Mathematical and Computational Models of Cancer and The Immune System

by

Diego Chowell-Puente

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved July 2016 by the
Graduate Supervisory Committee:

Carlos Castillo-Chavez, Co-Chair
Karen S. Anderson, Co-Chair
Joseph N. Blattman
Melissa A. Wilson Sayres
Carlo C. Maley

ARIZONA STATE UNIVERSITY

August 2016

ABSTRACT

The immune system plays a dual role during neoplastic progression. It can suppress tumor growth by eliminating cancer cells, and also promote neoplastic expansion by either selecting for tumor cells that are fitter to survive in an immunocompetent host or by establishing the right conditions within the tumor microenvironment. First, I present a model to study the dynamics of subclonal evolution of cancer. I model selection through time as an epistatic process. That is, the fitness change in a given cell is not simply additive, but depends on previous mutations. Simulation studies indicate that tumors are composed of myriads of small subclones at the time of diagnosis. Because some of these rare subclones harbor pre-existing treatment-resistant mutations, they present a major challenge to precision medicine. Second, I study the question of self and non-self discrimination by the immune system, which is fundamental in the field in cancer immunology. By performing a quantitative analysis of the biochemical properties of thousands of MHC class I peptides, I find that hydrophobicity of T cell receptors contact residues is a hallmark of immunogenic epitopes. Based on these findings, I further develop a computational model to predict immunogenic epitopes which facilitate the development of T cell vaccines against pathogen and tumor antigens. Lastly, I study the effect of early detection in the context of Ebola. I develope a simple mathematical model calibrated to the transmission dynamics of Ebola virus in West Africa. My findings suggest that a strategy that focuses on early diagnosis of high-risk individuals, caregivers, and health-care workers at the pre-symptomatic stage, when combined with public health measures to improve the speed and efficacy of isolation of infectious individuals, can lead to rapid reductions in Ebola transmission.

To my parents Rafael and Esther

ACKNOWLEDGMENTS

I would first like to express enormous gratitude to my advisors, Karen Anderson and Carlos Castillo-Chavez, for all I learned from them, for their support through my PhD studies, and for the many great opportunities they gave me. Carlos, for introducing to the field of mathematical biology and for providing me with the first opportunity to do research at Los Alamos National Laboratory in 2004. This is when I became fascinated with applying mathematics to problems in medicine and biology. Karen, for giving me the opportunity to work in her laboratory at the Biodesign Institute, and for teaching me so much about cancer biology and immunology.

I would like to thank Carlo Maley, Melissa Wilson Sayres, and Joe Blattman for agreeing to serve on my thesis committee, their mentorship and all their support. They also introduced me to many great collaborators. I owe special thanks to Jay Taylor for all the intellectually stimulating conversations, for teaching me so much about statistics, stochastic modeling, and population genetics. I learned so much from him and I admire his dedication to his work. I also thank Professor James Collins who introduced me and taught me a wide range of topics in evolutionary biology. Very special thanks to my brother Gerardo Chowell who has always encouraged me to do my best in my work and has showed me that dreams are possible to achieve. I have become a better scientist and person because of all of them.

I would also like to thank Sri Krishna for being a great collaborator and friend, and for introducing me to the fascinating field of immunology, Shayesteh Ferdosi and Oscar Patterson-Lomba for being very good friends and always being there. I always had great conversations with them about science, politics, movies, coffee, food, and life.

Along the way, I have been fortunate to have the support of many special people; in particular, Robert Walker, Moe Z. Win, Joshua LaBaer, Baltazar Espinoza-Cortes, and Andres Gomez-Lievano. I would also like to thank all the students, postdocs, and faculty at the SAL MCMSC, Anderson Lab, Maley Lab, Aktipis Lab, and Wilson Sayres Lab. My spe-

cial thanks to Margaret Murphy-Tillis, Renate Mittelmann, Sherry Woodley, Dawn Bies, and Sherie Casillas for always being able to find a solution to any problem.

Lastly, I would like thank my family. They all have been great role models throughout my life.

TABLE OF CONTENTS

Page

LIST OF TABLES ix

LIST OF FIGURES x

1 INTRODUCTION 1

CHAPTER

2 MODELING THE SUBCLONAL EVOLUTION OF CANCER CELL POPULATIONS 4

2.1 Introduction 5

2.2 Results 7

2.2.1 Subclonal Evolutionary Model of Cancer Cell Populations 7

2.2.2 Drift Dominates Early Neoplastic Dynamics 9

2.2.3 Extent of Intratumor Subclonal Variation at Detection 11

2.2.4 Differential Fitness Between Dominant and Minor Subclones 15

2.2.5 Resistant Subclones Carrying Driver Mutations Are Present at Low Frequency at Detection 17

2.3 Discussion 19

2.4 Materials and Methods 23

2.4.1 Computer Code Availability 23

2.4.2 High Speed Algorithm Considerations 23

2.4.3 Software Tools 23

2.4.4 Hadoop Cluster Configuration 24

2.4.5 Statistical Analysis 24

3 TCR CONTACT RESIDUE HYDROPHOBICITY IS A HALLMARK OF IMMUNOGENIC CD8$^+$ T CELL EPITOPES 26

3.1 Significance Statement 27

CHAPTER Page

3.2 Introduction 27
3.3 Results 29
3.3.1 Amino Acid Use Differs Between Immunogenic and Non- immunogenic Peptides 29
3.3.2 Hydrophobicity Bias in Selective TCR Contact Residues 30
3.3.3 Differential Hydrophobicity Can Predict Immunogenic CTL Epitopes 34
3.3.4 Hydrophobicity-based ANN Prediction Model 35
3.3.5 Prediction Validation by in Vivo Discovery of HIV-1 Gag Epitopes 37
3.3.6 Prediction of Immunodominant Epitopes 39
3.4 Discussion 39
3.5 Materials and Methods 45
3.5.1 Construction of Datasets 45
3.5.2 Amino Acid Scales 46
3.5.3 Position-based Hydrophobicity Analysis 46
3.5.4 Hydrophobicity-based ANN Prediction Model (ANN-Hydro) ... 46
3.5.5 Application of ANN-Hydro 47
3.5.6 Vaccines 47
3.5.7 Immunization of Mice 47
3.5.8 Peptides 47
3.5.9 T Cell Epitope Mapping 48
4 MODELING THE EFFECT OF EARLY DETECTION OF EBOLA VIRUS DISEASE 49
4.1 Introduction 49

CHAPTER Page

4.2 Model Formulation 52
4.3 Model Analysis 55
4.3.1 Basic properties 55
4.3.2 Equilibrium Analysis 56
4.3.3 Normalized Sensitivity Analysis on $\mathscr{R}_c$ 58
4.3.4 Impact of Early Detection and Isolation on the Value of $\mathscr{R}_c$ 61
4.4 Discussion 63
REFERENCES 65
APPENDIX
A MODELING THE SUBCLONAL EVOLUTION OF CANCER CELL POPULATIONS 76
B TCR CONTACT RESIDUE HYDROPHOBICITY IS A HALLMARK OF IMMUNOGENIC CD8$^+$ T CELL EPITOPES 82
B.1 Construction of Datasets 83
B.2 Amino Acid Frequency Analysis 83
B.3 Position-based Hydrophobicity Analysis 83
B.4 Rate Analysis of Predicted Peptides 84
B.5 Hydrophobicity-based ANN Prediction Model (ANN-Hydro) 84
B.6 Application of ANN-Hydro 84
B.7 Statistical Analysis of Predicted CTL Epitopes 85
B.8 *In vivo* discovery of HIV-1 Gag Epitopes 85

LIST OF TABLES

Table		Page
2.1	Simulations Performed	11
4.1	Model State Variables	53
4.2	Model Parameters	54
4.3	Percent Change	59
A.1	Statistics of the Number of Dominant Clones	77
A.2	Differential Fitness	77
A.3	Average Time From Treatment to Disease Recurrence	78
B.1	Schematic of ConsB and CN54	90
B.2	Summary of Identified Epitopes	91
B.3	Amino Acid Property Scales Used for Analyses	92
B.4	Hydrophobicity Comparison	93
B.5	HLA-A2 Restricted CTL Epitopes for Dengue Virus	94
B.6	Probabilities Assigned by the ANN-Hydro A2-Model	95
B.7	Ranking Comparison of all the Predicted Epitopes	97
B.8	Ranked List of top 20 Predicted Peptides	99

LIST OF FIGURES

Figure		Page
2.1	Schematic Representation	10
2.2	Intratumor Subclonal Variation	13
2.3	Relation Between Dominant Clones and Minor Subclones	14
2.4	Probability Density Function	15
2.5	Relation Between Dominant Clones and Minor Subclones	16
2.6	Resistant Subclones (A)	18
2.7	Resistant Subclones (B)	19
3.1	Probability Ratio	32
3.2	Hydrophobicity Comparison	33
3.3	Efficiency of Predicting Experimentally	35
3.4	ANN-Hydro Model Prediction Validation	38
3.5	Incorporating ANN-Hydro	41
4.1	Compartmental Model Showing the Transition Between Model States.	53
4.2	Impact of Early Detection of Pre-Symptomatic Individuals on the Value of $\mathscr{R}_c$	61
4.3	Impact of Isolating Infectious Individuals on the Value of $\mathscr{R}_c$.	63
B.1	Bias in Amino Acid Usage	87
B.2	Hydrophobicity Comparison	88
B.3	Workflow for CTL Epitope Prediction	89

Chapter 1

INTRODUCTION

This dissertation is aimed at understanding some fundamental processes in the human immune system, cancer, and cancer immunology. Cancer is a subclonal evolutionary process that is governed by the dynamic interplay of mutation, stochastic drift, and selection (Greaves and Maley, 2012; Marusyk et al., 2012; Nowell, 1976; Lipinski et al., 2016; Frank, 2007). In the field of cancer biology, the term driver mutation is used to refer to mutations under positive selection within a population of cells, and the term passenger mutation is used for mutations that are neutral or deleterious (Martincorena and Campbell, 2015; Vogelstein et al., 2013; McFarland et al., 2013), and increase in frequency due to hitchhiking alongside driver mutations. A common model of the evolutionary process explaining tumor growth envisions driver mutations causing clonal expansions that sweep through the tumor cell population and reach fixation (100% frequency) (Fearon et al., 1990; Bozic et al., 2010). If such a driver mutation did reach fixation, it would appear as a trunk mutation, present in all the tumor cells. However, tumor clonal architectures are often observed experimentally to be the consequence of a complex branched subclonal processes across multiple cancer types, with divergent subclones evolving simultaneously (Navin et al., 2010; Campbell et al., 2010; Anderson et al., 2011; Notta et al., 2011; Sottoriva et al., 2013; Campbell et al., 2008; Gerlinger et al., 2012; Nik-Zainal et al., 2012; Shain et al., 2015; Burrell et al., 2013; Burrell and Swanton, 2016; Zhao et al., 2016).

In Chapter 2, I present a stochastic model to study the dynamics of the subclonal evolution of cancer and to assess the extent to which heterogeneous subclonal expansions occur. By employing a supercomputer to simulate the process, we were able to keep track of all subclones that arise and die out, or are maintained and grow during the evolutionary pro-

cess. Under different combinations of the evolutionary parameter values, including those that have been previously estimated for two cancer types (glioblastoma multiforme and colorectal cancer), simulations lead to the conclusion that the distribution of sizes of subclones carrying driver mutations has a heavy right tail at time of tumor detection, with only 1-4 dominant clones present at $\geq 10\%$ frequency, composing most of the tumor cell population. By contrast, our model predicts that the vast majority of subclones will be present at $< 10\%$ frequency. In addition, we find that most of treatment-resistant subclones harboring driver mutations are often present at low frequency, below the detection limit of standard sequencing techniques. Moreover, our results suggest that the number of minor subclones is strongly correlated with the number of numerically dominant clones in a tumor.

The immune system of an organism combats invading pathogens, thereby protecting the host from disease. Jawed vertebrates, such as humans, have an adaptive immune system that enables them to mount pathogen-specific and tumor-specific immune responses. The importance of the adaptive immune response for human health is highlighted by the opportunistic infections that afflict individuals with compromised adaptive immune systems. The flexible adaptive immune system can also go awry, and many diseases (e.g., multiple sclerosis and type I diabetes) are the consequence of the adaptive immune system failing to discriminate between markers of self and non-self. Thus, CD8+ T cells must distinguish immunogenic epitopes from non-immunogenic self-peptides to respond effectively against an antigen without endangering the viability of the host. In Chapter 3, we interrogated the biochemical properties of 9,888 MHC class I peptides. We identified a strong bias towards hydrophobic amino acids (aas) at TCR contact residues within immunogenic epitopes of MHC allomorphs, which permitted us to develop and train a hydrophobicity-based artificial neural network (ANN-Hydro) to predict immunogenic epitopes. The immunogenicity model was validated in a blinded in vivo overlapping epitope discovery study of 364 peptides from 3 human immunodeficiency virus 1 (HIV-1) Gag protein variants. Apply-

ing the ANN-Hydro model on existing peptide-MHC algorithms consistently reduced the number of candidate peptides across multiple antigens and may provide a correlate with immunodominance. Hydrophobicity of TCR contact residues is a hallmark of immunogenic epitopes and marks a step towards eliminating the need for empirical epitope testing for vaccine development.

In Chapter 4, we study the effect of early detection in the context of Ebola virus disease. We evaluated the potential effect of early diagnosis of pre-symptomatic individuals in west Africa. We used a simple mathematical model calibrated to the transmission dynamics of Ebola virus in west Africa. The baseline model includes the effects of contact tracing and effective isolation of infectious individuals in health-care settings. Our findings suggest that a strategy that focuses on early diagnosis of high-risk individuals, caregivers, and health-care workers at the pre-symptomatic stage, when combined with public health measures to improve the speed and efficacy of isolation of infectious individuals, can lead to rapid reductions in Ebola transmission.

Chapter 2

MODELING THE SUBCLONAL EVOLUTION OF CANCER CELL POPULATIONS

Abstract

Genome sequencing has revolutionized our understanding of the process of somatic evolution in cancer. Increasing evidence shows that tumor clonal architectures are often the consequence of a complex branched subclonal process. Yet, little is known about the expected dynamics and the extent to which these divergent subclonal expansions occur. Here, we study intratumor subclonal heterogeneity and its impact on treatment-resistance by developing and implementing more than 88,000 realizations of a stochastic evolutionary model simulating the process. Under different combinations of the population genetic parameter values, including those that have been previously estimated for colorectal cancer and glioblastoma multiforme, our results show that the distribution of sizes of subclones carrying driver mutations has a heavy right tail at the time of tumor detection, with only 1-4 dominant clones present at $\geq 10\%$ ffrequency, composing most of the tumor cell population. In contrast, our model also predicts that the vast majority of subclones will be present at $\leq 10\%$ frequency. We find that these minor, and often undetectable, subclones can harbor treatment-resistant mutations. In our analysis, the number of dominant clones ($\geq 10\%$) in a tumor has a strong correlation with the number of minor subclones. Model predictions are consistent with empirical data on the number of dominant clones at detectable levels across different cancer types. Our results contribute to explaining why tumors with greater numbers of detectable clonal populations are associated with poorer prognosis in multiple types of cancer.

2.1 Introduction

Cancer, a subclonal evolutionary process, is governed by the dynamic interplay of mutation, stochastic drift, and selection (Greaves and Maley, 2012; Marusyk et al., 2012; Nowell, 1976; Lipinski et al., 2016; Frank, 2007). Although most mutations that steadily accumulate in our cells are neutral or weakly deleterious, a fraction of these mutations in a gene or a regulatory element can confer a selective advantage to the cell by increasing its fitness (Martincorena and Campbell, 2015; Shendure and Akey, 2015; Lynch, 2016; Hanahan and Weinberg, 2011), and in cancers can result in increased survival of a clone (Martincorena and Campbell, 2015; Genovese et al., 2014; Fisher, 1958). In the field of cancer biology, the term driver mutation is used to refer to mutations under positive selection within a population of cells, and the term passenger mutation is used for mutations that are neutral or deleterious (Martincorena and Campbell, 2015; Vogelstein et al., 2013; McFarland et al., 2013), and increase in frequency due to hitchhiking alongside driver mutations. A common model of the evolutionary process explaining tumor growth envisions driver mutations causing clonal expansions that sweep through the tumor cell population and reach fixation (100% frequency) (Fearon et al., 1990; Bozic et al., 2010). If such a driver mutation did reach fixation, it would appear as a trunk mutation, present in all the tumor cells. However, tumor clonal architectures are often observed experimentally to be the consequence of a complex branched subclonal processes, with divergent subclones evolving simultaneously (Navin et al., 2010; Campbell et al., 2010; Anderson et al., 2011; Notta et al., 2011; Sottoriva et al., 2013; Campbell et al., 2008; Gerlinger et al., 2012; Nik-Zainal et al., 2012; Shain et al., 2015; Burrell et al., 2013; Burrell and Swanton, 2016; Zhao et al., 2016). While a few clones may dominate the composition of a tumor, minor and often undetectable subclones can dominate the clinical course of disease progression and recurrence (Landau et al., 2013; Schmitt et al., 2015; Maley et al., 2006; Chowell et al., 2016; Bozic

and Nowak, 2014; Landau et al., 2015). For example, in patients with chronic lymphocytic leukemia (CLL) who received chemotherapy, the presence of subclones harboring one or more cancer-driver genes in the original leukemia impacted prognosis and clinical outcome (Landau et al., 2013, 2015). More recently, similar findings have been observed across diverse cancer types (Zhang et al., 2014; Morris et al., 2016; Andor et al., 2016). Moreover, several studies have also revealed that tumor cells corresponding to the relapsed clone were often present as a minor subclone in the primary tumor before the initiation of therapy, which suggest that mutations contributing to recurrence are selected for during treatment (Mullighan et al., 2008; Hyo-eun et al., 2015; Roche-Lestienne et al., 2003; Morrissy et al., 2016; Diaz Jr et al., 2012; Ding et al., 2012; Roche-Lestienne et al., 2002, 2008). A Big Bang model of tumor evolution has been recently observed in an analysis of 349 samples from 15 independently derived colorectal tumors (Sottoriva et al., 2015). This model shows that colorectal tumors grow as a single expansion generating several intermixed subclones where both clonal and most observable subclonal mutations occur early during transformation (Sottoriva et al., 2015). These subclones are defined from the founding clone and are established by the additional mutations they acquired, which may not be present in the bulk of the tumor. Many subclones carrying driver mutations can remain rare and undetectable because their abundance falls below the detection limit of standard genome or exome sequencing techniques (Campbell et al., 2008; Schmitt et al., 2015; Barber et al., 2015; Tirosh et al., 2016; Wang et al., 2014; Gerstung et al., 2012). Current standard sequencing methods have low sensitivity and high false positive rate when detecting mutations below 10% frequency in the DNA extracted from the tumor sample (Barber et al., 2015). Although the Bing Bang model anticipates uniformly high levels of subclonal mutations throughout the neoplasm, a quantitative assessment of this subclonal diversity across cancers, and theoretical expectations are needed. Here, to gain insight into the dynamics of the subclonal evolution of cancer and to assess the extent to which heterogeneous subclonal expansions

occur, we develop a computational model. By employing a supercomputer to simulate the process, we were able to keep track of all subclones that arise and die out, or are maintained and grow during the evolutionary process. Under different combinations of the evolutionary parameter values, including those that have been previously estimated for two cancer types (glioblastoma multiforme and colorectal cancer), simulations lead to the conclusion that the distribution of sizes of subclones carrying driver mutations has a heavy right tail at time of tumor detection, with only 1-4 dominant clones present at $\geq 10\%$ frequency, composing most of the tumor cell population. By contrast, our model predicts that the vast majority of subclones will be present at $< 10\%$ frequency. In addition, we find that most of treatment-resistant subclones harboring driver mutations are often present at low frequency, below the detection limit of standard sequencing techniques. Moreover, our results suggest that the number of minor subclones is strongly correlated with the number of numerically dominant clones in a tumor. We finally discuss empirical data that supports our findings, the implications of our model predictions, the strengths of the model, and its limitations.

2.2 Results

2.2.1 Subclonal Evolutionary Model of Cancer Cell Populations

Previous dynamical models developed to study tumor evolution (Bozic et al., 2010; Waclaw et al., 2015; Durrett et al., 2011; Beerenwinkel et al., 2007) assume that each driver mutation affects the fitness of a tumor cell lineage equally (with the exception of ref. (Durrett et al., 2011). They also assume that the fitness effect of a driver mutation is independent of the other driver mutations carried by the cell. However, these epistatic interactions are a very central aspect of the dynamics of adaption of asexual populations (Lang et al., 2013; Breen et al., 2012), and should be relevant to asexual tumor populations as well (Sprouffske et al., 2012). Moreover, as it is a computationally intensive task, the majority of cancer

evolution models do not study the extent of heterogeneous subclonal expansions that can occur simultaneously during the neoplastic process. Clonal heterogeneity is particularly important because multiple subclones carrying favorable (and potentially even resistance) mutations can emerge and compete with each other, increasing variation in evolutionary outcomes, as shown previously in laboratory evolution experiments (Lang et al., 2013; Hegreness et al., 2006; Levy et al., 2015).

Our computational model is based on a branching evolutionary process (Bozic et al., 2010; Haccou et al., 2005), where we model selection as an epistatic process (Breen et al., 2012; Kryazhimskiy et al., 2011). In the model, a subclone is defined as a subpopulation of cells that descended from another clone but then diverged by accumulating another driver mutation. Similar to the framework presented in (Bozic et al., 2010), each simulation is initiated with a single cell carrying a single driver mutation (the potential founder of a primary tumor) that provides a selective growth advantage over normal neighboring cells.

At each time step, a cell may either die or divide, and it can further acquire an additional driver mutation in one of the daughter cells at a rate μ_d. For each driver mutation occurring during cell division, we sample a selection coefficient from an exponential distribution of mean $\bar{s}$ (Hegreness et al., 2006; Orr, 2003), and update the cell fitness f. We assume that each subsequent driver mutation increases the probability of cell division, defined as $b = 1/2 f = 1/2[w_{wt} + d(1 - \Pi_{i=1}^{n}(1 - s_i))]$ (Kryazhimskiy et al., 2011; Nagel et al., 2012; Kryazhimskiy et al., 2014; Wiser et al., 2013). The fitness change in a cell produced by a driver is thus dependent of the other driver mutations and the temporal order at which they occur in the cell; consequently, each cell lineage has its own fitness trajectory during the neoplastic process. The parameter w_{wt} represents the fitness of the wild-type cell in which the first driver occurs, and without loss of generality it is assumed to be 1. The parameter d is a measure of the maximum possible fitness gain through adaptation; we assume $d = 1$

(Nagel et al., 2012). The parameters $s_1, s_2, ..., s_n$ characterize the fitness effects associated with each of n driver mutations that a cell lineage carries.

By employing a supercomputer (see Materials and Methods for details), we are able to keep track of all subclones that arise and die out, or are maintained and grow during the evolutionary process. Because of the limited quantitative knowledge of parameter values across cancers, we test a range of values for μ_d and $\bar{s}$. The ranges we explored are centered on values obtained from the literature. The values selected for the driver mutation rate, μ_d, are: $1 \times 10^{-8}, 1 \times 10^{-7}, 1 \times 10^{-6}$, and 1×10^{-5} alterations per cell division (Shendure and Akey, 2015; Lynch, 2016; Bozic et al., 2010; Sottoriva et al., 2015; Beerenwinkel et al., 2007; Vermeulen et al., 2013). And the values chosen for the mean, $\bar{s}$, of the exponential distribution of fitness effects are: 0.1, 0.01, and 0.005 (Bozic et al., 2010; Beerenwinkel et al., 2007; Vermeulen et al., 2013).

2.2.2 Drift Dominates Early Neoplastic Dynamics

A necessary step in neoplastic initiation is that the first mutated cell lineage survives stochastic drift to result in a clone growing at the expense of its normal neighbors. The growth of the first clone is important in increasing the number of cells in which a second driver mutation could occur, and subsequently, another clone could emerge from the cell with the second driver and so on until a tumor is formed (Fig. 2.1) (Vogelstein and Kinzler, 2015). To quantify the effect of stochastic drift in neoplastic initiation and to generate at least 100 clinically detectable simulated tumors (tumor cell populations reaching $\approx 10^9$ cells) for each combination of the chosen parameter values, we ran 88,265 simulations of the process (Table 2.1). Overall, we observe that out of the total number of realizations executed, only 1,432 became clinically detectable tumors, despite each simulation being initiated with a driver mutation. Thus, on average, $\approx 98\%$ of all the initiating mutated cell

lineages carrying a driver mutation die out early (Table 2.1), which is in line with theoretical expectations. This result highlights the importance of drift affecting neoplastic initiation.

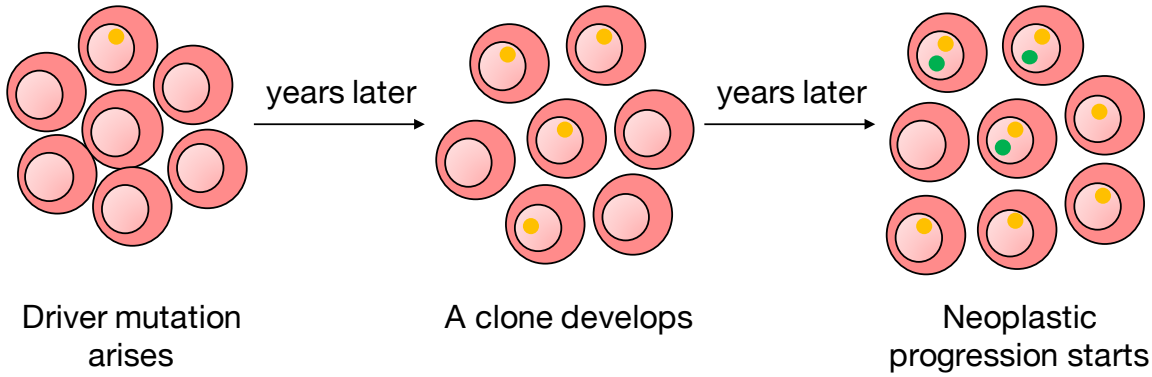


Figure 2.1: Schematic representation of the expansion of a single mutated cell into a clonal population under the influence of a driver mutation (yellow circle), and the potential conversion of the clone into cancer through subsequent driver mutations (green circle).

To test how this model works with parameter values from a known cancer type, we use estimates from colorectal cancer. It has been experimentally estimated that colorectal cancer cells divide every 4 days (Bozic et al., 2010; Jones et al., 2008). Assuming this cell division time in the simulations, we find that the average expected time from onset to clinical detection of the simulated tumors ranges from 1.64 years to 27.97 years, depending on the values for $\bar{s}$ and μ_d (Table 2.1). Additionally, using the parameter values $\bar{s} = 0.005$ and $\mu_d = 1 \times 10^{-5}$ per cell division, which have been estimated for colorectal cancer (Bozic et al., 2010), our model predicts that an average of 18.28 years would take for a colorectal tumor to develop (Table 2.1). This estimate is concordant with previous reports (Bozic et al., 2010; Jones et al., 2008). By having generated a total of 1,432 clinically detectable simulated tumors under a wide range of parameter values, we can determine the general contribution of each of the parameters to initiation and neoplastic progression. We find, consistent with a previous report (Beerenwinkel et al., 2007), that selection has a larger effect on neoplastic initiation than the driver mutation rate (Table 2.1). Moreover, we find that the average expected time from initiation to detection of a tumor increases with decreasing

the average fitness effect of driver mutations and with decreasing the driver mutation rate (Table 2.1).

Table 2.1: Number of simulations performed for each combination of parameter values ($\bar{s}$, μ_d). The mean time in generations and in years from initiation to clinical detection of a simulated tumor is also shown. The generation time assigned in the simulations is $T = 4$ days

$\bar{s}$	μ_d	Number of realizations	Number of simulations that reached 10^9 cells	Percentage of simulations that reached 10^9 cells	Mean number of generations to detection	Mean time to detection (years)
0.1	1×10^{-8}	10155	162	1.6%	1596.66	17.50
0.1	1×10^{-7}	1948	112	5.7%	475.08	5.21
0.1	1×10^{-6}	748	134	17.9%	158.54	1.74
0.1	1×10^{-5}	748	111	14.8%	147.50	1.62
0.01	1×10^{-8}	6867	125	1.8%	1807.63	19.80
0.01	1×10^{-7}	6866	113	1.6%	1406.75	15.41
0.01	1×10^{-6}	6866	120	1.7%	1263.80	13.85
0.01	1×10^{-5}	6865	115	1.7%	1018.40	11.16
0.005	1×10^{-8}	11951	102	0.9%	2552.70	27.97
0.005	1×10^{-7}	11751	112	1.0%	2546.85	27.91
0.005	1×10^{-6}	11750	126	1.1%	2046.78	22.43
0.005	1×10^{-5}	11750	100	0.9%	1668.07	18.28
		88265	1432	1.6%		

2.2.3 Extent of Intratumor Subclonal Variation at Detection

To gain insight into the extent of subclonal populations carrying driver mutations within a tumor at the time of detection and to determine how the different evolutionary parameters impact the subclonal composition, we analyzed all the 1,432 detectable tumors generated by our model. In most tumors, we find that the number of dominant clones, defined here as a subclone present at $\geq 10\%$ frequency in a tumor, ranges from 1 to 4; however, the average number of dominant clones in a tumor tend to increase with decreasing the average fitness effect of driver mutations and with increasing the driver mutation rate (Fig.2.2). Across all

the 1,432 detectable simulated tumors, the average number of dominant clonal populations is 1.47. Of note, the predicted range by the model on the number of numerically dominant clones in a tumor is largely concordant with the findings reported in two recent pan-cancer analyses of intratumor heterogeneity (Morris et al., 2016; Andor et al., 2016). Importantly, we find that even though only a few dominant clones compose the majority of the cancer cell population in a simulated tumor (range 90.6% - 99.5%; (Fig.2.2)), the number of minor subclones present at $< 10\%$ frequency is substantial (Fig.2.2); and this number increases with decreasing the average fitness effect of drivers and with increasing the driver mutation rate (Fig.2.2). For example, given the parameter values $\bar{s} = 0.005$ and $\mu_d = 1 \times 10^{-5}$, which have been estimated for glioblastoma multiforme and colorectal cancer (Bozic et al., 2010), the model predicts that, on average, 1.8 dominant clones and 2,705 minor subclonal populations carrying driver mutations compose the tumor.

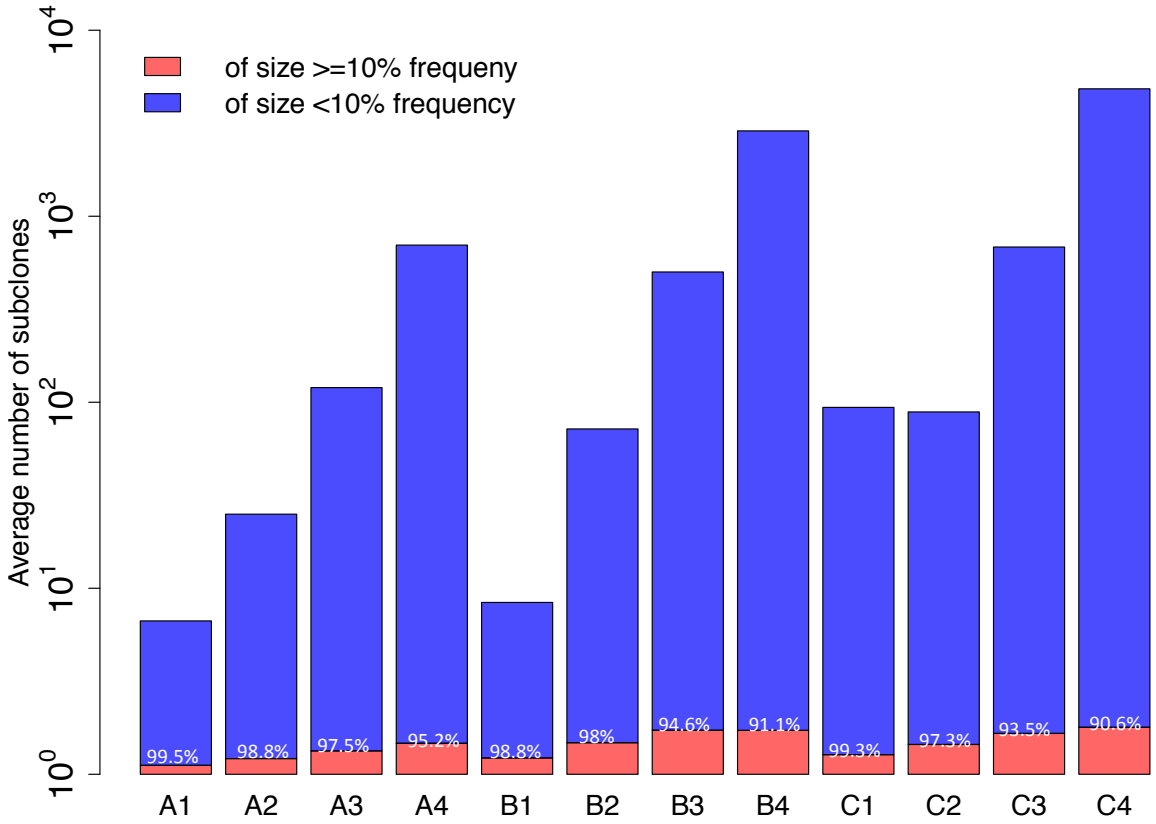


Figure 2.2: Intratumor subclonal variation. Bar plots of the log10 average number of dominant clones present at $\geq 10\%$ frequency in the simulated tumors (red) and the log 10 average number of minor subclones that are present at $< 10\%$ frequency (blue). The notation bellow each bar identifies the combination of parameter values used, and it reads as follows A: $\overline{s} = 0.1$; B: $\overline{s} = 0.01$; C: $\overline{s} = 0.005$; a: $\mu_d = 1 \times 10^{-8}$; b: $\mu_d = 1 \times 10^{-7}$; c: $\mu_d = 1 \times 10^{-6}$; d: $\mu_d = 1 \times 10^{-5}$. The value shown over each red bar represents the average percentage tumor cell population composed by the dominant clones. Note that the extent of intratumor subclonal variation is greatly affected by both the mean, $\overline{s}$, of the exponential distribution of fitness effects associated with the driver mutations and the driver mutation rate, μ_d.

Since both selection and the driver mutation rate impact the extent of intratumor subclonal variation (Fig.2.2), we find that there is a strong, statistically significant correlation between the average number of dominant clones and the average number of minor subclones across all simulated tumors (Pearson $r = 0.91, P < 0.001$; (Fig.2.3)). This result suggests that the number of detectable clonal populations by standard sequencing techniques may simultaneously reflect the extent of undetectable minor subclones carrying driver mutations in a tumor.

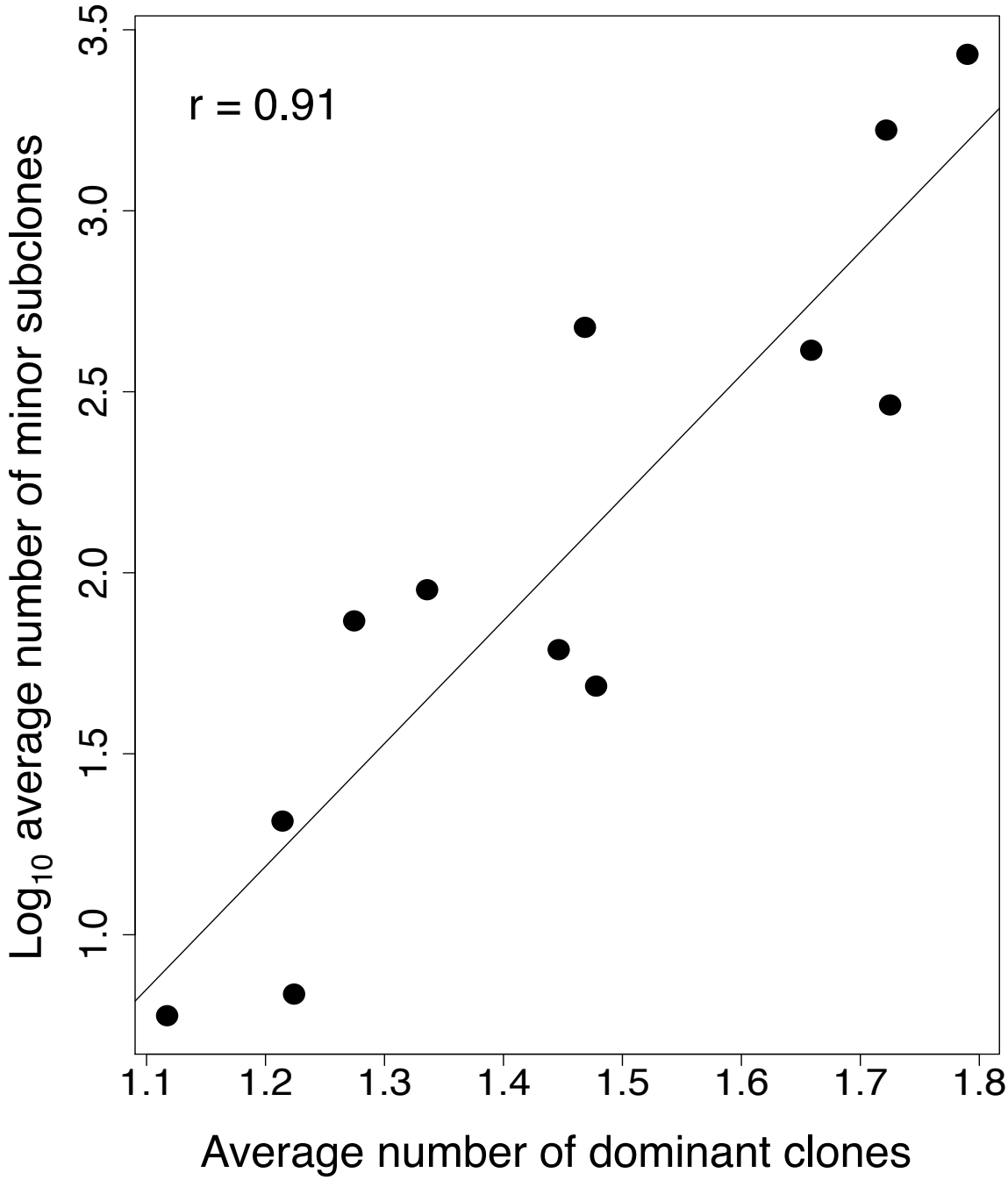


Figure 2.3: Relation between the average number of dominant clones and the log 10 average number of minor subclones in the simulated tumors for each combination of parameter values used as in (A)

For each combination of parameter values, we then computed the probability density function for the subclone sizes present in the simulated tumors (Fig.2.4). Overall, we find

that the distribution of subclone sizes harboring driver mutations has a heavy right tail, with only a few clones present at $\geq 10\%$ frequency in the tumor, and with most subclones present at frequencies as low as 10^{-7} (Fig.2.4).

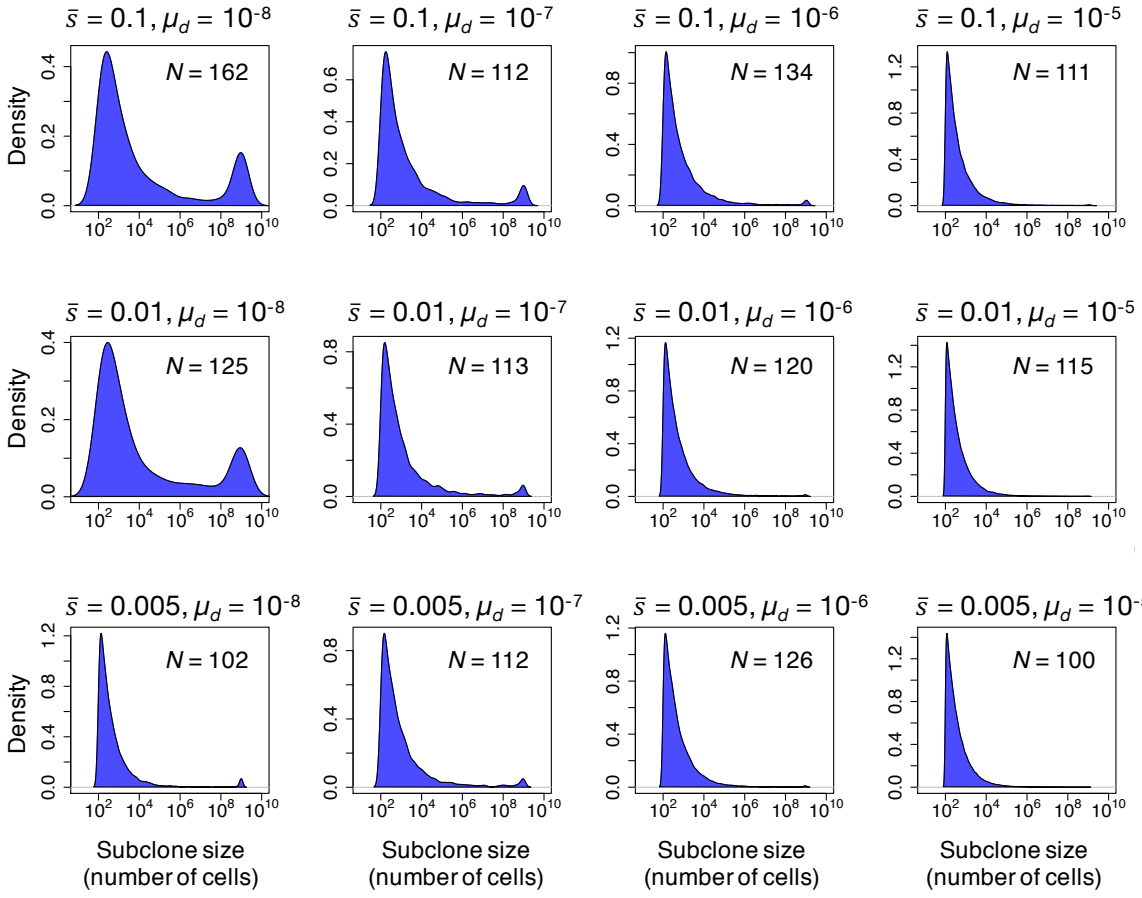


Figure 2.4: Probability density function for the log 10 sizes of subclones carrying driver mutations present in the clinically detectable simulated tumors for each combination of parameter values; N is the number of tumors generated by the model. The distribution has a heavy right tail with only a few dominant clones present at $\geq 10\%$ frequency, composing most of the tumor cell population (Fig.2.2), and with the majority of subclones present at frequencies below the detection limit of standard sequencing techniques.

2.2.4 Differential Fitness Between Dominant and Minor Subclones

A key aspect of our computational model is that it can simulate subclonal heterogeneity- each evolving cellular lineage has its own fitness, which is dependent on the fitness effects conferred by the driver mutations and their epistatic interactions. For illustrative purposes,

we show the subclonal composition and their corresponding fitness in two clinically detectable simulated tumors using the parameter values $\bar{s} = 0.01$ and $\mu_d = 1 \times 10^{-5}$ (Fig.2.5 A); and $\bar{s} = 0.005$ and $\mu_d = 1 \times 10^{-5}$ (Fig.2.5 C).

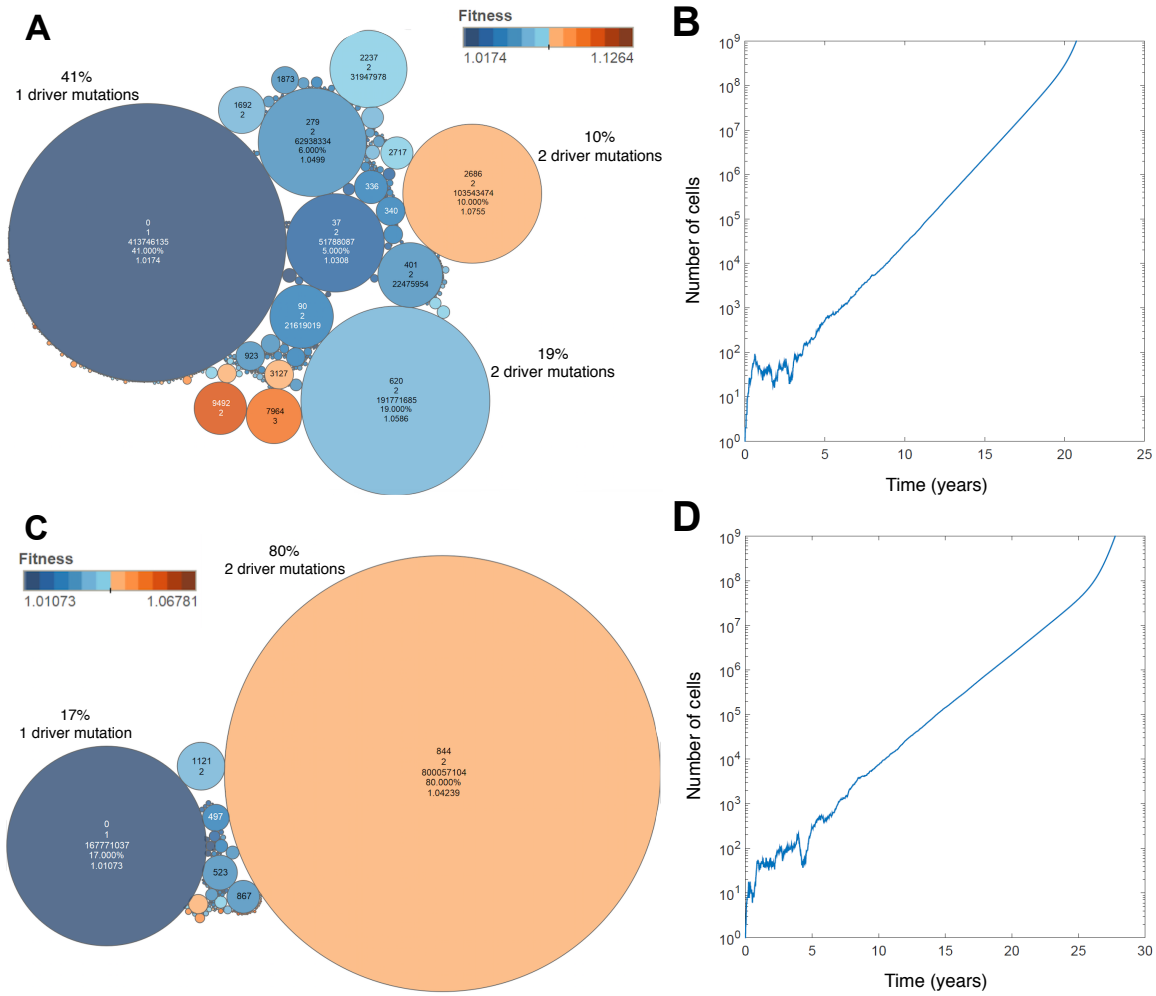


Figure 2.5: Variability in subclonal composition, differential fitness among subclones, and population dynamics of progression of two simulated tumors. (A and B) Parameter values used are: $\bar{s} = 0.01$ and $u_d = 1 \times 10^{-5}$. (C and D) Parameter values used are: $\bar{s} = 0.005$ and $\mu_d = 1 \times 10^{-5}$. Each circle represents a subclone carrying a certain number of driver mutations. The color of a circle corresponds to the fitness of the subclone. The size of a circle is proportionate to the number of cells composing the subclone. In both cases, the generation time used is $T = 4$ days. Note that the relative fitness of some minor subclones is greater than the fitness of the dominant clones present at $\geq 10\%$ frequency. And, as would be expected, there is substantial intra- and inter- subclonal variation between the two simulated tumors.

The corresponding population dynamics of both simulated tumors are shown as well (Fig.2.5 B and D, respectively). We observe that the simulated tumor presented in Fig.2.5

A has 3 dominant clones present at 41%, 19%, and 10% frequency in the tumor, carrying 1-2 driver mutations. On the other hand, the simulated tumor shown in Fig.2.5 C has only two dominant clones preset at 80% and 17% frequency, harboring 2 and 1 driver mutations, respectively. Additionally, we also observe that, as would be expected, there is substantial intra- and inter- subclonal variation in both cases (Fig.2.5 A and C). More importantly, we find that, on average, the relative fitness of minor subclones is greater than the fitness of the dominant clonal populations (Table A.2) and (Fig.2.5 A and C). These results show that, even though cells acquiring subsequent driver mutations, conferring a fitness advantage, cannot sweep through the population within the timescale of tumor growth. And consequently, a substantial number of minor and often fitter subclones harboring driver mutations remain in the tumor at very low frequency.

2.2.5 Resistant Subclones Carrying Driver Mutations Are Present at Low Frequency at Detection

Populations can adapt to novel environments in two different ways- selection acting on pre-existing genetic variants and selection acting on novel mutations (Barrett and Schluter, 2008). Adaptation from standing genetic variation is probably to be faster than from novel mutations, not only because beneficial mutations are immediately available in the new environment, but also because they may start at higher frequencies (Barrett and Schluter, 2008). We thus tested whether subclones carrying driver mutations could also carry at least one resistance mutation in the clinically detectable simulated tumors. To test this, we used all the chosen values for the mean, $\bar{s}$, of the exponential distribution of fitness effects associated with the driver mutations, and the following values for the driver mutation rate $\mu_d : 1 \times 10^{-5}$ and 1×10^{-6} alterations per cell division (Bozic et al., 2010). We assume that multiple different mutations can independently cause resistance, and a resistance mutation hence occurs at a rate of 1×10^{-8} (Komarova et al., 2014; Komarova and Wodarz, 2005)

in a cell that coincidentally acquire a driver mutation during cell division. Additionally, we assume that the resistance mutation is neutral. Under these assumptions and parameter values, we find that the majority of resistant subclones carrying driver mutations are often present at low frequency, below the detection limit of standard DNA sequencing techniques (Fig.2.6).

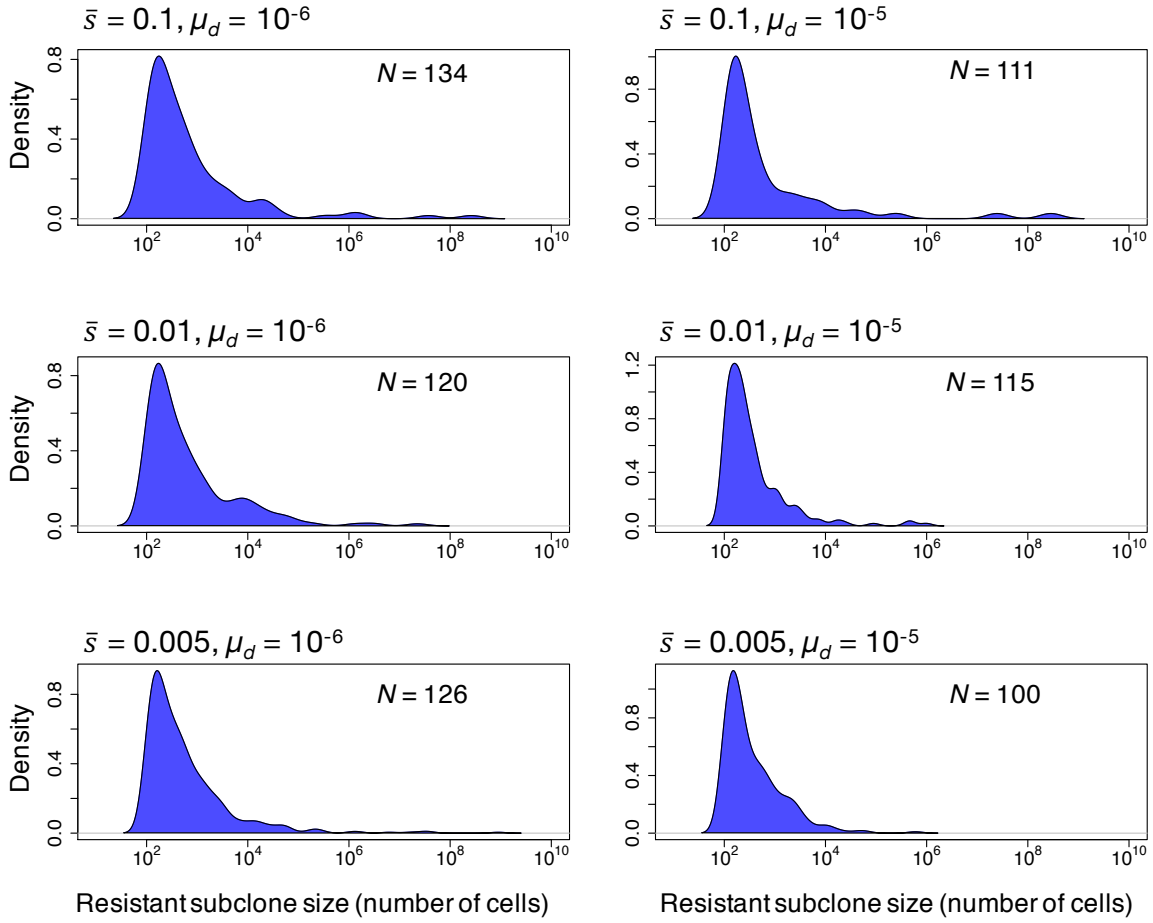


Figure 2.6: Resistant subclones in the clinically detectable simulated tumors. Probability density function for the log10 sizes of treatment-resistant subclones carrying driver mutations within in the primary simulated tumor using the combinations of parameter values.

We then calculated the number of independent resistant subclones within each simulated tumor (Fig.2.7). We find that the maximum number of independent resistant subclones in a particular simulated tumor is 18, and the minimum is 0 (Fig.2.6). Overall, we find that the number of independent resistant subclones in a simulated tumor increases with

decreasing the average fitness effect of the drivers and with increasing the driver mutation rate (Fig.2.7). Therefore, these results along with other studies (Mullighan et al., 2008; Hyo-eun et al., 2015; Roche-Lestienne et al., 2003; Morrissy et al., 2016; Diaz Jr et al., 2012; Ding et al., 2012; Roche-Lestienne et al., 2002, 2008; Siravegna et al., 2015) suggest that the extent of standing subclonal variation in a tumor may be the most important factor for the evolution of acquired resistance to treatment.

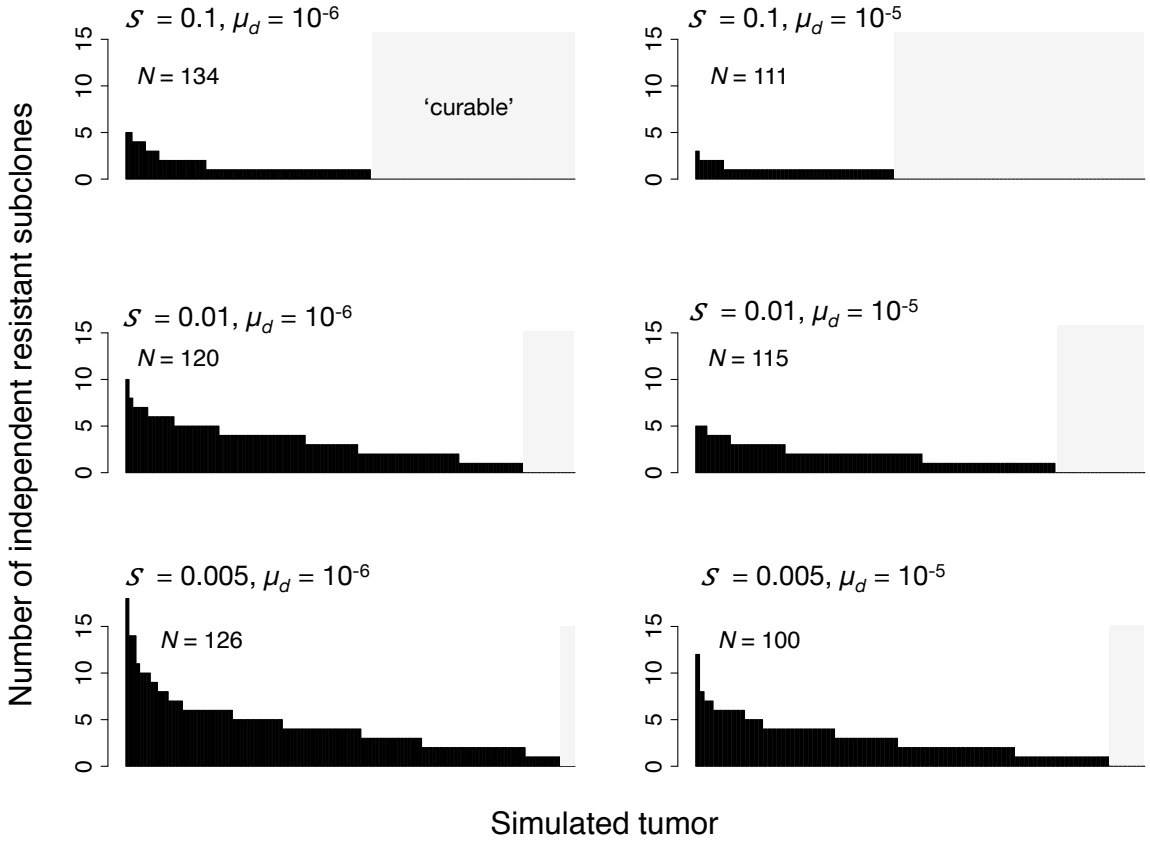


Figure 2.7: Resistant subclones in the clinically detectable simulated tumors. Number of independent resistant subclones in each primary tumor. Each bar represents a simulated tumor. For each set of parameter values used, N is the number of tumors generated by the model. It is assumed that a resistant mutation occurs at a rate of 1×10^{-8} in a cell that coincidentally acquire a driver mutation during cell division. Resistant mutations are assumed to be neutral.

2.3 Discussion

In this study, we have developed and implemented a stochastic evolutionary modeling framework to study the subclonal dynamics of cancer progression and to assess the extent to

which heterogeneous subclonal expansions occur. First, we show that drift substantially affects neoplastic initiation (Table 2.1). This result is expected from the theory of population genetics. Moreover, we find that selection has a much larger effect on neoplastic initiation than the driver mutation rate (Table 2.1). These results are in agreement with a report where the authors quantified the competitive benefit of Apc loss, Kras activation, and P53 mutations in intestinal tumor initiation (Vermeulen et al., 2013). They showed that most mutated cells are lost and that clones in colitis-affected intestines carrying P53 mutations are particularly favored in tumor initiation. Additionally, we show that the average expected time from onset to clinical detection of a tumor increases with decreasing the average fitness effect associated with the driver mutations and with decreasing the driver mutation rate (Table 2.1). Altogether, we hypothesize that the fitness effects conferred by driver genetic alterations in certain sporadic childhood cancers should be greater than those associated with drivers in sporadic adult cancers. For example, in chronic myeloid leukemia, the fitness effect conferred by the chromosome translocation juxtaposing the BCR and ABL genes may be quite large, and consequently, this single genetic alteration may be all that is required for tumor initiation and progression (Michor et al., 2006; Tomasetti et al., 2015).

A statistical strength of this study is that we were able to simulate over 1,400 clinically detectable tumors, from which we could analyze and gain important biological insights. Across all the combinations of the chosen evolutionary parameter values, our analysis show that the distribution of sizes of subclones carrying driver mutations has a heavy right tail at the time of tumor detection, with only 1-4 dominant clones present at $\geq 10\%$ frequency, composing most of the tumor cell population (Fig.2.2). In contrast, our model predicts that the vast majority of subclones will be present at $< 10\%$ frequency in the tumor (Fig.2.2). We also find that the number of these minor subclones present at $< 10\%$ increases with decreasing the average fitness effect of drivers and with increasing the driver mutation rate (Fig.2.2). Additionally, we find that there is a strong, statistically significant correlation

between the average number of numerically dominant clones and the average number of minor subclones across all simulated tumors (Fig.2.3).

The distribution of subclones sizes inferred by our model is qualitatively similar to that found in a recent study where the authors used ultra-deep sequencing technology to detect mutations in specific known cancer genes in biopsies of sun-exposed eyelids from different individuals (Martincorena et al., 2015). Moreover, the predicted range by the model on the number of dominant clonal populations is largely concordant with two recent pan-cancer analyses of intratumor heterogeneity (Morris et al., 2016; Andor et al., 2016). In one of these studies, the authors analyzed over 3,300 tumors and integrated SNP array copy number and whole exome sequencing mutational data from The Cancer Genome Atlas (TCGA) to infer the number of clonal populations present in multiple tumor types (Morris et al., 2016). They found that 92% of all the tumors analyzed had between 1 and 4 clonal populations at detectable levels by standard sequencing methods (Morris et al., 2016).

By employing ultra-deep sequencing technology to study intratumor subclonal diversity in patients with CLL, it has been shown the presence of rare subclones at frequencies down to the limit of detection for depth of this specific sequencing method, 10-4 (Campbell et al., 2008). Our model predicts that minor subclones carrying driver mutations can be present in a tumor at lower frequencies, 10-7 (Fig.2.4). Importantly, we find that these often undetectable minor subclones can harbor treatment-resistant mutations (Fig.2.5 A). These results are in line with previous reports showing that tumor cells corresponding to the relapsed clone were often present as a rare subclone in the diagnostic tumor before the initiation of treatment (Mullighan et al., 2008; Hyo-eun et al., 2015; Roche-Lestienne et al., 2003; Morrissy et al., 2016; Diaz Jr et al., 2012; Ding et al., 2012; Roche-Lestienne et al., 2002, 2008; Siravegna et al., 2015). We also find that the number of independent resistant subclones in a simulated tumor depends on the population genetic parameter values-

it increases with decreasing the average fitness effect of the driver mutations and with increasing the driver mutation rate (Fig.2.7).

Given that cancer is the result of a very complex process, our approach has some limitations. First, we have not taken space into consideration in our computational model, which may restrict the expansion of certain subclonal populations during the neoplastic process. Thus, our modeling framework may be more suitable for cancers that develop in the absence of spatial constraints. Second, the host immune system has not been taken into account, which has been demonstrated to have an important role for both cancer suppression and promotion (Schreiber et al., 2011). Third, future studies should extend this work and consider non-cell autonomous interactions, as well as "common good" factors, which are likely to be a strong influence on subclonal dynamics (Marusyk et al., 2014). Despite these caveats, our model captures some essential features of the dynamics of subclonal evolution of cancer progression. The subclonal dynamics predicted by our model are consistent with the "Big Bang" model (Sottoriva et al., 2015), where clonal driver mutations and most detectable subclonal drivers occurred relatively early during tumor growth. This result is in contrast with the traditional clonal selection model, where sequential stepwise accumulation of drivers leads to fitter clones that sweep through the population (Fearon et al., 1990; Bozic et al., 2010).

In conclusion, our modeling framework serves as a proof-of-concept to test that a substantial number of subclonal populations carrying driver mutations will remain rare and undetectable within a tumor because their abundance falls below the detection limit of standard genome sequencing methods. Additionally, these minor and often undetectable subclones can also harbor treatment-resistant mutations, which present a major challenge for personalized medicine and clinical management. Altogether, our findings explain why tumors with greater numbers of detectable clonal populations are associated with poorer

prognosis across multiple cancer types (Landau et al., 2013; Zhang et al., 2014; Morris et al., 2016).

These results suggest that driver mutations that have been identified in individual tumors through standard genome sequencing (Lawrence et al., 2013) are likely to constitute only the "tip of the iceberg", with several driver mutations that may impact the evolutionary dynamics of cancer progression never rising above very low frequencies, until the time point when treatment starts.

2.4 Materials and Methods

2.4.1 Computer Code Availability

The computer code 'tumorsim.scala' is available at:
https://github.com/WilsonSayresLab/TumorHeterogeneity.

2.4.2 High Speed Algorithm Considerations

We used a hierarchical data structure to store common attributes for all cells within the same subclonal population.

2.4.3 Software Tools

The following open-source platforms and programming languages were used for tumor simulation, monitoring and analysis: Apache Hadoop (HortonWorks 2.6.0) - software framework written in Java for distributed storage and processing of very large data sets over commodity hardware clusters. Apache Hive (1.2.1 spark HiveMetastoreConnection version 1.2.1, interactive hive-cli-0.14) - external data warehousing stacked on Hadoop, provides simulation monitoring, data summarization, query and analysis. Apache Scala (2.10.5) functional programming language that utilizes the JVM (Java Virtual Machine) for plat-

form independency, controls tumor simulation logic. Apache Spark (1.6.0 with a min of 1.4.0) - distributed computing framework originally developed at UC Berkeley AMPLab, stacked on top of Scala and distributed over Hadoop, tracks tumor array memory across multiple machines. Bash (Sun AMD64 Linux $2.6.32 - 504.el6.x86_64$) Bash is a UNIX shell and command language; various scripts used for monitoring, analysis and data export to spreadsheets or other visualizations. Java/JDK (Oracle 1.7) - Spark, Scala and Hive are converted to run in the java virtual machine for platform independency. Microsoft Excel (2013) - for aggregates and table visualizations. Module (3.2.10) - Bash environment management scripting un/loader. Tableau (public 9.1 to 9.3) - for visualization of subclonal composition of simulated tumors. YARN (2.2.4.2-2) - Yet Another Resource Manager, manages Hadoop data and hardware resources.

2.4.4 Hadoop Cluster Configuration

The computation and data intensive piece includes a 44 node HDP 2.3 cluster on Dell PowerEdge 720xd servers. Each of the 40 worker nodes has 128GB ram, 2x Intel E5-2640 6 core processors and 22TB of disk. The cluster backbone network consists of 10Ge HA top of rack switching combined with Intel x520 10Ge NICs in each server. Although the tumor simulator can run parallel jobs utilizing multiple resources, the demands upon the hadoop NameNode (worker, memory, disk resource managment) are quite exhaustive; therefore, it is suggested to run sequential jobs on a single node for as many images as needed to emulate parallelization.

2.4.5 Statistical Analysis

We created scripts on RStudio (Version 0.99.891) to analyze the data sets, perform statistical analysis, and generate most of the figures (with the exception of the figures dis-

playing subclonal composition of simulated tumors). All the R scripts are available upon request.

Chapter 3

TCR CONTACT RESIDUE HYDROPHOBICITY IS A HALLMARK OF IMMUNOGENIC CD8$^+$ T CELL EPITOPES

This chapter is published:

Diego Chowell, Sri Krishna, Pablo D. Becker, Clement Cocita, Jack Shu Xuefang Tan, Philip D. Greenberg, Linda S. Klavinskia, Joseph N. Blattman, and Karen S. Anderson[1]

Proceedings of the National Academy of Sciences 112.14 (2015): E1754-E1762

Abstract

Despite the availability of MHC-binding peptide prediction algorithms, the development of T cell vaccines against pathogen and tumor antigens remains challenged by inefficient identification of immunogenic epitopes. CD8+ T cells must distinguish immunogenic epitopes from non-immunogenic self-peptides to respond effectively against an antigen without endangering the viability of the host. Since this discrimination is fundamental to our understanding of immune recognition, and critical for rational vaccine design, we interrogated the biochemical properties of 9,888 MHC class I peptides. We identified a strong bias towards hydrophobic amino acids (aas) at TCR contact residues within immunogenic epitopes of MHC allomorphs, which permitted us to develop and train a hydrophobicity-based artificial neural network (ANN-Hydro) to predict immunogenic epitopes. The immunogenicity model was validated in a blinded in vivo overlapping epitope discovery study of 364 peptides from 3 human immunodeficiency virus 1 (HIV-1) Gag protein variants. Applying the ANN-Hydro model on existing peptide-MHC algorithms consistently reduced the number of candidate peptides across multiple antigens and may provide a correlate

[1]L.S.K, J.N.B, and K.S.A are senior co-authors

with immunodominance. Hydrophobicity of TCR contact residues is a hallmark of immunogenic epitopes and marks a step towards eliminating the need for empirical epitope testing for vaccine development.

3.1 Significance Statement

The design of effective T cell vaccines against pathogens and tumor antigens is challenged by the highly inefficient identification of the subset of peptides from a given antigen that effectively stimulate an immune response. Here, we report that the relative hydrophobicity of TCR contact residues is markedly enriched in immunogenic MHC class I epitopes in both human and murine MHCs, and in both self and pathogen-derived immunogenic epitopes. Incorporating hydrophobicity into T cell epitope prediction models increases the efficiency of epitope identification, which will manifest in time and cost of T cell vaccine development. Amino Acid hydrophobicity may represent a biochemical basis by which T cells discriminate immunogenic epitopes within the background of self-peptides.

3.2 Introduction

The interaction of $CD8^+$ T cells with peptide-MHC complexes (pMHCs) is a key event in the development of cell-mediated immunity (Grakoui et al., 1999). MHC class I (MHC-I) molecules typically present 8-11 aa peptides derived predominantly from proteasomal degradation of intracellular proteins, either self-peptides or infection-derived antigens (Blum et al., 2013). T cell receptors (TCRs) from $CD8^+$ T cells bind antigenic pMHC molecules, triggering a downstream signaling cascade that leads to T cell activation, differentiation, and ultimately to cytolysis of target cells presenting the same epitope (Hennecke and Wiley, 2001). Vaccines and immunotherapies for the treatment of infection and cancer seek to incorporate cytotoxic T cell (CTL) epitopes, but defining such epitopes remains a costly and arduous process (Purcell et al., 2007). Understanding the molecular basis

of TCR-pMHC recognition will aid discovery of immunogenic epitopes in infectious and autoimmune disease.

During thymic development, $CD8^+$ T cells undergo both positive and negative selection to acquire the ability to discriminate antigenic peptides from self-peptides (Hogquist et al., 1994). Costimulatory signals can enhance this discrimination (Medzhitov and Janeway, 2002), but a primary event that triggers $CD8^+$ T cell activation is the non-covalent pMHC-TCR interaction. Proteasomal cleavage patterns and binding affinities of peptides to different MHCs have been extensively studied (Falk et al., 1991; Rammensee et al., 1999; Kubo et al., 1994). In contrast, the biochemical basis of immunogenic epitopes are less well-defined (van der Merwe and Dushek, 2011). T cell epitope discovery is complicated by the codominance and polymorphism of MHC alleles, diversity of antigens (both infectious and self-antigens), limited mass spectrometry-based confirmation of MHC-bound peptides, and a scarcity of experimentally confirmed immunogenic epitopes within the infectious and self-proteome (Purcell et al., 2007). As a result, T cell epitope prediction algorithms have focused on aa binding affinity for specific MHC-motifs and the proteins proteasomal cleavage pattern to identify candidate T cell epitopes (Honeyman et al., 1998; Moutaftsi et al., 2006; Nielsen et al., 2007; Tenzer et al., 2005). Although computational tools have improved over the past decade, they have not been trained to predict immunogenicity. The major limitation when using such prediction algorithms is the presence of a significant number of binders from a given antigen that will never lead to an immune response (Newell et al., 2013). Thus, immunogenic CTL epitopes fulfill additional criteria that go beyond antigen processing and MHC-binding.

Here, we sought to identify the biochemical criteria that define immunogenicity within the subset of MHC-I binding peptides. Using a curated repository of MHC-I epitopes from the Immune Epitope Database (IEDB) (Vita et al., 2010), we evaluated the biochemical properties of aas that discriminate between immunogenic epitopes and non-immunogenic

self-peptides. We found a strong bias towards hydrophobicity in aa residues of immunogenic CTL epitopes that is highly selective for exposed TCR contact residues. Using these criteria, we trained an artificial neural network (ANN) model to identify immunogenic CTL epitopes from a data set and empirically assessed our prediction model for 3 human immunodeficiency virus 1 (HIV-1) Gag protein variants in a murine model of immunogenicity. We demonstrate the utility of this ANN model, which has the potential to significantly enhance the efficiency of T cell epitope discovery.

3.3 Results

3.3.1 Amino Acid Use Differs Between Immunogenic and Non- immunogenic Peptides

CTLs recognize immunogenic epitopes from a background of poorly immunogenic self-peptides. To understand the biochemical basis of differences between these two classes of peptides, we retrieved all known MHC class I-binding peptides reported as T cell reactive (hereafter immunogenic) and self-peptides from MHC-ligand elution experiments with no known immunogenicity (hereafter non-immunogenic) from IEDB. Any eluted peptide that was immunogenic (either pathogen derived or self-antigen derived) was excluded to generate two mutually exclusive datasets that avoid any potential bias. Out of the 34,586 total retrieved peptides from IEDB, 5,035 8-11mer non-redundant peptides were reported to be immunogenic and 4,853 were non-redundant non-immunogenic and were used for further analysis. Frequency distributions of aas in 8-11mer immunogenic and non-immunogenic peptides showed significant variability in aa composition (Fig. B.1 A).

To identify overrepresentation of certain aas in immunogenic epitopes, we computed a probability ratio for each aa. We then performed a correlation analysis between the probability ratio of each aa and three major biochemical properties using independent numeric scales: hydrophobicity (Kyte-Doolittle) (Kyte and Doolittle, 1982), polarity (Grantham)

(Grantham, 1974), and side chain bulkiness (Zimmerman) (Zimmerman et al., 1968) (Table B.3). We found a strong, statistically significant correlation between probability ratios and hydrophobicity values (Spearman $\rho = 0.71, P = 4.24 \times 10^{-4}$; (Fig. 3.1 A)). Similarly, we also found a negative correlation between probability ratios and polarity of aas (Spearman $\rho = -0.77, P = 6.97 \times 10^{-5}$; (Fig. 3.1 B)), with highly polar aas being underrepresented in immunogenic epitopes. No significant correlation was observed with aa side chain bulkiness (Fig. 3.1 C). Most of the overrepresented and strongly bulky aas were also strongly hydrophobic. Cysteine, a non-polar hydrophobic aa was an outlier in the immunogenic dataset. Excluding cysteine did not change significantly our results (Fig. B.1 D).

Two potential sources of bias in our analyses were the variation in peptide length of MHC-I peptides and the dominance of human leukocyte antigen (HLA) A2 epitopes within existing databases. Therefore, we performed the analyses on the 9mer epitopes (Fig. B.1 B) and on HLA class I restricted peptides excluding HLA-A2 peptides (Fig. B.1 C). Within these subsets, the overrepresentation of non-polar, hydrophobic aas in immunogenic epitopes was maintained.

3.3.2 Hydrophobicity Bias in Selective TCR Contact Residues

We first compared the mean hydrophobicity of each residue between immunogenic and non-immunogenic peptides using the Kyte-Doolittle numeric hydrophobicity scale. Immunogenic 9mer epitopes were significantly more hydrophobic than non-immunogenic 9mer peptides at each residue ($P < 1.6 \times 10^{-5}$; (Fig. 3.2 A) and (Table B.4)). We observed similar results in 10mer peptides ($P < 2 \times 10^{-7}$ at every residue; (Fig. B.2 A)), and within HLA-A2 excluded 9mer and 10mer subsets (Figs. B.2 B and C). Because the immunogenic dataset is biased to pathogen-derived immunogenic epitopes, we performed similar analyses between immunogenic self-epitopes and non-immunogenic self-peptides ($P < 1 \times 10^{-4}$ at all residues, except P5 and P6; (Fig. B.2 D)). We further compared im-

munogenic HLA-A2 restricted 9mer epitopes derived from pathogens with those derived from self-antigens and observed no significant difference in hydrophobicity ($P > 0.05$ at each aa residue except P6, $P = 0.04$; (Fig. B.2 G)) revealing that T cells that escape thymic deletion recognize self-peptides with a hydrophobicity profile that is virtually the same as that of pathogen-derived epitopes. Lastly, to evaluate if there is potential bias created by using peptide immunization experiments, we did the same analysis using immunogenic epitopes identified using whole "Organism" as the immunogen ($P < 0.01$ at all residues except P1, P5 (Fig. B.2 E)). Thus, our results demonstrate a preference for hydrophobicity in immunogenic epitopes across antigenic sources (self and pathogen) and MHC molecules (HLA-A2 and non-HLA-A2).

The location of anchor residues and TCR contacts have been mapped for many MHC peptides (Rudolph et al., 2006). If the observed bias toward non-polar hydrophobic aas within immunogenic epitopes affects TCR affinity, we predicted that it would be selective for TCR contact residues. We analyzed the mean hydrophobicity along the peptide for the highest represented MHC epitopes within the database: HLA-A2 (Fig. 3.2 B), and murine MHC H-2D^b and H-2K^b (Fig. 3.2 C and D). HLA-A2 restricted 9mer peptides are anchored at residues P2 and P9, with P6 as an auxiliary anchor. We observed no statistical difference in hydrophobicity between the anchor residues of immunogenic and non-immunogenic peptides ($P2, P = 0.9; P9, P = 0.08$; (Fig. 3.2 B)). The observed difference in hydrophobicity was at specific TCR contact residues P4, P7 and P8 ($P4, P = 6.3 \times 10^{-12}; P7, P = 5 \times 10^{-13}; P8, P < 2.2 \times 10^{-16}$). In contrast, the auxiliary anchor P6 was more hydrophobic in non-immunogenic peptides ($P = 3.1 \times 10^{-7}$). Similar results were found for HLA-A2 restricted 10mer peptides (Fig. B.2 F).

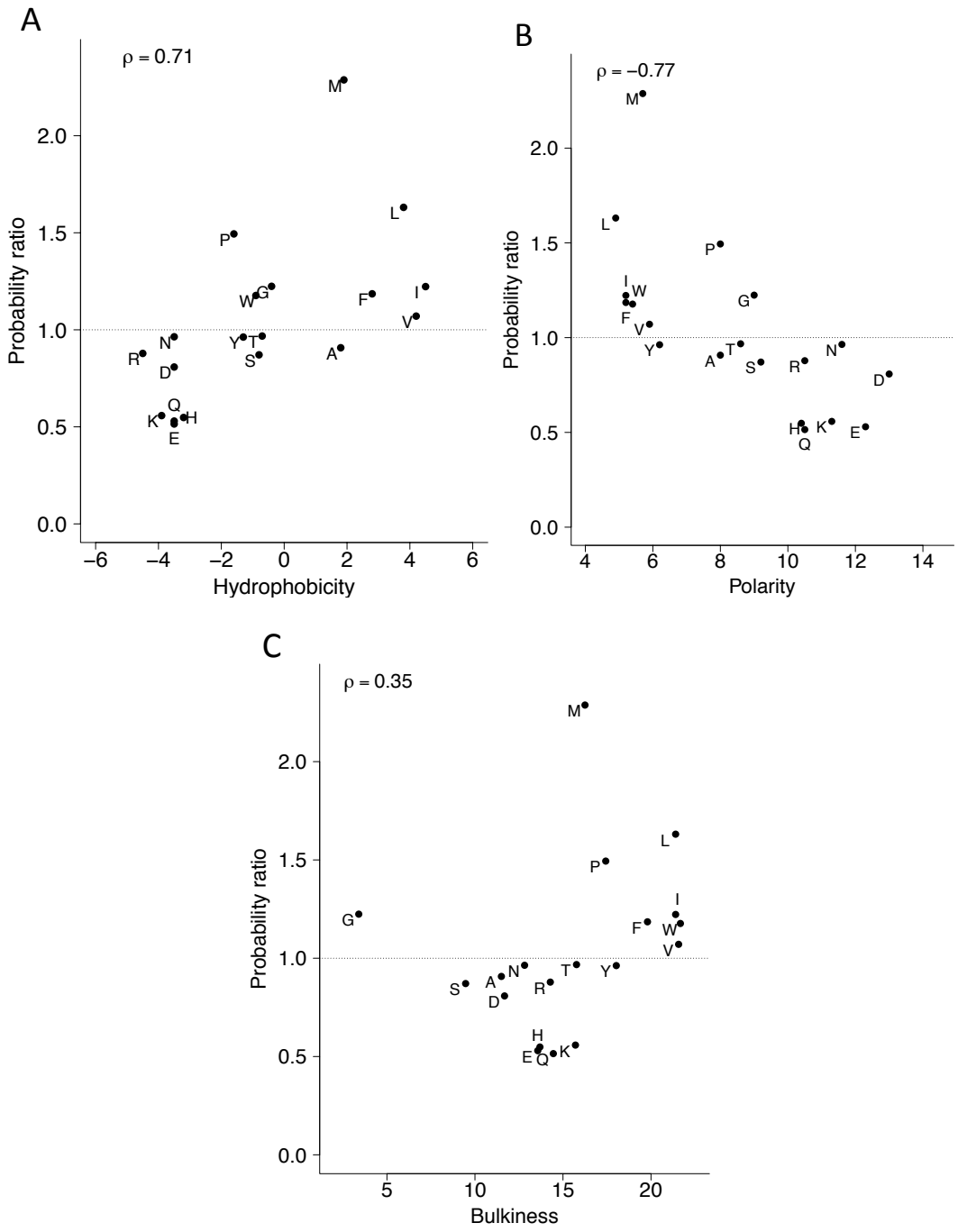


Figure 3.1: Probability ratio $[P(xIimmunogenic)/P(Inonimmunogenic)]$ of each amino acid as a function of its corresponding biochemical property. Each probability of each amino acid was computed from the frequency distribution of immunogenic epitopes and nonimmunogenic peptides. Biochemical properties analyzed were (A) hydrophobicity (Kyte and Doolittle, 1982), (B) polarity (Grantham, 1974), and (C) side-chain bulkiness (Zimmerman et al., 1968). A probability ratio > 1 indicates overrepresentation of the amino acid in the immunogenic dataset. The overrepresented outlier cysteine (C) was omitted for scale. Spearman correlations coefficients (ρ) are shown

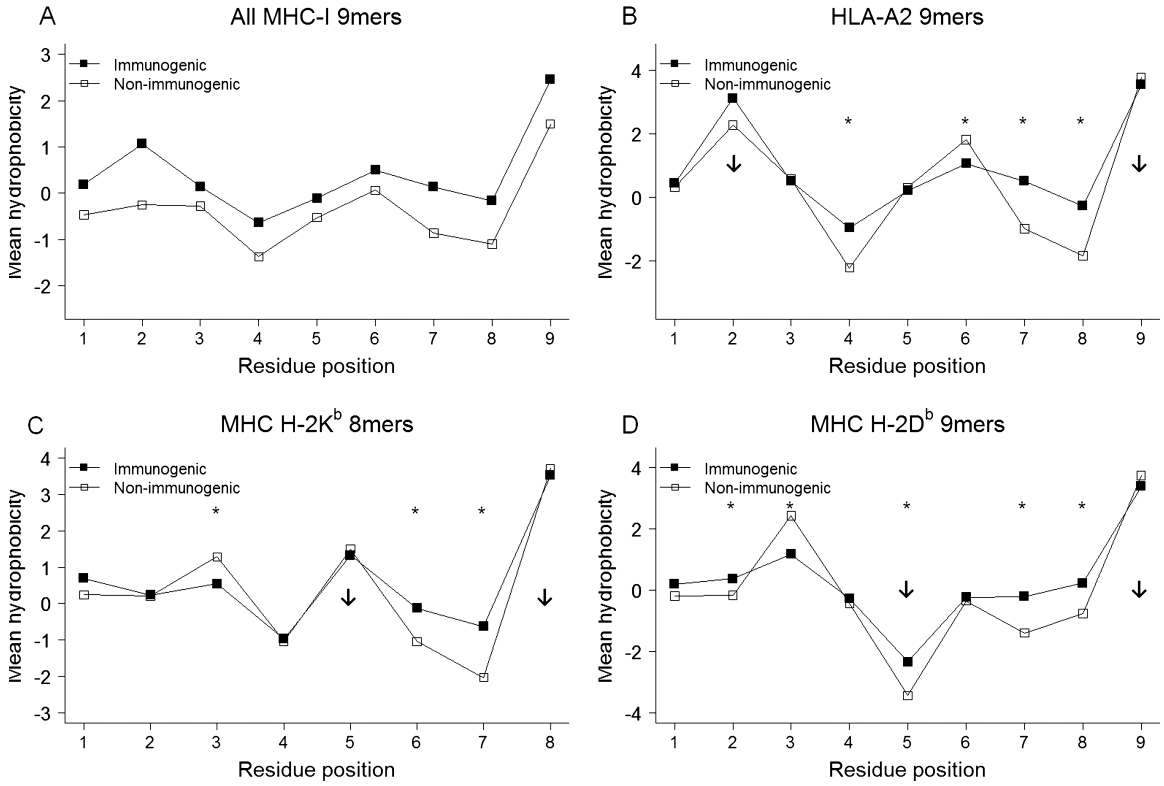


Figure 3.2: Hydrophobicity comparison at each residue position between immunogenic and non-immunogenic MHC-I peptides. Each peptide sequence in the dataset was transformed into a numeric sequence based on amino acid hydrophobicity, and the mean hydrophobicity at each position was calculated. (A) All immunogenic and nonimmunogenic MHC-I 9-mers; every residue had $P < 1.610^5$. (B) HLA-A2restricted immunogenic and nonimmunogenic 9-mers. (C) Murine MHC H-2D^{b}restricted immunogenic and nonimmunogenic 9-mers. (D) Murine MHC H-2K^{b}restricted immunogenic and nonimmunogenic 8-mers. Down-arrows in BD indicate anchor residues based on specific MHC motifs. $^*P < 0.008$ in that residue position. P values are listed in *SI* Appendix, (Table B.3).

To determine if the difference in hydrophobicity was species-specific, we evaluated the subset of known mouse MHC H-2K^b restricted 8mer peptides. Again, we observed a marked increase in relative hydrophobicity for the TCR contact residues P6 and P7 of immunogenic epitopes ($P6, P = 7 \times 10^{-5}; P7, P = 1.1 \times 10^{-6}$) but no difference in anchor residues ($P5, P = 0.67; P8, P = 0.15$; (Fig. 3.2 C)). As observed with HLA-A2, the auxiliary anchor residue P3 was more hydrophobic in non-immunogenic peptides ($P = 0.005$). Finally, we analyzed mouse MHC H-2D^b restricted 9mer peptides and observed that P7 and P8 TCR contact residues were more hydrophobic in immunogenic epitopes

($P7, P = 1.1 \times 10^{-4}$; $P8, P = 0.001$; (Fig. 3.2D)), with no difference in anchor residue P9 ($P = 0.127$). One exception was the anchor residue P5, which was more hydrophobic in immunogenic epitopes ($P = 4.9 \times 10^{-10}$). This discrepancy might be due to the presence of other potential anchors at P5 (apart from Asn) within immunogenic dataset. Hence we demonstrate that the observed bias towards relative hydrophobic aas in immunogenic epitopes is selective for TCR contact residues.

3.3.3 Differential Hydrophobicity Can Predict Immunogenic CTL Epitopes

While MHC-binding is necessary for antigen presentation, it is not sufficient to stimulate an immune response. We predicted that hydrophobicity could be incorporated into existing binding algorithms to improve prediction of CTL epitopes. To test this hypothesis, we used the IEDB-consensus binding prediction tool to generate peptide predictions for HLA-A2 restricted peptides (9, 10mers) for two viral proteins: polyprotein from dengue virus type 1 (DENV1) and tegument protein pp65 from cytomegalovirus (CMV). Using mean hydrophobicity of aas in TCR contact residues (all residues except anchors: P2, P6, and P9 or P10), each predicted peptide was re-ranked with decreasing TCR contact hydrophobicity values (Fig. 3.3). The rate at which experimentally defined HLA-A2 restricted CTL epitopes (Table B.5) were identified was increased using hydrophobicity-based predictions compared to IEDB-consensus binding predictions (Fig. 3.3 A and B). As a negative control, we performed re-ranking of top predictions from the 2 proteins using mean hydrophobicity of just anchor residues (Fig. 3.3 C and D). The rate of prediction of HLA-A2 restricted CTL epitopes was similar to IEDB-consensus binding predictions, confirming that relative hydrophobicity impacts immunogenicity and not HLA-binding. These results suggest that using TCR contact hydrophobicity could improve prediction of immunogenic epitopes.

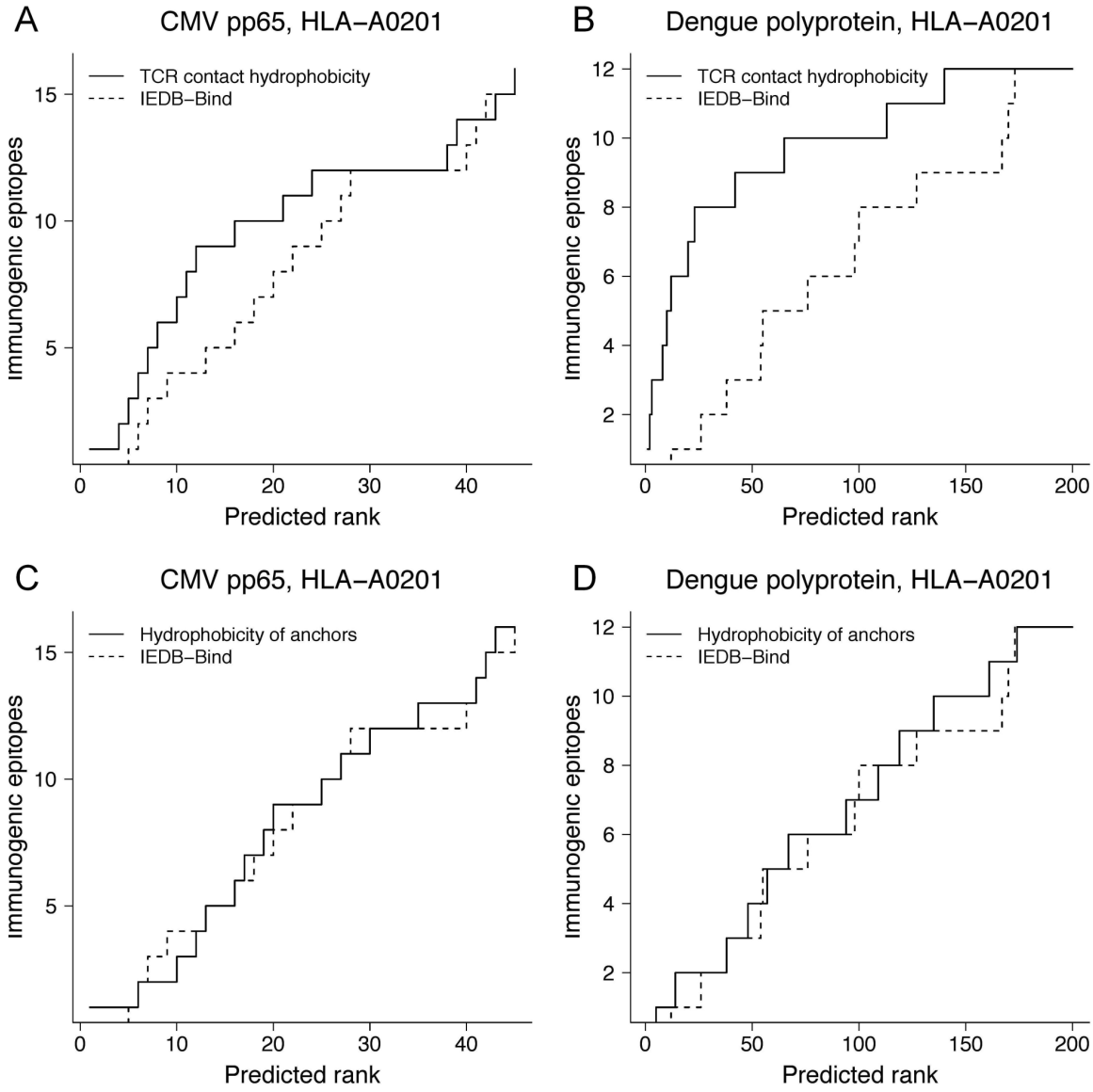


Figure 3.3: Efficiency of predicting experimentally defined HLA-A0201 restricted immunogenic epitopes using mean hydrophobicity of TCR contact residues (straight lines) compared to IEDB consensus binding tool (IEDB-Bind) are shown (dashed lines). Tegument protein pp65 from cytomegalovirus (CMV) and polyprotein from dengue virus type 1 were used for predictions. (A-B) Predicted peptides from the IEDB-Bind were re-ranked using mean hydrophobicity of TCR contact residues. (C-D) Predicted peptides from the IEDB-Bind were re-ranked using mean hydrophobicity of anchor residues.

