QUANTIFYING THE KINETICS OF IMMUNE RESPONSES TO CANCER: APPLICATIONS IN SERUM PROFILING AND SINGLE-CELL BIOLOGY

A Dissertation

Presented to

the Faculty of the Department of Chemical and Biomolecular Engineering

University of Houston

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

in Chemical and Biomolecular Engineering

by

Jay R T. Adolacion

August 2019

QUANTIFYING THE KINETICS OF IMMUNE RESPONSES TO CANCER: APPLICATIONS IN SERUM PROFILING AND SINGLE-CELL BIOLOGY

Jay R T. Adolacion

Approved:

Chair of the Committee,
Navin Varadarajan, Associate Professor,
Chemical and Biomolecular Engineering

Committee Members:

Richard C. Willson, Professor,
Chemical and Biomolecular Engineering

Patrick C. Cirino, Associate Professor,
Chemical and Biomolecular Engineering

Mehmet Sen, Assistant Professor,
Biology and Biochemistry

Chandra Mohan, Professor,
Biomedical Engineering

Suresh K. Khator, Associate Dean,
Cullen College of Engineering

Michael P. Harold, Professor and Chair,
Chemical and Biomolecular Engineering

Acknowledgements

When I first began this journey, I just thought of it as a way out — to start a new life far away from home. I wasn't sure what I was getting myself into but, back then, all I was thinking was it must be way better than how things were during that time. Arriving here, I found myself homesick and wondering if I made the right decision. Little did I know that I would discover something I love doing — data analysis — in a subject area that I least liked back in high school — biology. I wouldn't have been able to make it this far if not for the support of the many people I'm indebted to.

First, I would like to thank Dr. Rizalinda de Leon for a lot of things — for vouching for me to receive the fellowship, for being one of my guarantors, for her motherly advice whenever I doubt myself. I don't know what you see in me, but you've always encouraged me to be my best. To my Master's advisor, Dr. Maria Lourdes Dalida, I've always found your zest for life very contagious. Thank you for teaching me to take life a little less seriously and to take it on with a smile. And to my students who sent me off with a journal of your thoughts and wishes for me, I'm grateful as your words have been a source of comfort during difficult times.

I would like to thank my Ph.D. advisor, Dr. Navin Varadarajan, who told me I have a knack at data analysis and for essentially shaping my career in this area through the different project opportunities you provided. To my Ph.D. co-advisor, Dr. Richard Willson, thank you for sharing your insights and experiences; I have always admired your magnanimity. My thanks to Dr. Patrick Cirino, Dr. Mehmet Sen, and Dr. Chandra Mohan for accepting to be in my panel and for the helpful comments and feedback regarding my work.

I also want to extend my gratitude to Dr. David Peabody and Dr. Bryce Chackerian from the University of New Mexico who allowed me to come to their laboratory and

train on panning with their VLP library platform. During my two-month stay there, I was assigned to Dr. Jerri Caldeira, whose invaluable mentorship I'm very grateful for.

To my colleagues in the Varadarajan lab and in the Willson lab, thank you for all the help and support you've given. I especially would like to give a shout-out to Dr. Katerina Kourentzi, Dr. