed a similar saturation mutagenesis approach, EMPIRIC, to accurately map the impact of all mutations on a region of yeast Hsp90 (Hietpas et al., 2012). EMPIRIC, a site saturation mutagenesis technique, provides fitness measurements of all possible single amino acid substitution in selected regions of a protein. EMPIRIC uses cassette ligation method, bulk competition and NGS to measure fitness impact of mutations in a protein. EMPIRIC has since been used to probe experimental yeast fitness and to generate a comprehensive map of fitness landscapes of many viral proteins that provide predictions of protein evolution and evolution of drug resistance. EMPIRIC was initially used to measure distribution of fitness effects on a region of yeast Hsp90. Since then it has been used to measure environmental impact on Hsp90 function, for example, to measure the overall robustness of Hsp90 under shifted environmental conditions (Flynn et al., 2020), it was used to study the impact of expression levels of Hsp90 on yeast fitness (Jiang et al., 2013). EMPIRIC has also been used to determine the fitness landscape of yeast ubiquitin under different environmental conditions (Mavor et al., 2016), in quantifying the effects of all mutations in ubiquitin on yeast growth rate and E1 activation (Roscoe et al., 2013; Roscoe & Bolon, 2014).

EMPIRIC has also been used extensively in studying viral evolution and drug resistance in viruses and oncogenes, investigating fitness landscapes in regions of IAV HA protein (Canale et al., 2018), IAV NA protein (Jiang et al., 2016), fitness landscape of HIV-1 protease (Boucher et al., 2019), and HIV-1 envelope CD4 binding loop (Duenas-Decamp et al., 2016).
More recently EMPIRIC was used to screen for SARS-CoV-2 main protease (Mpro) inhibitor resistance potential and provide a complete dataset that can be used for inhibitor design and prediction of Mpro resistance evolution (Flynn et al., 2022). In cancer, EMPIRIC has been used in identifying clinically relevant mutations in oncogenes BRAFV600E and BCR-ABL1 (Ma et al., 2017). BRAFV600E is found significantly in human cancers, and at a high frequency in melanoma, and BCR-ABL1 is seen in about 20-30% of adults with acute lymphoblastic leukemia. These and ongoing studies using EMPIRIC have provided broad insights on protein evolution, adaptation and prediction of resistance evolution.

Understanding biophysical and biochemical functions in high-throughput by harnessing display technologies

Different expression systems are used to express proteins and peptides on the cell-surface. George P. Smith was the first to display small proteins on the surface of bacteriophage (Smith, 1985) various protein libraries have been displayed on bacteriophages. Phage-display libraries of recombinant antibodies for affinity maturation have yielded numerous successes (Griffiths et al., 1994). However, certain limitations restrict the use of the phage-display approach, for example, it’s proven difficult to achieve high enough affinity for use in tumor therapy (Schier et al., 1996), and there can be size limitations for the screening of the peptide libraries (Little, 1993). Other limitations of
Phage display include limited ability to express eukaryotic proteins and relatively difficult elution process. Bacterial systems are also frequently used for surface expression of antibodies, protein engineering, vaccine development and biosensor applications (Georgiou et al., 1997). E.coli surface expression allows for larger molecule expression, an advantage compared to the phage display. In addition, using flow cytometry for screening provides an easier screening process compared with panning on immobilized antigen in the case of phage display. Similar to the phage, the disadvantage of using bacterial systems is the limitation of expressing eukaryotic proteins in a prokaryotic host.

In contrast to bacteria and phage, yeast provides protein folding and secretory machinery similar to that of mammalian cells, making yeast surface display (YSD) a powerful tool for over the past two decades. The strengths of YSD include its ability to process complex eukaryotic proteins, and quantitative flow cytometric screening.

Different yeast essential cell-surface proteins are used as display systems, for example, agglutinin (Agα1 and Aga1) and flocculin (Flo1, etc.) and other GPI-anchored proteins. Yeast also allows surface expression of complex oligomeric eukaryotic proteins to oligomerize and export to the surface such as heterodimer surface expression of class II MHC (Boder et al., 2005) and homooligomeric surface expression of streptavidin fused with Flo1p (Furukawa et al., 2006).
The mating-type-specific agglutinin yeast display system developed by Boder and Wittrup, uses Aga2 as a surface display protein. Aga2 is a 69 amino acid binding subunit that is linked by disulfide linkages to the 725 amino acid core subunit Aga1 (Boder and Dane Wittrup 1997). Aga1 anchors the whole assembly to the yeast cell wall (Cappellaro et al., 1991, 1994; Roy et al., 1991). Yeast display has offered many applications including selecting and screening antibodies to various antigen targets, for example, screening high-affinity antibodies within mutant libraries (Boder and Wittrup 1998; Boder and Dane Wittrup 1997). YSD has been used in affinity maturation of anti-fluorescein single-chain antibody (scFV) (Boder et al., 2000), anti-lysozyme scFV (VanAntwerp & Dane Wittrup, 1998). Other examples include YSD used to increase the affinity anti-CD3 diphtheria toxin antibody using random mutagenesis (Wang et al. 2007), to increase affinity of anti-TNF antibody using combinatorial mutagenesis (Rajpal et al., 2005), and to increase affinity maturation of neutralizing antibodies against botulinum neurotoxins using molecular evolution (Razai et al., 2005). Yeast display systems have also been a popular application for biosensors, for example in glucose biosensor (Wang et al. 2013).

Another application of YSD is in affinity maturation of T-cell receptors (TCRs), studies on TCRs state their potential in disease diagnostic and therapeutic areas (Buonpane et al., 2005, 2007; Holler et al., 2000). More recently, YSD has been used in antibody screening against West Nile virus NS1 glycoprotein, and generated 22 new NS1-specific monoclonal antibodies (Chung et al., 2006) and more epitope specific antibodies to epidermal growth
factor. YSD was also used to screen a breast cancer specific cDNA library to detect tumor specific new antigens (Wadle et al., 2005)

In summary, mutational scanning in parallel with NGS holds tremendous potential to determine fitness landscapes for analysis of protein sequence-structure-function relationships, protein function, substrate specificity, epistasis and drug resistance evolution.

**Influenza A Virus  Life cycle**

**Structure**

Influenza viruses belong to the orthomyxaviride family, which contain viruses that infect both humans and animals. These are typically pleomorphic and have single-stranded negative-strand RNA as a genome, the genome of all orthomyxaviride is segmented which allows for gene reassortment (Payne 2017). The family comprises Influenza viruses A, B, C, Thogotovirus, Quaranjavirus, and Isavirus. Influenza A (IAV) and Influenza B (IAB) cause mild to serious infections in humans. IAV infect humans, horses, dogs, bats, avian, swine species, making IAV a significant risk factor of causing zoonotic infections (Webster, 2002). There are 8 single stranded RNA segments in IAV that encode for structural and functional units of the virus (Hoffmann et al., 2000). Functional units include, polymerase basic 2 PB2, polymerase basic 1 PB1, polymerase acidic PA, nonstructural proteins NS1 and NS2, these are encoded by segments 1, 2, 3, and 8, segment 8 has alternate open reading frames for NS1 and NS2 (Briedis & Lamb, 1982). The structural units are hemagglutinin HA, nucleoprotein NP, neuraminidase NA, and Matrix proteins -
M1 and M2 are encoded by segments 4, 5, 6, and 7, segment 7 has alternate reading frame for proteins M1 and M2 (Lamb & Choppin, 1981). There are different types of IAV that have caused pandemics and epidemics over the past few decades, the classification is based on the type of HA and NA variants on the surface of the virus. (Kash & Taubenberger, 2015). There are 18 subtypes of HA and 11 subtypes of NA (Tong et al., 2013).

**IAV infection**

IAV infects the respiratory tract via binding its HA surface protein to the sialic acid (SA) receptors on the host cells (Kwong et al., 2002). For humans, IAV specifically binds SA receptors that have a2-6 SA linkages but at a lower frequency. Avian species carry a 2-3 linkages and swine carry both 2-3 and 2-6 linkages, IAVs that infect avian species have evolved high affinity towards a2-3 SA linkages, and swine tracheal epithelial cells can get co-infected with human and avian IAV both, making them a mixing vessel for pandemic strains. (Couceiro et al., 1993; Shi et al., 2014).

HA is encoded as a polypeptide that forms a HA trimer, HA1 subunit SA binding domain and HA2 subunit contains fusion peptide, HA binding to SA receptors receptor mediated endocytosis through which the virus enters the host cell (Wiley & Skehel, 1987). Inside of the virus, each of the 8 gene segments are coated with the polymerase heterotrimers (Rna dependent rna pol RdRp) at the 5’ and 3’ untranslated regions, the rest of the gene segment is coated with NP proteins, together this complex makes the viral
ribonucleoprotein (vRNP) (Stubbs & te Velthuis, 2014). The vRNP is then transported into the nucleus where the viral transcription and replication occurs (Jackson et al., 1982)

Upon entering into the nucleus, the RdRp synthesizes two positive sense RNA molecules, i) mRNA, which makes viral proteins. Transcription of viral mRNA is dependent on cellular mRNA 5' cap-snatching by the viral RdRp (Krug, 1981) and ii) cRNA to transcribe more copies of negative-sense, genomic vRNA (Eisfeld et al., 2015). Poly-adenylated viral mRNAs are then trafficked into the cytoplasm from the nucleus which is mediated by NS2. mRNAs are then translated into viral proteins by host machinery, some of these newly synthesized proteins, namely, viral NP and polymerase are then trafficked back into the nucleus to aid in further round of replication and transcription, while proteins M1 and NS (NS1/2) are trafficked back into the nucleus to aid in the export of RNPs into the cytoplasm (Moeller et al., 2012; O’Neill, 1998). Viral budding involves accumulation of the surface proteins HA, NA and M2 as well as the other viral components at the plasma membrane. M1 provides a docking site for the vRNPs by binding to the HA and NA cytoplasmic tails, while M2 facilitates membrane curvature. The random packaging, where the genomic segments are packaged randomly into a virion (Bancroft & Parslow, 2002) vs the specific packaging model, where, segment-specific packaging signals incorporate the 8 segments into the virion (Liang et al., 2005) both have been hypothesized to explain the packaging of the viral segments. Studies have identified specific packaging signals in the 5’
and 3' regions of some segments; this further confirms the later model (Fujii et al., 2003). For the release of the nascent viral particles, NA functions as a sialidase enzyme, by enzymatically cleaving sialic acids from host cell surface receptors and from nascent virions, the HA and NA on nascent virions, contain sialic acids as part of the glycosylation processes within the host cell. This step ensures that the viral progeny does not bind back to the host cell via HA, thus enabling efficient viral release (Gottschalk, 2006; Palese & Schulman, 1974).

NA plays an important step in IAV infection and thus is a major target for antivirals. Not only does NA function on viral release but it also influences viral entry by cleaving Sialic acids on the decoy receptors (McAuley et al., 2019).

**Function and significance of Surface protein – NA**

NA is a type II transmembrane protein, it assembles as a tetramer forming about 40–50 NA spikes on the surface. (Varghese et al., 1983), The four monomers form the cytoplasmic tail, the transmembrane region, the stalk, and the catalytic head. The N-terminal hydrophobic transmembrane domain attaches NA to the viral envelope (Bos et al., 1984), and is also involved in NA translocation from the endoplasmic reticulum to the cell surface (Barman & Nayak, 2000). The stalk renders stability to the tetramer. (Varghese et al., 1983)
The head domain contains a functional catalytic site on the surface of each monomer facing sideward, which enables NA to cleave sialic acids from nearby membrane glycoproteins (Burmeister et al., 1992; Varghese et al., 1998). These catalytic sites include a large number of charged residues. The residues that make up the active site are highly conserved and interact directly with the substrate SA (Arg118, Asp151, Arg152, Arg224, Glu276, Arg292, Arg371, and Tyr406) (Colman et al., 1983). Other residues that do not interact with the substrate but are important for structural stability, called framework residues are Glu119, Arg156, Trp178, Ser179, Asp198, Ile222, Glu227, His274, Glu277, Asn294, and Glu425 (reported as N2 numbering) (Colman et al., 1983).

**NA inhibitors and drug resistance**

Since the 1940s, inactivated influenza vaccines have been very commonly used as a prevention. Traditionally these are combinations of strains, either trivalent or quadrivalent, and are administered annually. For the current influenza season of 2021-2022, the FDA recommendation includes, Influenza A H1N1, H3N2 and an Influenza B strain for trivalent or Influenza A H1N1, H3N2 and two from Influenza B strain for quadrivalent vaccines. Vaccines become less effective if the same strains do not circulate in the next flu season (Potter, 2001). In addition to this, in a pandemic it becomes exceedingly difficult to generate a specific combination while the pandemic
strains evolve at a higher rate (Moscona, 2005). Moreover, many patients do not have access to vaccines, or are reluctant towards the usage.

An alternative therapy is small molecule inhibitors of IAV, these are potent inhibitors against M2 ion channels and NA. M2 channel blockers, amantadine and rimantadine work by preventing viral genome release into the host cell cytoplasm by inhibiting the acidification of the endosome (Jing et al., 2008). In the influenza season of 2004-2005 rimantadine and amantadine had limited usage mainly due to the emergence of drug resistance mutations (Duwe, 2017). NA competitive inhibitors, Oseltamivir and Zanamivir, bind the active site pocket of the NA, inhibiting the SA processing and viral particle release. Inhibition of the particle release leads to aggregation of virions on the host cell surface, limiting the next round of infection and faster recovery. These have been mostly prescribed as an antiviral treatment for influenza since the 2004-2005 season. Zanamivir, a sialic acid analogue, was the first NA inhibitor. The positively charged guanidino moiety of zanamivir binds with the negatively charged amino acids in the active site pocket of the NA, which makes it less bioavailable and hence is administered orally via inhalation (Glaxo Wellcome Inc. RELENZA (zanamivir for inhalation). North Carolina: Glaxo Wellcome Inc., Apr 2000). Another inhibitor, Oseltamivir, acts similar to Zanamivir, but has a hydrophobic group that binds with a hydrophobic area within the active site pocket, this leads to poor absorption of the drug and to overcome the hampered bioavailability it is usually administered as the phosphate salt (Dreitlein et al., 2001). Moreover, the binding of the
hydrophobic moiety induces conformational change of the framework residue E277, opening up a potential to virus evolution under oseltamivir treatment.

Given the different chemical structures of Zanamivir and Oseltminvir, distinct mutations affect each inhibitor. Generally, patients with virus isolates showing oseltamivir resistance are treated with Zanamivir. Resistance to Zanamivir occurs very rarely while resistance to oseltamivir occurs at a higher percentage.

Some studies have shown that mutations may occur in HA in response to NAI binding; these occurred mostly in the HA receptor binding site, although these studies were done in vitro (Blick et al., 1998). Mutations in NA have been predominantly in the active site catalytic and framework residues (Gubareva et al., 1997; Staschke et al., 1995). These mutations are predominantly detected in vitro by screening patient isolated viruses in an enzyme inhibition assay. Mutations are specific to the subtype of influenza virus. Mostly these mutations are observed at residues 119, 152, 274, and 292 of the enzyme’s active site (Gubareva et al., 2000). For N2 viruses, the most common mutation observed in vivo and in vitro is Arg292Lys, additionally, Glu119Val (E119V) has also been observed (Gubareva et al., 1997).

Despite of their potency, drug resistance mutations to NA have also been on a steep rise, with only about only 12.3% of H1N1 viruses being oseltamivir-resistant in the 2007-2008 season and 98.5% in the 2008-2009 season (Dharan et al., 2009).
Scope of this thesis

In this dissertation, I investigate the interdependence of mutations in viral proteins, with the goal of understanding the effects of mutational dependencies on protein activity and drug resistance. In chapter II, I use the EMPIRIC approach developed in our lab to study the fitness landscape of HIV-1-1 PR substrate cleavage sites, in an attempt to understand how changes to amino acid sequences in the cleavage site affect protease cleavage. The naturally diverse cleavage sites make an excellent model to study epistasis. Specifically, I study the functional effects of all possible mutations in a 12-amino acid region of three different cleavage sites in the Gag polyprotein. I observe that there are commonalities between all three cleavage sites with respect to preference for hydrophobic, aromatic and negatively charged amino acids, although the positions where we see preference differ between each cleavage site. This observation led to further investigation of the biophysical properties of amino acids and contextual determinants that may mediate cleavage. The analysis suggests that positions distal from scissile bond had very little biophysical underpinnings, while the position close to scissile bond showed complex biophysical underpinnings. I also observe that contextuality is a strong determinant of the cleavage. In summary, I use a systematic mutagenesis approach to investigate how viral proteins are able to cleave diverse substrates. An in-depth analysis of interactions of mutations would help understand potential epistasis that can have important consequences in the context of evolution of drug resistance. This system can also be used in understanding fundamentals
of other viral systems that have a similar basis of protease cleavage. Fitness landscapes of cleavage sites in addition to future structural studies would also provide a valuable dataset in designing protease or maturation inhibitors with reduced potential of resistance evolution.

In chapter III, I investigate the epistatic potential of a permissive mutation in NA of IAV. I use single point mutations that affect viral fitness at previously reported sites to investigate potential epistasis between drug resistance mutations. Drug resistance mutations described here are mutations proximal to the active site that show fitness defects in the absence of drug pressure (Jiang et al., 2016). Through a combination of biochemical and viral replication studies with or without a permissive mutation, I observe that the permissive mutation imparts negative epistasis on most of the studied DRMs, except for H275Y, a known major Oseltamivir resistance mutation. The positive epistasis is mediated through an increase in expression and activity of NA. Understanding the mechanism of permissive mutations in compensating fitness defects of DRMs provide potential avenues for surveillance and prediction of drug resistance.

In summary these findings highlight the importance of studying mutational dependencies of proteins in understanding their function and resistance evolution potential.
Chapter II - Sequence dependencies and biophysical features both govern cleavage of diverse cut-sites by HIV protease

This chapter has been published previously:

Neha Samant, Gily Nachum, Tenzin Tsepal, Daniel N.A. Bolon. “Sequence dependencies and biophysical features both govern cleavage of diverse cut-sites by HIV protease” Protein Science. 2022;31(7):e4366. doi.org/10.1101/2022.04.18.488666

Abstract

The infectivity of HIV-1 requires its protease cleave multiple cut-sites with low sequence similarity. The diversity of cleavage sites has made it challenging to investigate the underlying sequence properties that determine binding and turnover of substrates by PR. We engineered a mutational scanning approach utilizing yeast display, flow cytometry, and deep sequencing to systematically measure the impacts of all individual amino acid changes at 12 positions in three different cut-sites (MA/CA, NC/p1, and p1/p6). The resulting fitness landscapes revealed common physical features that underlie cutting of all three cut-sites at the amino acid positions closest to the scissile bond. In
contrast, positions more than two amino acids away from the scissile bond exhibited a strong dependence on the sequence background of the rest of the cut-site. We observed multiple amino acid changes in cut-sites that led to faster cleavage rates, including a preference for negative charge five and six amino acids away from the scissile bond at locations where the surface of protease is positively charged. Analysis of individual cleavage sites using full-length matrix-capsid proteins indicate that long-distance sequence context can contribute to cutting efficiency such that analyses of peptides or shorter engineered constructs including those in this work should be considered carefully. This work provides a framework for understanding how diverse substrates interact with HIV-1 protease and can be extended to investigate other viral proteases with similar properties.

Introduction
The interdependence of mutations, also known as epistasis, can provide valuable insights into biochemical mechanisms and are also critical to understanding evolution. In terms of mechanism, mutational dependencies have revealed concerted motions in proteins that govern allostery and protein evolution (Starr et al., 2018; Stiffler et al., 2015), provided physical maps to improve the prediction of protein structure (Rocklin et al., 2017; Rollins et al., 2019), and identified mutations that increase thermodynamic stability (Araya et al., 2012). In evolution, epistasis can have dramatic impacts on the rate and pattern of substitutions (Pollock et al., 2012; Sarkisyan et al., 2016).

While the value of understanding epistasis is clear, the prevalence, magnitude, and evolutionary impacts of mutational dependencies are controversial (Mackay, 2013). For example, some studies (Pollock et al., 2014) support the idea that the preferences of amino acids at a position in a protein change during evolution whereas others (Ashenberg et al., 2013) indicate that key aspects of these preferences remain largely conserved during evolution. Part of the controversy regarding the impacts of epistasis stems from a dependence on the resolution of fitness measurements. According to well established population genetic theory, the magnitude of fitness effects that contribute to selection in natural evolution is approximately the inverse of the effective population size (Ohta, 1973). For example, most microbes exhibit effective population sizes of greater than a million meaning that fitness effects of one in a million would contribute to natural selection. The patterns of epistasis discerned from analyses of sequence patterns in natural evolution will be determined by minute fitness effects. In contrast,
epistasis from experimental measures of fitness will be determined by their
resolution that is typically on the order of one in a hundred to one in a
thousand. Investigation of experimentally assessed epistasis across studies
ranging from viruses to animals indicated the greatest level of epistasis in
viruses (Sanjuán et al., 2006). This study suggests that compact genomes
(e.g., as in HIV-1) tend to experience high levels of epistasis. While some of
the features contributing to patterns of epistasis in different organisms have
been investigated, the structural and physical underpinnings of epistasis
remain poorly understood for most proteins and organisms.

Here, we investigate the structural and physical underpinnings of
epistasis in cleavage sites of HIV-1 protease (PR). PR cut-sites provide a rich
system for investigating epistasis because they exhibit dramatic sequence
variation (Figure 2.1A). PR recognizes and cleaves its substrate Gag and
Gag-Pol polyproteins at twelve different sites that are required to generate
mature viral proteins and infectious viruses. The twelve cleavage sites are
highly variable in amino acid sequence (Figure 2.1A). Despite the varied
sequence of cut-sites, structural studies have shown that they occupy a
conserved shape in the PR active site (Ozen et al., 2011; Potempa et al.,
2018)

Figure 2.1.
Figure 2.1. Yeast display approach to examine physical underpinnings of cut-sites for HIV-1 protease. (A) Logo plot showing the frequency of each amino acid in an alignment of all PR cut-sites in the proteome of the NL4-3 HIV-1 strain. The WT cut-sites analyzed in this work (MA/CA, NC/p1 and p1/p6) are shown below. (B) Outline of yeast-display approach to measure the impacts of all point mutations on the function of cut-sites. (C) Correlation between full biological replicates of yeast display measures of functional effects in the MA/CA cut-site. (D) Correlation between
yeast display estimates of cutting function and previously reported measures using a
gel-based assay for a panel of variants in the MA/CA and p1/p6 sites [40].
Correlations were estimated for both linear fits (shown in black), and non-linear fits
(shown in green). The grey points were excluded from fits to avoid potential lever arm
effects on linear fits to data points with extreme values.

PR is a 99 amino acid protein that functions as a homodimer with one
active site per dimer (Ozen et al., 2011). A catalytic aspartate at position 25
activates a water molecule for peptide-bond hydrolysis of substrates. The
active site surrounding the catalytic aspartate is a groove that directly contacts
the four amino acids immediately before and after the cleavage site (Ozen et
al., 2011; Prabu-Jeyabalan et al., 2000). The amino acids adjacent to the
cleavage site are referred to as position 4-3-2-1/1'-2'-3'-4 (cleavage site
indicated by /). While positions 4 to 4’ directly contact the active site (Ozen et
al., 2011), recent studies indicate that substrate positions beyond P4 and P4’
 can contribute to substrate binding (Laco, 2015; Prabu-Jeyabalan et al.,
2000).

While all cut-sites must be cleaved by PR in order to generate
infectious virions, the turnover rate (kcat) of sites vary over a wide range: from
2×10−5 sec-1 for the NC/TFP site to 7 sec-1 for the MA/CA site (Könnyű et
al., 2013; Lee et al., 2012). The wide range of substrate cleavage rates led to
the suggestion that the order of cleavage may be important for viral
maturation (Wiegers et al., 1998). Recent studies (Lee et al., 2009, 2012)
indicate that the cutting rate of many sites can be altered dramatically without
compromising fitness. However, the relationship between viral fitness and cutting efficiency has not been directly assessed for most cleavage sites.

Protease inhibitors (PIs) potently inhibit replication of the wildtype HIV-1, and the virus responds through the accumulation of mutations in both PR and some cut-sites (Mammano et al., 2000; Schneider-Nachum et al., 2021). High level resistance to PI requires the accumulation of multiple mutations in PR, typically involving 10-30 amino acid changes in this 99 residue protein (Wu et al. 2003; Zhang et al. 1997). To our knowledge, all PR variants that cause resistance to PI's exhibit dramatically reduced enzyme activity (Croteau et al., 1997; Mahalingam et al., 1999; Ragland et al., 2014). Many studies have found that cleavage sites in Gag co-evolve with drug resistant mutations in PR, indicating that the mutations in the cleavage sites may compensate for reduced enzyme activity of PR (Bally et al., 2000; Fehér et al., 2002; Kolli et al., 2006, 2014; Mammano et al., 1998; Özen et al., 2012, 2014; Zahra et al., 2021; Zennou et al., 1998). Cut-site compensatory evolution has most commonly been associated with the NC/p1 and p1/p6 sites.

While there is extensive evidence that cut-sites are critical for viral replication and subject to selection pressure by PI's, the wide variation in cut-site sequences has made it challenging to elucidate the features that underlie cut-site recognition and processing by PR. Prior studies have found evidence supporting both context-dependent effects of mutants (Kolli et al., 2009; Laco, 2015) and context-independent features (Pettit et al., 1991), but it
is not clear how differences in experimental approaches may contribute to these conclusions. To address this challenge, we analyzed how all possible single amino acid changes in three different cleavage sites (MA/CA, NC/p1, and p1/p6) impact cleavage rate by PR in the same assay. The resulting cut-site fitness landscapes reveal both general biophysical features that underlie cutting efficiency in all three cleavage sites as well as features that depend on the rest of the sequence background of the cleavage site.

Results and Discussion

Experimental approach to investigate features underlying cleavage efficiency

The sequence of the different sites that HIV-1 PR must cleave in order to generate an infectious virus vary dramatically, which has made it challenging to understand the sequence rules that govern cleavage. The cut-sites encoded by a single viral genome (NL4-3) are shown in Figure 2.1A. All residues adjacent to the scissile bond vary, with no position showing even 50% sequence identity. The two closest residues on either side of the scissile bond exhibit the least amino acid variation, but even these sites include amino acids that vary dramatically in physical properties including size, hydrophobicity, and hydrogen bond potential. Positions three or more sites from the scissile bond exhibit similar highly diverse amino acids.
To study both general and sequence-dependent features, we chose to analyze the fitness landscape at 12 positions in three different cut-sites (MA/CA, NC/p1, and p1/p6). The sequence of these cut-sites have two identical amino acids within the 12 amino acid region of study (N at position 2 in MA/CA and p1/p6, and Q at position 3 in MA/CA and NC/p1). This level of identity (2 of 36 comparisons = 5.5%) closely matches expectations for comparisons of random amino acid sequences (1/20 = 5%). We were motivated to investigate features of NC/p1 and p1/p6 that may underlie their observed co-evolution with drug-resistant mutations in PR. In addition, we were interested in MA/CA because it is one of the most efficiently cleaved cleavage sites (Lee et al., 2012). We chose to analyze a 12 amino acid region to explore the possibility of selection beyond the 8 positions that have been extensively characterized in complex with PR by crystallography.

To measure fitness landscapes, we engineered a yeast-display approach similar to prior work (Rocklin et al., 2017) to quantify cleavage efficiency in high throughput (Figure 2.1B). We generated plasmids encoding a fusion of Aga2 followed by an HA-tag, a short flexible glycine-rich region, 12 residues of each cut-site (MA/CA, NC/p1, or p1/p6), another short flexible glycine-rich region, and a C-terminal Myc-tag. The Aga2 protein serves to direct the fusion protein to the yeast surface where it is accessible to bulk solution. We used fluorescently labeled antibodies directed at the HA-and myc-tags to distinguish yeast with different levels of cleaved and uncleaved surface-displayed protein. Using yeast displaying only a wildtype (WT) cut-site, we identified conditions that resulted in extensive but not complete
cleavage. We reasoned that these conditions would provide high sensitivity by flow cytometry to mutations that either increased or decreased cutting efficiency. We generated site-saturation plasmid libraries at each of the 12 positions for each cut-site and used our flow-cytometry based approach to separate populations based on level of cleavage. We analyzed each FACS population using deep sequencing and used read counts to estimate the frequency of each variant in each sort window. Based on these measurements, we calculated the average fluorescence of each variant. To estimate relative cleavage, we normalized fluorescence estimates to WT, such that 0 represents a cut site and 1 represents the extent of cutting of a WT cleavage site.

Based on the experimental setup, we expect our readout to be sensitive to changes in both $k_{cat}$ and $K_M$. Based on the reported range of yeast display efficiency ($10^4$-$10^6$ molecules per cell) (Boder et al., 2000) and our conditions that included $10^6$ yeast in a 0.1 mL reaction, we estimate total substrate concentration was 0.16-16 nM. Previously reported $K_M$'s (Könnyū et al., 2013; Lee et al., 2012) for MA/CA (150 μM), NC/p1 (170 μM) and p1/p6 (1.2 mM) are well above these estimates of substrate concentration and well above the concentration of PR enzyme in our reactions (1 μM). Under these conditions, enzyme theory indicates that both the likelihood of binding PR to substrate ($K_S$) and hydrolysis of bound substrate ($k_{cat}$) should impact cutting efficiency in our yeast display experiments. Of note, the concentration of the Gag protein containing these cleavage sites in HIV virions is estimated at about 3 mM (Lee et al., 2012). Competition for PR to bind to multiple sites in
Gag causes both $k_{cat}$ and $KS$ to contribute to the relative rate at which cut-sites in Gag are predicted to be processed in virions (Lee et al., 2012). Enzymatic modeling indicates that the changes in cutting efficiency in our yeast display experiments should trend with the efficiency that these cut-sites are processed in virions.

To assess the reproducibility of our approach to measure the protein fitness landscape of a cut-site, we performed a full biological replicate including separate yeast transformations for the MA/CA library (Figure 2.1C). The relative function of MA/CA cut-site variants was strongly correlated ($R^2=0.97$) with the most noticeable distinction between replicates occurring for variants with very low levels of relative cleavage. In addition, independent measurements of WT synonyms at different positions (green symbols in Figure 2.1C) were tightly clustered. Of note, stop codons in our library generate a displayed protein that lacks the C-terminal myc-tag and thus provide a representation of full cutting (red symbols in Figure 2.1C). We scaled our measurements of relative function so that 0 represents a fully defective variant, and 1 represents the level of cutting of WT under the experimental conditions. One advantage of bulk competitions such as our yeast display cutting competition is that each variant is exposed to identical conditions including PR concentration, temperature, pH, etc. This internal experimental consistency should contribute to the strong reproducibility that we observed for replicate measurements.
We compared our yeast display estimates of relative cleavage to measurements of cleavage efficiency \((k_{cat}/K_{M})\) made for a panel of individual cut-site variants for MA/CA and p1/p6 that were analyzed in isolation using a gel readout (Kolli et al., 2009). As shown in Figure 2.1D, the relative cleavage rates that we determined correlate roughly with measurements from gel assays for both MA/CA and p1/p6, indicating that both studies capture some shared and fundamental aspects that determine cleavage efficiency. We examined both linear correlations as well as a non-linear correlation with the form of the Michaelis-Menten equation that can account for a threshold in detection of rapidly cutting variants in the yeast display assay. Of note, the correlation with individual measurements from either fitting approach \((R^2 = 0.41\) or \(0.53\) for MA/CA) is smaller than the correlation between experimental replicates for MA/CA \((R^2 = 0.97,\) Figure 2.1C), suggesting that distinctions between experimental setups of the yeast display and gel-based assays may impact relative cutting efficiency. Both our yeast display setup as well as the gel-based assays were analyzed using engineered globular protein substrates with triple glycines adjacent to the cut-site to minimize long-range context dependencies. However, the globular proteins differed between each study. Our substrate included Aga2 to facilitate yeast display, while the studies of Lee et al. utilized GST fused to a truncated MA-CA. The different fused proteins may contribute to variations between the two studies. Of note, amino acid changes that resulted in faster than WT cutting function by yeast display exhibited a wide range of impacts in gel assays (Figure 2.1D), suggesting that findings of faster than WT cutting should be considered cautiously. The rough
overall agreement between our study and those of Lee et al. indicates that the sequence of the cut-site itself makes a large contribution to cutting efficiency.

**Comparing fitness landscape for the three cut-sites**

To search for patterns of mutant impacts on the three cut-sites, we generated heatmap representations (Figure. 2.2A) of our yeast-display measurements (Supplementary Table S1). To facilitate comparison of variants that cut faster than WT in the heatmaps, an additional scaling of these variants was performed so that the maximum value was identical between each cut-site experiment. From an unfocused perspective the three sites exhibited similar levels of overall selection as all sites had many mutations that severely reduced cutting function (red squares in Figure. 2.2A) and many mutations that increased cutting function (blue squares in Figure. 2.2A). The average selection on each amino acid across all positions in all three cut-sites (far right column in Figure. 2.2A) slightly favored aromatic and negatively charged amino acids, and slightly disfavored amino acids that were either positively charged or polar and uncharged. Aromatic amino acids are enriched at protein-protein interfaces (Meyer et al., 2003; Thornton et al., 1988) indicating that they have a general propensity to favor binding interactions. The interactions of polar amino acids at protein interfaces tend to be more specific than aromatic interactions because they depend on complementary hydrogen bonding and/or charge (Bolon et al., 2001). The slight average preference for negatively charged amino acids but not uncharged or positively charged
amino acids, suggests the potential for either complementary interactions with positive charge on PR, and/or charge interactions that alter the conformation of our engineered substrate.
Figure 2.2. Impacts of amino acid changes in different cleavage sites. (A) Heatmap representation of yeast-display estimates of cutting function. To facilitate comparison across different substrates, scales were normalized so that the fastest cut variants and slowest cut variants of each cut-site are similar colors. (B) Scatter plots comparing rank order of amino acid impacts in different cut-sites across positions 6-6'.
To further investigate the effects of mutations in different cut-sites, we compared the relative function of all mutations in each cut-site (Figure. 2.2B). We observed correlation coefficients ranging between 0.31 and 0.38, indicating that only a fraction of the observed variation in cutting function was determined by sequence independent impacts of amino acid changes. Amino acid changes that caused very poor cutting in one cut-site tended to have similar impacts in other cut-sites. In contrast, amino acid changes with even modest measured cutting function in one cut-site exhibited a wide variety of effects in other cut-sites. Considering the heatmap (Figure. 2.2A), we observed stronger and more prevalent decreases in relative function for mutations close to the scissile bond than those further away. Our results also showed that aromatic amino acids could increase relative function in all three cut-sites. However, the location of aromatics that increased activity differed between the cut-sites, indicating some level of sequence dependence. While visual inspection revealed overall trends, detailed physical and sequence relationships were challenging to extract by visual inspection alone.
Analysis of common physical underpinnings of fitness effects

To quantitatively search for characteristics that may underly the impacts we observed on relative cutting function, we correlated cutting at each site in each substrate with a set of physical properties of amino acids previously developed for this purpose (Abriata et al., 2015). The strongest correlation ($R^2=0.75$) with cutting effects that we observed was with hydrophobicity at position 6 in the NC/p1 substrate (Figure. 2.3A). The impacts of mutations at this position in the NC/p1 substrate are modest and span from 0.6 to 1.4. These observations indicate that binding at position 6 in NC/p1 could be mediated by a hydrophobic patch on PR that can accommodate different size and shaped side chains (e.g. a flat and/or flexible hydrophobic surface). Of note, we are not claiming that the correlation in Figure 2.3A is statistically significant that would require an analysis of multiple hypothesis testing. Rather, we are highlighting how rarely we observed strong correlations between physical properties and relative cutting.
Figure 2.3. Correlation of physical properties with relative function of positions in different cut-sites. (A) Strongest correlation of property with any position/cut-site was for hydrophobicity at position 6 in NC/p1. (B) Distribution of correlation coefficients between simple physical properties and positions in each cut-site. (C) Heatmap
representation of the correlation between physical properties and position/cut-sites. (D&E) Strong correlation of correlations between position 1’ in MA/CA and p1/p6 (D) and position 6’ in MA/CA and NC/p1. (F) Correlation of correlations for the same position in different cut-sites

Very few physical parameters exhibited even modest correlations ($R^2 > 0.5$) with the observed functional impacts of cut-site mutations at any position/substrate (Figure. 2.3B). In addition, the strongest correlations between a physical property and a specific site in one substrate generally did not show similar patterns at the same site in the other substrates. For example, while hydrophobicity correlated well with fitness effects at position 6 of the NC/p1 substrate it showed much weaker correlation with cutting effects in MA/CA and p1/p6 substrates. The physical properties that showed the strongest correlation tend to be highly dependent on the sequence background of the cut-site.

Next, we considered the possibility that multiple biophysical parameters with complex interdependencies might underly cutting efficiency. In this case, we expect to see relatively weak correlations between a position and any simple physical property. However, if the complex physical underpinnings at a position were similar across multiple substrates, we would expect to see a correlation of correlations. For example, we would expect similar patterns of physical preferences at a position for multiple substrates. We observed a striking correlation of correlations for a few positions including position 1’ of the MA/CA and p1/p6 substrates ($R^2 = 0.99$, Figure. 2.3D), and position 1 of
MA/CA and p1/p6 \( (R^2=0.92, \text{Figure 2.3E}) \). With the exception of position 6’ in MA/CA and NC/p1, all the strongest correlation of correlations that we observed were close to the scissile bond (Figure 2.3F). Of note, the high correlations of correlations are skewed towards the C-terminal side of the substrate \( (p1, p1’, p2’, p3’) \) for reasons that we do not understand. The correlation of correlations suggests that multiple physical features with complex interdependencies underly cutting at positions close to the scissile bond. Further from the scissile bond, the physical features that mediate cutting efficiency appear to differ between different cleavage sites.

**Distribution of functional effects (DFE)**

To assess the general impacts of mutations at cut-sites, we examined the distribution of functional effects for each cut-site (Figure 2.4). All three cut-sites showed similar patterns with many amino acid changes causing either large decreases or large increases in cutting relative to WT. For all three cut-sites, there were very few amino acid changes that resulted in WT-like cutting function. Strong defects have been commonly observed in almost all previous analyses of either random or systematic amino acid changes in proteins (Boucher et al., 2016; Jiang et al., 2013; Serohijos et al., 2012). These general findings are consistent with the cooperative and sensitive nature of the forces that govern protein conformation. Amino acid changes at many/most positions can alter main-chain structure sufficient to severely disrupt function. Most previous analyses of the distribution of effects
of amino acid changes in proteins have observed many mutations with little to no effect (WT-like), and very few mutations with increased function relative to WT (Boucher et al., 2016; Jiang et al., 2013; Serohijos et al., 2012), even under stress conditions (Flynn et al., 2020; Romero et al., 2015) consistent with natural selection for WT sequences with near-optimal function. In contrast, the distribution of effects that we observe for cut-sites indicates that the WT sequences in our engineered setup are at an intermediate level of potential function. The intermediate level of WT function may have multiple contributing factors that include: balancing natural selection where an intermediate cutting function results in the most effective natural viral fitness; and/or our engineered yeast display setup may artificially alter the way mutations impact cutting function.
Figure 2.4. Distribution of mutant effects in cut-sites. (A) Distribution of functional effects measured by yeast-display for point mutants in the MA/CA, NC/p1, and p1/p6 cleavage sites. Point mutations within two standard deviations of WT synonyms are indicated in light gray, while variants cut slower are in black and variants cut faster are in dark gray. (B) Fraction of WT-like, slower-, and faster-cut variants for each cut-site.
Variants with increased cutting function

Motivated by the unusually high number of variants that increased cutting function, we analyzed them in further detail (Figure 2.5). For all three cleavage sites, amino acid changes that increased function were consistently observed in both experimental replicates (Figure 2.5A), as expected based on the reproducibility of the yeast display assay. While all three cleavage sites showed a similar number of variants with increased cutting function, the MA/CA site had the most. We explored the physical properties of MA/CA amino acid changes that increased cutting function. Consistent with the average trends across all three cleavage sites (Figure 2.2), the variants with faster cutting of MA/CA contained changes that were predominantly to aromatic and negatively charged amino acids (Figure 2.5B). The faster cutting amino acid changes in MA/CA tended to occur away from the scissile bond and were most abundant at positions 5, 4, 4’, and 5’ (Figure 2.5C). The structure of a PR-substrate complex shows that five Lysine and Arginine residues on PR that are in proximity of the last visualized substrate atoms at positions 4 and 4’ (Figure 2.5D). An electrostatic representation of the surface of PR (Figure 2.5E) indicates that these lysine and arginine residues generate positive charge that is located adjacent to both the 4 and 4’ positions of substrate. Of note, the structure of the PR dimer is highly symmetrical such that the local environment adjacent to positions 4 and 4’ are nearly identical. Faster cutting of MA/CA substrates in our yeast display assay may be driven by electrostatic complementarity with the positively charged surface region of PR adjacent to positions 4 and 4’ of substrate.
Figure 2.5

(A) Variants that cut faster than WT

(B) Fraction of faster cutting MA-CA variants

(C) Fraction of faster cutting MA-CA variants by position

(D) Molecular structure

(E) Molecular structure

(F) Average fraction MA-CA remaining (N=3)

(G) Relative function by gel assay
**Figure 2.5.** Examining variants that were cut fast by yeast display assay. (A) The number of variants that were cut faster than WT by more than two standard deviations of WT synonyms. (B) The type of amino acids changes in variants of the MA/CA cleavage site that resulted in faster than WT function. (C) The position of amino acid changes in the MA/CA cleavage site that were cut faster than WT. (D&E) Structural representation of PR with bound MA/CA peptide substrate based on 1KJ4.PDB [14]. (D) Positions 4-4′ of substrate are shown in yellow with Cα atoms as spheres. Side-chains of lysine and arginine residues of PR that are located in the proximity of the end of the active site are shown in ball and stick representation with carbon atoms colored green. (E) Surface representation indicating electrostatic charge. (F&G) Analysis of the cutting function of a panel of variants in full-length MA-CA substrate using a gel-based assay. Error bars in panel G represent the standard deviation of three assays.
Motivated by the wide range of function seen in gel-based analyses of MA/CA substrates that were fast cutting in our yeast display assay, we investigated an additional panel of MA/CA variants using a gel-based approach. Using purified PR, we measured the cutting of full-length MA-CA protein (Figure 2.5F&G). The individual variants showed a wide range of cutting in the gel assay using MA-CA as substrate, but none of them were significantly faster than WT. The faster cutting of these variants in the yeast display assay but the WT-like or slower cutting in the gel assay indicate that substrate residues far from the scissile bond can have large impacts on cutting rate. As further discussed in conclusions, our results suggest that the native MA-CA substrate is highly selected to cut rapidly including long-range interactions from sequences far from the scissile bond. By removing these long-range interactions in our engineered yeast display substrates we may have provided artificial opportunities for cut-site mutations to increase cutting rate.

**Comparison of yeast display findings with natural variation in HIV-1**

To explore cut-site diversity in circulating HIV-1, we extracted and analyzed sequences of the MA/CA, NC/p1, and p1/p6 cut-sites from the Los Alamos database (http://www.hiv.lanl.gov/). To facilitate the analysis of all three cleavage sites, we scaled variants that cut faster than WT such that 2 represents full cutting under the experimental conditions. A combined analysis of all three cleavage sites (Figure 2.6A) provided similar trends to each site individually (data not shown). Common polymorphisms that were at a frequency of $3 \times 10^{-4}$ or greater formed two clusters based on the yeast display
measures of function, one cluster was faster than WT and one was broad and overlapped with WT (Figure 2.6A). We chose to consider amino acid changes observed at a frequency of $10^{-3}$ or greater, corresponding to more than 10 observations, in order to eliminate the majority of null alleles such as stop codons that have been observed in similar HIV-1 sequence datasets (Boucher et al., 2019). The functional distribution of polymorphisms is similar to the distribution of all mutations with less than a 50% defect in function. In contrast, very few of the polymorphisms showed strong decreases in yeast display function despite this being a common outcome among the comprehensive variants we analyzed. Random sampling of amino acid changes led to a significantly greater number of variants with strong functional defects compared to natural polymorphisms ($p<0.01$, Figure 2.6B). These observations indicate that our yeast display assay captures fundamental aspects that determine strong defects in the function of cut-sites and that purifying selection acts to purge these variants from circulating viral populations. In contrast, random sampling led to a similar number of variants with increased function relative to WT according to our yeast-display measurements. These observations are consistent with a lack of selection for faster cutting sites and/or an enrichment of faster cutting measurements caused by artifacts in our yeast display approach.
Figure 2.6. Comparison of functional effects with sequence variation in circulating isolates. (A) Overlay of a violin plot in grey of the functional effects measured by yeast display of all possible point mutations and black dots representing common polymorphisms observed four or more times in a dataset of roughly 12,000 sequenced isolates. (B) Comparison of the number of common polymorphisms with either hi or lo relative function with randomly sampled sets of variants. Error bars indicate the standard deviation from 1,000 random simulations.
Conclusions
This study highlights how interpreting biochemical analyses in the light of evolution is both important for understanding biology (Dobzhansky, 1973) and challenging. One of the biggest challenges in this work was considering potential differences between high-throughput measurements that we made using a yeast-display approach, and the cutting that occurs in viruses and how it relates to fitness. Translating to viral fitness is particularly challenging because cutting in viruses can occur intramolecularly (Pettit et al., 2004), and the relationships between cutting efficiency and fitness remain poorly defined for most cut-sites (Lee et al., 2012). Our display approach fused cleavage sites to a globular protein with cut-sites bracketed by multiple glycines, as similar approaches had been shown to mimic the cutting of biologically relevant substrates (Laco, 2015). While we observed strong correlation of our yeast-display results with cutting of endogenous substrates for cut-site variants with WT-like or lower function, we found that variants that cut faster than WT by yeast-display had a wide array of function when measured using a full-length substrate (Figures 1.1&1.5). In both the yeast display measurements and gel-based analyses, substrate concentrations were well below previously determined Michaelis-Menten values (Lee et al., 2012), such that both assays should report on impacts on enzyme proficiency (kcat/Km). In generating conclusions based on our results, we carefully consider that many sites that were cut faster than WT by yeast-display were due to features in the yeast-display substrates that differ from full-length substrates.
Somewhat ironically, this caveat supports one of our main conclusions, that the function of cut-site variants has strong sequence dependencies. Our observations that cut-site variants that were faster than WT by yeast display had WT or lower function when analyzed in a natural substrate context demonstrates that long range sequence dependencies beyond positions 6-6’ contribute to function. Of note, biochemical analyses indicate that regions of substrate far from the scissile bond can contribute to binding to protease (Prabu-Jeyabalan et al., 2000). Our findings are consistent with a model where our yeast display construct disrupts or eliminates contacts with protease compared to the natural MA-CA substrate (e.g., Figure 1.5). In this model (Figure 1.7), the artificial disruption of contacts in the yeast display setup provides an opportunity for amino acid changes at positions towards the end and past the active site to recoup binding energy with protease. However, in the natural MA/CA substrate, the effect of these amino acid changes has varied impacts because it disrupts the natural interactions of MA/CA with protease.
Figure 2.7. Model of potential distinctions between functional impacts of mutations in engineered cut-sites compared to their native context. Long-range contacts of engineered substrates with PR are unlikely and may provide opportunities for cut-site proximal mutations to increase cutting rate that are not available in native substrates where long-range contacts have been selected.
We also observe strong sequence dependencies throughout most 6-6’ positions in most cut-sites (e.g., Figure 1.2). Importantly, the comparison between the functional effects of amino acid changes in different cleavage sites where all measurements were made using yeast display are inherently normalized for the impacts of long-range sequence dependencies because they all are in the same (yet artificial) long-range background. Our analyses indicate that positions 6-3 and 4’-6’ that are distal to the scissile bond had very little to no consistent biophysical underpinnings (Figure 2.3).

In contrast, we observe consistent but complex physical underpinnings of function at positions adjacent to the scissile bond (Figure 2.3). We did not find any simple biophysical parameter that correlated strongly with the functional impact of mutations at any position in any cut-site that we tested. However, we do find that the correlations of physical parameters correlates strongest between cut-sites at positions 1,1’, and 2’. This observation suggests that similar physical properties with complex interdependencies mediate cut-site efficiency at positions adjacent to the scissile bond.

This work provides a new and generalizable approach to investigating the relationship between viral proteases and their substrates. We show that long-range and short-range sequence interdependencies are prevalent for HIV-1 cut-site efficiency, but that consistent and complex biophysical properties have large impacts at sites adjacent to the scissile bond. We also
demonstrate the importance of careful investigation of experimental setup compared to natural substrates. In future efforts, it will be interesting to examine if similar or distinct physical and sequence-dependent features mediate other viral proteases.
Methods

Engineered libraries of cut-site variants

Cut-site variants were engineered in the pCTCON2 yeast-display plasmid (Stiffler et al., 2015). The pCTCON2 plasmid encodes a galactose-inducible promoter driving the expression of a fusion of the Aga2 gene followed by an HA-tag and then a Myc-tag. The fusion protein is directed to the yeast surface by Aga2. We engineered pCTCON2 to encode three glycines and a unique PstI restriction site after the HA-tag and a unique Nhel site and three glycines before the Myc-tag. Next, we used cassette ligation to introduce cut-site variants between the PstI and Nhel restriction sites. To generate point mutant libraries, we used cassettes encoding individual NNK codons (where N indicates a mixture of A,C,G,T and K a mixture of G,T). For each of the MA/CA, NC/p1, and p1/p6 sites, separate cassette ligations were performed with NNK at each of the 12 positions under investigation. Point mutant libraries for each site were generated by mixing libraries from each position (6 to 6’). The resulting plasmids encode Aga2-HA-GGG-CS-GGG-Myc, where CS indicates the cut-site from position 6 to 6’. The wildtype sequence for each cut-site was based on the sequence of NL4-3.
**Yeast surface display**

Yeast strain EBY100 (Stiffler et al., 2015) was used in all yeast-display experiments. We used the lithium acetate method (Gietz et al., 2007) to introduce plasmids into EBY100 cells. For cut-site libraries, 1 μg of plasmid was used. After transformation with cut-site library plasmids, cells were allowed to recover in 5 mL of synthetic dextrose media at room temperature for 16 hours. Cells were then collected by centrifugation and washed extensively in synthetic dextrose media to remove extracellular plasmid. Cells were then grown in 100 mL of casamino acid (CAA) media with dextrose. CAA-D (CAA with dextrose) media lacks tryptophan, which selects for yeast transformed with the pCTCON2 plasmid that enables tryptophan production. A portion of cells were plated to determine the number of transformed cells. Cells were grown in CAA-D media at 30 °C until they reached saturation and then mixed with glycerol to a concentration of 25% and frozen at -80 °C.