3.3.4 Hydrophobicity-based ANN Prediction Model

The relative contribution of each aa residue to immunogenicity varies between MHC allomorph and is motif-dependent (Fig. 3.2) and (Table B.4). Furthermore, the immuno-

genicity of a peptide might result from nonlinear interactions between different TCR contact residues. Artificial neural networks (ANN) are designed to handle such nonlinearity (Honeyman et al., 1998) . Therefore, we developed and trained an ANN-based prediction model of immunogenicity using aa hydrophobicity (ANN-Hydro) with the goal of improving existing CTL epitope prediction algorithms for H-2D^b and HLA-A2. Each peptide sequence in the H-2D^b and HLA-A2 datasets was transformed into a corresponding numeric sequence based on the hydrophobicity value of aas and were used as the trainings sets for the two ANN-Hydro models (Fig. B.3). An initial assessment of the trained ANN-Hydro model for HLA-A2 assigned a good probability of immunogenicity to 54/64 ($> 80\%$) experimentally defined HLA-A2 restricted epitopes from 3 recent studies (Newell et al., 2013; Assarsson et al., 2008; Weiskopf et al., 2011) ($P < 0.001$, compared to the distribution of probabilities of immunogenicity of 64 randomly generated 9mer peptides) (Table B.6). We then developed an epitope discovery strategy incorporating the ANN-Hydro model to predict a previous set of experimentally validated H-2D^b and HLA-A2 epitopes from 5 pathogen and 5 tumor antigens (Table B.7). The IEDB-consensus MHC-binding prediction algorithm was used to obtain a list of predicted peptides for each antigen, which were each assigned a normalized binding score (SB). Since T cell epitopes are a subset of predicted peptides that bind to MHC molecules, normalized scores (SI) based on probabilities of immunogenicity obtained by ANN-Hydro were assigned to each peptide (Fig. B.3). We then defined a total score (S) as S = SB SI for the rate of identifying CTL epitopes from the list of predicted H-2D^b and HLA-A2 peptides from each antigen. The total score is therefore dependent on the contribution of both scores, reflecting two critical aspects: binding and immunogenicity (Fig. B.3). Our strategy of re-ranking by prioritization of high-binding and high-immunogenic peptides over other predicted peptides ((Fig. B.3), Materials and Methods), scored 42 out of the 43 H-2D^b and HLA-A2 9mer epitopes within the top 20 ranked peptides (Table B.7). In contrast individual prediction algorithms, ranked

the same epitopes up to rank 133 (Table B.7). Therefore, the ANN-Hydro model can be used in conjunction with IEDB-consensus to improve the efficiency of prediction of CTL epitopes.

3.3.5 Prediction Validation by in Vivo Discovery of HIV-1 Gag Epitopes

To comprehensively evaluate the predictive capacity of our approach for CTL epitope discovery and to correlate immunodominance, we interrogated 3 HIV-1 Gag variant proteins: Consensus B (ConsB), 96ZM651.8 (ZM96), and 97/CN54 (CN54) (Fig. 3.4). With no prior knowledge of Gag-specific CTL epitopes, our model was used to generate a list of ranked H-2D^b restricted peptides, of which the top 20 predictions for each interrogated Gag sequence are shown (Table B.8). To validate our predicted epitopes in vivo, B6 mice were immunized independently against each of the three different Gag variants and the peptide specificity of effector CD8$^+$ T cell responses analyzed using overlapping peptide pools (Table B.1). Deconvolution and truncation experiments allowed us to define a unique dominant H-2D^b-restricted epitope within each Gag protein (SI9 for ConsB, QL11 for CN54, RT9 for ZM96), as well as shared subdominant epitopes: Db-restricted RT9, AI9, YI9 and Kb-restricted VL8 (Fig. 3.4 F-H) and (Table B.2). Comparison of empirically defined epitopes to predictions made using ANN-Hydro revealed that H-2D^b restricted 9mer CTL epitopes for HIV-1 CN54 Gag and ZM96 Gag correlated with ANN-Hydro model epitope sequences predicted within the top 15 ranked peptides; and for ConsB Gag within the top 11 ranked peptides (Table B.7). In striking contrast, prediction of the identified Gag epitopes by individual prediction algorithms was more varied, with predictions up to rank number 46, depending on the binding or processing algorithm used. Although the IEDB consensus-binding and NetMHCpan algorithms predicted the identified Gag epitopes within the top 6 ranked peptides, the performance of these algorithms (unlike the ANN-Hydro model) was highly variant with the antigen selected (variance ranges from 66.72 to 220.27; (Ta-

ble B.7)). In sum, the ANN-Hydro model predicted 52 out of 53 experimentally validated H-2D^b and HLA-A2 9mer epitopes from 13 different antigens within the top 20 ranked peptides (Fig. 3.5), corresponding to a 98% success rate in identifying immunogenic epitopes. Moreover, this predictive improvement was reflected in lower variability of epitope identification, a variance of 37.72 using ANN-Hydro as opposed to 66.72 by IEDB alone ($P < 0.05$, F-test).

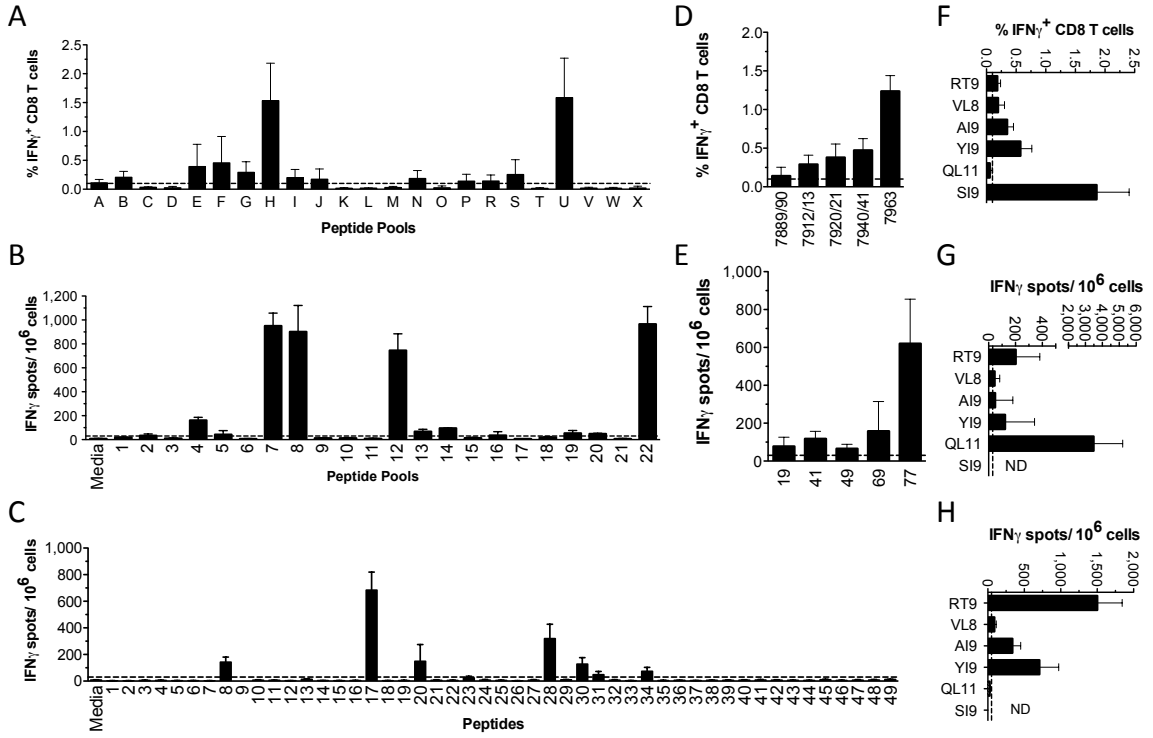


Figure 3.4: ANN-Hydro model prediction validation by *in vivo* discovery of HIV-1 Gag epitopes. Predictions for H-2D^b epitopes were made for 3 HIV-1 Gag proteins using the ANN-Hydro model and then a blinded epitope discovery study was performed in vivo. The top 20 predicted peptides for each protein using the model is shown in Table B.7. B6 mice were immunized with AdHu5 vaccines expressing the ConsB, CN54, or ZM96 Gag, and CD8$^+$ T cell responses determined by intracellular IFN- or IFN- ELISPOT after *ex vivo* stimulation with peptide pools of 15mer peptides (overlapping by 11mer) spanning the entire Gag sequence (ConsB or CN54, (Fig. 3.5 A and B)) or with a complete set of overlapping 20mer peptides spanning ZM96 (Fig. 3.4 C). Positive responses to pools were deconvoluted by stimulation with individual 15mer peptides from the positive pools (ConsB or CN54, (Fig. 3.5 D and E)). Minimal epitopes were identified by stimulation with truncated peptides and are shown (Fig. 3.4 F-H).

3.3.6 *Prediction of Immunodominant Epitopes*

The probabilities of immunogenicity assigned by ANN-Hydro were interrogated with respect to epitope immunodominance using three antigens with a clear vertical epitope hierarchy (identified by *ex vivo* experimental data) (Oukka et al., 1996; van der Most et al., 1998) and this study). The epitope hierarchy defined experimentally in LCMV-GP, Flu-NP, and ZM96 Gag showed robust correlation with the probabilities of immunogenicity assigned by ANN-Hydro ($r > 0.94, P < 0.05$; (Fig. 3.5 B-D)). In contrast, predicted MHC-binding assigned by IEDB-consensus showed no correlation with epitope immunodominance in LCMV-GP and ZM96-Gag (Fig. 3.5 E and G). Epitope immunodominance in Flu-NP correlated with both predicted ANN-Hydros probability and predicted MHC-binding (Fig. 3.5 C and F). As a further correlate, seven of 13 epitopes predicted in lower rankings by ANN-Hydro along with IEDB-consensus were modest immunogens derived from LCMV-GP, LCMV-NP, ZM96, CN54, and Consensus Gag (Table B.7). Therefore, efficient pMHC-TCR affinity may contribute towards epitope immunodominance. By using ANN-Hydro, epitope predictions were consistently less variable, and improved the prediction of immunodominant CTL epitopes.

3.4 Discussion

At present, there is no consensus on the molecular mechanisms by which CD8+ T cells discriminate immunogenic antigens within the background of poorly immunogenic self-peptides. Understanding this discrimination has implications in rational vaccine design and the identification of antigenic targets of malignant and autoimmune diseases. While several theories have been proposed to explain the concept of self/non-self discrimination (Pradeu and Carosella, 2006), the present study is the first attempt to provide a biochemical explanation for this fundamental phenomenon. We show that relative aa hydrophobicity

within immunogenic epitopes reveal an antigenic pattern that could be recognized by TCRs. We leveraged these findings to design an immunogenicity model, which was trained and validated using experimentally defined epitopes. ANN-Hydro consistently reduced variable standard prediction outputs across multiple antigens, demonstrating an important step forward in reducing the empirical element of T cell epitope prediction.

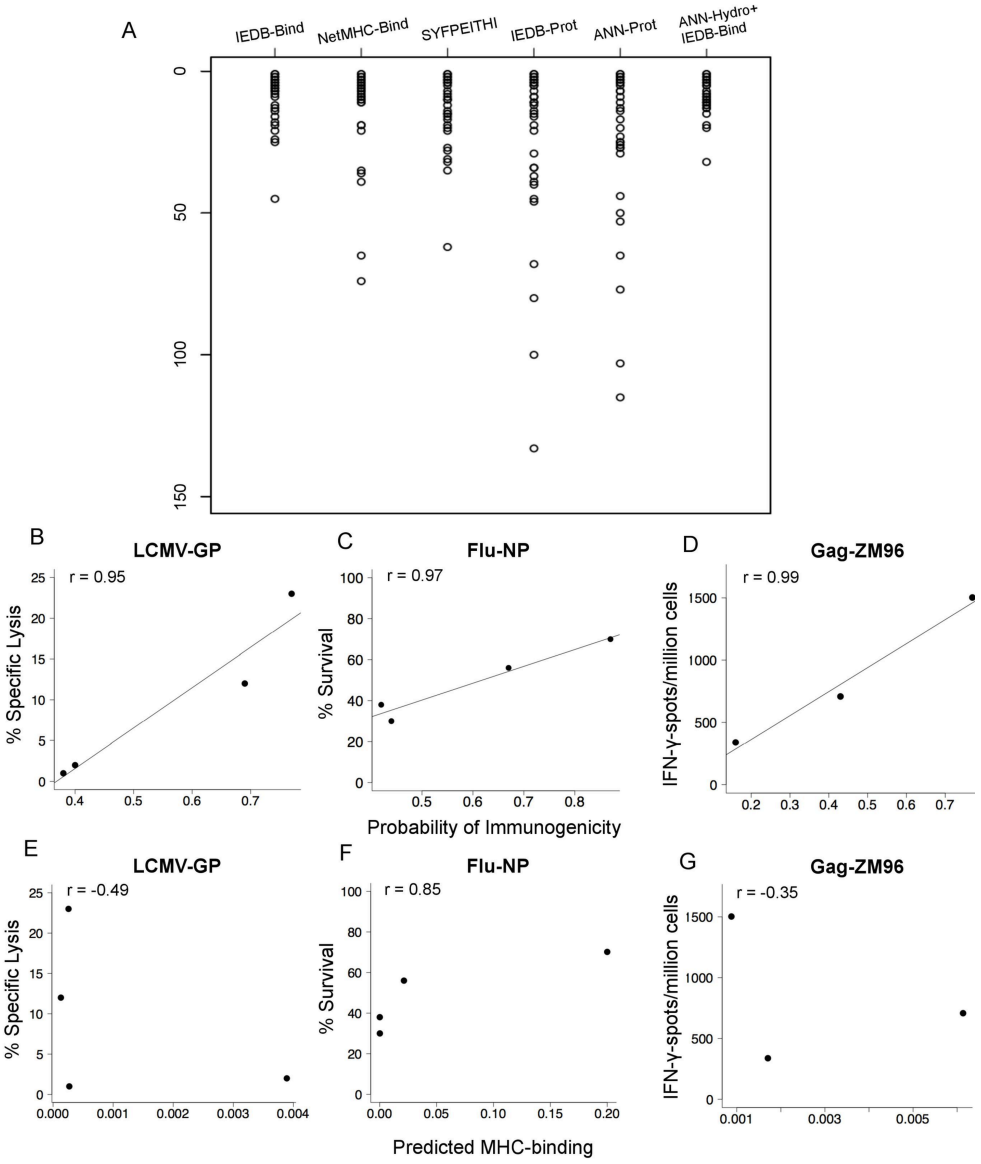


Figure 3.5: Incorporating ANN-Hydro in the IEDB-consensus binding tool improves epitope prediction. (A) Ranked epitopes are shown by scatter plots for 26 H-2D^b CTL epitopes from 8 well-described antigens: LCMV-GP, LCMV-NP, AdV.E1B, Flu-NP, Flu-NA, and 3 HIV-1 Gag variants (ConsB, ZM96, CN54); and for 27 HLA-A2 CTL epitopes from 5 tumor antigens: Melan-A, Wt-1, gp100, TRAG-3, and p53. Prediction algorithms used are: IEDB-Bind: IEDB consensus binding tool; NetMHC-Bind: NetMHCpan binding tool; SYFPEITHI: SYFPEITHI epitope prediction tool; IEDB-Prot: IEDB recommended processing prediction; ANN-Prot: IEDB processing predictions using ANN. Epitopes and their corresponding predicted ranks by prediction algorithms are shown in Table B.7. (B-D) Epitope immunodominance as a function of probability of immunogenicity for LCMV-GP, Flu-NP, and Gag-ZM96. (E-G) Epitope immunodominance as a function of predicted MHC-binding (IEDB-consensus) for LCMV-GP, Flu-NP, and Gag-ZM96. Immunodominance was determined from percentage-specific lysis of target cells *ex vivo* (B and E, 9mer versions of SGV11 and CSA10 were used) (van der Most et al., 1998), percentage-survival of peptide-primed mice upon lethal challenge of virus (C and F) (Oukka et al., 1996), or IFN--spots/million cells upon *ex vivo* peptide stimulation post-vaccination with antigen (D and G) (this study).

The majority of antigens within the immunogenic dataset used in this study are derived from intracellular pathogens, such as viruses, which have been shown to favor a lower G+C genomic content, reflected in their aa usage (Calis et al., 2010). Strongly hydrophobic aas (e.g. L, I, V, F, M) are characterized by low G+C codons while hydrophilic aas are not (Khrustalev and Barkovsky, 2011). This suggests the possibility that pathogens, in general, have a higher usage of hydrophobic aas that could be exploited for TCR recognition. A second possibility is that antigen presentation inherently favors hydrophobic regions within a protein. A recent study demonstrated that exposing hydrophobic domains significantly enhances the rate of proteasomal degradation and MHC presentation (Huang et al., 2011). Moreover, immunogenic CTL epitopes are also positionally biased towards the center of their source antigens (Kim et al., 2013), consistent with the fact that cytosolic proteins with a central hydrophobic core are the major substrates of proteasomal degradation. Thus, protein hydrophobicity can enhance both antigen presentation and immunogenicity, perhaps an evolutionary adaptation of hydrophobicity driven by damage-associated molecular patterns (Seong and Matzinger, 2004). TCRs are estimated to recognize on average about 5 non-anchor residues of a presented peptide because of the angle of peptide contact (Hennecke and Wiley, 2001; Burroughs et al., 2004). For 3 pMHC allomorphs analyzed by hydrophobicity in this study, only 4-5 positions on the peptide were significantly different between immunogenic and non-immunogenic peptides (Fig. 3.2), similar to published pMHC-TCR structures (Rudolph et al., 2006). This hydrophobicity difference is relative, not absolute. Certain aa positions in the peptide may be hydrophilic (e.g. P4 in HLA-A2 9mers, (Fig. 3.2 B)). However, even in such inherently hydrophilic residues in the peptide, immunogenic epitopes are less hydrophilic (more hydrophobic). Covering exposed hydrophobic residues on the peptide by a TCR may be a thermodynamically favorable process, facilitating the pMHC-TCR interaction as noted in retrospect, by a recent study (Birnbaum et al., 2014). TCR-engagement of pMHC complexes may be enhanced by water-exclusion from the im-

munological synapse or by increased K_{on} rates of the TCR-pMHC complex by relatively hydrophobic aas.

In the absence of a good understanding of the biochemical composition of peptide ligands that result in T cell activation, current strategies for epitope discovery either rely on the unbiased synthesis of a large number of overlapping peptides, or use MHC-binding/antigen processing algorithms to select candidate peptides. While the former is an expensive and laborious process, the latter results in a large number of false positive peptides that are not immunogenic. Advances in the development of combinatorial technologies have allowed the rapid identification and characterization of antigen-specific T cells (Newell and Davis, 2014). However, even such novel technologies rely on binding predictions to create candidate peptides lists that require extensive empirical validation. For instance, 77 candidate good-binders for HLA-A2 from the rotavirus proteome were chosen for recombinant pMHC tetramer production based on their MHC-binding capability, but only 6 (four being 9mer epitopes) were confirmed to be immunogenic epitopes (Newell et al., 2013). Therefore, T cell antigen discovery studies need strategies that improve the efficiency of epitope prediction. ANN-Hydro assigned high probabilities of immunogenicity to 80% of the HLA-A2 9mer epitopes described in the three proteome wide studies. Of note, three of the four rotavirus 9mer epitopes from the data set scored a probability of immunogenicity greater than 0.8 (Table B.6). In the HIV-1 Gag study, over 364 overlapping peptides were tested in vivo from the Gag variants (length: 500aa) for epitope discovery. Using the ANN-Hydro model combined with SB scores narrows the validation discovery process down to 11-15 peptides per Gag protein to be tested. Similarly, applying ANN-Hydro also improved predictions of immunogenic H-2D^b and HLA-A2 epitopes from 10 independent antigens compared to individual prediction algorithms. Thus, models such as ANN-Hydro adds an extra dimension (immunogenicity) to MHC-binding for CTL epitope prediction and could be used to significantly reduce the variability associated with standard predic-

tion algorithms, and the time and cost of experimental validation (Fig. 3.5), (Table B.7). With the advent of tumor exome sequencing in immune therapy settings, we anticipate that immunogenicity models such as ANN-Hydro will be critical in identifying immunogenic neo-antigens for tumor immune therapies (Snyder et al., 2014; Yadav et al., 2014).

The ANN-Hydro model differs from existing MHC-binding/antigen-processing prediction algorithms in two respects: First, the ANN-Hydro was trained on a relative hydrophobicity scale, which facilitates the model to discover complex numeric relationships between different aa residues. Second, the dataset used for training was immunogenic epitopes and non-immunogenic self-peptides, which do not differ in binding motifs but only in immunogenicity. While some high-binding epitopes (e.g. SI9 from ConsB) are readily predicted by all algorithms, other epitopes (e.g. the immunodominant dominant RT9 from ZM96, LL9 from LCMV-GP) are predicted at variable rankings by different algorithms (Table B.7). In comparison, ANN-Hydro rescued these epitopes by virtue of their probability of immunogenicity. Although ANN-Hydro marks a step forward in efficiently predicting 9mer epitopes, it is currently limited in terms of predicting longer or shorter epitopes, exemplified by the 11mer epitope (QL11) deduced by epitope mapping from the CN54 Gag protein. To improve longer or shorter epitope predictions, larger representative datasets are required for training. Nonetheless, the model predicted a 9mer version of this epitope ranked at 35 and 44, which is consistent with presentation of nested length peptides (Riemer et al., 2010). A second limitation of the current model is its applicability to predict epitopes for other HLA class I alleles. In theory, the ANN-Hydro model could be applied to predict CTL epitopes for any MHC class I allele, but large representative datasets are required for training the model for representative MHC allomorphs. We anticipate that advances in mass spectrometry-based MHC peptide discovery will result in more extensive training databases for predicting longer and shorter epitopes from a broader selection of HLA class I molecules (Riemer et al., 2010; Tan et al., 2011).

While immunogenicity models have been developed by others for prediction of CTL epitopes (Calis et al., 2013; Harndahl et al., 2012), they considered only the impact of pMHC stability and positional significance along the peptide for immunogenicity. In contrast, a crucial feature of our approach is the use of ligand-eluted non-immunogenic self-peptides as the comparator set. Because binding and antigen processing are required for all epitopes, we built upon existing algorithms for immunogenic pMHC predictions. "Layering" the immunogenicity model on top of existing prediction algorithms enabled us to predict epitopes with increased effectiveness than standalone predictions. Importantly, the empirical evaluation of our immunogenicity model and epitope prediction approach without a *priori* knowledge of the immunodominant HIV-1 Gag epitopes *in vivo* gives strong support for these results. In summary, integrating aa hydrophobicity into pMHC prediction algorithms should significantly enhance the success of epitope discovery. The biological mechanism underlying TCR preferences for non-polar hydrophobic residues remains to be evaluated.

3.5 Materials and Methods

For full details of methods and construction of data sets see *SI* Appendix.

3.5.1 Construction of Datasets

All MHC-I peptides used in this study and design of the ANN-Hydro prediction model were retrieved from IEDB (Vita et al., 2010) (www.iedb.org, last accessed: 08/11/2013). Epitopes with a positive T cell response represent the immunogenic epitope group. The non-immunogenic self-peptide group represent cell surface ligand eluted MHC-I self-peptides that have been antigenically processed and MHC-bound. Additional curation and exclusion criteria resulted in a final dataset with 5,035 8-11mer immunogenic epitopes and 4,853 8-11mer non-immunogenic peptides (See *SI* for further details.

3.5.2 Amino Acid Scales

These were derived from Expasys ProtScale (http://web.expasy.org/protscale/) (Gasteiger et al., 2005), specifically, Hydrophobicity scale (Kyte and Doolittle) (Kyte and Doolittle, 1982), Polarity (Grantham) (Grantham, 1974), and Bulkiness (Zimmerman) (Zimmerman et al., 1968). The scales are relative, e.g., negative to positive values in the hydrophobicity scale correspond to a relative hydrophobicity increase between aas (Table B.3).

3.5.3 Position-based Hydrophobicity Analysis

We transformed our datasets of immunogenic and non-immunogenic peptides into numeric arrays using the R statistical software (RDevelopment, 2012). Separate numeric arrays were generated for immunogenic and non-immunogenic 8, 9 and 10mers. Mean hydrophobicity of immunogenic and non-immunogenic peptides at each position was calculated and were compared residue-by-residue through Wilcoxon rank-sum tests to quantify statistical significance.

3.5.4 Hydrophobicity-based ANN Prediction Model (ANN-Hydro)

The R neuralnet package was used to design and train the two ANN-Hydro models on H-2D^b and HLA-A2 restricted 9mer peptides known to be immunogenic ($n = 204$ and $n = 374$, respectively) or non-immunogenic ($n = 232$ and $n = 201$, respectively). Each peptide sequence in the respective H-2D^b and HLA-A2 datasets were transformed into a corresponding numeric sequence based on aa hydrophobicity using R statistical software. A three-layer fully connected feed-forward ANN was comprised by nine input neurons, one hidden layer with three neurons, and one output variable (Fig. B.3).

3.5.5 Application of ANN-Hydro

For each H-2D^b and HLA-A2 restricted epitope prediction, we used IEDB-consensus to generate a list of epitope predictions. Each peptide was assigned a normalized binding score (SB) and a subset of these predicted peptides was then selected by defining a SB -threshold of 0.1 for antigen length $>$ 100 aas and a SB -threshold of 0.2 for antigen length $\leq$ 100 aas. Independently, probabilities of immunogenicity were obtained by applying ANN-Hydro to this subset of binding predictions. Normalized scores (SI) were then assigned based on the probabilities of immunogenicity (Fig. B.3). The list of predicted peptides was ranked based on a total score, S = SB SI, ranging from lowest score to the highest score. The lower the total score of a predicted peptide, the higher its probability of being an immunogenic epitope. See *SI* for details.

3.5.6 Vaccines

Recombinant Adenovirus type 5 (rAdHu5) vectors encoding codon optimized HIV-1 Gag from Cons B, strain 96ZM651.8 (ZM96) and strain 97CN54 (CN54) (Bachy et al., 2013) are described in *SI Materials and Methods*.

3.5.7 Immunization of Mice

C57BL/6 mice were immunized with 10^9 virus particles. All animal studies were conducted in accordance with UK Home Office regulations and Kings College London ethics committee.

3.5.8 Peptides

15mer peptides spanning HIV-1 CN54 Gag and a 20mer set of peptides spanning HIV-1 ZM96 were provided by the UK Centre for AIDS Reagents. 15mer peptides spanning HIV-

1 Cons B Gag were provided by the NIH AIDS Reagent Reference Program. Truncated HIV-1 Gag peptides were purchased from Proimmune.

3.5.9 T Cell Epitope Mapping

Spleen cells were re-stimulated either with media alone or with peptides, either in pools or individually (each at 1M final concentration) and IFN- production was detected by intracellular cytokine staining or by ELISPOT assay as previously described (Bachy et al., 2013). Cons B and CN54 Gag epitopes were deconvoluted to individual 15mers from peptide pools, and truncated versions of the 15mer peptides were synthesized and tested. For ZM96 Gag, 49 individual 20mer peptides were tested. Reactive peptide sequences were confirmed against the corresponding 15mer peptide to the reactive sequence and 9mer peptides were synthesized and tested.

Chapter 4

MODELING THE EFFECT OF EARLY DETECTION OF EBOLA VIRUS DISEASE

This chapter is an extension of the work published in the paper:
Modelling the effect of early detection of Ebola
Diego Chowell, Carlos Castillo-Chavez, Sri Krishna, Xiangguo Qiu, Karen S. Anderson
The Lancet Infectious Diseases 15.2 (2015): 148-149

Abstract

The most recent Ebola outbreak in West Africa highlighted critical weaknesses in the medical infrastructure of the affected countries, including effective diagnostics tools, sufficient isolation wards, and enough medical personnel. Here, we develop and analyze a mathematical model to assess the impact of early diagnosis of pre-symptomatic individuals on the transmission dynamics of Ebola virus disease in West Africa. Our findings highlight the importance of implementing integrated control measures of early diagnosis and isolation. The mathematical analysis shows a threshold where early diagnosis of pre-symptomatic individuals, combined with a sufficient level of effective isolation, can lead to an epidemic control of Ebola virus disease.

4.1 Introduction

The Ebola viral strains are re-emerging zoonotic pathogens and members of the Filoviridae family consisting of five distinct species: *Bundibugyo, Cotes d'Ivoire, Reston, Sudan*, and *Zaire* with a high case-fatality rate in humans (Kugelman et al., 2015). Filoviruses are long filamentous enveloped, non-segmented, single-stranded viruses, consisting of a negative-sense RNA genome (Beeching et al., 2014). Each Ebola species genome encodes

seven linearly arranged genes: nucleoprotein (NP), polymerase cofactor (VP35), matrix protein (VP40), glycoprotein (GP), replication-transcription protein (VP30), matrix protein (VP24), and RNA-dependent RNA prolymerase (L) (Beeching et al., 2014). While there are no proven effective vaccines or effective antiviral drugs for Ebola, containing an outbreak relies on contact tracing and on early detection of infected individuals for isolation and care in treatment centers (Beeching et al., 2014). The most recent Ebola outbreak in West Africa, which began in December 2013, due to the Zaire strain, demonstrated several weaknesses in the medical infrastructure of the affected countries, including the urgent need of effective diagnostics, which have a fundamental role in both disease control and case management.