Before performing yeast-display assays, a single aliquot for each cleavage site library was thawed and used to inoculate 50 mL of CAA-D media. These cultures were grown in a shaking incubator at 30 °C for 24 hr, and then diluted into 50 mL of fresh CAAD to an OD600 of about 0.1. For the next 6-8h, yeast growth was monitored until the cells had undergone 1-2 doublings. Cells were then collected by centrifugation and washed 3 times with CAA-RG media (CAA media with 1% raffinose and 1% galactose). Cells were resuspended in 50 mL CAA-RG media to an OD600 of about 0.5, and shaken at 30 °C for a
further 16 hr. As a control of non-displaying cells, a subset of the culture was grown in CAA-D media.

**Protease treatment, labeling, and FACS of yeast display cells**

Yeast display cells were collected by centrifugation. Cells were washed with protease assay buffer (PAB). PAB contained 50 mM sodium acetate at pH 6.0, and 100 mM sodium chloride. 106 cells were resuspended in 100 μL PAB and purified PR was added to 1 μM. Purified HIV-1 PR was a kind gift from the Schiffer lab at UMass Chan Medical School. Samples were incubated in a shaking incubator at 30 °C for one hour for MA/CA, 30 °C for 75 minutes for NC/p1, and 30 °C for 40 minutes for p1/p6 libraries. The protease reaction was stopped by collecting cells by centrifugation and washing them three times in TBSB. A control sample of yeast were left untreated with protease. Following protease treatment, cells were labeled with antibodies to the HA-tag (Alexa488 conjugated, Cell Signaling Technology #2350) and the myc-tag (Alexa647 conjugated, Cell Signaling Technology #2233). Antibody labeling was performed according to the manufacturer’s recommendations using 1:100 dilutions of each antibody and a 1 hour incubation at 23 °C in TBSB. Labelled cells were washed and suspended in TBSB at a density of 106 cells/mL for FACS.

Cells were sorted on a FACS Aria II cell sorter (BD Biosciences) at the University of Massachusetts Chan Medical School Flow Cytometry Core. Cells untreated with protease were single labeled and used to setup forward-and side-scatter windows that omitted dead and/or aggregated cells.
Control cells untreated with protease that were double labeled were used to setup the voltage for robust detection of both labels. Protease treated samples were sorted into four windows: one window encompassing the profile of uncleaved cells, one window encompassing the profile of fully-cleaved cells, and two intermediate windows representing different levels of partial cleavage. A total of at least 150,000 cells were sorted for each window. Sorted cells were grown in 50 mL of CAA-D media on a shaking incubator at 30 °C for 24 hours. The cultures from sorted cells were collected by centrifugation, washed with TBSB and stored as pellets at -80 °C.

**Preparation of DNA and next-generation sequencing**

Plasmid DNA was isolated and analyzed essentially as previously described (Flynn et al., 2020). Briefly, cell pellets were resuspended in buffer P1 (Zymo Research) and treated with zymolyase (Zymo Research) to remove the cell wall. The cells were then lysed with buffer P2 (Zymo Research), neutralized with buffer P3 (Zymo Research), and plasmid was purified using a Zymo-Spin IIN spin column (Zymo Research). Purified samples were then amplified using primers bracketing the cut-site sequence. The forward PCR primer included a P5 sequence and the reverse primer included a P7 sequence in order to facilitate Illumina sequencing. We included a 6-base barcode on the P7-containing primer in order to distinguish samples from different cut-sites and sort windows. PCR products were purified using a Zymo-Spin IIN column following the manufacturers recommendations (Zymo Research). Each PCR
product was quantified by qPCR using the Kapa Sybr Fast qPCR Master Mix (Kapa Biosystems) on a Bio-Rad CFX instrument. All PCR products were diluted to 4nM and then pooled for next-generation sequencing. The pooled sample was sequenced on a NextSeq500 sequencer that was kindly made available by the Rando lab at UMass Chan Medical School. A single-end 100 base sequencing run was used to sequence both the cut-site and the indexes that distinguish each sample.

**Estimation of functional effects from sequencing results**

We obtained a total of about $3.8 \times 10^8$ sequence reads. Reads were checked for exact match to constant regions between the cut-site and the index. Sequences that passed this quality test and had PHRED scores of 30 or above for all bases encoding the cut-site were then parsed into separate files based on the index read. The sequence of the cut-site in these reads were then analyzed and the counts of all individual amino acid changes tabulated. In each sort window, we calculated the frequency of each variant ($F_i$) as the sequence counts of that variant divided by the total sequencing counts for the window. In order to account for the asymmetric distribution of variants across the FACS windows, we calculated the relative window size ($W_j$) as the fraction of cells from the protease treated library sample in each window. We then quantified the activity ($A$) of each variant using an approach similar to a center of mass calculation using the following equation:
\[ A = \sum_{j=1}^{4} \frac{F_i \times W_j \times V_j}{\sum_{j=1}^{4} F_i \times W_j} \]

Where \( V_j \) is the value associated with each window. We used values of 1-4 for \( V_j \) representing the window with the least cutting to the window with the most cutting. These activity estimates range from 1 for a variant that was uncut to 4 for a variant that was fully cut. To facilitate comparison between the effects of amino acid changes in each cut-site, we estimated relative function (RF) by normalizing the activity measurements with linear transformations such that 0 represents an uncut variant (based on the average of the five most defective variants in each library) and 1 represents the wildtype cut-site. For the heatmap analyses in Figure 2.2 and the comparison to diversity in circulating variants in Figure 2.6, we performed a second linear normalization on variants with faster than wildtype RF such that the average stop codon had a value of 2. This normalization was done to account for varied sensitivity to detect faster cutting variants due to distinctions in the fraction of cutting of the wildtype for each cut-site library.

**Gel-based analyses of the cutting of full-length MA/CA variants**

Full-length MA-CA was cloned into pET21 with an in-frame 6xHis tag at the C-terminus. Cut-site variants were generated by site-directed mutagenesis. Plasmid variants were transformed into BLR(DE3) pRIL bacteria that were grown in 2xYT media to an OD600=0.8 and induced with 1 mM IPTG for 3 hours. Cells were collected by centrifugation and stored at -80 °C. Cell pellets
were thawed and resuspended in 25 mM HEPES pH 7.5 with 500 mM sodium chloride. Cells were lysed by sonication and cell debris pelleted by centrifugation. Soluble MA-CA protein was purified from the supernatant by nickel affinity chromatography. Purified protein was dialyzed into 25 mM HEPES pH 7.5 with 10 mM sodium chloride and further purified by anion exchange chromatography. Purified protein was dialyzed into 25 mM HEPES pH 7.5 with 10 mM sodium chloride. The concentration of MA-CA protein was determined based on absorbance at 280 nm. Cutting assays were performed at 25 °C in 50 mM sodium acetate at pH 6 with 100 mM sodium chloride. Cutting assays contained 250 nM of purified PR and 10 μM of MA-CA variants.

Timepoints samples were taken and the reaction stopped by the addition of 2% SDS and heating to 95 °C for 2 minutes. Timepoint samples were run on a denaturing polyacrylamide gel and stained with Coomassie brilliant blue. The density of the MA-CA band was quantified using an Amersham Imager 600. All gel-based cutting assays were performed in triplicate.

**Analysis of cut-site variation in sequenced isolates**

We analyzed the cleavage site sequences of MA/CA, NC/p1 and p1/p6 in roughly 12,000 HIV-1 subtype-B isolates in the Los Alamos HIV database (http://hiv-web.lanl.gov). The majority of sequenced isolates (roughly 11,000) were from drug-treated individuals. Variants were defined as changes in amino acid with respect to the sequence of NL4-3. We identified 71 amino acid variations across all three cleavage sites that we analyzed (MA/CA,
NC/p1, p1/p6) that were observed at a frequency of 10^{-3} or greater. Roughly, equivalent percentage of these 71 amino acid changes were from each cut-site (38% for MA/CA, 29% for NC/p1, and 33% for p1/p6). Bootstrap analyses were performed by randomly sampling 71 amino acid changes from the combined set of functional scores for all three cut-sites with faster cutting sites rescaled so that 2 represents full cutting under the experimental conditions (Figure 1.6A). One thousand random simulations were performed and the average and standard deviations of sites with high functional scores (>1.8) and low functional scores (<0.2) were shown in Figure 1.6B.

**Acknowledgements**

This study benefited tremendously from useful input and discussions with Prof. Celia Schiffer and Dr. Julia Flynn. This work was supported by funds from grant R01GM112844 from the National Institutes of Health to D.N.B. The authors declare that they have no conflicts of interest with this work.
Chapter III - Identification of a Permissive Secondary Mutation That Restores the Enzymatic Activity of Oseltamivir Resistance Mutation H275Y

This chapter has been published previously:


This work was in collaboration with Jennifer P Wang, MD. It involves biochemical analysis of a panel of mutations that were previously identified as compensatory mutations in neuraminidase (NA). My contributions to this work include the experimental design using a modern strain of IAV, A/California/04/2009, including processing of viral DNA for site directed mutagenesis, cell transfection experiments for generating replicate data, and performing NA expression and activity experiments. I performed all of the cell culture and in-vitro experiments pertaining to this modern strain including, biochemical analysis and flow cytometry experiments, and contributed to the manuscript writing of this data set. Li Jiang and Ping Liu performed all other in-vitro and cell culture experiments using A/WSN/33 strain of IAV, including the viral assessments of A/WSN/33 in different cell lines. Li Jiang and Daniel N. Bolon prepared the manuscript.
Abstract

Many oseltamivir resistance mutations exhibit fitness defects in the absence of drug pressure that hinders their propagation in hosts. Secondary permissive mutations can rescue fitness defects and facilitate the segregation of resistance mutations in viral populations. Previous studies have identified a panel of permissive or compensatory mutations in neuraminidase (NA) that restore the growth defect of the predominant oseltamivir resistance mutation (H275Y) in H1N1 influenza A virus. In prior work, we identified a hyperactive mutation (Y276F) that increased NA activity by approximately 70%. While Y276F had not been previously identified as a permissive mutation, we hypothesized that Y276F may counteract the defects caused by H275Y by buffering its reduced NA expression and enzyme activity. In this study, we measured the relative fitness, NA activity, and surface expression, as well as sensitivity to oseltamivir, for several oseltamivir resistance mutations, including H275Y in the wild-type and Y276F genetic background. Our results demonstrate that Y276F selectively rescues the fitness defect of H275Y by restoring its NA surface expression and enzymatic activity, elucidating the local compensatory structural impacts of Y276F on the adjacent H275Y.
Importance

The potential for influenza A virus (IAV) to cause pandemics makes understanding evolutionary mechanisms that impact drug resistance critical for developing surveillance and treatment strategies. Oseltamivir is the most widely used therapeutic strategy to treat IAV infections, but mutations in IAV can lead to drug resistance. The main oseltamivir resistance mutation, H275Y, occurs in the neuraminidase (NA) protein of IAV and reduces drug binding as well as NA function. Here, we identified a new helper mutation, Y276F, that can rescue the functional defects of H275Y and contribute to the evolution of drug resistance in IAV.

Keywords: H275Y, Y276F, influenza A virus, neuraminidase

Influenza A virus (IAV) causes seasonal epidemics and recurrent pandemics (Molinari et al., 2007). IAV has two surface glycoproteins, hemagglutinin (HA) and neuraminidase (NA), which codetermine the subtype of IAV (e.g., H1N1 or H3N2). HA binds to sialic acid on the surface of host cells and mediates endocytosis of viral particles, while NA cleaves sialic acid to enable efficient release of newly budded viral particles. Disrupting the function of either NA or HA is effective at limiting IAV infections (Chen et al., 2019; García-Sastre, 2010; Ohuchi et al., 2006; Ping et al., 2010; Rott, 1992).

Oseltamivir is an orally administered NA competitive inhibitor (NAI). Oseltamivir remains the most widely used therapeutic against IAV and
continues to be stockpiled in preparation for future pandemics. However, numerous oseltamivir resistance mutations have been detected in viruses isolated from patient samples (McKimm-Breschkin, 2012; Nguyen et al., 2012). The predominant oseltamivir resistance mutation in the H1N1 subtype is the histidine-to-tyrosine mutation at position 275 of NA (H275Y in the N1 numbering system used in this work). H275Y causes multiple H1N1 strains to become insensitive to oseltamivir (Abed et al., 2003; Hurt et al., 2012). Notably, oseltamivir-resistant H1N1 strains with H275Y circulated globally with high frequency (.50%) between 2007 and 2009 and have continued to exist in human hosts with a 1 to 10% frequency (Hurt et al., 2009, 2011; Meijer et al., 2009; Storms, 2012). Understanding the fitness of oseltamivir resistance mutations is important for predicting their contribution to future IAV genetic variation and the likely future effectiveness of NAIs.

The majority of oseltamivir resistance mutations cause fitness defects in the absence of drug pressure. Previously, we performed a high-throughput mutational scan that quantified the fitness effect of all single nucleotide mutations in the active site of NA, which identified fitness defects in the absence of drug for known oseltamivir resistance mutations (Jiang et al., 2016). For example, I223M, H275Y, and N295S all showed ;30% fitness defects without drug pressure. Of note, both H275Y and N295S impaired the replication and transmission capability of IAV (Abed et al., 2003, 2006; Herlocher et al., 2004; Ives et al., 2002), consistent with the fitness defects observed in our screen. Most oseltamivir resistance mutations accumulate at the active site of NA and disrupt binding to substrate as well as drug, thus
reducing enzyme activity. Many of the resistant mutations, including H275Y, also decrease net enzyme activity by perturbing the expression level of NA (Bloom et al., 2010; Russell et al., 2008). Reduced enzyme activity of NA in turn leads to attenuated infectivity and/or transmission in the absence of drug pressure. In sum, the fitness cost of oseltamivir resistance mutations hinders their propagation.

Permissive or compensatory mutations can compensate for the fitness cost of oseltamivir resistance mutations. Secondary mutations are able to compensate for the reduced net enzyme activity and restore the replication efficiency. Studies by Bloom and colleagues found that R194G, R222Q, and V234M were able to increase the surface expression of NA with H275Y in the seasonal H1N1 strains and rescue fitness (Bloom et al., 2010). R222Q was also shown to improve the binding affinity (Km) and catalytic efficiency (Vmax) of NA with H275Y (Abed et al., 2011). Three additional permissive mutations identified from phylogenic analyses of natural isolates were also found to buffer the fitness defect of H275Y, though to a lesser extent (Duan et al., 2014). These permissive mutations exhibited high frequency in H1N1 when H275Y was prevalent from 2007 to 2009 (Abed et al., 2011; Duan et al., 2014), providing a plausible explanation for competent growth and transmission of H275Y. Although the vast majority of circulating viruses in 2012 to 2013 were NAI sensitive, approximately 99% of current circulating pandemic H1N1 2009 (pH1N1) carry two other mutations, V241I and N369K, which are permissive for the H275Y substitution (Butler et al., 2014; Meijer et al., 2014; Takashita et al., 2015). Experiments in ferret models also
demonstrated that V241I and N369K rescued the replication and transmission defect of a H1N1 strain with H275Y (Abed et al., 2013; Butler et al., 2014). High-throughput screening also identified E215D as an additional mutation compensating for the fitness defect of H275Y in several primary isolates (Wu et al., 2013). These observations indicate the continued potential for the emergence of a highly fit H1N1 with resistance to NAIs depending on the prevalence of permissive mutations. Further investigation of the mechanism of permissive mutations is thus necessary to understand the evolutionary potential of IAV in the presence of selection pressure from NAIs.

In this study, we determined the potential of Y276F as a permissive mutation to rescue the fitness defect of H275Y. Y276F is a hyperactive mutation that enhances the enzyme function of NA from A/WSN/33 by approximately 70% (Jiang et al., 2016). We hypothesized that Y276F may restore the fitness defect of oseltamivir resistance mutations by increasing net NA activity. We determined the experimental fitness, enzyme activity, surface expression, and sensitivity to oseltamivir for a panel of oseltamivir resistance mutations with or without Y276F. In the A/WSN/33 strain, we found that Y276F restores the fitness, enzyme activity, and surface expression for H275Y, but not other resistance mutations, indicating a specific epistatic interaction between Y276F and H275Y instead
Figure 3.1. Plaque size assessment of the experimental fitness of oseltamivir-responsive mutations with or without Y276F. (A) Representative images from replicate plaque assays used to estimate fitness. All variants and replicates were performed in parallel. (B) The fitness based on plaque size on MDCK cells of NA with mutations in an otherwise WT background (black) or with Y276F (gray). The plaque size of mutant virus was normalized to that of the WT virus. Error bars show standard deviations (n = 3). **, P < 0.05, based on Student's t test. (C) Fitness of double mutants from independent expectations (light blue) or from plaque size measurements on MDCK cells (dark blue). Error bars represent standard deviations (SD) (n = 3, with propagated errors for independent expectations). *, P < 0.1, based on Student's t test.
of a general buffering role of Y276F. In addition to the A/WSN/33 strain, we also found that Y276F rescues net enzyme activity of H275Y in a more modern strain (A/California/04/09 NA). Taken together, the results of our study show that Y276F can restore the enzyme activity and fitness of H275Y NA.

Results

**Y276F rescues the fitness defect of H275Y.** We sought to determine the relative fitness of the oseltamivir resistance mutations in the wild type (WT) or Y276F mutated genetic background. In the absence of Y276F, all of the oseltamivir resistance mutations exhibited 30 to 50% fitness defects (Figure. 3.1A and B; Table 3.1), consistent with our previous high-throughput screening. In contrast, Y276F provided a fitness advantage compared to WT of about 20% (P < 0.05), also consistent with our previous high-throughput studies (Jiang et al., 2016). Of note, Y276F has not been observed at high frequency in H1N1 isolates, indicating distinctions between lab and natural evolution that are further detailed in Discussion. Y276F did not change the fitness of most resistant mutations in the absence of drug pressure. Among the resistant mutations we analyzed (I223M, I223T, I223L, H275Y, and N295S), H275Y was the only one for which fitness was clearly increased by Y276F (Figure. 3.1B). Our measurements of fitness for single and double
mutations provided an opportunity to investigate epistatic effects among Y276F and oseltamivir resistance mutations. We quantified epistasis by comparing the fitness effects of double mutants to the effects of each mutation individually. We considered the fitness effects of each mutant independent if the double mutant was equal to the product of the fitness effects of single mutants. In contrast, positive epistasis means that the double mutant is more fit than predicted based on the fitness effects of the single mutants, and negative epistasis is when the double mutant is less fit than predicted. Most of the drug-resistant
Table 3.1 Relative plaque size measurements for A/WSN/33 viruses with different NA variants on MDCK cells

<table>
<thead>
<tr>
<th>Experimenta</th>
<th>No. of plaques</th>
<th>Size (μM)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>WT-1</td>
<td>18</td>
<td>58.5</td>
<td>16.6</td>
</tr>
<tr>
<td>WT-2</td>
<td>26</td>
<td>54.9</td>
<td>15.8</td>
</tr>
<tr>
<td>WT-3</td>
<td>30</td>
<td>64.3</td>
<td>11.2</td>
</tr>
<tr>
<td>Y276F-1</td>
<td>23</td>
<td>61.9</td>
<td>18.13</td>
</tr>
<tr>
<td>Y276F-2</td>
<td>19</td>
<td>80.9</td>
<td>19.21</td>
</tr>
<tr>
<td>Y276F-3</td>
<td>37</td>
<td>73.3</td>
<td>18.69</td>
</tr>
<tr>
<td>H275Y-1</td>
<td>12</td>
<td>35.2</td>
<td>7.45</td>
</tr>
<tr>
<td>H275Y-2</td>
<td>23</td>
<td>48.4</td>
<td>10.98</td>
</tr>
<tr>
<td>H275Y-3</td>
<td>12</td>
<td>45.6</td>
<td>9.32</td>
</tr>
<tr>
<td>I223L-1</td>
<td>26</td>
<td>37.7</td>
<td>13.45</td>
</tr>
<tr>
<td>I223L-2</td>
<td>26</td>
<td>43.9</td>
<td>10.25</td>
</tr>
<tr>
<td>I223L-3</td>
<td>26</td>
<td>45.9</td>
<td>14.29</td>
</tr>
<tr>
<td>I223M-1</td>
<td>13</td>
<td>33.8</td>
<td>10.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>-------</td>
<td>------</td>
<td>------</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I223M-2</td>
<td>24</td>
<td>34.4</td>
<td>10.51</td>
</tr>
<tr>
<td>I223M-3</td>
<td>26</td>
<td>27.4</td>
<td>7.49</td>
</tr>
<tr>
<td>I223T-1</td>
<td>20</td>
<td>48.3</td>
<td>13.01</td>
</tr>
<tr>
<td>I223T-2</td>
<td>17</td>
<td>49.7</td>
<td>16.22</td>
</tr>
<tr>
<td>I223T-3</td>
<td>37</td>
<td>46.9</td>
<td>15.34</td>
</tr>
<tr>
<td>N295S-1</td>
<td>32</td>
<td>43.2</td>
<td>10.71</td>
</tr>
<tr>
<td>N295S-2</td>
<td>23</td>
<td>49.2</td>
<td>14.01</td>
</tr>
<tr>
<td>N295S-3</td>
<td>32</td>
<td>46.5</td>
<td>11.65</td>
</tr>
<tr>
<td>Y276F-H275Y-1</td>
<td>30</td>
<td>56.4</td>
<td>18.39</td>
</tr>
<tr>
<td>Y276F-H275Y-2</td>
<td>29</td>
<td>60.0</td>
<td>21.82</td>
</tr>
<tr>
<td>Y276F-H275Y-3</td>
<td>29</td>
<td>58.0</td>
<td>11.67</td>
</tr>
<tr>
<td>Y275F-I223L-1</td>
<td>15</td>
<td>32.9</td>
<td>10.6</td>
</tr>
<tr>
<td>Y275F-I223L-2</td>
<td>32</td>
<td>40.2</td>
<td>11.04</td>
</tr>
<tr>
<td>Y275F-I223L-3</td>
<td>27</td>
<td>43.4</td>
<td>12.29</td>
</tr>
<tr>
<td>Y275F-I223M-1</td>
<td>8</td>
<td>28.4</td>
<td>6.86</td>
</tr>
<tr>
<td>Y275F-I223M-2</td>
<td>9</td>
<td>27.2</td>
<td>6.39</td>
</tr>
<tr>
<td>Y275F-I223M-3</td>
<td>12</td>
<td>35.3</td>
<td>13.17</td>
</tr>
<tr>
<td>Y275F-I223T-1</td>
<td>26</td>
<td>49.7</td>
<td>13.1</td>
</tr>
</tbody>
</table>
Y275F-I223T-2 | 26 | 50.4 | 12.43
Y275F-I223T-3 | 35 | 50.7 | 14.61
Y275F-N295S-1 | 26 | 41.7 | 11.77
Y275F-N295S-2 | 38 | 45.1 | 10.04
Y275F-N295S-3 | 38 | 47.4 | 10.94

aNA variant and repeat number.

mutants showed negative epistasis with Y276F (Figure. 3.1C). In contrast, Y276F shows a pattern of positive epistasis with H275Y. These analyses indicate that the impacts of Y276F on fitness depend on the genetic background. We further investigated the impacts of H275Y and Y276F on viral fitness by assessing viral replication kinetics by reverse transcription-quantitative PCR (RT-qPCR). Fitness assessed by viral replication in MDCK cells is similar to what was estimated by plaque size (Figure. 3.2A). The RT-qPCR studies showed statistically significant decreases in fitness for the H275Y enzyme compared to WT and increases for the Y276F enzyme relative to WT at 24 h post infection. We also saw statistically significant positive epistasis between H275Y and Y276F from the replication data (Figure. 3.2B). RT-qPCR data follow the same trends as the plaque size data in MDCK cells, with less measurement variation and therefore greater ability to discern statistically significant differences. We also
examined the fitness effects of H275Y and Y276F in A549 cells by measuring viral replication kinetics by RT-qPCR (Figure. 3.2C and D). The A549 cell line is derived from a human lung carcinoma and thus is more representative of cells infected by IAV in humans. A549 cells and MDCK cells differ in the fraction of sialic acid-galactose (SA-Gal) linkages in the \( \alpha2,3 \) and \( \alpha2,6 \) configurations. We observed statistically significant differences in viral expansion in A549 cells that follow the same trends as in MDCK cells. These results suggest that the fitness rescue of H275Y by Y276F is observed in both host cells.
Figure 3.2

(A) Relative abundance of influenza A/WSN/33 WT, Y276F, H275Y, and H275Y/Y276F NA from MDCK cell culture supernatants collected 24 h following infection with an MOI of 0.01. (B) Epistasis analysis of H275Y/Y276F. Predicted expansion based on independent effects of each individual mutation is lower than experimental measurement of this variant (P = 0.01, single-tail t test with errors propagated for independent prediction). (C) Relative abundance of
influenza A/WSN/33 WT, Y276F, H275Y, and H275Y/Y276F NA in A549 cell culture supernatants collected 24 h following infection with an MOI of 0.1. Viral copies in supernatants were quantified by RT-qPCR of influenza A virus M segment, and the viral copy number of each was normalized to the average value of WT (5.2 10^8 or 2.1 10^8 copies/mL for two independent MDCK cell experiments; 1.2 10^8 or 1.7 10^8 copies/mL for two independent A549 cell experiments). Error bars in panels A and C represent the standard deviations for 5 wells from two independent experiments. *, P < 0.05; **, P < 0.01; ***, P < 0.001; ****, P < 0.0001 (ordinary one-way ANOVA). (D) Viral growth kinetics of WSN/33 variants in A549 cells. Each point is the average value of IAV RNA quantified by RT-qPCR for replicate wells from a representative experiment.
Y276F restores the enzymatic activity and surface expression of H275Y NA. Having established that Y276F can rescue the fitness defect of the H275Y enzyme, we examined the mechanism of this rescue. We first investigated how these mutations impacted enzyme activity in A/WSN/33 viral particles. Consistent with previous results, all of the oseltamivir resistance mutations showed reduced NA activity in the WT genetic background, indicating that the fitness defect of these mutations resulted from defects in enzyme activity (Figure. 3.3A). Conversely, Y276F alone exhibited an increase
Figure 3.3

Figure 3.3  Enzyme activity of NA with oseltamivir resistance mutations with or without Y276F. (A) Enzyme activity of variants in a WT genetic background (black) or with Y276F (gray). The enzyme activity of NA was determined by a fluorometric assay (MUNANA) and normalized to the viral titer. Relative NA activities were determined by normalizing to WT. Error bar indicate SD (n = 3). **, P < 0.05, based on Student's t test. (B) Relative surface expression of WT, H275Y, Y276F, and H275Y/Y276F NA determined by expressing NA in 293T cells. Relative expression was determined based on mean fluorescence from flow cytometry normalized to WT. SD are shown (n = 5). ***, P < 0.005, and ****, P < 0.0001, based on an unpaired Student's t test. (C) Percent total NA enzymatic activity for WT, H275Y, Y276F, and H275Y/ Y276F NA, determined by expressing NA in 293T cells, followed by MUNANA assay and normalization to WT. Error bar, show SD (n = 5). ****, P < 0.0001, based on unpaired Student's t test.
in NA activity of approximately 70% compared to WT (Figure. 3.3A). The enzyme activity of most drug-resistant mutations with Y276F was slightly worse than or similar to that seen with these mutations in a WT background. Thus, Y276F failed to rescue the enzyme activity of most drug-resistant NA mutants. The ability of Y276F, but not most drug resistance mutations studied here, to increase enzyme activity of WT provides an explanation for the observed negative epistasis among these mutations. The total enzyme activity of NA can be modulated by changing its quantity (surface expression) or proficiency (e.g., kcat/Km), so we sought to distinguish these two possibilities. We determined the relative surface expression and enzyme activity of A/WSN/33 NA variants (WT, H275Y, Y276F, and H275Y/Y276F) expressed in 293T cells. H275Y resulted in roughly 50% less surface expression of NA than WT (Figure. 3.3B) and enzyme activity that was further reduced (Figure. 3.3C), indicating that H275Y also reduced enzyme proficiency. Y276F had slightly less surface expression than WT but slightly more enzyme activity. This finding indicates that Y276F NA is a more proficient enzyme than WT in the A/WSN/33 strain. The H275Y/Y276F double mutant showed about 20% more surface expression of NA than WT and enzyme activity similar to that of WT, indicating slightly reduced enzyme proficiency. The surface expression and activity of Y276F NA did not differ with statistical significance (P > 0.1) from those of H275Y/Y276F NA. Together, these results indicate that Y276F restored the function of H275Y NA primarily through enhancing surface expression.
**Y276F has little effect on oseltamivir resistance.** We then examined the sensitivity of IAV to oseltamivir in the WT or Y276F genetic background. We determined 50% enzyme inhibitory concentration (IC50) values for oseltamivir inhibition of NA enzyme activity (Fig. 3.4 and Table 3.2). Drug resistance mutations all exhibited increased IC50s relative to WT, consistent with the ability of these variants to avoid inhibition by oseltamivir. On its own, Y276F exhibited an oseltamivir profile similar to that of WT. All drug resistance mutations also exhibited similar oseltamivir profiles with or without Y276F, indicating that Y276F does not have a strong impact on the affinity of NA for oseltamivir.

**Y276F rescues H275Y defects in a modern strain of IAV.** H275Y has been observed at low frequency (1 to 3%) in modern circulating IAV strains (Takashita et al., 2015), and we were interested to know if the Y276F mutation would also be permissive in a more modern and clinically relevant strain than A/WSN/33. We evaluated the effect of H275Y and Y276F in the A/California/04/2009 strain by expressing single and double mutants of NA in 293T cells. H275Y showed roughly 50% less surface expression of NA compared to WT NA (Figure. 3.5A) and a similar reduction in enzyme activity (Figure. 3.5B), indicating similar enzyme proficiency. Y276F increased surface expression by about 10% relative to WT.
Figure 3.4 Inhibition of NA variants by oseltamivir. (A and B) Enzyme activity of viruses with increasing doses of oseltamivir (dose-response curve) in the WT genetic background (A) or with Y276F (B). (C) Dose-response curve for WT and H275Y in the WT or Y276F background.
and enzyme activity by about 15%, indicating that enzyme proficiency was slightly increased. The H275Y/Y276F double mutant showed a 5% reduction in surface expression compared to WT, indicating that the Y276F mutation largely rescued the expression defect of H275Y NA. The H275Y/Y276F double mutant exhibited enzyme activity similar to WT and clearly higher activity than H275Y. These results indicate that Y276F restored the enzyme activity for H275Y in the A/California/04/2009 strain primarily by improving the surface expression of NA.

**Discussion**

In this study, we found that Y276F can rescue the fitness defect of H275Y by restoring surface expression and net enzyme activity of NA from A/WSN/33 and A/California/ 04/09 NA bearing the H275Y drug resistance mutation. Moreover, the H275Y/Y276F double mutant remained insensitive to oseltamivir inhibition in the A/WSN/33 strain.
Table 3.2. IC50s of substitutions with Y276 (WT) or F276