The Ebola virus is transmitted as a result of direct contact with bodily fluids containing the virus (Fauci, 2014). The virus enters via small skin lesions and mucus membranes where it is able to infect macrophages and other phagocytic innate immune cells leading to the production of a large number of viral particles (Beeching et al., 2014). The macrophages, monocytes, and dendritic cells infected in the early stage of the disease serve to spread the virus throughout the organs, particularly in the spleen, liver, and lymph nodes (Beeching et al., 2014). Consequently, critically ill patients display intensive viremia (McElroy et al., 2014). Recognizing signs of Ebola viral disease is challenging because it causes common non-specific symptoms such as fever, weakness, diarrhea, and vomiting, and the incubation period typically lasts 5 to 7 days (Fauci, 2014). Therefore, functioning laboratories and effective point-of-care diagnostic tests are critically needed in order to minimize transmission, allow better allocation of scarce healthcare resources, and increase the likelihood of success of antiviral treatments as they are developed (Fink, 2015).

There is an ongoing effort in place to improve Ebola diagnostics, primarily to detect the disease early. In our current state, the cost and difficulty of testing limit diagnostic facilities to small mobile laboratories or centralized facilities with turnaround times measured in days rather than in a few hours, meaning that diagnosis is largely used to confirm disease. Ebola diagnosis can be achieved in two different ways: measuring the host-specific immune response to infection (e.g. IgM and IgG antibodies) and detection of viral particles (e.g. ReEBOV Antigen Rapid Test Kit for VP40), or particle components in infected individuals (e.g. RT-PCR or PCR). The most general assay used for IgM and IgG antibody detection are direct ELISA assays. Considering the physiological kinetics of the humoral immune system as well as impaired antigen-presenting cell function as a result of viral hemorrhagic fever, antibody titers are low in the early stages and often undetectable in severe patients prior to death (Ippolito et al., 2012). This leaves polymerase chain reaction (PCR) for antigen detection as a viable option for early diagnostic assays. PCR is a chemical reaction that amplifies pieces of a virus's genes floating in the blood by more than a millionfold, which makes detection of pre-symptomatic individuals likely identifiable. Indeed, a research article published in 2000, illustrates the power of this technology to detect Ebola virus in humans in the pre-symptomatic stage (Leroy et al., 2000). In this study, 24 asymptomatic individuals who had been exposed to symptomatic Ebola patients were tested using PCR. Eleven of the exposed patients eventually developed the infection. Seven of the 11 tested positive for the PCR assay. And none of the other 13 did.

In this chapter, we extend the work presented in (Chowell et al., 2015). Here, we have developed and analyzed a mathematical model to evaluate the impact of early diagnosis of pre-symptomatic individuals on the transmission dynamics of Ebola virus disease in West Africa, under the assumption that the disease is maintained possibly at very low levels due to the deficiencies in health systems and our incomplete understanding of Ebola infection

as illustrated by the case of Pauline Cafferkey. Therefore, eliminating Ebola may require a more sustained and long-term control effort.

4.2 Model Formulation

The total population is assumed to be classified into six mutually exclusive subgroups: susceptible $S(t)$, non-detectable latent $E_1(t)$, detectable latent $E_2(t)$, infectious $I(t)$, isolated $J(t)$, and recovered $R(t)$ individuals. Table 4.1 shows the state variables and their physical meaning. The transition between all these states is shown in Fig. 4.1. And model parameters and their description are presented in Table 4.2. Parameter values have been obtained from previous studies (Chowell and Nishiura, 2014; Fasina et al., 2014).

It is assumed that individuals are recruited (either through birth or migration) into the susceptible class at a rate Λ and die naturally with rate μ. Susceptible individuals get infected due to successful contacts with infectious or not perfectly isolated infected individuals at rate λ. As a consequence, they become latent undetectable, who develop their state of infection to become latent detectable at rate κ_1. We assume that the latent detectable class represent individuals whose viral load is above the detection limit of the PCR-based diagnostic test (Leroy et al., 2000; Qiu et al., 2014) . Latent detectable individuals either are diagnosed and get isolated with probability f_T or develop symptoms to become infectious, who sequentially either get isolated at rate α, or are removed from the system by recovery or Ebola-induced death at rate γ. It is assumed here that Ebola-induced deaths occur for the infectious individuals with probability q_1. Similarly, isolated individuals leave their class at rate γ_r, by either dying due to Ebola with probability q_2, or they recover and become immune. It is assumed that isolation is partially effective so that successful contacts with susceptible individuals may lead to infection with probability r; this parameter is a measure

of isolation effectiveness of infectious individuals. Thus, the force of infection is given by

$$\lambda(t) = \frac{\beta[I(t) + (1-r)\ell J(t)]}{N(t) - rJ(t)}. \tag{4.1}$$

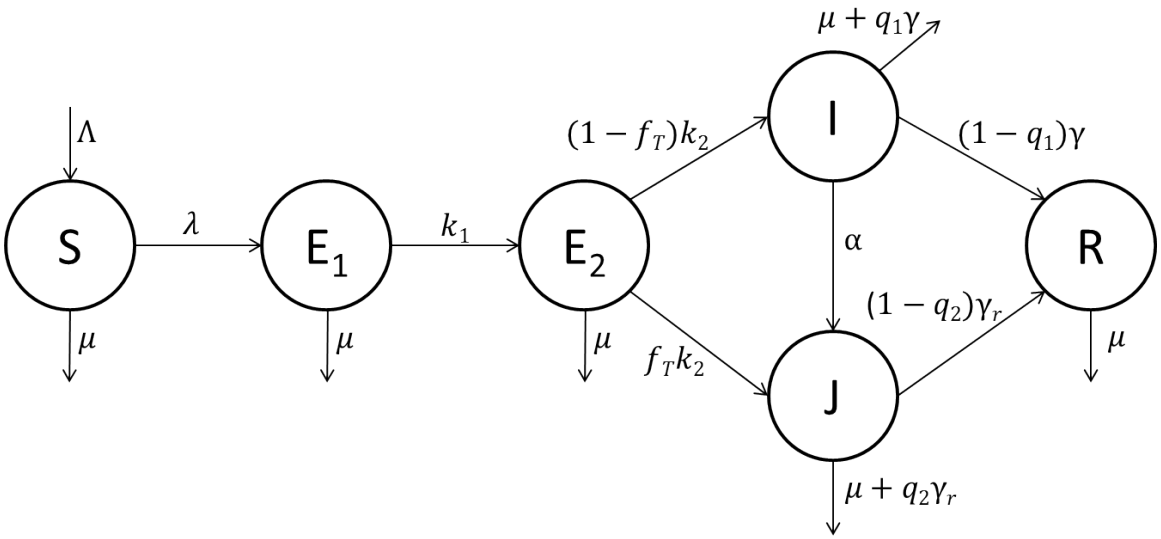


Figure 4.1: Compartmental Model Showing the Transition Between Model States.

Table 4.1: Definition of model states.

Variable	Description
$S(t)$	Number of susceptible individuals at time t
$E_1(t)$	Number of latent undetectable individuals at time t
$E_2(t)$	Number of latent detectable individuals at time t
$I(t)$	Number of infectious individuals at time t
$J(t)$	Number of isolated individuals at time t
$R(t)$	Number of recovered individuals at time t

Table 4.2: Definition of model parameters.

Parameter	Value	Unit	Description
Λ	17182	$\frac{population}{day}$	Recruitment rate
β	0.3335	day^{-1}	Mean transmission rate
μ	4.98×10^{-5}	day^{-1}	Natural death rate
κ_1	$1/4$	day^{-1}	Transition rate from undetectable to detectable latent state
κ_2	$1/3$	day^{-1}	Exit rate of latent detectable individuals by either becoming infectious or moving to isolation state
γ	$1/6$	day^{-1}	Removal rate of infectious individuals by either recovery or Ebola-induced death
γ_r	$1/7$	day^{-1}	Removal rate of isolated individuals by either recovery or Ebola-induced death
α	$1/5$	day^{-1}	Rate at which infectious individuals get isolated
f_T	$0.25 \in [0,1]$	–	Fraction of latent detectable individuals who are diagnosed and get isolated
q_1	0.7	–	Probability that an infectious individual dies due to Ebola
q_2	0.63	–	Probability that an isolated individual dies due to Ebola
r	$0.35 \in [0,1]$	–	Effectiveness of isolation
ℓ	$0.5 \in [0,1]$	–	Relative transmissibility of isolated individuals with respect to infectious individuals

The assumptions mentioned above lead to the following model of equations

$$\begin{aligned}
\frac{dS}{dt} &= \Lambda - \lambda S - \mu S, \\
\frac{dE_1}{dt} &= \lambda S - (\kappa_1 + \mu) E_1, \\
\frac{dE_2}{dt} &= \kappa_1 E_1 - (\kappa_2 + \mu) E_2, \\
\frac{dI}{dt} &= (1 - f_T)\kappa_2 E_2 - (\alpha + \gamma + \mu) I, \\
\frac{dJ}{dt} &= f_T \kappa_2 E_2 + \alpha I - (\gamma_r + \mu) J, \\
\frac{dR}{dt} &= (1 - q_1)\gamma I + (1 - q_2)\gamma_r J - \mu R
\end{aligned} \tag{4.2}$$

where

$$N(t) = S(t) + E_1(t) + E_2(t) + I(t) + J(t) + R(t)$$

is the total population size at time t. On adding all equations of system (4.2) together, we get

$$\frac{dN}{dt} = \Lambda - \mu N - q_1\gamma I - q_2\gamma_r J. \tag{4.3}$$

4.3 Model Analysis

4.3.1 Basic properties

Since model (4.2) imitates the dynamics of human populations, all variables and parameters should be non-negative. Thus, following the approach shown in appendix A of (Thieme, 2003), we show the following result.

Theorem 1. *The variables of model (4.2) are non-negative for all time.*

Lemma 1. *The closed set*

$$\Omega = \left\{(S, E_1, E_2, I, J, R) \in \mathbb{R}^6_+ : \frac{\Lambda}{\mu + q_1\gamma + q_2\gamma_r} \leq S + E_1 + E_2 + I + J + R \leq \frac{\Lambda}{\mu}\right\}$$

is positively invariant for model (4.2) and is absorbing.

Proof: Equation (4.3) implies that

$$\frac{dN}{dt} \leq \Lambda - \mu N, \tag{4.4}$$

$$\frac{dN}{dt} \geq \Lambda - (\mu + q_1\gamma + q_2\gamma_r)N. \tag{4.5}$$

It follows from (4.4) that

$$N(t) \leq \frac{\Lambda}{\mu} + \left(N(0) - \frac{\Lambda}{\mu}\right) e^{-\mu t} \tag{4.6}$$

and from (4.5) that

$$N(t) \geq \frac{\Lambda}{\mu + q_1\gamma + q_2\gamma_r} + \left(N(0) - \frac{\Lambda}{\mu + q_1\gamma + q_2\gamma_r}\right) e^{-(\mu + q_1\gamma + q_2\gamma_r)t}. \tag{4.7}$$

If we assume $N(0) > \Lambda/\mu$, then $dN/dt < 0$ and therefore (based on inequality (4.6)), $N(t)$ decreases steadily until reaching Λ/μ when t tends to ∞. Similarly, if we assume $N(0) < \Lambda/(\mu + q_1\gamma + q_2\gamma_r)$, then $dN/dt > 0$ and therefore (based on inequality (4.7)), $N(t)$ increases steadily until reaching a maximum at $\Lambda/(\mu + q_1\gamma + q_2\gamma_r)$ when t tends to ∞. It remains to check the case if $N(0)$ lies in the phase between $\Lambda/(\mu + q_1\gamma + q_2\gamma_r)$ and Λ/μ. To this end, both inequalities (4.6) and (4.7) are combined together to get

$$\frac{\Lambda}{\mu + q_1\gamma + q_2\gamma_r} + \left(N(0) - \frac{\Lambda}{\mu + q_1\gamma + q_2\gamma_r}\right) e^{-(\mu + q_1\gamma + q_2\gamma_r)t} \leq N(t) \leq \frac{\Lambda}{\mu} + \left(N(0) - \frac{\Lambda}{\mu}\right) e^{-\mu t}.$$

On taking the limit when t tends to ∞, we find that $N(t)$ remains within the same phase. Thus, the set Ω is positively invariant and absorbing.

4.3.2 Equilibrium Analysis

Ebola-free equilibrium and the control reproduction number $\mathscr{R}_c$

It is easy to check that model (4.2) has the Ebola-free equilibrium

$$E_0 = \left(\frac{\Lambda}{\mu}, 0, 0, 0, 0, 0\right)' \tag{4.8}$$

where the prime " $'$ " means vector transpose.

The basic reproduction number, $\mathscr{R}_0$, is a measure of the average number of secondary cases produced by a typical infectious individual during the entire course of infection in a completely susceptible population and in the absence of control interventions (Brauer et al., 2001; Anderson et al., 1992). On the other hand, the control reproduction number, $\mathscr{R}_c$, quantifies the potential for infectious disease transmission in the context of a partially susceptible population due to the implementation of control interventions. When $\mathscr{R}_c > 1$, the infection may spread in the population, and the rate of spread is higher with increasingly high values of $\mathscr{R}_c$. If $\mathscr{R}_c < 1$, infection cannot be sustained and is unable to generate an epidemic. For our model, $\mathscr{R}_c$ is computed using the next generation matrix approach shown

in (Van den Driessche and Watmough, 2002). Accordingly, we compute the matrices $\mathbf{F}$ (for the new infection terms) and $\mathbf{V}$ (for the transition terms) as

$$\mathbf{F} = \begin{pmatrix} 0 & 0 & \beta & (1-r)\ell\beta \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \quad , \quad \mathbf{V} = \begin{pmatrix} \kappa_1+\mu & 0 & 0 & 0 \\ -\kappa_1 & \kappa_2+\mu & 0 & 0 \\ 0 & -(1-f_T)\kappa_2 & \alpha+\gamma+\mu & 0 \\ 0 & -f_T\kappa_2 & -\alpha & \gamma_r+\mu \end{pmatrix}.$$

Thus, the control reproduction number is given by

$$\begin{aligned} \mathscr{R}_c &= \rho(\mathbf{F}\mathbf{V}^{-1}) = \frac{\kappa_1\kappa_2\beta[(1-f_T)(\mu+\gamma_r)+(1-r)\ell(\alpha+f_T(\gamma+\mu))]}{(\kappa_1+\mu)(\kappa_2+\mu)(\alpha+\gamma+\mu)(\gamma_r+\mu)} \\ &= \frac{\kappa_1\kappa_2\beta}{(\kappa_1+\mu)(\kappa_2+\mu)(\alpha+\gamma+\mu)}\left[1-f_T+(1-r)\ell\left(\frac{\alpha}{\gamma_r+\mu}+f_T\frac{\gamma+\mu}{\gamma_r+\mu}\right)\right] \\ &= \mathscr{R}_0\left[1-\frac{\alpha}{(\alpha+\gamma+\mu)}\right]\left[1-f_T+(1-r)\ell\left(\frac{\alpha}{\gamma_r+\mu}+f_T\frac{\gamma+\mu}{\gamma_r+\mu}\right)\right] \end{aligned} \tag{4.9}$$

where ρ is the spectral radius (dominant eigenvalue in magnitude) of the matrix $\mathbf{F}\mathbf{V}^{-1}$ and

$$\mathscr{R}_0 = \frac{\kappa_1\kappa_2\beta}{(\kappa_1+\mu)(\kappa_2+\mu)(\gamma+\mu)} \tag{4.10}$$

is the basic reproduction number for the model.

The local stability of the Ebola-free equilibrium, E_0, for values of $\mathscr{R}_c < 1$ is established based on a direct use of Theorem 2 in (Van den Driessche and Watmough, 2002). We summarize our result in the following lemma.

Lemma 2. *The Ebola-free equilibrium E_0 of model (4.2) is locally asymptotically stable if and only if $\mathscr{R}_c < 1$.*

Ebola-endemic equilibrium

On putting the derivatives in the left hand side of (4.2) equal zero and solving the resulting algebraic system with respect to the variables $\bar{S}, \bar{E}_1, \bar{E}_2, \bar{I}, \bar{J}$, and $\bar{R}$, we obtain

$$\begin{aligned}
\bar{S} &= \frac{\Lambda}{\bar{\lambda}+\mu}, \\
\bar{E}_1 &= \frac{\Lambda}{\bar{\lambda}+\mu}\cdot\frac{\bar{\lambda}}{\kappa_1+\mu}, \\
\bar{E}_2 &= \frac{\kappa_1}{\kappa_2+\mu}\cdot\frac{\Lambda}{\bar{\lambda}+\mu}\cdot\frac{\bar{\lambda}}{\kappa_1+\mu}, \\
\bar{I} &= \frac{(1-f_T)\kappa_2}{\alpha+\gamma+\mu}\cdot\frac{\kappa_1}{\kappa_2+\mu}\cdot\frac{\Lambda}{\bar{\lambda}+\mu}\cdot\frac{\bar{\lambda}}{\kappa_1+\mu}, \\
\bar{J} &= \frac{\kappa_1}{\kappa_2+\mu}\cdot\frac{\Lambda}{\bar{\lambda}+\mu}\cdot\frac{\bar{\lambda}}{\kappa_1+\mu}\cdot\frac{\kappa_2}{\gamma_r+\mu}\left[f_T+(1-f_T)\frac{\alpha}{\alpha+\gamma+\mu}\right], \\
\bar{R} &= \frac{1}{\mu}[(1-q_1)\gamma I+(1-q_2)\gamma_r J]
\end{aligned} \tag{4.11}$$

where

$$\bar{\lambda} = \frac{\beta(I+(1-r)\ell\bar{J})}{\bar{N}-r\bar{J}} \tag{4.12}$$

is the equilibrium force of infection. On substituting from (4.11) into (4.12) and simplifying (with the assumption that $\lambda \neq 0$), we get

$$\bar{\lambda} = \frac{\mu(\mathscr{R}_c-1)}{1-Term} \tag{4.13}$$

where

$$Term = \frac{\kappa_1\kappa_2[q_1(1-f_T)\gamma(\gamma_r+\mu)+(r\mu+q_2\gamma_r)(f_T(\gamma+\mu)+\alpha)]}{(\kappa_1+\mu)(\kappa_2+\mu)(\alpha+\gamma+\mu)(\gamma_r+\mu)}.$$

Hence, the Ebola-endemic equilibrium is unique and we show the following lemma.

Lemma 3. *Model (4.2) has a unique endemic equilibrium that exists if and only if* $\mathscr{R}_c > 1$.

4.3.3 Normalized Sensitivity Analysis on $\mathscr{R}_c$

In considering the dynamics of the Ebola system (4.2), we conduct normalized sensitivity analysis on $\mathscr{R}_c$ to determine the impact of parameter perturbations on the transmission

dynamics of the system. By computing the normalized sensitivity indices, we consider the percent change in the output with respect to a percent change in the parameter input. Those parameters with the largest magnitude of change impact the compartment model the most; the sign indicates whether the change produces an increase or a decrease on $\mathscr{R}_c$.

The normalized sensitivity indices for $\mathscr{R}_c$ are calculated by taking the partial derivative of $\mathscr{R}_c$ with respect to each parameter and multiply the derivative with the ratio of the parameter to $\mathscr{R}_c$. This value represents the percent change in $\mathscr{R}_c$ with respect to a 1% change in the parameter value (Caswell, 2001).

Table 4.3: Percent change in $\mathscr{R}_c$ with respect to a 1% change in the parameter value, for a low and a high isolation effectiveness r, and a low and a high value of f_T, while keeping the other parameter values as presented in Table 4.2.

	Parameter	β	r	ℓ	γ_r	γ	α	f_T
$f_T = 0.25$	% change for $r = 0.35$	1%	-0.23%	0.423%	-0.423%	-0.382%	-0.195%	-0.119%
	% change for $r = 0.95$	1%	-1.014%	0.053%	-0.053%	-0.445%	-0.501%	-0.306%
$f_T = 0.75$	% change for $r = 0.35$	1%	-0.402%	0.747%	-0.747%	-0.167%	-0.086%	-0.471%
	% change for $r = 0.95$	1%	-3.521%	0.185%	-0.185%	-0.383%	-0.431%	-2.373%

We use the parameters values from (Table 4.2) to study the sensitivity of $\mathscr{R}_c$ to each parameter. We compute normalized sensitivity analysis on all parameters, but we just consider the impact of parameters that are the most sensitive: $\beta, r, \ell, \gamma_r, \gamma, \alpha$, and f_T. The other parameters (μ, κ_1, and κ_2) have a very low impact, namely less than 0.001%. The numerical simulations to the sensitivity of $\mathscr{R}_c$ with respect to each of the most sensitive parameters

are given in (Table 4.3), for two different levels of isolation effectiveness ($r = 0.35$ and $r = 0.95$) and two values of f_T ($f_T = 0.25$ and $f_T = 0.75$), which is the fraction of pre-symptomatic individuals diagnosed and isolated. The other parameter values are kept as shown in Table 4.2.

In the case of high isolation effectiveness ($r = 0.95$), simulations show that both the removal rate, γ_r, of isolated individuals and the relative transmissibility parameter ℓ of isolated individuals with respect to infectious individuals are the least sensitive parameters (with 0.053% change of $\mathscr{R}_c$), while the parameter of isolation effectiveness, r, is the most sensitive one, where a 1% increase in r causes a 1.014% reduction in the value of $\mathscr{R}_c$. Also, the rate at which infectious individuals get isolated, α, and the fraction of pre-symptomatic individuals detected and isolated, f_T, impact negatively on the level of $\mathscr{R}_c$, where a 1% percent increase in the value of f_T causes approximately a 0.31% decline in the value of the reproduction number $\mathscr{R}_c$. Thus, as pre-symptomatic individuals are diagnosed and as isolation is highly effective, the number of available infectious individuals who are capable of transmitting Ebola decreases and therefore, the reproduction number decreases. Also, the removal (by recovery or Ebola-induced death) rate γ of infectious individuals affects negatively on $\mathscr{R}_c$. Hence, for the case of highly effective isolation, the parameters concerning early diagnosis and isolation have a significant impact on the reproduction number.

This percent impact of the parameters on $\mathscr{R}_c$ remains so as long as isolation is highly effective. However, if the effectiveness of isolation is low, in the sense that all parameter values are kept the same except the value of the parameter r, which is reduced to 0.35, then we get the results presented in Table 4.3. In this case, both the relative transmissibility ℓ and the removal rate of isolated individuals, γ_r, are the second most sensitive parameters, after β which is the most impactful one. Also, ℓ became more sensitive than r. The implication is that, when isolation is less effective, there exists the possibility for isolated people to make successful contacts with susceptible individuals and therefore the possibility of causing

new infections increases. This causes an increase in the reproduction number. Also, it is noted that the effect of f_T and α is reduced, which means that diagnosing and isolating infected individuals becomes a weak strategy if the effectiveness of isolation is low.

On repeating the previous analyses, but this time for a higher value of f_T ($f_T = 0.75$), we obtain the results shown in Table 4.3. In comparison to the scenario when $f_T = 0.25$, the simulations show that increasing the fraction of pre-symptomatic individuals who are diagnosed and isolated, f_T, increases the percent impact of the parameters r, ℓ, γ_r, and f_T, and decreases the percent impact of the parameters γ and α, on the value of the control reproduction number $\mathscr{R}_c$.

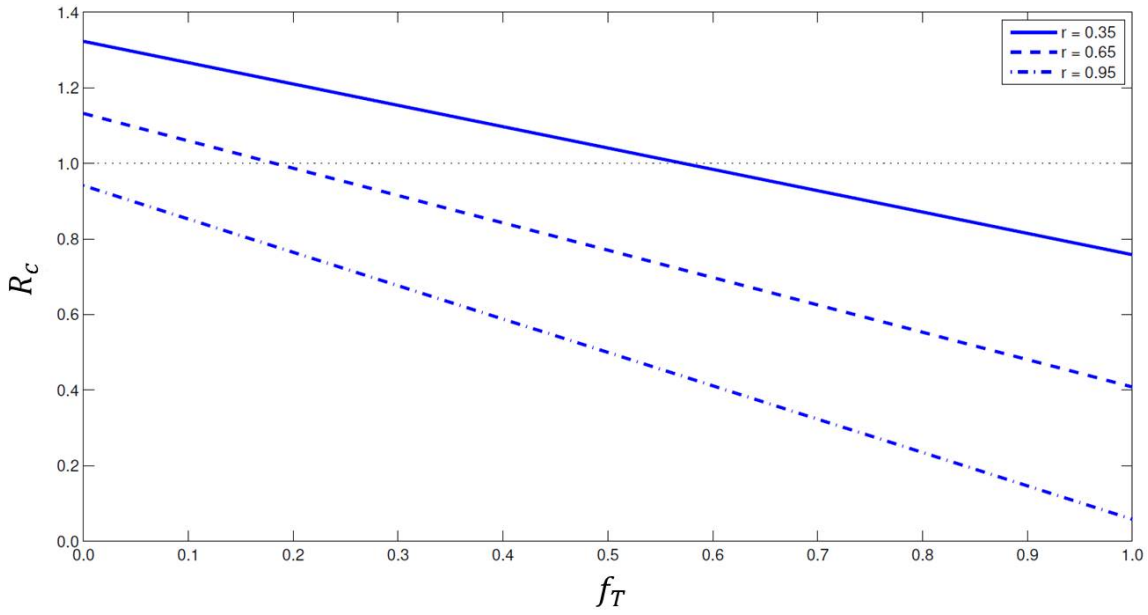


Figure 4.2: Impact of Early Detection of Pre-Symptomatic Individuals on the Value of $\mathscr{R}_c$.

4.3.4 Impact of Early Detection and Isolation on the Value of $\mathscr{R}_c$

To study the impact of early detection of pre-symptomatic individuals and isolation on the reproduction number, we first depict $\mathscr{R}_c$ as a function of f_T, for different levels of isolation effectiveness r. Fig. 4.2 shows that the control reproduction number declines as the proportion, f_T, of pre-symptomatic individuals, who get diagnosed and isolated, increases.

Simulations are done using parameter values from (Table 4.2), but for three different values of r. It further shows that the curve corresponding to a low and an intermidate value of isolation effectivenes r (e.g. $r = 0.35$ for the solid curve and $r = 0.65$ for the dashed curve) hits $\mathscr{R}_c = 1$ at some critical value of f_T (say $f_T^\star$), while for the high value of r ($r = 0.95$), it never hits the critical threshold $\mathscr{R}_c = 1$, as the curve is totally below the critical threshold. This indicates that for a high effectiveness of isolation, the control reproduction number is less than one and therefore the infection dies out. Analytically, the exact form of $f_T^\star$ is

$$f_T^\star = \left[1 + (1-r)\ell\frac{\alpha}{\gamma_r+\mu} - \frac{1}{\mathscr{R}_0}\left(1+\frac{\alpha}{\gamma+\mu}\right)\right] / \left[1 - \frac{(1-r)\ell(\gamma+\mu)}{\gamma_r+\mu}\right]. \tag{4.14}$$

The critical proportion $f_T^\star$ represents the minimum proportion of pre-symptomatic individuals who are detected and get isolated to ensure an effective control of Ebola. This critical value remains feasible as long as the following inequality holds

$$(1-r)\ell < \frac{\gamma_r+\mu}{(\gamma+\mu)\mathscr{R}_0}. \tag{4.15}$$

If we keep all parameters fixed except r, then condition (4.15) could be rewritten in a more convenient form

$$r > 1 - \frac{\gamma_r+\mu}{\ell(\gamma+\mu)\mathscr{R}_0}. \tag{4.16}$$

This gives the minimum level of effectiveness of isolation required to obtain an isolation and early diagnosis-based control strategy for Ebola tranmission.

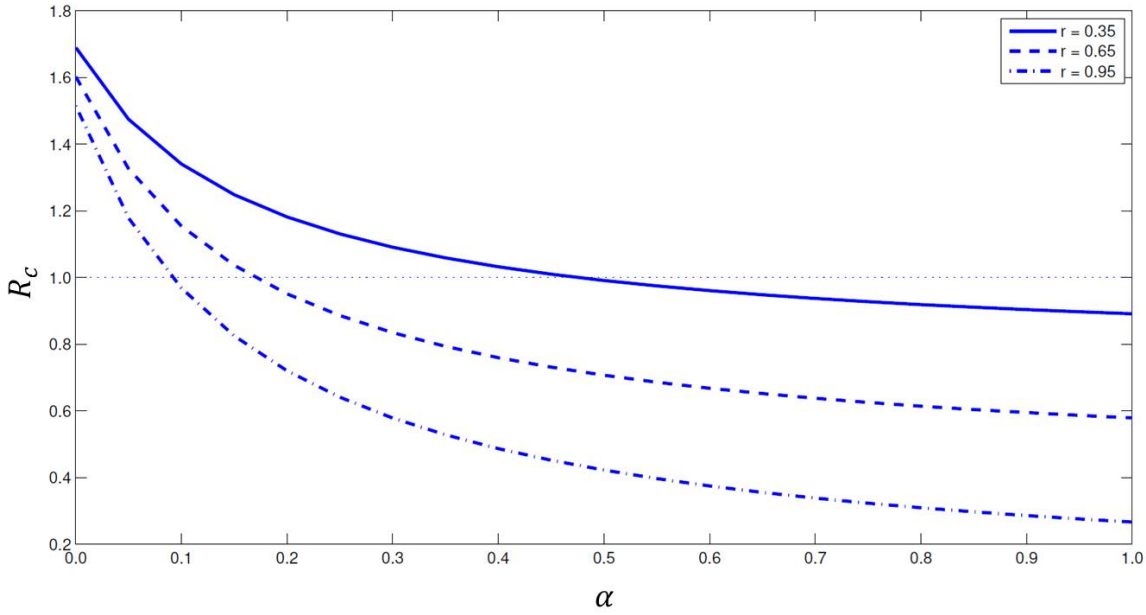


Figure 4.3: Impact of Isolating Infectious Individuals on the Value of $\mathscr{R}_c$.

Now, we could also ask a similar question on the role of isolating infectious individuals to contain Ebola transmission. Fig. 4.3 shows the impact of changing the rate at which infectious individuals get isolated, α, on $\mathscr{R}_c$, for the same three different levels of isolation effectivenes, as used above. The analysis shows that it is possible to control the epidemic if and only if $\alpha > \alpha^\star$, where

$$\alpha^\star = \frac{[(1-f_T)(\gamma_r+\mu)(\gamma+\mu)+(1-r)\ell f_T(\gamma+\mu)^2]\mathscr{R}_0-(\gamma_r+\mu)(\gamma+\mu)}{(\gamma_r+\mu)-\ell(1-r)\mathscr{R}_0(\gamma+\mu)} \tag{4.17}$$

and with the implementation of condition (4.15).