<table>
<thead>
<tr>
<th>Substitution</th>
<th>IC50 (μM) (fold change)</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Y276</td>
<td>F276</td>
<td></td>
</tr>
<tr>
<td>WT</td>
<td>0.33 (1)</td>
<td>0.21 (0.65)</td>
<td></td>
</tr>
<tr>
<td>I223L</td>
<td>3.2 (9.9)</td>
<td>3.0 (9.3)</td>
<td></td>
</tr>
<tr>
<td>I223M</td>
<td>3.4 (10)</td>
<td>5.0 (15)</td>
<td></td>
</tr>
<tr>
<td>I223T</td>
<td>1.1 (3.4)</td>
<td>1.4 (4.2)</td>
<td></td>
</tr>
<tr>
<td>H275Y</td>
<td>150 (460)</td>
<td>77 (230)</td>
<td></td>
</tr>
<tr>
<td>N295S</td>
<td>29 (90)</td>
<td>17 (52)</td>
<td></td>
</tr>
</tbody>
</table>
**Figure 3.5** Impacts of H275Y and Y276F in pH1N1 A/California/04/2009 strain. (A) The relative surface expression of NA variants (WT, H275Y, Y276F, and H275Y/Y276F) was assessed based on expression in 293T cells and flow cytometry. (B) Enzyme determined by MUNANA assay and normalization to WT. Error bars show SD (n = 5). ***, P = 0.004, and ****, P < 0.0001, based on unpaired Student's t test.
These results indicate that Y276F may serve as a permissive mutation of H275Y, allowing robust growth of H275Y comparable to that of WT. In particular, Y276F itself exhibited no fitness defect, which may facilitate its expansion in IAV populations. Consistent with this idea, Y276F was identified, albeit at low frequency, in human isolates in 2009 and 2010 (Jiang et al., 2016). Moreover, Y276F did not affect sensitivity of the H275Y enzyme to oseltamivir, so H275Y/Y276F may confer resistance to oseltamivir treatment without diminishing viral fitness. In a survey of H1N1 sequences in the Influenza Research Database (Squires et al., 2012), Y276F was observed in six of 20,433 full-length NA sequences obtained from human isolates since 1933. The dearth of Y276F in the database indicates that it has not attained high frequency, by selection or otherwise, during this period in circulating viral populations. In addition, we do not know of any observations of Y276F in laboratory selection experiments, suggesting that it has not arisen to high frequency in experimental evolution studies. Multiple factors could underlie the low frequency of Y276F, despite its apparent fitness benefit in A/WSN/33 in laboratory experiments. These include discrepancies between selection pressures in vitro versus in vivo, negative epistasis with other sites in proximity (as suggested by our results), or interference with other linked sites experiencing purifying selection (Hill et al., 1966; Pénisson et al., 2017). Importantly, as our biochemical analyses demonstrate that Y276F rescues H275Y (i.e., rendering it neutral rather than deleterious) in both A/WSN/33 and A/California/ 04/2009 in the absence of oseltamivir, the double mutant would be expected to be segregating only at low frequency under
mutation-drift equilibrium in the absence of treatment, consistent with the database observation. Furthering this logic, as H275Y is strongly beneficial in the presence of oseltamivir regardless of the presence or absence of Y276F on the genetic background, there would similarly not be a particular expectation of observing the double mutant at high frequency in populations experiencing treatment. However, if H275Y were to be brought to high frequency by positive selection in response to oseltamivir treatment, and treatment were ceased, it would then experience purifying selection (as the mutant would be less fit than the WT in the absence of drug) and begin rapidly decreasing in frequency. If Y276F were to occur on the H275Y background in this rather specific temporal period, the double mutant would be expected to be governed by genetic drift rather than purifying selection and as such might maintain H275Y at appreciable frequency longer than expected. This scenario would render the population more susceptible to rapid resistance evolution should drug treatment resume. In our experiments, Y276F compensated only for the fitness defect of H275Y, not for oseltamivir resistance mutations at residue 223 or 295. Intriguingly, Y276F did not significantly impact the experimental fitness or NA activity of other resistance mutations. If Y276F and resistant mutations at positions 223 and 295 acted independently, these resistance mutations would be more fit with Y276F; however, observations rather
Figure 3.6 Position of Y276 relative to drug resistance mutations in NA. Residue 223 (magenta), 275 (red), 276 (cyan), and 295 (orange) are highlighted as sticks. The inhibitor oseltamivir is also shown as sticks and is colored yellow. Position 276 is close to 275 but distal from positions 223 and 295. (B) Residues 275 to 277, 303, and oseltamivir are highlighted as sticks and colored based on atom type (C, green; O, red; N, blue). The molecular images of NA in both panels were generated from PDB ID 3CL0.
indicate negative epistasis among these mutations. We investigated the structural basis of the specific compensatory function of Y276F for H275Y. Among the positions with resistant mutations that we studied, residue 275 was in the closest structural proximity to residue 276 (Figure. 36). Inspecting the local conformation of residue 275 to 277 (Figure. 3.6B) reveals that the H275Y substitution sterically forces the glutamate at position 277 closer to oseltamivir, as we previously described (Prachanronarong et al., 2016), which alters the active site and disrupts drug binding. The WT tyrosine at position 276 forms a hydrogen bond with the main-chain carbonyl oxygen at position 303. The Y276F substitution generates space in the structure that may permit the repositioning of both residues 276 and 275. Thus, it appears that the Y276F substitution may locally reorganize the active site. Further experimental work could aid in understanding the physical interactions governing epistasis between Y276F and drug resistance mutations in NA. In conclusion, we identified Y276F as a permissive mutation of H275Y that can restore NA expression level and enzymatic activity. The appearance of Y276F in circulating strains of IAV may increase the expected frequency of H275Y. The likely effect of this would be to move from mutation selection equilibrium to mutation drift equilibrium and thus facilitate oseltamivir resistance.
Materials and Methods

Generation of single and double mutations by site-directed mutagenesis. Plasmids encoding the WT NA gene and the other seven gene segments of the H1N1 A/WSN/33 strain in the pHW2000 vector were kindly provided by R. Webster (St. Jude Children’s Research Hospital, Memphis, TN). Mutations were introduced using site-directed mutagenesis and confirmed by Sanger sequencing. Plasmids for expressing NA in 293T cells were based on pJB992, kindly provided by J. Bloom (University of Washington, Seattle, WA). In this plasmid, a cytomegalovirus (CMV) promoter drives the expression of NA with a C-terminal V5 epitope, used for surface staining, followed by an internal ribosome entry site green fluorescent protein (IRES-GFP), used for calculating transfection efficiency (Hooper & Bloom, 2013). The NA genes were cloned between EcoRI and NotI restriction sites, and mutations were introduced by site-directed mutagenesis.

Cell culture. 293T and MDCK cell lines were obtained from the American Type Culture Collection (Manassas, VA). The 293T cell line was maintained in Opti-MEM I reduced-serum medium (Gibco, Grand Island, NY) supplemented with 5% fetal bovine serum (HyClone, Logan, UT), 100 U/mL penicillin, and 100 mg/mL of streptomycin at 37°C with 5% carbon dioxide. The MDCK cell line was maintained in minimal essential medium (MEM) supplemented with 10% fetal bovine serum, 2 mM L-glutamine, 10 mM sodium pyruvate, 1 nonessential amino acid, 100 U/mL penicillin, and 100 mg/mL of streptomycin.
at 37°C with 5% carbon dioxide. All cell culture reagents were from Corning (Manassas, VA) unless otherwise indicated.

**Fitness analyses.** The experimental fitness of virus stocks was assessed based on plaque size as previously described (Jiang et al., 2016). Briefly, serial dilutions were performed on the viral samples, followed by 1 h of binding on confluent MDCK cells. Unbound virus was washed away with phosphate-buffered saline (PBS), and the cells were overlaid with 0.5% agar in Dulbecco’s modified Eagle medium-nutrient mixture F-12 supplemented with penicillin-streptomycin, L-glutamine, bovine serum albumin, HEPES, sodium bicarbonate, and acetylated trypsin. After the agar solidified, the plates were incubated for 48 h. Cells were fixed and stained with anti-H1 primary antibody (MAB8261; MilliporeSigma) and visualized with a horseradish peroxidase-conjugated secondary antibody. All variants were analyzed in parallel to minimize experimental variability. For each variant, a well with clearly separated plaques was chosen for size analysis (Figure 3.1A). The diameter of all individual plaques in the chosen well were measured (average of 24 plaques per well measured; range, 8 to 38). For fitness estimates, the diameter of each variant was normalized to the diameter of the parental WT virus. Full experimental replicates were performed in triplicate for each variant and WT. To estimate independent fitness effects for H275Y/Y276F, we multiplied the normalized plaque size of H275Y and Y276F and propagated errors. To determine experimental viral growth, MDCK or A549 cells were seeded in 6-well plates and infected with virus variants at a multiplicity of infection (MOI) of 0.01 for MDCK cells or 0.1 for A549 cells.
After inoculation, cells were washed and replenished with growth medium. Supernatant samples were collected at 1, 24, and 48 h postinfection and stored at 280°C. After thawing of supernatant samples, viral RNA was isolated in TRIzol LS following the manufacturer’s instructions (Thermo Fisher). Viral RNA was quantified by RT-qPCR. TaqMan primers and probe sequences targeting IAV-M were as follows: FLUAM-7-F, CTTCTAACC GAGGTCGAAACGTA; FLUAM-161-R, GGTGACAGGATTGGTCTTGTCTTTA; and FLUAM-49-P6, TCAGGCCCCC TCAAAGCCGAG. Quantification of IAV RNA was performed with the QuantiFast pathogen RT-PCR kit (Qiagen). P values were determined for the data at 24 h postinfection (Figure. 3.2A and C) using ordinary one-way analysis of variance (ANOVA) in GraphPad Prism 9.

**Generation of recombinant influenza viruses.** Viral mutants were recovered from plasmids as previously described (Jiang et al., 2016). Briefly, equal numbers of 293T and MDCK cells were seeded in 6-well plates and transfected the next day with 1 mg of NA plasmid (WT or mutants) as well as 1 mg of plasmids carrying the other seven gene segments using TransIT-LT1 reagent (Mirus, Madison, WI). At 6 h posttransfection, cell growth medium was replaced with fresh Opti-MEM I reduced-serum medium. At 30 h posttransfection, tosylsulfonyl phenylalanyl chloromethyl ketone (TPCK)-trypsin (Sigma-Aldrich, St. Louis, MO) was added to cell growth medium to a final concentration of 0.5 mg/mL. At 72 h posttransfection, supernatant containing viral particles was harvested and centrifuged at 1,200 rpm for 15 min. Supernatants were stored at 280°C. Plaque assays were performed to determine the titer (PFU/mL) of each sample as previously
described (Renzette et al., 2013). The radius of all well-separated plaques in each well was measured using a Nikon SMZ1500 microscope, and the mean of all measurements was calculated as an estimate of relative fitness.

**NA enzymatic activity in vitro.** Enzyme activity of NA was determined using fluorogenic 29-(4-methylumbelliferyl)-a-D-N-acetylneuraminic acid (MUNANA; Life Technologies, Carlsbad, CA) according to a previously published protocol (Jiang et al., 2016). The relative fluorescence was normalized to the titer of the virus stock in order to estimate fluorescence activity per infectious virion. The data were then normalized to the WT. To estimate sensitivity to oseltamivir, MUNANA assays were performed in the presence of a range of oseltamivir concentrations to generate a dose-response curve following a previously published protocol (Jiang et al., 2016). For each viral variant, the fluorescence signals in increasing concentrations of oseltamivir were normalized to the signal without drug. IC50s were estimated by fitting normalized signal to a standard binding equation.

**Surface enzymatic activity and expression of NA on 293T cells.** Surface enzymatic activity and expression of WT, H275Y, Y276F, and H275Y/Y276F NA were estimated by transiently expressing A/WSN/33 and A/California/04/09 NA on the surface of 293T cells according to a previously published protocol with modifications (Bloom et al., 2011; Butler et al., 2014). Briefly, 1 mg plasmid harbouring WT or mutant NA was transfected into 293T cells using TransIT-LT1. 293T cells were harvested 22 to 24 h posttransfection and resuspended in nonlysis buffer. Cells (2%) were subjected to MUNANA
assays as described for recombinant viruses to estimate enzyme activity. The rest of the cells were used for flow cytometry to estimate the surface expression of NA after staining with anti-V5 antibody conjugated to allophycocyanin (APC; Abcam, Cambridge, MA). Transfection efficiency was estimated from flow cytometry as the fraction of GFP-positive cells. The enzyme activity was normalized to transfection efficiency. The surface expression of NA was estimated as the mean fluorescence signal of APC. Relative surface expression and enzyme activity were determined by normalization to WT.

**Acknowledgment**

This work was supported by the Office of the Assistant Secretary of Defense for Health Affairs, through the Peer Reviewed Medical Research Program under award no. W81XWH-15-1-0317. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the Department of Defense.
Chapter VI – Discussion

Summary:

The overall goal of this research is to understand the role of sequence on viral protein function and evolution. The two viral systems studied in this thesis are HIV-1 and IAV. In Chapter II, we used deep mutational scanning to experimentally determine the impact of all amino acid substitutions to the protease cleavage site on the proteolytic cleavage in order to identify potential of epistasis within the cleavage site residues and explore biophysical parameters that mediate substrate recognition and specificity. In Chapter III, we identify potential epistasis between previously identified drug resistance and permissive mutations in the regions of NA in Influenza A virus. In summary, using high-throughput mutational scanning and biochemical assays we dissect the evolutionary potential of viral proteins.

Biophysical features and sequence contextuality mediate HIV-1 PR cleavage

We selected the HIV-1 protease substrate as our model system for several reasons. HIV-1 protease is a well-studied enzyme; it is able to cleave diverse sequences as compared to other retroviruses, for example, SARS CoV-2 also has a diverse amino acid sequence at the cleavage site, but has strong preference for Glu at P1, while HIV-1 PR does not show any strong preference for any single amino acid. Numerous structural and mechanistic studies on protease have implicated protease specificity to be shape
dependent despite the sequence diversity (Prabu-Jeyabalan et al., 2000, 2002). Aside from the structure-based studies, computational approaches including neural networks and sequence-based machine learning methods have also provided insights on HIV-1 substrate specificity. Basically machine learning algorithms predict whether a peptide is cleavable or non-cleavable based on previous experimental datasets (Kim et al., 2008; Kontijevskis et al., 2007; You Liwen et al., 2005). Despite the high prediction rate for some of the algorithms, the conclusions have been inconsistent as they are based on the previously compiled small datasets. More recently a larger dataset which included 752 experimentally tested peptides, and two biophysical properties, hydrophobicity or polarity and size were used to predict cleavage. This study supports the substrate envelope hypothesis that the protease recognises a 3D conformation rather than a specific sequence. Regardless of their ability to predict cleaved vs non-cleaved substrates, the computational approaches do not sample how well a peptide is cleaved. Specificity studies using short peptides have attempted to understand efficiency of cleavage using mutational analysis approach of individual sites (Bagossi et al., 2005; Beck et al., 2002; Eizert et al., 2008; Tözsér et al., 1991; Tritch et al., 1991). A more recent work by Potempa and coworkers generated a dataset of close to a 100 variants of different cleavage sites, using site directed mutagenesis and reported the relative processing rates for each by the WT PR. Although this study reports substrate specificity for multiple cleavage sites, the number of variants sampled are not comprehensive (Potempa et al., 2018).
Despite the prevalence of computational methods modeling substrate recognition and specificity, the apparent success of the structure-based studies including the substrate envelope hypothesis, and a handful of mutational analysis on the substrate, there is a dearth of saturation mutagenesis modeling substrate recognition and specificity at every position of a cleavage site.

Saturation mutagenesis approach facilitates large-scale prediction of protein function and has been used to predict substrate specificity of different proteases (Chen et al., 2020; Melnikov et al., 2014; Wrenbeck et al., 2017). For example, Schlick and coworkers investigated HIV-1 PR specificity at positions P4-P2 for p6/PR cleavage site. They used saturation mutagenesis at these positions and a bacterial screening system to screen for substrate specificity. Although, their findings were consistent with previous studies and provided a dataset of how well a peptide is cleaved, their screening approach has limitations with respect to assay maximum, or unequal distribution of stop codons and frame-shifts, in addition, the assay does not distinguish between cleavage by HIV-1 PR and other bacterial enzymes (Schlick & Skern, 2008).

In this work, we investigate the substrate specificity at every position of HIV-1 PR cleavage site using EMPIRIC, we also study the underlying sequence dependencies and biophysical features that mediate cleavage. The traditional protease substrate cleavability studies have used (i) shorter peptides spanning P4-P4’ residues, that provide information on only the residues that bind in the active site pocket, and do not necessarily look into
the long range residues that may affect cleavage, or (ii) random mutagenesis testing a few mutations or positional scanning on a few positions that do not broadly sample the complete cleavage site. We performed a comprehensive analysis of all positions of an extended cleavage site spanning positions P6-P6'. We used yeast-display as a method to express the mutant libraries of cleavage site peptide sequences and analyzed cleavage using FACS and deep mutational scanning. The strong correlation of the data using this method with previously reported point mutation peptide assays suggest that this method can be broadly applied to other viral protease-substrates systems to understand substrate specificity and effect of potential of epistasis on the protease function.

Using this method, we have determined the fitness landscape of three different Gag cleavage sites, MA/CA, NC/p1 and p1/p6. In this manner, we generate a comprehensive map of all possible substitutions in each position and compare the impacts of mutations between the three different cleavage sites. The fitness maps reveal that all the three cleavage sites show potential for improvement. The overall preference is for aromatic, hydrophobic and negatively charged amino acid substitutions; however, the positions at which these amino acids are preferred differ between each cleavage site. At positions P1 and P1', only WT or aromatic amino acids are functional, the flanking residues are more tolerant of amino acid substitutions, even showing distinct biophysical properties. We observed that the preference for an amino acid at any position could not be directly associated with any single physical property, rather we observed very rarely any correlation with any distinct
Physical property and cutting function. The physical properties that did show highest correlation were dependent on the position and the cleavage site and are at positions closer to the scissile bond, suggesting that these preferences are sequence interdependent. Additionally, comprehensive and biochemical analysis reveal that residues distal to the scissile bond, P5, P5', P6 and P6' show varied function when analyzed in the context of full-length Gag substrate.

Overall, two conclusions were drawn from this work. First, the biophysical properties have large impacts on positions close to the scissile bond and the interplay of long-range and short-range residues of cleavage sites is prevalent for cutting function. Our conclusion of long-range epistasis further provides evidence that experimental setup of peptide vs natural substrate can have large impacts on cutting function.

Permissive mutation hold potential for evolution of oseltamivir resistance

In Chapter III we identify potential epistatic interactions between previously identified drug resistance mutations in the proximal regions of NA in IAV. We identify positive epistasis between previously known drug resistance mutation and a hyperactive mutation that has not previously been associated with oseltamivir resistance. Previously, Li et al examined the impact of all possible mutations in regions proximal to the NA active site in the presence and absence of oseltamivir. These experiments were performed in
influenza A/WSN/33 a lab adapted H1N1 strain. The fitness landscape identified drug adaptive mutations; one of them H275Y, is a major oseltamivir resistance mutation, and another, Y276F, is a compensatory mutation that has not yet been associated with oseltamivir resistance. We investigated the epistatic effect amongst Y276F and the five drug adaptive mutations, including H275Y. From experiments based on viral fitness gauging plaque sizes and replication kinetics, we determined that Y276F shows positive epistasis with H275Y and negative epistasis with others, in two different host cell lines. We further determined that Y276F improved surface expression and NA enzyme activity impaired by the H275Y mutant, not only in the A/WSN/33 strain but also in a more modern and clinically relevant IAV strain, A/California/04/2009. Y276F has not yet attained high frequency in the circulating viruses, however, it may aid in stabilizing H275Y in the circulating viral population thus facilitating oseltamivir resistance. Overall, our conclusion from these studies is that studying epistasis is useful in understanding the evolutionary potential of viruses in the presence of continued therapies and that compensatory mutations contribute towards drug resistance.
Advantages and shortcomings of using EMPIRIC combined with yeast display approach to probe viral protein function

EMPIRIC has been extensively applied in understanding protein evolution (Bank et al., 2015). One of the major advantages of EMPIRIC compared to other random mutagenesis methods is relatively simple experimental setup, in addition, commercial sources of gene synthesis can also be used in making site saturation. EMPIRIC provides a comprehensive and highly reproducible approach that screens for all possible substitution in any given region, including in viral proteins, where the reproducibility of the viral expansion has been higher than other high throughput approaches, $R^2 > 0.9$ for IAV ((Jiang et al., 2016). EMPIRIC focuses on the impact of individual amino-acid changes in a protein, and currently is limited by its ability to examine double or higher-order mutations. Evidence suggests that multiple mutations can have large impacts on enzyme activity and drug resistance. (Zhang et al. 2012). EMPIRIC could be further optimized to examine higher-order mutations, but considerable experimental efforts would be required to maintain the reproducibility and throughput of the method.

We have used a yeast-display based next-generation sequencing approach to determine the fitness landscape of the cleavage sites in high throughput. Next-generation sequencing yields hundreds of thousands of reads (Fowler et al., 2010; Hietpas et al., 2012; Melamed et al., 2013) and by pairing it with nucleic acid barcodes to track cleavage amongst different mutant variants we have found a very small amount of noise in our estimation
of variant frequency. Numerous approaches of selections have been utilized for quantifying fitness landscapes, for example, in vitro affinity, experiments on growth effects in yeast, bacteria, viruses and mammalian cells; for our assay, we have harnessed yeast-display based FACS method as a mutational scanning readout. A very similar approach was used in quantification of protein stability (Rocklin et al., 2017). In addition to providing important insights on protein engineering, epistasis, evolutionary potential and the development of drug resistance in viruses and mammalian diseases such as cancer, mutational scanning paired with the next-generation sequencing technology also has wide applicability in clinical settings, metagenomic studies and infectious disease diagnosis and monitoring. As seen during the course of the SARS-CoV-2 pandemic, deep mutational scanning and NGS methods are proving to be pivotal in viral surveillance and management.

**Future directions**

In this dissertation, I used the EMPIRIC approach to investigate biophysical features and sequence dependencies that mediate cleavage by HIV-1 protease. Two inferences can be drawn from studying the fitness landscape of the three cleavage sites, first, that there is room for improvement in the tested cleavage sites, and there is a complex relationship between the biophysical properties, sequence context and cleavage. Further crystallographic studies investigating structural constraints of cleavage site mutations in combination with WT PR may explain the observed position specific preferences of amino
acids. And second, positions further away from the scissile bond, that is, P5 P6 and P5', P6' show evidence that selection at distal positions also affect cleavage. Our gel-based assay showed WT-like or slightly decreased cutting of a few variants that showed improved cutting in the yeast based assay. We suspect that this effect may be specific to the assay settings, different pH conditions and a full-length Gag vs 12-amino acid peptide substrate. Previous publications lack consistency in the proteolysis assay conditions, especially pH of the reaction, as peptide substrates are often used. Optimizing and using similar pH conditions in the gel-based assay will help compare the data from both the assays as pH variations can affect protease cleavage. In addition, the yeast Aga protein may provide an artificial context to the peptide cleavage, although we assured that long range sequence context would not alter cleavage by placing glycine residues on the either side of the cleavage site, yet it may have steric or organizational impact which was not tested here. Using a full length MA-CA polyprotein in the yeast assay and using a 12-amino acid peptide in the gel-assay may prove to be an important control to dissect the possibility of the long range contextuality. We also identified several general patterns common between the cleavage sites, for example, highly tolerant P4', Glu and Asp preferred at distal positions, gel assay to test those patterns would help confirm those results.

In parallel with this research, I am currently investigating the effect of sequence changes on the activity of drug resistance protease variants. The goal here is to investigate potential co-evolution in cleavage sites and drug resistant (DR) PRs, as discussed further in the Appendix I section below. In
addition to this, analyzing the rest of the cleavage sites from Gag and Gag-pol will be interesting to explore as it may provide insight on substrate specificity if a certain pattern can be discerned. For example, the most interesting cleavage sites to investigate would be the autoproteolytic site and CA/p2 because there is a preference for freeing the N-terminus of PR in terms of impact on PR activity compared to freeing the C-terminus (Louis et al., 1999). Determining the impact of all mutations in TF/PR cleavage site on PR activity may provide information about the relationship between autoproteolytic efficiency and cutting at other sites as well as may further explain the observed preference for N-terminus freeing, eventually shedding light on overall substrate processing. Recent development of the first generation maturation inhibitor Bevirimat, which specifically targets CA/p2 cleavage site suggests that Gag is the target of inhibition rather than the PR (Salzwedel et al., 2007). Bevirimat efficacy in clinical trials was impeded by the resistance viruses containing preexisting polymorphisms in p2 (Adamson et al., 2009). From a basic biological standpoint, fitness landscapes of cleavage sites, and studies on natural polymorphism, and on in-vitro and clinically observed PI resistant variants will provide dataset for predicting drug resistance and viral evolution.

Our yeast-display based assay can also be adapted to analyze the functional constraints and drug resistance potential of other viruses, such as Sars-CoV-2. For example, EMPIRIC studies on Sars-CoV-2 Mpro revealed residues in Mpro that are highly mutation tolerant, avoiding targeting these residues during inhibitor design would minimize the risk of potential resistance
development. In conjunction with this, a fitness landscape of Mpro cleavage sites can provide guidelines for co-evolution of protease-substrate mutations and interaction of residues. The hot spots of resistance mutations identified can then be used to perform additional structural studies to understand molecular and biochemical mechanisms.

From our fitness landscape, we have observed multiple residues that improve protease cutting; one of the major challenges here is that we do not fully understand how this improved cutting function relates to viral fitness. For example, mutations that improve cutting function can cause changes in the order of cleavage of the substrate, and likely negatively affect viral fitness. As there is a non-linear function-fitness relationship (Heinrich and Rapoport 1974, Jiang, Mishra et al. 2013) the effects of improved cutting on viral fitness would need to be determined by performing viral replication experiments to test for viral fitness and infectivity.

Broad impact:

A classical question in biology is how does sequence determine the physical property and function of any protein? The overarching goal of this dissertation is to understand how changes in amino acid sequence changes protein function or specificity. I utilize a high throughput mutational scanning approach to map the mutational dependencies and identify the role of large-effect or small-effect mutations on retroviral enzyme activity.
Epistasis: Interaction of mutations

Complex biophysical interactions among residues affect protein properties like structure, substrate specificity, function and numerous other characteristics and are prevalent in evolution. Viral proteins are subjected to a higher mutational rate than other cellular proteins; consequently, these viruses circulate with proteins harboring adaptive mutations. Understanding the collective constraints of biophysical properties and epistasis governing protein function is critical to understanding viral evolution and drug resistance.

Permissive mutation

Permissive mutations can cause large functional impacts when combined with other mutations. These can allow a protein to circumvent the deleterious effects of mutations that cause fitness defects in isolation. The long-term evolutionary potential of mutations that cause fitness defects is highly dependent on permissive mutations that offset the deleterious effects of these primary mutations. Investigating fitness landscapes to identify mutational hotspots, followed by experiments generating double or higher order mutations around these hotspots would provide mechanistic insights on epistasis and predicting evolution of drug resistance in viruses.
Appendix: Investigating co-evolution of HIV-1 cleavage sites and drug resistance proteases.

The following work is an extension of the data presented in this dissertation. While this is a relevant analysis, it is an independent experiment and it will not be submitted for publication in its current state. This research requires further in depth analysis and additional structural and functional studies to elucidate the effects of cleavage site mutations on drug resistance PR variants.

Introduction

The mutations in the cleavage sites of Gag have been associated with many protease drug resistance mutations under the inhibitor therapy (Bally et al., 2000; Koch et al., 2001; Zhang et al., 1997). Mutations in the cleavage sites improve viral replication capacity and substrate processing (Dam et al., 2009; Nijhuis et al., 2007). Many studies investigate the effect of cleavage site mutations in the presence of an inhibitor, or using protease containing one or two drug resistance mutations. What effect a multidrug resistance protease has on the substrate processing remains understudied. Here we aim to analyze the impact of all possible mutations in three cleavage sites of Gag polyprotein on cleavage by DR PR variants.

Out of the nine FDA approved protease inhibitors: Saquinavir (SQV), Indinavir (IDV), Ritonavir (RTV), Nelfinavir (NFV), Amprenavir (APV), Lopinavir (LPV), Atazanavir (ATV), Tipranavir (TPV), and Darunavir (DRV), Darunavir, approved in 2006, is the most potent of the inhibitors. These share similar
structural and binding patterns and despite their potency; we continue to see highly mutated PRs in the clinical isolates.

The substrate envelope hypothesis that forms the basis of substrate recognition is based on the volume and shape of the substrate. Often, drug resistance mutations arise at the site where the inhibitor protrudes the substrate envelope (King, et al., 2004). Efforts have been made to synthesize PIs using the substrate envelope hypothesis as a constraint, and show lesser drug resistance profile. (Ali, et al., 2006; Ali, et al., 2010; Nalam, et al., 2010; Nalam and Schiffer, 2008).

Some of these inhibitors developed by Nalam and coworkers are highly potent DRV analogues (Nalam et al., 2013). The inhibitors, named UMass1-10, have been studied using structural, in vitro pharmacokinetics studies, and in vivo passaging experiments to select for DR PRs against them. However, the effect of resistant PRs selected against these inhibitors on mutant substrate processing remains to be investigated. How a mutant PR processes a mutant cleavage site can be assessed using the yeast display method to detect cutting and EMPIRIC to create mutant cleavage site library.

Here, we performed a comprehensive mutational analysis of three Gag cleavage sites, MA/CA, NC/p1 and p1/p6, using PRs selected against UMass2 and DRV inhibitors, namely UMass2 PR and DRV PR (Nalam et al., 2013). UMass2 PR and DRV PR were selected because they share catalytic efficiency, Kcat/km similar to that of WT PR; these proteases harbor nine and ten mutations respectively, shown in Table 3.1. As a control, we also used PR
containing a single mutation; I84V, I84V is a major drug resistance mutation and shows resistance to all the nine FDA approved inhibitors (Wensing et al. 2014).

**Table 5.1.** Panel of drug resistant mutant proteases

<table>
<thead>
<tr>
<th>Protease variant</th>
<th>No. of Mutations</th>
<th>Mutations</th>
<th>kcat/Km (µM-1s-1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>WT PR</td>
<td>None</td>
<td>I84V</td>
<td>17.1</td>
</tr>
<tr>
<td>I84V PR</td>
<td>1Mut</td>
<td>I84V</td>
<td>9.8</td>
</tr>
</tbody>
</table>

The work described here focuses on mutant landscapes without inhibitor selection because we aim to investigate cutting efficiency of DR PRs and the data here provides preliminary information regarding cutting efficiency, as well as interdependencies in the cleavage sites when treated with DR PRs.
Results and Discussion

The cleavage sites evolve mutations in response to drug resistance observed in the PR, (Dam et al., 2009; Doyon et al., 1996; Mammano et al., 1998). To investigate whether mutational impacts on cleavage sites are dependent on specific mutations in PR, we performed EMPIRIC on three cleavage sites, MA/CA, NC/p1 and p1/p6 and analyzed the impact of mutations on DR PR cleavage.

FACS analysis with high resolution with increased cutting

In Chapter II, we performed similar studies to analyze the impact on mutations in cleavage sites on WT PR. While using WT PR, we optimized the proteolytic cleavage assay conditions to get partial cleavage of WT cleavage site sequences in order to obtain good resolution on library variants showing increased and decreased cutting efficiency. The time required for partial cleavage in our yeast assay was one hour for MA/CA, 75 minutes for NC/p1, 40 minutes for p1/p6.

In this study, in an attempt to aim for experimental consistency across the board with WT PR and DR PRs we did not optimize the PR assay for partial cleavage (approximately 50% cleavage) of the WT cleavage site. Instead we performed the assays using the same parameters as for the WT PR, including, pH, temperature, concentration of DR PRs and proteolysis time (one hour for MA/CA, 75 minutes for NC/p1, 40 minutes for p1/p6), as discussed in Chapter II Methods. We anticipate that to aim for approximately
50% cleavage of WT sequence of each cleavage site with each DR PRs, we would need to optimize conditions further, especially the incubation time of the proteolytic cleavage assay.

The resulting flow cytometry analysis of WT sequences of all the three cleavage sites, using I84V PR, UMass2 PR and DRV PR resulted in differential cleavage. For MA/CA WT sequence we were able to achieve some level of partial cleavage with drug resistant proteases. However, WT sequences of NC/p1 and p1/p6 exhibited extremely poor to no cleavage with the drug resistant proteases, suggesting that mutations in the PR have different effects on different cleavage sites (Figure 5.1.)
Figure 5.1

Figure 5.1. Resolution on faster cutting mutations depend on how much WT is cut

Flow cytometry analysis of MA/CA and NC/p1 cleavage sites with WT PR, UMass2 PR and DRV PR with an overlay of no PR control cells in red and PR treated cells in blue. The four windows, W1-W4 represent the sorting windows or gates, numbers indicate percentage of cells in each window.
The data discussed below is mainly for the NC/p1 cleavage site.

**Distribution of functional effects (DFE)**

To analyze the general impact of mutations in a cleavage site with each DR PR, we analyzed the distribution of fitness effects (DFE) of each cleavage site library. In contrast to DFE with WT PR where we observe a broad distribution of variants, we observe a narrow distribution of variants with DR PRs (Figure 5.2). We see many mutations causing large decrease and WT-like function and very few mutations causing large increase in function relative to the WT. The assay conditions setup with DR PRs cause a very poor cutting of the WT cleavage site sequence, resulting in the narrow distribution. For this reason we were not able to directly compare the functional impacts of variants between the WT and DR PRs. However, the assay conditions provide an opportunity for a high resolution on beneficial mutations with increased cutting efficiency. We term the variants causing very large impacts in cutting function as the hyperactive mutations.
Figure 5.2

DFE of NC/p1 variants across drug resistance proteases. The three DR PR variants show narrow distribution of functional scores with a maximum score for beneficial mutants of up to 30 for DRV PR and 25 and 20 for I84V PR and UMass2 PR. This again supports the high resolution on beneficial mutations.
Analysis of hyperactive mutations

In order to study the effects of hyperactive mutations, we analyzed variants causing more than 10-fold cutting relative to the WT from all the three DR PRs. We found a total of 39 hyperactive mutations and 15 (38%) of those were common to all three DR PRs. Representation of common mutation is less likely to be random, due to the significantly greater number of variants (38%) being hyperactive than by random sampling. Since these mutations had the largest effects on activity, we further investigated their location with a goal of understanding whether the positions are proximal or distal to the scissile bond. Most of the hyperactive mutations were found at P4’ position. The WT residue at P4’ is Lysine and almost any substitution at this position is beneficial for cleavage, similar to what we observe with WT PR (Figure 2.2A). In future it will be interesting to analyze structural and biochemical determinants of tolerance at the P4’ position.
Figure 5.3. Identifying variants that were hyperactive in all the three DR PRs. Venn diagram showing common and unique mutations in DR PRs in NC/p1 cleavage site.
Table 5.2. Common and unique hyperactive mutations

<table>
<thead>
<tr>
<th>Common mutations</th>
<th>Unique mutations</th>
<th>I84V PR</th>
<th>UMass2 PR</th>
</tr>
</thead>
<tbody>
<tr>
<td>NP1M</td>
<td>RP4I</td>
<td></td>
<td>GP3'L</td>
</tr>
<tr>
<td>NP1L</td>
<td>LP2'V</td>
<td></td>
<td>QP3L</td>
</tr>
<tr>
<td>NP1Y</td>
<td>KP4'E</td>
<td></td>
<td>QP3W</td>
</tr>
<tr>
<td>NP1F</td>
<td>KP4'D</td>
<td></td>
<td>KP4'R</td>
</tr>
<tr>
<td>LP2'E</td>
<td>KP4'S</td>
<td></td>
<td>RP4H</td>
</tr>
<tr>
<td>GP3'E</td>
<td>KP4'Q</td>
<td></td>
<td>QP3Y</td>
</tr>
<tr>
<td>KP4'M</td>
<td></td>
<td></td>
<td>QP3F</td>
</tr>
<tr>
<td>KP4'I</td>
<td></td>
<td></td>
<td>GP3'Q</td>
</tr>
<tr>
<td>KP4'L</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KP4'A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KP4'V</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KP4'Y</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KP4'W</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KP4'F</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IP5'R</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Next, we identified the unique hyperactive mutations for all the three DR PRs (Table 5.2.), surprisingly we noticed no unique hyperactive mutation with DRV PR, and most of the mutations with other DR PRs were located at P3, P3’ or P4, P4’ positions. As previous studies show that the PR recognizes conserved shape (Ozen et al., 2011; Prabu-Jeyabalan et al., 2000), we speculate that the hyperactive mutations likely make the cleavage site to be in extended conformation and fits better within the substrate envelope. Further studies using structural simulation will help elucidate the mechanism of observed functional effects.
Figure 5.4

Figure 5.4 Correlation of RF scores in different cleavage sites across P6-P6'.