4.4 Discussion

The Ebola epidemic has shown us major weaknesses not only in health systems in West Africa, but also in our global capacity to respond early to an outbreak with effective diagnositc capacities. After multiple outbreaks of infectious diseases, from severe acute respiratory syndrome (SARS) to Middle East respiratory syndrome coronavirus (MERS-CoV), we still do not have effective diagnostic tools to rapidly respond to a number of

potential epidemics. The main reason why we lack of such diagnostic preparedness against infectious diseases is because of the lack of a financed global strategy that can be implemented ahead, rather than during an epidemic. This strategy must primarily focus on two critical aspects: First, a continuous intereaction between the field to detect small outbreaks and collect samples, and reference laboratories with advanced sequencing tools to identify the pathogen. Second, the need of assay development for early diagnosis, their regulatory approval, and a plan of implementation in anticipation of an outbreak.

Here, motivated by some studies showing that PCR assay can detect Ebola virus in both humans and non-human primates during the pre-symptomatic stage (Leroy et al., 2000; Qiu et al., 2014), we have developed and analyzed a mathematical model calibrated to the transmission dynamics of Ebola virus disease in West Africa to evaluate the impact of early diagnosis of pre-symptomatic infections. In the absence of effective treatments and vaccines, our results show the importance of implementing integrated control measures of early diagnosis and isolation. Importantly, our analysis identifies a threshold where early diagnosis of pre-symptomatic individuals, combined with a sufficient level of effective isolation, can lead to an epidemic control of Ebola virus disease. Furthermore, the need to incorporate vital dynamics is justified by our still limited understanding of Ebola infection including whether or not Ebola virus may persist among recovered individuals. The use of $\mathscr{R}_c$ in this context reflect our view that control measures should be sustainable and not just in response to an outbreak.

REFERENCES

Anderson, K., Lutz, C., Van Delft, F. W., Bateman, C. M., Guo, Y., Colman, S. M., Kempski, H., Moorman, A. V., Titley, I., Swansbury, J., et al. (2011). Genetic variegation of clonal architecture and propagating cells in leukaemia. *Nature*, 469(7330):356–361.

Anderson, R. M., May, R. M., and Anderson, B. (1992). *Infectious diseases of humans: dynamics and control*, volume 28. Wiley Online Library.

Andor, N., Graham, T. A., Jansen, M., Xia, L. C., Aktipis, C. A., Petritsch, C., Ji, H. P., and Maley, C. C. (2016). Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nature medicine*, 22(1):105–113.

Assarsson, E., Bui, H.-H., Sidney, J., Zhang, Q., Glenn, J., Oseroff, C., Mbawuike, I. N., Alexander, J., Newman, M. J., Grey, H., et al. (2008). Immunomic analysis of the repertoire of t-cell specificities for influenza a virus in humans. *Journal of virology*, 82(24):12241–12251.

Bachy, V., Hervouet, C., Becker, P. D., Chorro, L., Carlin, L. M., Herath, S., Papagatsias, T., Barbaroux, J.-B., Oh, S.-J., Benlahrech, A., et al. (2013). Langerin negative dendritic cells promote potent cd8+ t-cell priming by skin delivery of live adenovirus vaccine microneedle arrays. *Proceedings of the National Academy of Sciences*, 110(8):3041–3046.

Barber, L. J., Davies, M. N., and Gerlinger, M. (2015). Dissecting cancer evolution at the macro-heterogeneity and micro-heterogeneity scale. *Current opinion in genetics & development*, 30:1–6.

Barrett, R. D. and Schluter, D. (2008). Adaptation from standing genetic variation. *Trends in ecology & evolution*, 23(1):38–44.

Beeching, N. J., Fenech, M., and Houlihan, C. F. (2014). Ebola virus disease. *BMJ*, 349:g7348.

Beerenwinkel, N., Antal, T., Dingli, D., Traulsen, A., Kinzler, K. W., Velculescu, V. E., Vogelstein, B., and Nowak, M. A. (2007). Genetic progression and the waiting time to cancer. *PLoS Comput Biol*, 3(11):e225.

Birnbaum, M. E., Mendoza, J. L., Sethi, D. K., Dong, S., Glanville, J., Dobbins, J., Özkan, E., Davis, M. M., Wucherpfennig, K. W., and Garcia, K. C. (2014). Deconstructing the peptide-mhc specificity of t cell recognition. *Cell*, 157(5):1073–1087.

Blum, J. S., Wearsch, P. A., and Cresswell, P. (2013). Pathways of antigen processing. *Annual review of immunology*, 31:443.

Bozic, I., Antal, T., Ohtsuki, H., Carter, H., Kim, D., Chen, S., Karchin, R., Kinzler, K. W., Vogelstein, B., and Nowak, M. A. (2010). Accumulation of driver and passenger mutations during tumor progression. *Proceedings of the National Academy of Sciences*, 107(43):18545–18550.

Bozic, I. and Nowak, M. A. (2014). Timing and heterogeneity of mutations associated with drug resistance in metastatic cancers. *Proceedings of the National Academy of Sciences*, 111(45):15964–15968.

Brauer, F., Castillo-Chavez, C., and Castillo-Chavez, C. (2001). *Mathematical models in population biology and epidemiology*, volume 40. Springer.

Breen, M. S., Kemena, C., Vlasov, P. K., Notredame, C., and Kondrashov, F. A. (2012). Epistasis as the primary factor in molecular evolution. *Nature*, 490(7421):535–538.

Burrell, R. A., McGranahan, N., Bartek, J., and Swanton, C. (2013). The causes and consequences of genetic heterogeneity in cancer evolution. *Nature*, 501(7467):338–345.

Burrell, R. A. and Swanton, C. (2016). Re-evaluating clonal dominance in cancer evolution. *Trends in Cancer*, 2(5):263–276.

Burroughs, N. J., de Boer, R. J., and Keşmir, C. (2004). Discriminating self from nonself with short peptides from large proteomes. *Immunogenetics*, 56(5):311–320.

Calis, J. J., Maybeno, M., Greenbaum, J. A., Weiskopf, D., De Silva, A. D., Sette, A., Keşmir, C., and Peters, B. (2013). Properties of mhc class i presented peptides that enhance immunogenicity. *PLoS Comput Biol*, 9(10):e1003266.

Calis, J. J., Sanchez-Perez, G. F., and Keşmir, C. (2010). Mhc class i molecules exploit the low g+ c content of pathogen genomes for enhanced presentation. *European journal of immunology*, 40(10):2699–2709.

Campbell, P. J., Pleasance, E. D., Stephens, P. J., Dicks, E., Rance, R., Goodhead, I., Follows, G. A., Green, A. R., Futreal, P. A., and Stratton, M. R. (2008). Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. *Proceedings of the National Academy of Sciences*, 105(35):13081–13086.

Campbell, P. J., Yachida, S., Mudie, L. J., Stephens, P. J., Pleasance, E. D., Stebbings, L. A., Morsberger, L. A., Latimer, C., McLaren, S., Lin, M.-L., et al. (2010). The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature*, 467(7319):1109–1113.

Caswell, H. (2001). Matrix population models second edition. *Sunderland, MA: Sinauer Associates*.

Chowell, D., Boddy, A. M., Mallo, D., Tollis, M., and Maley, C. C. (2016). When (distant) relatives stay too long: implications for cancer medicine. *Genome biology*, 17(1):1.

Chowell, D., Castillo-Chavez, C., Krishna, S., Qiu, X., and Anderson, K. S. (2015). Modelling the effect of early detection of ebola. *The Lancet Infectious Diseases*, 15(2):148–149.

Chowell, G. and Nishiura, H. (2014). Transmission dynamics and control of ebola virus disease (evd): a review. *BMC medicine*, 12(1):196.

Diaz Jr, L. A., Williams, R. T., Wu, J., Kinde, I., Hecht, J. R., Berlin, J., Allen, B., Bozic, I., Reiter, J. G., Nowak, M. A., et al. (2012). The molecular evolution of acquired resistance to targeted egfr blockade in colorectal cancers. *Nature*, 486(7404):537–540.

Ding, L., Ley, T. J., Larson, D. E., Miller, C. A., Koboldt, D. C., Welch, J. S., Ritchey, J. K., Young, M. A., Lamprecht, T., McLellan, M. D., et al. (2012). Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature*, 481(7382):506–510.

Durrett, R., Foo, J., Leder, K., Mayberry, J., and Michor, F. (2011). Intratumor heterogeneity in evolutionary models of tumor progression. *Genetics*, 188(2):461–477.

Falk, K., Rötzschke, O., Stevanović, S., Jung, G., and Rammensee, H.-G. (1991). Allele-specific motifs revealed by sequencing of self-peptides eluted from mhc molecules. *Nature*, 351(6324):290–296.

Fasina, F., Shittu, A., Lazarus, D., Tomori, O., Simonsen, L., Viboud, C., and Chowell, G. (2014). Transmission dynamics and control of ebola virus disease outbreak in nigeria, july to september 2014. *Euro Surveill*, 19(40):20920.

Fauci, A. S. (2014). Ebolaunderscoring the global disparities in health care resources. *New England Journal of Medicine*, 371(12):1084–1086.

Fearon, E. R., Vogelstein, B., et al. (1990). A genetic model for colorectal tumorigenesis. *Cell*, 61(5):759–767.

Fink, S. (2015). Ebola drug aids some in a study in west africa. *The New York Times*.

Fisher, J. (1958). Multiple-mutation theory of carcinogenesis. *Nature*.

Frank, S. A. (2007). *Dynamics of cancer: incidence, inheritance, and evolution*. Princeton University Press.

Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M. R., Appel, R. D., and Bairoch, A. (2005). *Protein identification and analysis tools on the ExPASy server*. Springer.

Genovese, G., Kähler, A. K., Handsaker, R. E., Lindberg, J., Rose, S. A., Bakhoum, S. F., Chambert, K., Mick, E., Neale, B. M., Fromer, M., et al. (2014). Clonal hematopoiesis and blood-cancer risk inferred from blood dna sequence. *New England Journal of Medicine*, 371(26):2477–2487.

Gerlinger, M., Rowan, A. J., Horswell, S., Larkin, J., Endesfelder, D., Gronroos, E., Martinez, P., Matthews, N., Stewart, A., Tarpey, P., et al. (2012). Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *New England Journal of Medicine*, 366(10):883–892.

Gerstung, M., Beisel, C., Rechsteiner, M., Wild, P., Schraml, P., Moch, H., and Beerenwinkel, N. (2012). Reliable detection of subclonal single-nucleotide variants in tumour cell populations. *Nature communications*, 3:811.

Grakoui, A., Bromley, S. K., Sumen, C., Davis, M. M., Shaw, A. S., Allen, P. M., and Dustin, M. L. (1999). The immunological synapse: a molecular machine controlling t cell activation. *Science*, 285(5425):221–227.

Grantham, R. (1974). Amino acid difference formula to help explain protein evolution. *Science*, 185(4154):862–864.

Greaves, M. and Maley, C. C. (2012). Clonal evolution in cancer. *Nature*, 481(7381):306–313.

Haccou, P., Jagers, P., and Vatutin, V. A. (2005). *Branching processes: variation, growth, and extinction of populations*. Number 5. Cambridge University Press.

Hanahan, D. and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *cell*, 144(5):646–674.

Harndahl, M., Rasmussen, M., Roder, G., Dalgaard Pedersen, I., Sørensen, M., Nielsen, M., and Buus, S. (2012). Peptide-mhc class i stability is a better predictor than peptide affinity of ctl immunogenicity. *European journal of immunology*, 42(6):1405–1416.

Hegreness, M., Shoresh, N., Hartl, D., and Kishony, R. (2006). An equivalence principle for the incorporation of favorable mutations in asexual populations. *Science*, 311(5767):1615–1617.

Hennecke, J. and Wiley, D. C. (2001). T cell receptor–mhc interactions up close. *Cell*, 104(1):1–4.

Hogquist, K. A., Jameson, S. C., Heath, W. R., Howard, J. L., Bevan, M. J., and Carbone, F. R. (1994). T cell receptor antagonist peptides induce positive selection. *Cell*, 76(1):17–27.

Honeyman, M. C., Brusic, V., Stone, N. L., and Harrison, L. C. (1998). Neural network-based prediction of candidate t-cell epitopes. *Nature biotechnology*, 16(10):966–969.

Huang, L., Kuhls, M. C., and Eisenlohr, L. C. (2011). Hydrophobicity as a driver of mhc class i antigen processing. *The EMBO journal*, 30(8):1634–1644.

Hyo-eun, C. B., Ruddy, D. A., Radhakrishna, V. K., Caushi, J. X., Zhao, R., Hims, M. M., Singh, A. P., Kao, I., Rakiec, D., Shaw, P., et al. (2015). Studying clonal dynamics in response to cancer therapy using high-complexity barcoding. *Nature medicine*, 21(5):440–448.

Ippolito, G., Feldmann, H., Lanini, S., Vairo, F., Di Caro, A., Capobianchi, M. R., and Nicastri, E. (2012). Viral hemorrhagic fevers: advancing the level of treatment. *BMC medicine*, 10(1):31.

Jones, S., Chen, W.-d., Parmigiani, G., Diehl, F., Beerenwinkel, N., Antal, T., Traulsen, A., Nowak, M. A., Siegel, C., Velculescu, V. E., et al. (2008). Comparative lesion sequencing provides insights into tumor evolution. *Proceedings of the National Academy of Sciences*, 105(11):4283–4288.

Khrustalev, V. V. and Barkovsky, E. V. (2011). Percent of highly immunogenic amino acid residues forming b-cell epitopes is higher in homologous proteins encoded by gc-rich genes. *Journal of theoretical biology*, 282(1):71–79.

Kim, Y., Yewdell, J. W., Sette, A., and Peters, B. (2013). Positional bias of mhc class i restricted t-cell epitopes in viral antigens is likely due to a bias in conservation. *PLOS Comput Biol*, 9(1):e1002884.

Komarova, N. L., Burger, J. A., and Wodarz, D. (2014). Evolution of ibrutinib resistance in chronic lymphocytic leukemia (cll). *Proceedings of the National Academy of Sciences*, 111(38):13906–13911.

Komarova, N. L. and Wodarz, D. (2005). Drug resistance in cancer: principles of emergence and prevention. *Proceedings of the National Academy of Sciences of the United States of America*, 102(27):9714–9719.

Kryazhimskiy, S., Draghi, J. A., Plotkin, J. B., et al. (2011). In evolution, the sum is less than its parts. *science*, 332(6034):1160–1161.

Kryazhimskiy, S., Rice, D. P., Jerison, E. R., and Desai, M. M. (2014). Global epistasis makes adaptation predictable despite sequence-level stochasticity. *Science*, 344(6191):1519–1522.

Kubo, R. T., Sette, A., Grey, H. M., Appella, E., Sakaguchi, K., Zhu, N.-Z., Arnott, D., Sherman, N., Shabanowitz, J., and Michel, H. (1994). Definition of specific peptide motifs for four major hla-a alleles. *The Journal of Immunology*, 152(8):3913–3924.

Kugelman, J. R., Sanchez-Lockhart, M., Andersen, K. G., Gire, S., Park, D. J., Sealfon, R., Lin, A. E., Wohl, S., Sabeti, P. C., Kuhn, J. H., et al. (2015). Evaluation of the potential impact of ebola virus genomic drift on the efficacy of sequence-based candidate therapeutics. *MBio*, 6(1):e02227–14.

Kyte, J. and Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *Journal of molecular biology*, 157(1):105–132.

Landau, D. A., Carter, S. L., Stojanov, P., McKenna, A., Stevenson, K., Lawrence, M. S., Sougnez, C., Stewart, C., Sivachenko, A., Wang, L., et al. (2013). Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell*, 152(4):714–726.

Landau, D. A., Tausch, E., Taylor-Weiner, A. N., Stewart, C., Reiter, J. G., Bahlo, J., Kluth, S., Bozic, I., Lawrence, M., Böttcher, S., et al. (2015). Mutations driving cll and their evolution in progression and relapse. *Nature*, 526(7574):525–530.

Lang, G. I., Rice, D. P., Hickman, M. J., Sodergren, E., Weinstock, G. M., Botstein, D., and Desai, M. M. (2013). Pervasive genetic hitchhiking and clonal interference in forty evolving yeast populations. *Nature*, 500(7464):571–574.

Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., Carter, S. L., Stewart, C., Mermel, C. H., Roberts, S. A., et al. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499(7457):214–218.

Leroy, E. M., Baize, S., Volchkov, V., Fisher-Hoch, S., Georges-Courbot, M., Lansoud-Soukate, J., Capron, M., Debre, P., Georges, A., and McCormick, J. (2000). Human asymptomatic ebola infection and strong inflammatory response. *The Lancet*, 355(9222):2210–2215.

Levy, S. F., Blundell, J. R., Venkataram, S., Petrov, D. A., Fisher, D. S., and Sherlock, G. (2015). Quantitative evolutionary dynamics using high-resolution lineage tracking. *Nature*.

Lipinski, K. A., Barber, L. J., Davies, M. N., Ashenden, M., Sottoriva, A., and Gerlinger, M. (2016). Cancer evolution and the limits of predictability in precision cancer medicine. *Trends in cancer*, 2(1):49–63.

Lynch, M. (2016). Mutation and human exceptionalism: Our future genetic load. *Genetics*, 202(3):869–875.

Maley, C. C., Galipeau, P. C., Finley, J. C., Wongsurawat, V. J., Li, X., Sanchez, C. A., Paulson, T. G., Blount, P. L., Risques, R.-A., Rabinovitch, P. S., et al. (2006). Genetic clonal diversity predicts progression to esophageal adenocarcinoma. *Nature genetics*, 38(4):468–473.

Martincorena, I. and Campbell, P. J. (2015). Somatic mutation in cancer and normal cells. *Science*, 349(6255):1483–1489.

Martincorena, I., Roshan, A., Gerstung, M., Ellis, P., Van Loo, P., McLaren, S., Wedge, D. C., Fullam, A., Alexandrov, L. B., Tubio, J. M., et al. (2015). High burden and pervasive positive selection of somatic mutations in normal human skin. *Science*, 348(6237):880–886.

Marusyk, A., Almendro, V., and Polyak, K. (2012). Intra-tumour heterogeneity: a looking glass for cancer? *Nature Reviews Cancer*, 12(5):323–334.

Marusyk, A., Tabassum, D. P., Altrock, P. M., Almendro, V., Michor, F., and Polyak, K. (2014). Non-cell-autonomous driving of tumour growth supports sub-clonal heterogeneity. *Nature*, 514(7520):54–58.

McElroy, A. K., Erickson, B. R., Flietstra, T. D., Rollin, P. E., Nichol, S. T., Towner, J. S., and Spiropoulou, C. F. (2014). Ebola hemorrhagic fever: novel biomarker correlates of clinical outcome. *Journal of Infectious Diseases*, page jiu088.

McFarland, C. D., Korolev, K. S., Kryukov, G. V., Sunyaev, S. R., and Mirny, L. A. (2013). Impact of deleterious passenger mutations on cancer progression. *Proceedings of the National Academy of Sciences*, 110(8):2910–2915.

Medzhitov, R. and Janeway, C. A. (2002). Decoding the patterns of self and nonself by the innate immune system. *Science*, 296(5566):298–300.

Michor, F., Iwasa, Y., and Nowak, M. A. (2006). The age incidence of chronic myeloid leukemia can be explained by a one-mutation model. *Proceedings of the National Academy of Sciences*, 103(40):14931–14934.

Morris, L., Riaz, N., Desrichard, A., Şenbabaoğlu, Y., Hakimi, A. A., Makarov, V., Reis-Filho, J. S., and Chan, T. A. (2016). Pan-cancer analysis of intratumor heterogeneity as a prognostic determinant of survival. *Oncotarget*, 7(9):10051–10063.

Morrissy, A. S., Garzia, L., Shih, D. J., Zuyderduyn, S., Huang, X., Skowron, P., Remke, M., Cavalli, F. M., Ramaswamy, V., Lindsay, P. E., et al. (2016). Divergent clonal selection dominates medulloblastoma at recurrence. *Nature*, 529(7586):351–357.

Moutaftsi, M., Peters, B., Pasquetto, V., Tscharke, D. C., Sidney, J., Bui, H.-H., Grey, H., and Sette, A. (2006). A consensus epitope prediction approach identifies the breadth of murine tcd8+-cell responses to vaccinia virus. *Nature biotechnology*, 24(7):817–819.

Mullighan, C. G., Phillips, L. A., Su, X., Ma, J., Miller, C. B., Shurtleff, S. A., and Downing, J. R. (2008). Genomic analysis of the clonal origins of relapsed acute lymphoblastic leukemia. *Science*, 322(5906):1377–1380.

Nagel, A. C., Joyce, P., Wichman, H. A., and Miller, C. R. (2012). Stickbreaking: a novel fitness landscape model that harbors epistasis and is consistent with commonly observed patterns of adaptive evolution. *Genetics*, 190(2):655–667.

Navin, N., Krasnitz, A., Rodgers, L., Cook, K., Meth, J., Kendall, J., Riggs, M., Eberling, Y., Troge, J., Grubor, V., et al. (2010). Inferring tumor progression from genomic heterogeneity. *Genome research*, 20(1):68–80.

Newell, E. W. and Davis, M. M. (2014). Beyond model antigens: high-dimensional methods for the analysis of antigen-specific t cells. *Nature biotechnology*, 32(2):149–157.

Newell, E. W., Sigal, N., Nair, N., Kidd, B. A., Greenberg, H. B., and Davis, M. M. (2013). Combinatorial tetramer staining and mass cytometry analysis facilitate t-cell epitope mapping and characterization. *Nature biotechnology*, 31(7):623–629.

Nielsen, M., Lundegaard, C., Blicher, T., Lamberth, K., Harndahl, M., Justesen, S., Røder, G., Peters, B., Sette, A., Lund, O., et al. (2007). Netmhcpan, a method for quantitative predictions of peptide binding to any hla-a and-b locus protein of known sequence. *PloS one*, 2(8):e796.

Nik-Zainal, S., Van Loo, P., Wedge, D. C., Alexandrov, L. B., Greenman, C. D., Lau, K. W., Raine, K., Jones, D., Marshall, J., Ramakrishna, M., et al. (2012). The life history of 21 breast cancers. *Cell*, 149(5):994–1007.

Notta, F., Mullighan, C. G., Wang, J. C., Poeppl, A., Doulatov, S., Phillips, L. A., Ma, J., Minden, M. D., Downing, J. R., and Dick, J. E. (2011). Evolution of human bcr-abl1 lymphoblastic leukaemia-initiating cells. *Nature*, 469(7330):362–367.

Nowell, P. C. (1976). The clonal evolution of tumor cell populations. *Science*, 194(4260):23–28.

Orr, H. A. (2003). The distribution of fitness effects among beneficial mutations. *Genetics*, 163(4):1519–1526.

Oukka, M., Manuguerra, J. C., Livaditis, N., Tourdot, S., Riche, N., Vergnon, I., Cordopatis, P., and Kosmatopoulos, K. (1996). Protection against lethal viral infection by vaccination with nonimmunodominant peptides. *The Journal of Immunology*, 157(7):3039–3045.

Pradeu, T. and Carosella, E. D. (2006). On the definition of a criterion of immunogenicity. *Proceedings of the National Academy of Sciences*, 103(47):17858–17861.

Purcell, A. W., McCluskey, J., and Rossjohn, J. (2007). More than one reason to rethink the use of peptides in vaccine design. *Nature reviews Drug discovery*, 6(5):404–414.

Qiu, X., Wong, G., Audet, J., Bello, A., Fernando, L., Alimonti, J. B., Fausther-Bovendo, H., Wei, H., Aviles, J., Hiatt, E., et al. (2014). Reversion of advanced ebola virus disease in nonhuman primates with zmapp. *Nature*.

Rammensee, H.-G., Bachmann, J., Emmerich, N. P. N., Bachor, O. A., and Stevanović, S. (1999). Syfpeithi: database for mhc ligands and peptide motifs. *Immunogenetics*, 50(3-4):213–219.

RDevelopment, C. (2012). Team 2009: R: A language and environment for statistical computing. *Vienna, Austria. Internet: http://www. R-project. org*.

Riemer, A. B., Keskin, D. B., Zhang, G., Handley, M., Anderson, K. S., Brusic, V., Reinhold, B., and Reinherz, E. L. (2010). A conserved e7-derived cytotoxic t lymphocyte epitope expressed on human papillomavirus 16-transformed hla-a2+ epithelial cancers. *Journal of Biological Chemistry*, 285(38):29608–29622.

Roche-Lestienne, C., Deluche, L., Corm, S., Tigaud, I., Joha, S., Philippe, N., Geffroy, S., Laï, J.-L., Nicolini, F.-E., and Preudhomme, C. (2008). Runx1 dna-binding mutations and runx1-prdm16 cryptic fusions in bcr-abl+ leukemias are frequently associated with secondary trisomy 21 and may contribute to clonal evolution and imatinib resistance. *Blood*, 111(7):3735–3741.

Roche-Lestienne, C., Laï, J.-L., Darré, S., Facon, T., and Preudhomme, C. (2003). A mutation conferring resistance to imatinib at the time of diagnosis of chronic myelogenous leukemia. *New England Journal of Medicine*, 348(22):2265–2266.

Roche-Lestienne, C., Soenen-Cornu, V., Grardel-Duflos, N., Laı, J.-L., Philippe, N., Facon, T., Fenaux, P., and Preudhomme, C. (2002). Several types of mutations of the abl gene can be found in chronic myeloid leukemia patients resistant to sti571, and they can preexist to the onset of treatment. *Blood*, 100(3):1014–1018.

Rudolph, M. G., Stanfield, R. L., and Wilson, I. A. (2006). How tcrs bind mhcs, peptides, and coreceptors. *Annu. Rev. Immunol.*, 24:419–466.

Schmitt, M. W., Loeb, L. A., and Salk, J. J. (2015). The influence of subclonal resistance mutations on targeted cancer therapy. *Nature Reviews Clinical Oncology*.

Schreiber, R. D., Old, L. J., and Smyth, M. J. (2011). Cancer immunoediting: integrating immunitys roles in cancer suppression and promotion. *Science*, 331(6024):1565–1570.

Seong, S.-Y. and Matzinger, P. (2004). Hydrophobicity: an ancient damage-associated molecular pattern that initiates innate immune responses. *Nature Reviews Immunology*, 4(6):469–478.

Shain, A. H., Yeh, I., Kovalyshyn, I., Sriharan, A., Talevich, E., Gagnon, A., Dummer, R., North, J., Pincus, L., Ruben, B., et al. (2015). The genetic evolution of melanoma from precursor lesions. *New England Journal of Medicine*, 373(20):1926–1936.

Shendure, J. and Akey, J. M. (2015). The origins, determinants, and consequences of human mutations. *Science*, 349(6255):1478–1483.

Siravegna, G., Mussolin, B., Buscarino, M., Corti, G., Cassingena, A., Crisafulli, G., Ponzetti, A., Cremolini, C., Amatu, A., Lauricella, C., et al. (2015). Clonal evolution and resistance to egfr blockade in the blood of colorectal cancer patients. *Nature medicine*, 21(7):795–801.

Snyder, A., Makarov, V., Merghoub, T., Yuan, J., Zaretsky, J. M., Desrichard, A., Walsh, L. A., Postow, M. A., Wong, P., Ho, T. S., et al. (2014). Genetic basis for clinical response to ctla-4 blockade in melanoma. *New England Journal of Medicine*, 371(23):2189–2199.

Sottoriva, A., Kang, H., Ma, Z., Graham, T. A., Salomon, M. P., Zhao, J., Marjoram, P., Siegmund, K., Press, M. F., Shibata, D., et al. (2015). A big bang model of human colorectal tumor growth. *Nature genetics*, 47(3):209–216.

Sottoriva, A., Spiteri, I., Piccirillo, S. G., Touloumis, A., Collins, V. P., Marioni, J. C., Curtis, C., Watts, C., and Tavaré, S. (2013). Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. *Proceedings of the National Academy of Sciences*, 110(10):4009–4014.

Sprouffske, K., Merlo, L. M., Gerrish, P. J., Maley, C. C., and Sniegowski, P. D. (2012). Cancer in light of experimental evolution. *Current Biology*, 22(17):R762–R771.

Tan, C. T., Croft, N. P., Dudek, N. L., Williamson, N. A., and Purcell, A. W. (2011). Direct quantitation of mhc-bound peptide epitopes by selected reaction monitoring. *Proteomics*, 11(11):2336–2340.

Tenzer, S., Peters, B., Bulik, S., Schoor, O., Lemmel, C., Schatz, M., Kloetzel, P.-M., Rammensee, H.-G., Schild, H., and Holzhütter, H.-G. (2005). Modeling the mhc class i pathway by combining predictions of proteasomal cleavage, tap transport and mhc class i binding. *Cellular and Molecular Life Sciences CMLS*, 62(9):1025–1037.

Thieme, H. R. (2003). *Mathematics in population biology*. Princeton University Press.

Tirosh, I., Izar, B., Prakadan, S. M., Wadsworth, M. H., Treacy, D., Trombetta, J. J., Rotem, A., Rodman, C., Lian, C., Murphy, G., et al. (2016). Dissecting the multicellular ecosystem of metastatic melanoma by single-cell rna-seq. *Science*, 352(6282):189–196.

Tomasetti, C., Marchionni, L., Nowak, M. A., Parmigiani, G., and Vogelstein, B. (2015). Only three driver gene mutations are required for the development of lung and colorectal cancers. *Proceedings of the National Academy of Sciences*, 112(1):118–123.

Van den Driessche, P. and Watmough, J. (2002). Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission. *Mathematical biosciences*, 180(1):29–48.

van der Merwe, P. A. and Dushek, O. (2011). Mechanisms for t cell receptor triggering. *Nature Reviews Immunology*, 11(1):47–55.

van der Most, R. G., Murali-Krishna, K., Whitton, J. L., Oseroff, C., Alexander, J., Southwood, S., Sidney, J., Chesnut, R. W., Sette, A., and Ahmed, R. (1998). Identification of d b-and k b-restricted subdominant cytotoxic t-cell responses in lymphocytic choriomeningitis virus-infected mice. *Virology*, 240(1):158–167.

Vermeulen, L., Morrissey, E., van der Heijden, M., Nicholson, A. M., Sottoriva, A., Buczacki, S., Kemp, R., Tavaré, S., and Winton, D. J. (2013). Defining stem cell dynamics in models of intestinal tumor initiation. *Science*, 342(6161):995–998.

Vita, R., Zarebski, L., Greenbaum, J. A., Emami, H., Hoof, I., Salimi, N., Damle, R., Sette, A., and Peters, B. (2010). The immune epitope database 2.0. *Nucleic acids research*, 38(suppl 1):D854–D862.

Vogelstein, B. and Kinzler, K. W. (2015). The path to cancer–three strikes and you're out. *The New England journal of medicine*, 373(20):1895–1898.

Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A., and Kinzler, K. W. (2013). Cancer genome landscapes. *science*, 339(6127):1546–1558.

Waclaw, B., Bozic, I., Pittman, M. E., Hruban, R. H., Vogelstein, B., and Nowak, M. A. (2015). Spatial model predicts dispersal and cell turnover cause reduced intra-tumor heterogeneity. *arXiv preprint arXiv:1503.07116*.

Wang, Y., Waters, J., Leung, M. L., Unruh, A., Roh, W., Shi, X., Chen, K., Scheet, P., Vattathil, S., Liang, H., et al. (2014). Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature*, 512(7513):155–160.

Weiskopf, D., Yauch, L. E., Angelo, M. A., John, D. V., Greenbaum, J. A., Sidney, J., Kolla, R. V., De Silva, A. D., de Silva, A. M., Grey, H., et al. (2011). Insights into hla-restricted t cell responses in a novel mouse model of dengue virus infection point toward new implications for vaccine design. *The Journal of Immunology*, 187(8):4268–4279.

Wiser, M. J., Ribeck, N., and Lenski, R. E. (2013). Long-term dynamics of adaptation in asexual populations. *Science*, 342(6164):1364–1367.

Yadav, M., Jhunjhunwala, S., Phung, Q. T., Lupardus, P., Tanguay, J., Bumbaca, S., Franci, C., Cheung, T. K., Fritsche, J., Weinschenk, T., et al. (2014). Predicting immunogenic tumour mutations by combining mass spectrometry and exome sequencing. *Nature*, 515(7528):572–576.

Zhang, J., Fujimoto, J., Zhang, J., Wedge, D. C., Song, X., Zhang, J., Seth, S., Chow, C.-W., Cao, Y., Gumbs, C., et al. (2014). Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing. *Science*, 346(6206):256–259.

Zhao, B., Hemann, M. T., and Lauffenburger, D. A. (2016). Modeling tumor clonal evolution for drug combinations design. *Trends in Cancer*, 2(3):144–158.

Zimmerman, J., Eliezer, N., and Simha, R. (1968). The characterization of amino acid sequences in proteins by statistical methods. *Journal of theoretical biology*, 21(2):170–201.

APPENDIX A

MODELING THE SUBCLONAL EVOLUTION OF CANCER CELL POPULATIONS

Table A.1: Statistics of the number of dominant clones and minor subclones across all combinations of parameter values.

$\bar{s}$	μ_d	Mean number of dominant clones	Mean number of minor subclones	Minimum number of dominant clones	Minimum number of minor subclones	Maximum number of dominant clones	Maximum number minor subclones
0.1	1×10^{-8}	1.12	6	1	0	2	22
0.1	1×10^{-7}	1.21	20.61	1	0	2	98
0.1	1×10^{-6}	1.34	89.78	1	11	4	267
0.1	1×10^{-5}	1.47	476.43	1	112	4	1258
0.01	1×10^{-8}	1.25	6.86	1	0	4	34
0.01	1×10^{-7}	1.48	48.63	1	8	4	222
0.01	1×10^{-6}	1.73	290.74	1	97	4	862
0.01	1×10^{-5}	1.72	1670.8	1	730	4	3478
0.005	1×10^{-8}	1.27	73.63	1	1	4	2300
0.005	1×10^{-7}	1.45	61.32	1	7	4	150
0.005	1×10^{-6}	1.66	411.87	1	144	4	1322
0.005	1×10^{-5}	1.8	2705.32	1	1190	4	6734

Table A.2: Differential fitness between dominant and minor subclones

$\bar{s}$	μ_d	Average of maximum fitness values in dominant clones	Average of maximum fitness values in minor subclones
0.1	1×10^{-8}	1.04	1.07
0.1	1×10^{-7}	1.22	1.62
0.1	1×10^{-6}	1.32	1.77
0.1	1×10^{-5}	1.48	1.89
0.01	1×10^{-8}	1.04	1.08
0.01	1×10^{-7}	1.04	1.09
0.01	1×10^{-6}	1.06	1.11
0.01	1×10^{-5}	1.07	1.14
0.005	1×10^{-8}	1.02	1.04
0.005	1×10^{-7}	1.03	1.05
0.005	1×10^{-6}	1.03	1.08
0.005	1×10^{-5}	1.04	1.09

Table A.3: Average time from treatment to disease recurrence. The generation time assigned in the simulations is $T = 4$ days

$\bar{s}$	μ_d	Mean number of generations from treatment to recurrence	Mean time from treatment to recurrence (years)
0.1	1×10^{-8}	144.67	1.59
0.1	1×10^{-7}	33.33	0.37
0.1	1×10^{-6}	27.00	0.30
0.1	1×10^{-5}	28.67	0.31
0.01	1×10^{-8}	186.33	2.04
0.01	1×10^{-7}	83.00	0.91
0.01	1×10^{-6}	124.67	1.37
0.01	1×10^{-5}	135.33	1.48
0.005	1×10^{-8}	412.67	4.52
0.005	1×10^{-7}	163.00	1.79
0.005	1×10^{-6}	238.67	2.62
0.005	1×10^{-5}	241.67	2.65

These are the necessary steps to run the tumorsim.scala application, crash recovery, monitoring, and data extracts to Excel

Shipping DOS code to Linux: dos2unix with vi command sequence

<esc>:set fileformat=unix<enter><esc>:wq<enter>

Useful screen commands that allow the simulator to run in the background while logged out

```
### To start a persistent shell:
screen -S spark
screen -rd spark
screen -X -S spark quit
screen -wipe dead.pid

### If prompt disappears
stty icanon echo > /dev/null 2>&1
```

Initialize Spark environment

```
module load spark/1.6.0

### this is the contents of the spark environment configuration module file
#%Module1.0#####################################################
##########################################
## sge modulefile
##
proc ModulesHelp { } {

        puts stderr "\tAdds spark-1.6.0 to your PATH and MANPATH."
}

module-whatis   "Adds spark-1.6.0 binaries to your path."

set topdir/usr/hdp/2.2.4.2-2/spark-1.6.0
setenv          YARN_CONF_DIR   /etc/hadoop/conf
setenv          SPARK_HOME      $topdir
prepend-path    PATH            $topdir/bin
prepend-path    LD_LIBRARY_PATH $topdir/lib
```

Start the simulator

```
### with spark-shell
module load spark/1.6.0; tumorargs="help" spark-shell --master yarn --
packages com.databricks:spark-csv_2.10:1.3.0 --driver-memory 8G --
executor-memory 16G --executor-cores 1 --num-executors 1 --conf
spark.akka.frameSize=16  -i ~/tumor_20160602scala


### with spark-submit
spark-submit --class main.scala.Tumor --master yarn --packages
com.databricks:spark-csv_2.10:1.3.0  --deploy-mode client --driver-memory
8G --executor-memory 16G --executor-cores 1 --num-executors 1 --conf
spark.akka.frameSize=16 --conf spark.shuffle.spill=false --conf
spark.shuffle.compress=true ~/Tumor/target/Tumor-1.0-SNAPSHOT.jar
sequential 500000 500009

### help for optional opperands at the end look like the following:
Usage: <add|del|mult|repair> <start#> [<mult_end#>]
```

Crash recovery: manual partial tumor removal

```
hive -e "create database if not exists tumorsim location '/project/tumorsim'"
hive -e 'use tumorsim; show partitions tumors' | grep 284
hive -e "use tumorsim;alter table tumors drop if exists partition
(tumorid='AB9_284')"
hadoop dfs -ls /project/tumorsim/tumors/tumorid=AB9_284
hadoop dfs -ls /project/tumorsim/populations/tumorid=AB9_284
hadoop dfs -rmr /project/tumorsim/tumors/tumorid=AB9_284
hadoop dfs -ls /project/tumorsim/tumors | grep tumor | sort | tail
hadoop dfs -ls /project/tumorsim/tumors/*/* | grep tumor | sort | tail
hadoop dfs -ls /project/tumorsim/tumors/* | grep "finalsnapshot=Y" | sort |
tail

### to see hdfs partitions
hadoop dfs -ls /project/tumorsim/tumors/*/* | grep "finalsnapshot="  | sort |
tail
```

Export tumors from Hive to tab delimited files

Provided below is the Hive commands to create tab-delimited files that can be used to create Excel files used for import of data to Tableau. The commands work on individual tumors.

```
### detectable tumors at final snapshot with only the live subclones for a single tumor class
hive -e "set hive.execution.engine=tez;set hive.cli.print.header=true; select * from tumorsim.tumors_detectable where substr(tumorid,1,3) = 'AA9' and cells > 0 and finalsnapshot = 'Y' order by split(tumorid,'_')[1] , finalsnapshot , year, subclone" | sed 's/tumors_detectable.//g' > /home/project/tumorsim/tumors_detectable_final_live_AA9.xls

### detectables exports by tumorclass
hive -e "set hive.execution.engine=tez;set hive.cli.print.header=true; select * from tumorsim.populations_detectable where substr(tumorid,1,3) = 'AA9' order by split(tumorid,'_')[1] , generation" | sed 's/populations_detectable.//g' > /home/project/tumorsim/populations_detectable_AA9.xls


### full exports
hive -e "set hive.cli.print.header=true; select '_c0',* from tumorsim.populations_tabs order by tumorid , generation" | grep -E '^\_c0'| cut -c5- | sed 's/populations.//g' > /home/project/tumorsim/populations.txt
hive -e "set hive.cli.print.header=true; select '_c0',* from tumorsim.tumors_tabs order by tumorid , generation" | grep -E '^\_c0'| cut -c5- | sed 's/tumors.//g' > /home/project/tumorsim/tumors.txt

### copy as comma delimited as well;
cat tumors_detectable_final_live_AA9.xls | tr "\\t" "," > tumors_detectable_final_live_AA9.csv
```

Analytics written in Hive/SQL to create tumor class aggregates into a spreadsheet called tumor_evolution.xls

tumor_analysis.hive

Monitoring actively long running simulations written in Hive and Bash scripting

tumor_monitor.hive

APPENDIX B

TCR CONTACT RESIDUE HYDROPHOBICITY IS A HALLMARK OF IMMUNOGENIC $CD8^+$ T CELL EPITOPES

B.1 Construction of Datasets

All MHC class I peptides used in this study for analyses and the design of ANN-Hydro prediction model were retrieved from IEDB Vita et al. (2010) (www.iedb.org, last accessed: 08/11/2014). The IEDB is the largest curated dataset of MHC-I peptides identified from different primary research studies from over 334 different source organisms. We set the "Immune recognition context" as T cell response and selected "MHC class I" as the criteria for data retrieval. In total, there were 28,444 T cell epitopes reported to be immunogenic by T cell assays, including self and pathogenic epitopes and 6,142 peptides were reported to be positive by ligand elution analysis (either mass spectrometry or HPLC). To avoid redundancy and overrepresentation bias, we excluded all duplicate peptides, so that each peptide is present only once in the dataset. Positive CTL epitopes represent the immunogenic epitope group. Ligand eluted MHC-I self-peptides are generally eluted from cell surface and therefore they have been antigenically processed and MHC-bound. A vast majority of eluted self-peptides are derived from endogenous proteins. To completely separate immunogenic and non-immunogenic datasets, any immunogenic eluted self-peptide associated with autoimmunity or cancer was excluded. The remaining peptides were used as the non-immunogenic peptide dataset for our analyses. Additionally, we removed any pathogen derived non-self- eluted peptides from the eluted peptide dataset to generate mutually exclusive datasets. These unique peptides were further annotated for antigen name, peptide starting position, peptide ending position, and MHC restriction, which were required for inclusion. Peptides with "undetermined class I alleles" were also excluded. These filtering criteria resulted in a final dataset of 5,035 8-11mer immunogenic epitopes and 4,853 8-11mer non-immunogenic peptides.

B.2 Amino Acid Frequency Analysis

Overrepresentation of certain amino acids in immunogenic peptides was identified by calculating probability ratios of amino acids given by $P(x\,I\,immunogenic)/P(x\,I\,non-immunogenic)$, where $P(x\,I\,immunogenic)$ and $P(x\,I\,non-immunogenic)$ correspond to probability mass functions and x is an amino acid. Individual amino acid probability mass functions were calculated from their frequency distributions of immunogenic epitopes and non-immunogenic peptides. Spearman's rank correlations were quantified between probability ratios and biochemical properties (hydrophobicity, polarity, or bulkiness) of amino acids using the described amino acid scales.

B.3 Position-based Hydrophobicity Analysis

We transformed our datasets of immunogenic and non-immunogenic peptides into numeric arrays using the R statistical software RDevelopment (2012). Separate numeric arrays were generated for immunogenic and non-immunogenic 8, 9 and 10mers. Mean hydrophobicity of immunogenic and non-immunogenic peptides at each position was calculated and were compared residue-by-residue through Wilcoxon rank-sum tests to quantify statistical significance.

B.4 Rate Analysis of Predicted Peptides

An efficient prediction algorithm identifies consistently all possible CTL epitopes from a given protein in the fewest number of hits consistently. For each test protein, we created a subset with unique CTL epitopes retrieved either from the IEDB database. Each predicted peptide starting from rank one was queried for an exact match in the dataset of CTL epitopes. When there was an exact match, a positive hit was recorded. Graphical representations comparing the rate of predictions by the IEDB-consensus binding prediction algorithm and hydrophobicity-based predictions were generated (Fig. 3.3).

B.5 Hydrophobicity-based ANN Prediction Model (ANN-Hydro)

The R neuralnet package was used to design and train the two ANN-Hydro models on H-2D^b and HLA-A2 restricted 9mer peptides known to be immunogenic ($n = 204$ and $n = 374$, respectively) or non-immunogenic ($n = 232$ and $n = 201$, respectively). Each peptide sequence in the respective H-2D^b and HLA-A2 datasets were transformed into a corresponding numeric sequence based on the hydrophobicity value of amino acids. Training peptides were derived from IEDB and SYFPEITHIs epitope database Vita et al. (2010); Rammensee et al. (1999). A three-layer fully connected feed-forward ANN was comprised by nine input neurons, one hidden layer with three neurons, and one output variable (Fig.B.3).

Our ANN-Hydro prediction model is given by the following mathematical framework:

$$y(H) = f\left(w_0 + \sum_{i=1}^{3} w_i \cdot f(w_{0i} + W_i^T H)\right),$$

where w_0 denotes the intercept of the output neuron and w_{0i} the intercept of the i^{th} hidden neuron. Additionally, w_i denotes the synaptic weight corresponding to the synapse starting at the i^{th} hidden neuron and leading to the output neuron. $W_i = (w_{1i}, w_{2i}, \cdots, w_{9i})$ is the vector of all synaptic weights corresponding to the synapse leading to thei^{th} hidden neuron, and $H = (h(R_1), h(R_2), \cdots, h(R_9))$ the vector of all inputs, which corresponds to the numeric hydrophobicity representation of a 9mer peptide, where $h(R_i)$ is the hydrophobicity value of the amino acid R_i. Finally, the output variable $y(H)$ denotes the probability of a peptide being immunogenic (p-ANN-Hydro). Since the starting values for the weights are drawn from the standard normal distribution, the outputs were averaged over 60 realizations. The activation function $f(v)$ was chosen to be the sigmoid function $f(v) = 1/(1+e^{-v})$, and the sum of squared errors was used for the error function. The learning procedure was the resilient back-propagation with learning rate set to 0.01; a threshold set to 0.01 was defined for the partial derivatives of the error function

B.6 Application of ANN-Hydro

For each H-2D^b and HLA-A2 restricted epitope prediction, we used the MHC-binding prediction tool IEDB-consensus to generate a list of epitope predictions on which the immunogenicity model could be applied. We normalized prediction binding scores (percentile rank) using the expression $S_{B_i} = (\delta_i - \delta_{min})/(\delta_{max} - \delta_{min})$ where S_{B_i} represents the normalized score of a given peptide; δ_i, the assigned output score by IEDB-consensus; δ_{min}, the

minimum score assigned in prediction output by IEDB; and δ_{max} , the maximum score assigned in the entire prediction output by IEDB. To remove poor binding peptides from the list, a subset of predicted peptides was selected by defining a S_B-threshold of 0.2 for antigen length ≤ 100 aas and a S_B-threshold of 0.1 (10^{th} percentile of predicted binders) for antigen length > 100 aas. Independently, probabilities of immunogenicity were obtained by applying the ANN-Hydro model to this subset of binding predictions. Normalized scores (S_I) were then assigned based on these probabilities of immunogenicity. Within the spectrum of predicted binders, we prioritized epitope re-ranking based on both S_B and S_I scores with first priority given to high-immunogenicity high-binders (probability of immunogenicity ≥ 0.4 and $S_B \leq 0.05$; region I in (Fig.B.3)), followed by modest-immunogenicity high-binders (probability of immunogenicity < 0.4 and $S_B \leq 0.05$; region II in (Fig.B.3)), then high-immunogenicity modest-binders (probability of immunogenicity ≥ 0.4 and $S_B > 0.05$; region III in (Fig.B.3)), and modest-immunogenicity modest-binders (probability of immunogenicity < 0.4 and $S_B > 0.05$; region IV in (Fig.B.3)). For the antigens with length ≤ 100 aas, the S_B cutoff for the four regions was set to 0.1 and probability of immunogenicity threshold remained at 0.4. Predicted peptides in each section were re-ranked based on a total score defined as $S = S_B \cdot S_I$. Final ranked list was obtained by sequential appending of the re-ranked peptides from each region. The list of predicted peptides was ranked based on this total score ranging from lowest score to the highest score. The lower the total score of a predicted peptide, the higher its probability of being an immunogenic epitope. Workflow of the prediction strategy is shown in (Fig.3.3).

B.7 Statistical Analysis of Predicted CTL Epitopes

We used the F-test to quantify statistical significance ($P < 0.05$) of the variation of predicted rankings of T cell epitopes across different antigens between ANN-Hydro together with IEDB-consensus and IEDB-consensus alone

B.8 *In vivo* discovery of HIV-1 Gag Epitopes

Mice

C57BL/6 mice were obtained from Harlan Laboratories. All mice used were between 6 and 8 weeks of age. All animal study protocols were conducted in accordance with guidelines approved by the Institutional Animal Care and Use Committee at Kings College London and in full compliance with UK Home Office regulations under a project license to L.S.K.

Vaccine Immunization

Codon optimized HIV-1 gag plasmid DNA ZM96 from strain 96ZM651.8 (provided by B Hahn, through the Centre for AIDS Reagents [CFAR] UK) and codon optimized HIV-1 gag Consensus B plasmid DNA (provided by D Garber, Emory University, USA) were used to construct and propagate replication defective (E1, E3 deleted) recombinant Adenovirus type 5 (rAdHu5) vectors as described previously for the HIV-1 gag strain 97CN54 Bachy et al. (2013). Animals were immunized with 10^9 virus particles (vp) as determined by the DNA Pico-Green assay (Invitrogen) and administered either i.m. in the quadricep muscle (rAdHu5 Consensus B gag) or i.d. at the base of the tail (rAdHu5 ZM96 and rAdHu5 CN54).

Peptides

15mer peptides with an 11 amino acid overlap spanning the HIV-1 CN54 Gag protein and a 20mer set of peptides with 10 amino acid overlap spanning HIV-1 ZM96 were provided by CFAR, a set of 15mers with an 11 amino acid overlap spanning the HIV-1 Consensus subtype B Gag protein were provided from the NIH AIDS Reagent and Reference Program. 'Optimal' 9mer or 11mer peptides from HIV-1 CN54 Gag, ZM96 Gag and HIV-1 Consensus B were purchased from Proimmune.

T Cell Epitope Mapping by Intracellular Interferon Gamma Staining

Spleens were harvested 14 days after immunization, homogenized to single-cell suspensions, and RBCs were lysed using ACK lysis buffer (Lonza). Splenocytes were then used for *in vitro* re-stimulation, where 10^6 cells were incubated for 6 h at $37C$ with anti-CD28 ($2\mu g/ml$; BD Pharmingen), either alone (unstimulated control) or with peptides, either in pools or individually (each at $1\mu g$ peptide/ml), derived from Consensus B Gag. Brefeldin A ($10\mu g/ml$, Sigma-Aldrich) was added for the last 5 h of culture. After washing, cells were stained with anti-CD8 (clone 37.51, BD Biosciences) for 20 min, then fixed and permeabilized with the BD Cytofiix/Cytoperm Kit according to the manufacturers instructions, and then stained 30 min with anti-IFN-γ (clone XMG1.2, eBiosciences), washed and analyzed by flow cytometry. Consensus B epitopes were deconvoluted to individual 15mers from peptide pools, where each peptide is present in two independent pools within the matrix and reactive peptides confirmed in the second round against the 15mer peptide. Finally, based on the sequence of the reactive 15mer peptide, truncated versions of the 15mer peptides were synthesized and tested.

T Cell Epitope Mapping by ELISPOT Assay

14 days after immunization, splenocytes prepared (as detailed above) were re-stimulated *in vitro* with media alone, or with peptides, either in pools or individually (each at $1\mu M$ final concentration) derived from CN54 or ZM96 Gag on mouse anti-INF-γ antibody coated 96 well plates (U-Cytech) and incubated for 16 h at $37C$, 5% CO2. IFN-γ production was revealed according to the manufacturers instructions and IFN-γ spot forming cells (SFCs) enumerated using an immunospot image analyser (Bioreader 5000). In the first round, CN54 Gag epitopes were deconvoluted to individual 15mers from peptide pools, where each peptide is present in two independent pools within the matrix and reactive peptides confirmed in the second round against the 15mer peptide. Finally, based on the sequence of the reactive 15mer peptide, 9mer peptides were synthesized and tested. For ZM96 (due to the absence of a complete set of overlapping 15mer peptides), 49 individual 20mer peptides were tested. The reactive peptide sequences were confirmed against the corresponding 15mer peptide to the reactive sequence and then 9mer peptides synthesized and tested.

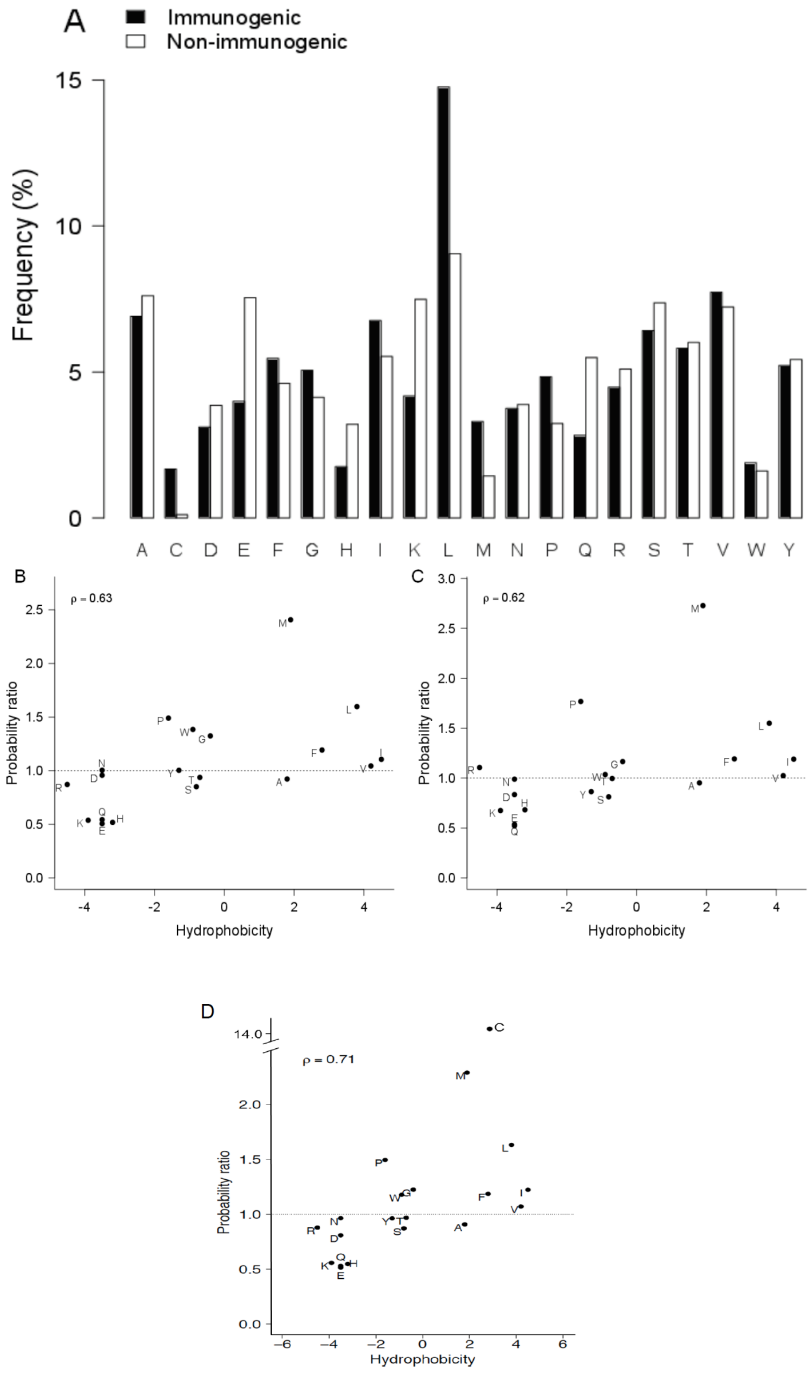


Figure B.1: Bias in amino acid usage between immunogenic and non-immunogenic MHC-I peptides. (A) Comparison of frequency distributions of amino acids between immunogenic and non-immunogenic datasets. (B) Probability ratio $(P(x\ I\ immunogenic)/P(x\ I\ non-immunogenic))$ of each amino acid as a function of its hydrophobicity, analyzed on just 9mer MHC-I peptides. (C) Probability ratio $(P(x\ I\ immunogenic)/P(x\ I\ non-immunogenic))$ of each amino acid as a function of its hydrophobicity, analyzed on 9mer HLA-I peptides excluding HLA-A2 restricted peptides. (D) Probability ratio of each amino acid as a function of its hydrophobicity; the plot shows cysteine (C) as an outlier in the immunogenic dataset.

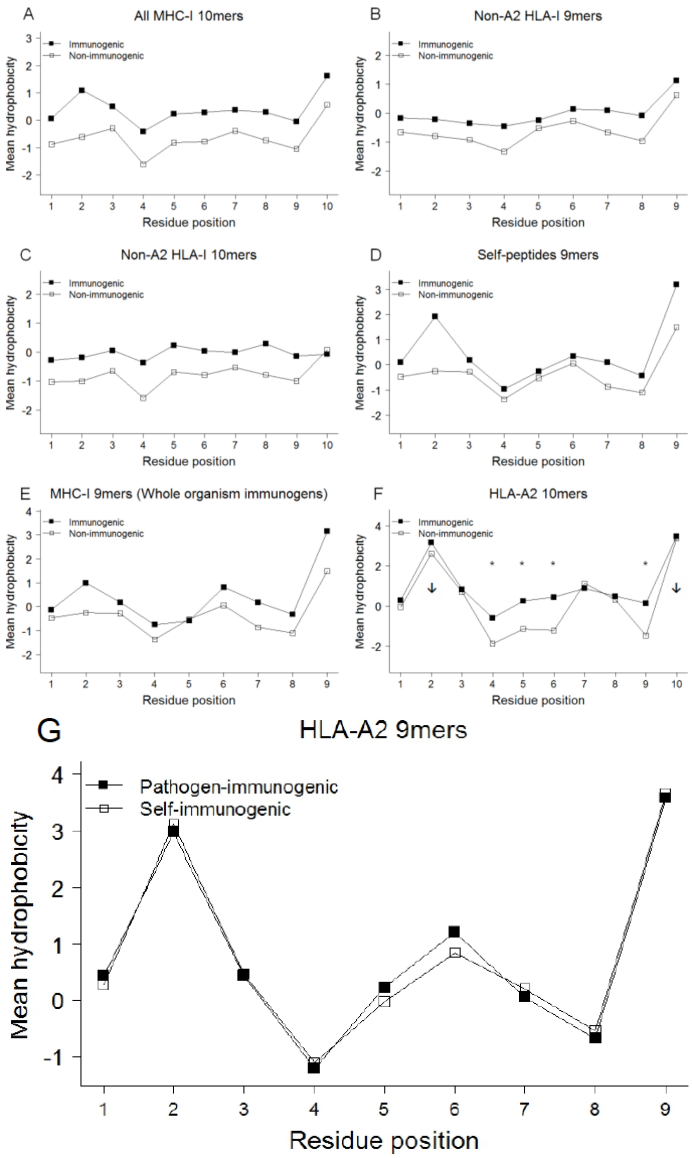


Figure B.2: Hydrophobicity comparison at each residue position between immunogenic and non-immunogenic MHC-I peptides, and immunogenic pathogen-derived and immunogenic self-epitopes. Each peptide sequence in the dataset was transformed into a numeric sequence based on amino acid hydrophobicity and the mean hydrophobicity at each position was computed. Unless indicated, analyzed peptides were not restricted to any MHC-motifs. (A) All immunogenic and non-immunogenic MHC-I 10mers; every single residue has $P < 2x10^{-7}$. (B) Human HLA-I immunogenic and non-immunogenic 9mers excluding HLA-A2 restricted peptides. (C) Human HLA-I immunogenic and non-immunogenic 10mers excluding HLA-A2 restricted peptides. (D) Immunogenic and non-immunogenic MHC-I 9mer self-peptides. (E) MHC-I 9mers peptides discovered using whole organism as immunogen as opposed to peptide-immunization experiments (non-immunogenic dataset same as (Fig.3.2A)) (F) Human HLA-A2 restricted immunogenic and non-immunogenic 10mers with arrows indicating anchor residues and stars for $P < 0.005$. (G) Human HLA-A2 restricted immunogenic pathogen-derived and immunogenic self 9mer epitopes. P-values for each figure were obtained using Wilcoxon rank-sum test and are shown in (Table B.4).

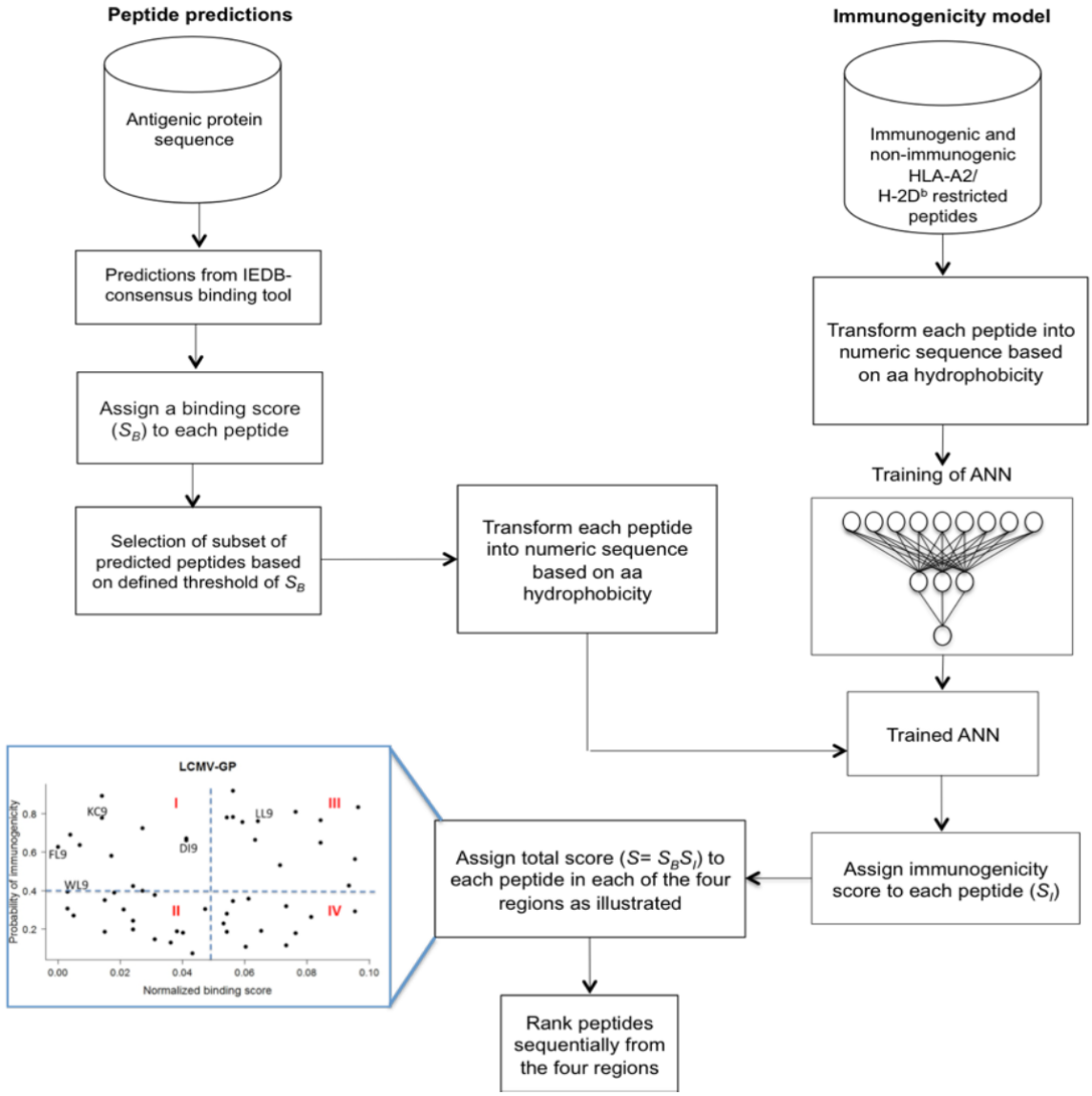


Figure B.3: Workflow for CTL epitope prediction using the ANN-Hydro model and the MHC-binding prediction tool IEDB-consensus. For training and application of the ANN-Hydro model for immunogenicity scores, each peptide sequence in the HLA-A2 and H-2D^b dataset was transformed into a corresponding numeric sequence based on the hydrophobicity value of amino acids. To obtain a list of candidates for MHC-bound peptides from a given antigen, IEDB-consensus binding algorithm was used and a normalized binding score (S_B) was assigned. The trained immunogenicity ANN model was applied on the same list of peptides independently to assign immunogenicity scores (S_I). After the subset of top binding peptides was selected, peptides from each region ranging from high-binding highly-immunogenic peptides to modest-binding low-immunogenic peptides (quadrants 1 through 4 in inset) were re-ranked based on total score $S = S_B \cdot S_I$. An example of epitope prediction is shown in the plot for experimentally defined H-2D^b restricted CTL epitopes from LCMV-GP. See Materials and Methods section Application of ANN-Hydro for full details.

Table B.1: Schematic of ConsB (top) and CN54 (bottom) 15mer peptide pools. Peptides were combined at $1\mu M$/each peptide such that each peptide occurs in only two pools numbered 7872 7994 for ConsB (top) or 7080.01-7080.121 for CN54 (indicated by 1121, bottom). Yellow highlight indicates positive response to peptide pool. Green highlight indicates positive response to individual 15mer peptide, and red indicates negative response to individual 15mer peptide.

	A	B	C	D	E	F	G	H	I	J	K	L
M	7872	7873	7874	7875	7876	7877	7878	7879	7880	7881	7882	7883
N	7884	7885	7886	7887	7888	7889	7890	7891	7892	7893	7894	7895
O	7896	7897	7898	7899	7900	7901	7902	7903	7904	7905	7906	7907
P	7908	7909	7910	7911	7912	7913	7914	7915	7916	7917	7918	7919
R	7920	7921	7922	7923	7924	7925	7926	7927	7928	7929	7930	7931
S	7932	7933	7934	7935	7936	7937	7938	7939	7940	7941	7942	7943
T	7944	7945	7946	7947	7948	7949	7950	7951	7952	7953	7954	7955
U	7956	7957	7958	7959	7960	7961	7962	7963	7964	7965	7966	7967
V	7968	7969	7970	7971	7972	7973	7974	7975	7976	7977	7978	7979
W	7980	7981	7982	7983	7984	7985	7986	7987	7988	7989	7990	7991
X	7992	7993	7994									

	P12	P13	P14	P15	P16	P17	P18	P19	P20	P21	P22
P1	1	2	3	4	5	6	7	8	9	10	11
P2	12	13	14	15	16	17	18	19	20	21	22
P3	23	24	25	26	27	28	29	30	31	32	33
P4	34	35	36	37	38	39	40	41	42	43	44
P5	45	46	47	48	49	50	51	52	53	54	55
P6	56	57	58	59	60	61	62	63	64	65	66
P7	67	68	69	70	71	72	73	74	75	76	77
P8	78	79	80	81	82	83	84	85	86	87	88
P9	89	90	91	92	93	94	95	96	97	98	99
P10	100	101	102	103	104	105	106	107	108	109	110
P11	111	112	113	114	115	116	117	118	119	120	121

Table B.2: Summary of identified epitopes. Responses to the RT9, VL8, AI9, and YI9 epitopes were observed for all three Gag protein variants, despite minor substitutions in the peptides. Overlapping sequences of individual peptides are shown. The QL11 epitope was only immunogenic for the CN54 Gag protein, but not ConsB or ZM96 Gag proteins, likely due to the A to E substitution at position 2. The SI9 epitope was only immunogenic in the ConsB Gag protein, as both CN54 and ZM96 had major deletions and substitutions in this sequence. MHC restriction was confirmed using MHC class I tetramer staining, and Gag amino acid positions are in reference to the HXB2 strain.

Source	Peptide	Amino Acid Sequence	Immunogenicity	Confirmed Epitope
Cons B	7889	QTGSEEL**RSLYNTVA**	+	H-2D^b
Cons B	7890	EEL**RSLYNTVAT**LYC	+	Gag76-84
CN54	7080.19	EEL**RSL**F**NTVAT**PYC	+	**RSLYNTVAT**
CN54	7080.20	**SL**F**NTVAT**P**Y**CVHTE	+	**RSLFNTVAT**
ZM96	7116.8	GTEEL**RSLYNTVAT**LYCVHE	+	**(RT9)**
Cons B	7912	EKAFSSPE**VIPMFSAL**	+	H-2K^b
Cons B	7913	SPE**VIPMFSAL**SEGA	+	Gag168-175
CN54	7080.41	F-SPE**VIPMFSAL**SEG	+	**VIPMFSAL**
CN54	7080.42	**VIPMFSAL**SEGATPQ	+	**VIPMFTAL**
ZM96	7116.17	EKAF-SPE**VIPMFT**A**L**SEGAT	+	**(VL8)**
Cons B	7920	GHQA**AMQMLKETI**NE	+	H-2D^b Gag197-205
Cons B	7921	**AMQMLKETI**NEEAAE	+	**AMQMLKETI**
CN54	7080.49	A**AMQ**I**LK**D**TI**NEEAA	+	**AMQILKDTI**
ZM96	7116.2	VGGHQA**AMQMLK**D**TI**NEEAA	+	**AMQMLKDTI** **(AI9)**
Cons B	7940	IVRM**YSPTSILDI**RQ	+	H-2D^b
Cons B	7941	**YSPTSILDI**RQGPKE	+	Gag277-285
CN54	7080.68	KIVRM**YSPTSILDI**K	+	**YSPTSILDI**
CN54	7080.69	M**YSPTSILDI**KQGPK	+	**YSPVSILDI**
ZM96	7116.28	NKIVRM**YSP**V**SILDI**KQGPK	+	**(YI9)**
Cons B	7949	ASQ**E**VKNWMTETLLV	-	H-2D^b
CN54	7080.77	QAT**QGVKNWMTDTL**L	+	Gag310-320
CN54	7080.78	**GVKNWMTDTL**LVQNA	+	**QGVKNWMTDTL**
ZM96	7116.32	Q**E**VKNWMTDTLLVQNA	-	**(QL11)**
Cons B	7963	EAM**SQVTNSATI**MMQ	+	H-2D^b
CN54	7080.91	AEAMSQ-TNSA-ILMQR	-	Gag368-376
ZM96	7116.37	RVLAEAMSQ-TNSVNILMQK	-	**SQVTNSATI** **(SI9)**

Table B.3: Amino acid property scales used for analyses Hydrophobicity scale (Kyte-Doolittle) Kyte and Doolittle (1982), Polarity (Grantham) Grantham (1974), and Bulkiness (Zimmerman) Zimmerman et al. (1968)

Amino acid	Hydrophobicity	Bulkiness	Polarity
Alanine (A)	1.8	11.5	8
Cysteine (C)	2.5	13.46	5.5
Aspartic acid (D)	-3.5	11.68	13
Glutamic acid (E)	-3.5	13.57	12.3
Phenylalanine (F)	2.8	19.8	5.2
Glycine (G)	-0.4	3.4	9
Histidine (H)	-3.2	13.69	10.4
Isoleucine (I)	4.5	21.4	5.2
Lysine (K)	-3.9	15.71	11.3
Leucine (L)	3.8	21.4	4.9
Methionine (M)	1.9	16.25	5.7
Asparagine (N)	-3.5	12.82	11.6
Proline (P)	-1.6	17.43	8
Glutamine (Q)	-3.5	14.45	10.5
Arginine(R)	-4.5	14.28	10.5
Serine (S)	-0.8	9.47	9.2
Threonine (T)	-0.7	15.77	8.6
Valine (V)	4.2	21.57	5.9
Tryptophan (W)	-0.9	21.67	5.4
Tyrosine (Y)	-1.3	18.03	6.2

Table B.4: Hydrophobicity comparison between immunogenic and non-immunogenic MHC class I peptides. *P*-values were calculated using Wilcoxon sum-ranked test using different datasets as described in main text.

Dataset	Sample size		Residue position in peptide									
	Immunogenic	Non Immunogenic	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
All MHC-I 9mers	2841	2660	$6.6x10^{-15}$	$< 2.2x10^{-16}$	$1.7x10^{-8}$	$< 2.2x10^{-16}$	$4.6x10^{-5}$	$1.6x10^{-5}$	$< 2.2x10^{-16}$	$< 2.2x10^{-16}$	$< 2.2x10^{-16}$	-
All MHC-I 10mers	1007	1013	$1.4x10^{-13}$	$< 2.2x10^{-16}$	$1.7x10^{-9}$	$< 2.2x10^{-16}$	$1.3x10^{-13}$	$1.7x10^{-15}$	$2x10^{-7}$	$3.3x10^{-12}$	$1.2x10^{-15}$	$< 2.2x10^{-16}$
NonA2 HLA-I 9mers	1318	1897	$1.4x10^{-6}$	$7.8x10^{-6}$	$7.6x10^{-8}$	$< 2.2x10^{-16}$	0.1561	$3x10^{-4}$	$2.6x10^{-13}$	$< 2.2x10^{-16}$	$3.5x10^{-14}$	-
NonA2 HLA-I 10mers	518	859	$3.4x10^{-7}$	$1.4x10^{-6}$	$1.4x10^{-6}$	$4.6x10^{-13}$	$8.8x10^{-8}$	$1.4x10^{-6}$	0.0023	$1.5x10^{-8}$	$2.9x10^{-8}$	0.2293
Self-peptides 9mers	565	2660	$2.1x10^{-4}$	$< 2.2x10^{-16}$	$1x10^{-4}$	$7.8x10^{-5}$	0.292	0.1423	$1.2x10^{-11}$	$2.2x10^{-7}$	$< 2.2x10^{-16}$	-
MHC-I 9mers (Immunogen=Organism)	323	2660	0.06781	$6.5x10^{-9}$	0.0071	$1x10^{-5}$	0.6156	$1.3x10^{-4}$	$1.5x10^{-8}$	$3.2x10^{-6}$	$< 2.2x10^{-16}$	-
HLA A2 9mers	936	321	0.6384	0.8929	0.8487	$6.3x10^{-12}$	0.2917	$3.1x10^{-7}$	$5x10^{-13}$	$< 2.2x10^{-16}$	0.08106	-
HLA A2 10mers	288	77	0.4029	0.2247	0.5054	0.00094	0.00087	$1x10^{-5}$	0.3368	0.9912	$5.5x10^{-6}$	0.09147
MHC H-2K^b 8mers	306	315	0.06678	0.9329	0.0055	0.2288	0.6742	$7x10^{-5}$	$1.1x10^{-6}$	0.1513	-	-
MHC H-2D^b 9mers	194	228	0.1825	0.0087	$2.1x10^{-4}$	0.3298	$4.9x10^{-10}$	0.7829	$1.1x10^{-4}$	0.0013	0.1273	-
	Immunogenic pathogenic	Immunogenic self										
HLA A2 9mers	1508	244	0.389	0.385	0.877	0.97	0.13	0.04	0.776	0.35	0.63	

Anchor residues
TCR contact residues

Table B.5: HLA-A2 restricted CTL epitopes for dengue virus 1 polyprotein and cytomegalovirus pp65 used in the rate analysis of predicted epitopes as shown in (Fig.3.3). All epitopes were retrieved from IEDB Vita et al. (2010).

Antigen	Epitope	Epitope length
CMV-pp65	NLVPMVATV	9
	MLNIPSINV	9
	VLGPISGHV	9
	RLLQTGIHV	9
	LMNGQQIFL	9
	ILARNLVPM	9
	SLILVSQYT	9
	SIYVYALPL	9
	VIGDQYVKV	9
	YLESFCEDV	9
	AMAGASTSA	9
	KYQEFFWDA	9
	GLSISGNLL	9
	RQYDPVAAL	9
	VAALFFFDI	9
	ALFFFDIDL	9
	KISHIMLDVA	10
	SDNEIHNPAV	10
	FTWPPWQAGI	10
	LLCPKSIPGL	10
Dengue-Polyprotein	VLMLVAHYA	9
	ILLMRTTWA	9
	MLLALIAVL	9
	TLYAVATTI	9
	QEGAMHTAL	9
	LPAIVREAI	9
	SRNSTHEMY	9
	AIVREAIKR	9
	YLPAIVREA	9
	TLLCLIPTV	9
	VLNPYMPSV	9
	LMMMLPATL	9
	VTYECPLLV	9
	MMMLPATLA	9
	IILEFFLMV	9
	KTDFGFYQV	9
	VQADMGCVV	9
	GLLFMILTV	9
	QLWAALLSL	9
	LLMRTTWAL	9
	CLMMMLPATL	10
	ELMRRGDLPV	10
	MLLILCVTQV	10
	FLMVLLIPEP	10
	TLMLLALIAV	10
	LMLLALIAVL	10
	IILEFFLMVL	10
	TLTAAVLLLV	10
	VLLLVTHYAI	10
	ITLLCLIPTV	10
	KVLNPYMPSV	10
	HQLWATLLSL	10
	YTPEGIIPTL	10
	SIILEFFLMV	10
	LSMGLITIAV	10
	NQLIYVILTI	10
	LMMMLPATLA	10
	TLMAMDLGEL	10
	FTMGVLCLAI	10

Table B.6: Probabilities assigned by the ANN-Hydro A2-model for HLA-A2 restricted 9mer CTL epitopes. Three recent epitope discovery studies Assarsson et al. (2008); Weiskopf et al. (2011); Newell et al. (2013) that were based on a proteome-wide screen of various viral antigens and self-epitopes were chosen for assessment of the predictive capacity of the 'A2-model'. Neo-epitopes were obtained from rotavirus Newell et al. (2013) dengue virus Weiskopf et al. (2011) and influenza A Assarsson et al. (2008) and other positive control epitopes from several antigens (pathogenic and self) were obtained from Newell et al. Newell et al. (2013). Any epitope that was present in the training set for ANN-Hydro was removed. A cutoff probability (p-ANN-Hydro) of 0.4 was set for a positive "hit".

Source	Epitope	Antigen	p-ANN-Hydro	Reference
Neo-epitopes				
Rotavirus	SLISGMWLL	Rota-VP2_4	0.89	Newell et al.
	TLLANVTAV	Rota-VP6_4	0.87	
	FLDSEPHLL	Rota-NSP1_2	0.84	
	LLNYILKSV	Rota-VP7_1	0.37	
Influenza-A (FluA)	QIAILVTTV	NA	0.90	Assarsson et al.
	GLIYNRMGA	M1	0.89	
	GILGFVFTL	Flu_1	0.80	
	FVEALARSI	PB1	0.58	
	VMNILLQYL	GAD	0.57	
	FVANFSMEL	PB1	0.54	
	TTYQRTRAL	NP	0.47	
	GLADQLIHL	HIV_7	0.47	
Dengue-Virus 2 (DENV-2)	GLLTVCYVL	NS2B	0.88	Weiskopf et al.
	RLITVNPIV	E	0.88	
	IMAVGMVSI	NS2B	0.85	
	IILEFFLIV	NS4A	0.79	
	ALSELPETL	NS4A	0.61	
	YLPAIVREA	NS3	0.50	
	KLAEAIFKL	NS5	0.44	
	AAAWYLWEV	NS3	0.27	
Positive Control epitopes				
Human herpesvirus 5 (CMV)	ALFFFDIDL	CMV_5	0.84	Newell et al.
	LMNGQQIFL	CMV_18	0.81	
	RIFAELEGV	CMV_22	0.81	
	QMWQARLTV	CMV_21	0.78	
	NLVPMVATV	CMV_1	0.75	
	VLEETSVML	CMV-IE1	0.72	
	FLMEHTMPV	CMV_8	0.61	
	IIYTRNHEV	CMV_13	0.54	
	SLLSEFCRV	CMV_23	0.52	
	ILSPLTKGI	CMV_15	0.46	
	VLAELVKQI	CMV_2	0.24	
Human herpesvirus 4 (EBV)	GLCTLVAML	EBV_2	0.87	Newell et al.
	YVLDHLIVV	EBV_1	0.84	
	YLQQNWWTL	EBV_5	0.79	
	CLGGLLTMV	EBV_4	0.79	
	YLLEMLWRL	EBV_3	0.28	
Influenza-A (FluA)	FLDIWTYNA	Flu_4	0.85	
	NMLSTVLGV	Flu_14	0.78	
	LLIDGTASL	Flu_12	0.68	

	FMYSDFHFI	Flu_5	0.63	
	MMMGMFNML	Flu_13	0.48	
	GMFNMLSTV	Flu_7	0.29	
	RLIDFLKDV	Flu_15	0.23	
Hepatitis B virus (HBV)	WLSLLVPFV	HBV_2	0.89	Newell et al.
	FLLSLGIHL	HBV_5	0.74	
	FLLTRILTI	HBV_1	0.70	
Human Immunodeficiency virus (HIV-1)	NVWATHACV	HIV_1	0.94	
	TLNAWVKVV	HIV_2	0.86	
	KLTPLCVTL	HIV_4	0.84	
	SLYNTVATL	HIV_5	0.61	
	ALVEMGHHA	HIV_8	0.44	
	ILKEPVHGV	HIV_9	0.38	
	LTFGWCFKL	HIV_6	0.36	
Mycobacterium *tuberculosis*	GLPVEYLQV	TB_1	0.85	
	KLIANNTRV	TB	0.79	
Plasmodium *falciparum*	YLNKIQNSL	CSP	0.79	
LCMV	YLVSIFLHL	LCMV	0.77	
Herpes simplex virus (HSV-1)	SLPITVYYA	HSV1/2	0.90	
RSV	KMLKEMGEV	RSV	0.29	
Self Antigens	ALWMRLLPL	pp-Insulin	0.83	
	YMCSFLFNL	EZH2	0.51	
	YMDGTMSQV	Tyrosinase	0.49	

Table B.7: Ranking comparison of all the predicted epitopes (Prevalidation and in vivo validation) used in this study as shown in (Fig.3.5 A). The predictions used are as follows: ANN-Hydro - ANN-hydrophobicity prediction model combined with normalized binding scores from prediction algorithms, IEDB-Bind -IEDB consensus binding tool, NetMHC-Bind - NetMHCpan binding tool, SYFPEITHI - SYFPEITHI epitope prediction tool, IEDB-Prot-IEDB recommended processing prediction, ANN-Prot - IEDB processing predictions using ANN. p-ANN-Hydro Probability of immunogenicity assigned by the corresponding (H-2D^b or A2) ANN-Hydro immunogenicity model. All references were obtained from IEDB Vita et al. (2010) and are indicated.

Antigen	Epitope	MHC	Length	IEDB bind	Syfpeithi	NetMHC.bind	IEDB.prot	ANN.prot	ANN Hydro + IEDB.bind	p-ANN-Hydro	Reference
LCMV-GP	FALISFLLL	H-2D^b	9	1	10	1	3	1	1	0.62	(11)
	WLVTNGSYL		9	3	1	2	5	4	11	0.4	(12)
	LIDYNKAAL		9	45	12	39	68	77	32	0.8	(13)
	KAVYNFATC		9/11	8	9	10	39	53	5	0.77	(11, 14)
	DEVINIVII		9	24	4	74	133	115	10	0.66	(13)
LCMV-NP	FQPQNGQFI		9	1	1	1	2	1	1	0.69	(14)
	SEVSNVQRI		9	7	2	7	37	50	12	0.14	(13)
Ad.v.T. antigen	VNIRNCCYI		9	1	1	1	1	1	13	0.31	(15)
	CSDGNCHLL		9	21	4	9	11	44	20	0.8	(15)
Flu-NP	ASNENMETM		9	1	1	1	1	2	1	0.87	(16)
	RLIQNSLTI		9	3	3	2	3	3	2	0.67	(16)
	GERQNATEI		9	18	2	36	100	103	8	0.42	(16)
	YRRVNGKWM		9	19	4	35	80	65	9	0.44	(16)
FluA-Neuraminidase	FCGVNSDTV		9	3	2	4	11	3	13	0.35	(17)
	ITYKNSTWV		9	4	8	3	4	2	1	0.52	(17)
	YRYGNGVWI		9	5	7	11	29	7	4	0.45	(17, 18)
Consensus Gag	SQVTNSATI		9	1	1	1	1	1	7	0.2	This study
	AMQMLKETI		9	4	9	6	4	2	9	0.19	
	YSPTSILDI		9	6	19	4	11	3	11	0.39	
	RSLYNTVAT		9	3	32	5	45	29	1	0.82	
ZM96 Gag	AMQMLKDTI		9	1	4	4	3	2	13	0.17	This study
	YSPVSILDI		9	5	12	3	9	1	3	0.43	
	RSLYNTVAT		9	4	28	5	46	27	2	0.77	
97CN54 Gag	AMQILKDTI		9	1	4	4	3	2	13	0.16	This study
	YSPTSILDI		9	5	19	2	9	3	15	0.36	
	RSLFNTVAT		9	2	35	3	40	23	2	0.76	
Melan-A	ALMDKSLHV	A2	9	1	3	1	1	1	1	0.65	(19)
	GILTVILGV		9	2	1	2	2	2	2	0.86	(19)
	ILTVILGVL		9	7	5	5	4	5	8	0.9	(19)
	AAGIGILTV		9/10	9	4	7	19	17	9	0.86	(19)
Wt-1	SLGEQQYSV		9	1	3	3	3	3	1	0.79	(20)
	RMFPNAPYL		9	2	10	2	2	2	2	0.53	(20)
	ALLPAVPSL		9	3	1	1	1	1	3	0.44	(20)
	DLNALLPAV		9	6	2	8	16	27	4	0.9	(20)
	VLDFAPPGA		9	7	31	7	15	13	8	0.87	(20)
	KLGAAEASA		9	9	16	9	21	26	5	0.94	(20)
	NLGATLKGV		9	12	4	10	12	14	12	0.76	(20)
	CMTWNQMNL		9	13	27	11	7	9	13	0.69	(20)
	RVPGVAPTL		9	25	17	21	11	11	19	0.81	(20)
gp100	RLMKQDFSV		9	1	21	1	2	2	1	0.83	(21)

	MLGTHTMEV	A2	9	2	15	2	3	1	2	0.65	(21)
	KTWGQYWQV		9	5	62	3	4	3	5	0.5	(21)
	YLEPGPVTA		9	16	20	19	34	25	11	0.93	(21)
TRAG-3	GLIQLVEGV		9	1	1	1	1	1	1	0.43	(22, 23)
	HACWPAFTV		9	9	10	6	9	20	9	0.82	(22, 23)
	SILLRDAGL		9	6	2	8	5	7	5	0.92	(22, 23)
	ILLRDAGLV		9	3	3	2	4	4	3	0.7	(22, 23)
	ALSKFPRQL		9	4	5	4	3	2	4	0.34	(22, 23)
p53	RMPEAAPPV		9	1	1	8	2	2	1	0.51	(24)
	LLGRNSFEV		9	2	2	4	1	4	2	0.46	(24)
	VVPCEPPEV		9	13	14	21	14	11	9	0.77	(25)
	YQGSYGFRL		9	5	5	65	3	3	10	0.41	(24)
	KTCPVQLWV		9	14	15	19	34	25	11	0.77	(24)

Table B.8: Ranked list of top 20 predicted peptides for each of the Gag variant using the ANN-Hydro combined with normalized binding scores from predictions. (S_B) - Binding score, p-ANN-Hydro - probability of immunogenicity obtained by applying ANN-Hydro model to each peptide, (S) - Total score. This list was ranked based on total score S ranging from lowest score to the highest score within each section (I through IV) classified based on p-ANN-Hydro and S_B (see section Application of ANN-Hydro).

ConsB Gag predictions				
Rank	Epitope	Binding score (S_B)	p-ANN-Hydro	Total score (S)
1	**RSLYNTVAT**	**0.006**	**0.82**	**0.001**
2	ATPQDLNTM	0.033	0.77	0.008
3	QVSQNYPIV	0.049	0.83	0.008
4	RFAVNPGLL	0.025	0.66	0.008
5	RMYSPTSIL	0.039	0.72	0.011
6	KARVLAEAM	0.021	0.40	0.013
7	**SQVTNSATI**	**0.000**	**0.20**	**0.000**
8	SQVSQNYPI	0.004	0.12	0.004
9	**AMQMLKETI**	**0.008**	**0.19**	**0.006**
10	GWMTNNPPI	0.008	0.14	0.007
11	**YSPTSILDI**	**0.020**	**0.39**	**0.012**
12	RSLFGNDPS	0.023	0.27	0.017
13	ASVLSGGEL	0.022	0.07	0.020
14	KALGPAATL	0.036	0.36	0.023
15	AAMQMLKET	0.039	0.33	0.026
16	VQNANPDCK	0.053	0.84	0.009
17	SALSEGATP	0.054	0.84	0.009
18	LLVQNANPD	0.051	0.76	0.012
19	ASLRSLFGN	0.096	0.79	0.020
20	SLYNTVATL	0.061	0.64	0.022

ZM96 Gag predictions				
Rank	Epitope	Binding score (S_B)	p-ANN-Hydro	Total score (S)
1	MSQTNSVNI	0.000	0.64	0.000
2	**RSLYNTVAT**	**0.001**	**0.78**	**0.000**
3	**YSPVSILDI**	**0.004**	**0.43**	**0.002**
4	YMIKHLVWA	0.016	0.77	0.004
5	KVSQNYPIV	0.032	0.83	0.005
6	ATPQDLNTM	0.028	0.75	0.007
7	RMYSPVSIL	0.040	0.80	0.008
8	VQNANPDCK	0.048	0.83	0.008
9	LLVQNANPD	0.046	0.81	0.009
10	RFALNPGLL	0.027	0.67	0.009
11	KARVLAEAM	0.016	0.42	0.009
12	NFLQNRPEP	0.042	0.61	0.017
13	**AMQMLKDTI**	**0.000**	**0.18**	**0.000**
14	KSLFGSDPL	0.000	0.08	0.000
15	KALGPGATL	0.026	0.36	0.017
16	KIVRMYSPV	0.026	0.18	0.021
17	IMKQLQPAL	0.039	0.34	0.026
18	VKNWMTDTL	0.033	0.18	0.027
19	AWMTSNPPI	0.033	0.10	0.030
20	WMTSNPPIP	0.092	0.92	0.007

CN54 Gag predictions				
Rank	Epitope	Binding score (S_B)	p-ANN-Hydro	Total score (S)
1	MSQTNSAIL	0.002	0.85	0.0003
2	**RSLFNTVAT**	**0.002**	**0.76**	**0.0004**
3	YMLKHLVWA	0.018	0.72	0.005
4	KVSQNYPIV	0.032	0.83	0.005
5	ATPQDLNTM	0.028	0.78	0.006
6	SALSEGATP	0.049	0.84	0.008
7	RFALNPGLL	0.027	0.70	0.008
8	VQNANPDCK	0.048	0.82	0.008
9	LLVQNANPD	0.046	0.78	0.010
10	RMYSPTSIL	0.034	0.70	0.010
11	NFLQNRPEP	0.042	0.65	0.015
12	SALQTGTEE	0.042	0.57	0.018
13	**AMQILKDTI**	**0.000**	**0.16**	**0.000**
14	KAKVLAEAM	0.012	0.35	0.008
15	**YSPTSILDI**	**0.015**	**0.36**	**0.010**
16	RALGPGASI	0.020	0.24	0.015
17	KSLFGNDPS	0.025	0.28	0.018
18	IMKQLQSAL	0.037	0.33	0.025
19	VKNWMTDTL	0.033	0.20	0.027
20	WMTSNPPVP	0.077	0.92	0.006