Correlation between relative function scores of all NCp1 variants for each DR PR.

Black line indicates linear fit.
Correlation of relative functional scores reveal that moderate sequence dependencies mediate cleavage

To further investigate the effects of all mutations in NC/p1 cleavage sites on cutting efficiency of all three DR PRs, we compared the RF scores of all mutations across each PR (Figure 5.4). We observed moderate correlation of the RF scores between the three DR PRs, with correlation coefficients being moderately-high, ranging between 0.52 and 0.64. Here, the correlation coefficients provide an estimate of interdependencies of mutations between cleavage site sequence and PR sequence. Where, higher correlation coefficients predict higher independence, for example, the amino acids exhibit similar impact on cutting function when cleaved by any PR variant, suggesting selection of mutation is independent of sequence, while lower correlation coefficients indicate that selection of mutations is interdependent on the sequence of PR and the cleavage site. Here, we speculate that NC/p1 shows moderate interdependency with cleavage by DR PR variants. However, inferences of contextuality will be more conclusive upon further studies.

Future directions

In summary, we identified a handful of mutations that exhibit large functional impacts on cleavage by DR PR. These mutations are common to all DR PRs studied here, and show activity of more than 10 fold relative to the WT. We further aim to generate a fitness landscape of all three cleavage sites with WT
PR using conditions that will aim for poor cutting of the WT cleavage sites, in order to directly compare our datasets. Further structural and biophysical analysis may reveal potential co-evolution between cleavage sites and DR PRs.

**Material and methods**

The cleavage site library construction, yeast surface display, mutant protease treatment, labeling and FACS, sequencing analysis, estimation of functional scores of cleavage site variants and analysis of cut-site variation in sequence is as described in the material and methods section of Chapter II.

**Plasmid library preparation:**

Systematic libraries using NNK codon were generated in pCTCON2 yeast display plasmid for a 12-amino acid window of MA/CA, NC/p1 and p1/p6 cleavage sites. The library sequence is a fusion protein expressed on the yeast cell surface. The yeast surface protein Aga2 has HA-tag on its C-terminus followed by three glycines and a unique Pstl RE site, followed by the 12-amino acid library sequence(P6-P6’), a unique Nhel RE site, three glycines and a Myc-tag. Resulting plasmid contains, Aga2-HA-GGGPstl-CS library-NhelGGG-Myc.

We used EBY100 yeast surface display strain and introduced plasmid libraries for each cleavage site. After transformation, cells were grown in
CAAD media at 30°C until saturation and frozen with glycerol to 25% concentration at -80°C until needed for FACS assay.

**FACS Assay:**

Library of each cleavage site was thawed quickly for 5 minutes at 37°C, and grown in CAAD media at 30°C for 24 hours or until saturation OD600 of 1. Cells were then diluted into 50 mL of fresh CAAD media to an OD600 of about 0.1 and grown at 30°C was about 6-8h, until they reached an OD600 of 0.2. Cells were then induced in CAA-RG media (CAA media with 1% raffinose and 1% galactose) at a dilution of OD600 0.5, and shaken at 30°C until they reached an OD600 of 0.2 for about 16 hours.

**Protease assay**

10⁶ cells were collected by centrifugation and washed with PAB, purified DR PRs; 1 uM I84V PR, 1uM UMass2 PR, and 1uM DRV PR were added separately to each cleavage site library displaying yeast cells. Purified HIV-1 PRs were a kind gift from the Schiffer lab at UMass Chan Medical School. Samples were incubated in a shaking incubator at 30 °C for one hour for MA/CA, 30 °C for 75 minutes for NC/p1, and 30 °C for 40 minutes for p1/p6 libraries, and the reaction was stopped by washing the cells three times in TBSB. Cells were then labeled with antibodies to the HA-tag (Alexa488 conjugated, Cell Signaling Technology #2350) and the myc-tag (Alexa647 conjugated, Cell Signaling Technology #2233) at 1:100 dilutions of each
antibody and were incubated for 1 hour incubation at 23 °C in TBSB. Labeled cells were washed and suspended in 5ml TBSB at a density of 106 cells/mL for FACS. At least 150,000 cells were sorted for each window. Sorted cells from each window were grown separately in 50 mL of CAA-D media on a shaking incubator at 30 °C for 24-36 hours. The cultures from sorted cells were collected by centrifugation, washed with TBSB and stored as pellets at -80 °C.

Preparation of DNA for NGS and estimation of functional scores from NGS were exactly as described in the Chapter II, Methods.
Appendix II

Protocol: Yeast-based Flow cytometry assay

Experimental Design
The yeast display provides an in-vitro assessment of enzymatic cleavage. Even though the in-vitro assay conditions are not perfect mimics of physiological conditions, they provide a good estimate of cleavage efficiency.

Protease enzymatic cleavage conditions
pH conditions - We performed a pH titration using protease assay buffer (PAB) at pH 5.5, 6.0, 6.5, 7.0 to be able to aim for proper proteolytic cleavage, antibody binding and to maintain optimum conditions for yeast. Proteolysis at lower or higher pH did not affect the cleavage of WT substrate, although it may interfere with antibody binding and proper double positive yeast display (Data not shown).

Antibody labeling - We tested a few antibodies to obtain proper double positive expression in the flow cytometry analysis, we found that the conjugated antibodies against HA-tag (Alexa488 conjugated, Cell Signaling Technology #2350) and the myc-tag (Alexa647 conjugated, Cell Signaling Technology #2233) worked the best for our assay. However, appropriate fluorophores should be chosen for different experimental setup and according to the choice of flow cytometer
Yeast growth media -

In our experience, the cas amino acids media provided optimum growth and surface display to the yeast strain as compared to the synthetic dextrose media.

CAAD - Cas amino acid dextrose  1000 mL

- 20g Dextrose (D-glucose)
- 6.7g Yeast Nitrogen base
- 5.0g CAA
- 5.40g Na2HPO4 anhydrous
- 8.56g NaH2PO4.H2O
- Add deionized H2O to a final volume of 1000ml
- Filter sterilize

CAA -R/G - Cas amino acid galactose/raffinose (1% Gal and 1% Raf)

* Only autoclave the buffer and H2O

This medium must be stored at 4°C until needed, and can be stored at RT overnight before use.

* CAA media occasionally will have undissolved salt flakes.
PROCEDURE 1

Thaw frozen aliquots of cleavage site yeast library at 37 °C for 5 mins and grow overnight (about 24 h) in 50 ml of CAAD media in an incubator shaker set to 30 °C and 180 r.p.m, up to saturation, (OD600 = 1, about 10⁷ cells per ml).

Inoculate fresh 50 ml CAAD media with OD600 = 0.02 cells for 5-7 hours, until cells reach an OD600 = 0.1-0.2.

Pellet about OD600 = 0.02 cells from the library culture at 2,500g for 5 min and wash with TBSB. Resuspend the pellet in 50 ml CAA-R/G media for 16-18 hours until reach OD600 = 0.1-0.2. Grow OD600 = 0.02 cells from the library culture in CAAD for 10-12 hours as non-display yeast control.

Note: Cells in CAAD grow faster than in CAA-R/G

For each cleavage site library, collect at least 1.5 x 10⁶ display cells in a microfuge tube, wash twice with ice cold TBSB, and resuspended in 100 μL of TBSB. Add 1uM of purified protease and incubate for 1 hour for MA/CA, 75 minutes for NC/p1 and 40 minutes for p1/p6 in an incubator shaker at 30 °C.

Note: Tape the tubes horizontally to the shaker (on a plastic lid or similar flat surface) for consistent mixing of cells and protease.

Proteases used in the experiment:

The concentration of PRs used is 1uM and the volume is calculated catalytic efficiency is considered proportional to the WT PR efficiency Table 5.3.
Table 5.3. Mutant proteases catalytic efficiency

<table>
<thead>
<tr>
<th>Protease</th>
<th>kcat/Km (µM-1s-1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>WT</td>
<td>17.1</td>
</tr>
<tr>
<td>I84V PR</td>
<td>9.8</td>
</tr>
<tr>
<td>UMASS 2 PR</td>
<td>10.7</td>
</tr>
<tr>
<td>DRV PR</td>
<td>9.1</td>
</tr>
</tbody>
</table>

Wash the cells twice with 1 mL of ice cold TBSB. Resuspend the pellet in 100µl TBSB and label yeast cells with HA-tag (Alexa488 conjugated, Cell Signaling Technology #2350) and the myc-tag (Alexa647 conjugated, Cell Signaling Technology #2233) at a dilution of 1:100. Incubate at RT on a tube rotator for 1 hour and shield from light. Wash the labeled cells three times with 1ml of ice cold TBSB and resuspend in at least 500ml of TBSB before flow cytometry. The concentration of cells in the suspension should not exceed $10^8$ cells per ml, as higher concentration of suspension can lead to aggregation and clogging of the flow cytometer apparatus.

Before sorting, set up a flow cytometry analysis using control yeast samples. Negative control samples with non-display cells, single fluorophores, and double positives as shown in Figure 6.1 (C-F). Gate on forward and side scatter Figure 6.1 (A-B) and properly compensate to reduce crosstalk between the fluorophores.
**Figure 6.1.** Flow cytometry controls

A) Gated population  B) Single cell population

C) Display off HA and Myc  D) HA only  E) Myc only  F) HA and Myc

**Figure 6.1** The yeast cells are first gated based on the forward scatter and side scatter (A) to differentiate the cells from debris, and (B) to exclude doublets from the population. The single cells are then further analyzed using controls for uninduced yeast cells (C) that do not display cleavage site library on the cell surface and are stained with HA and Myc both, for single staining; (D) HA only, (E) Myc only, or for double staining (F) using both HA and Myc.
Sorting windows or gates shown in each plot divide the cells based on distribution of variants. (F) No protease treated control, cells present in the sorting gate 1 shown in red, represent the yeast cells displaying all the library variants, and cells in the sorting gate 4 shown in purple represent the stop codons from the library.

Next, we assessed the codon abundance in the library via FACS and deep sequencing and found that all the codon occurred in an equal abundance in our library Figure 6.2 A.

In order to quantify codon abundance from the library distribution, we gated the library in vertical gates Figure 6.2 B. Each sort window is sorted separately and the cells are collected in a fresh 50 ml CAAD media to grow overnight or until exponential phase, OD600 = 0.1-0.2

Wash cells approximately 20 ml of cells with OD600 = 0.1 with H2O and store the pellet at -80 oC freezer

*Note: Repeat a similar process for each library with protease treatment.*

Thaw frozen pellet and use Zymo plasmid kit as per the instructions to extract plasmid DNA.

Set up a test PCR with all the samples using a 5' P5 primer and any TrueSeq Index primer (barcode) Table 6.1.
Table 6.1. Plasmid PCR amplification parameters

<table>
<thead>
<tr>
<th>Cycle number</th>
<th>Denature</th>
<th>Anneal</th>
<th>Extend</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>95 °C, 2 min</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2-20</td>
<td>95 °C, 30 s</td>
<td>60 °C</td>
<td>72 °C, 1 min per kb</td>
</tr>
</tbody>
</table>

Confirm the PCR product on 1% agarose gel and column purify each sample separately and elute in a 20ul elution buffer.

Prepare an equimolar blend of the samples and column purify again.

Setup a qPCT to quantify the concentration of the pooled library.

Prepare the pooled library for a NextSeq run as per the instructions.
**Figure 6.2** Codon distribution in the library

A) The heatmap representation of MA/CA library based on counts of each codon in the library. B) FACS plot of MA/CA library with no protease treatment indicate that all variants are present at similar abundance in the library.

Figure 6.2. The no protease treated display cells were sorted using vertical sort windows in order to obtain high resolution and capture all possible cells from the library.

Quantification of codon abundance in the library. A) The heatmap representation of MA/CA library based on counts of each codon in the library. B) FACS plot of MA/CA library with no protease treatment indicate that all variants are present at similar abundance in the library.
Bibliography


https://doi.org/10.1038/ng1482


Barré-Sinoussi, F., Chermann, J. C., Rey, F., Nugeyre, M. T., Chamaret, S.,
Gruest, J., Dauguet, C., Axler-Blin, C., Vézinet-Brun, F., Rouzioux, C.,
retrovirus from a patient at risk for acquired immune deficiency syndrome

Identification of Efficiently Cleaved Substrates for HIV-1 Protease Using a

substrate selectivity. *Current Drug Targets. Infectious Disorders, 2*(1),
37–50.

Topics in Microbiology and Immunology* (pp. 177–218).
https://doi.org/10.1007/978-3-642-80145-7_6

proteinase using MA/CA junction peptides substituted at the P3-P1’
positions. In *Archives of Biochemistry and Biophysics* (Vol. 290, Issue 1,

Billich, S., Knoop, M. T., Hansen, J., Strop, P., Sedlacek, J., Mertz, R., &
Moelling, K. (1988). Synthetic peptides as substrates and inhibitors of
human immune deficiency virus-1 protease. *The Journal of Biological
Chemistry, 263*(34), 17905–17908.

Blick, T. J., Sahasrabudhe, A., McDonald, M., Owens, I. J., Morley, P. J.,


Bolon, D. N., & L., S. (2001). Polar Residues in the Protein Core of
Escherichia coli Thioredoxin Are Important for Fold Specificity.

*Biochemistry, 40*(34), 10047–10053.


https://doi.org/10.1002/j.1460-2075.1994.tb06799.x

https://doi.org/10.1002/j.1460-2075.1991.tb04984.x


https://doi.org/10.1002/(sici)1097-0134(199601)24:1<51::aid-prot4>3.0.co;2-r


https://doi.org/10.1128/jvi.80.3.1340-1351.2006


https://doi.org/10.3390/v2071411


Doyon, L., Croteau, G., Thibeault, D., Poulin, F., Pilote, L., & Lamarre, D.


Herlocher, Louise; Truscon, M., Elias, R., Yen, S., Roberts, H.-L.,; Ohmit, N.


https://doi.org/10.1073/pnas.100133697


https://doi.org/10.1073/pnas.080078297


Ives, ; Carr, J. A., ; Mendel, J. A., ; Tai, D. B., ; Lambkin, C. Y., ; Kelly, R., ;


Virology, 88(13), 7145–7154.


https://doi.org/10.1002/prot.21415


https://doi.org/10.1074/jbc.m112.339374


https://doi.org/10.1128/jvi.79.16.10348-10355.2005


https://doi.org/10.3389/fmicb.2019.00039

action and mechanisms of resistance. *Influenza and Other Respiratory Viruses*, 7(NA), 25–36.


O’Neill, R. E. (1998). The influenza virus NEP (NS2 protein) mediates the


https://doi.org/10.1016/b978-0-12-803109-4.00001-5


https://doi.org/10.1128/jvi.78.16.8477-8485.2004


epistasis, weak evidence against it. *Proceedings of the National Academy of Sciences of the United States of America*, 111(15), E1450 – NA.


Biology (Vol. 351, Issue 1, pp. 158–169).
https://doi.org/10.1016/j.jmb.2005.06.003


Sarkisyan, K. S., Bolotin, D. A., Meer, M. V., Usmanova, D. R., Mishin, A. S., Sharonov, G. V., Ivankov, D. N., Bozhanova, N. G., Baranov, M. S.,


https://doi.org/10.1016/s0006-291x(67)80055-x


https://doi.org/10.1006/jmbi.1996.0004


https://doi.org/10.1002/(sici)1099-1352(199812)11:1/6<10::aid-jmr381>3.0.co;2-h


https://doi.org/10.1186/1742-4690-9-29


https://doi.org/10.1038/303035a0


Zennou, Mammano, V., Paulous, F., Mathez, S., Clavel, D., & François. (1998). Loss of Viral Fitness Associated with Multiple Gag and Gag-Pol Processing Defects in Human Immunodeficiency Virus Type 1 Variants Selected for Resistance to Protease Inhibitors In Vivo. *Journal of*
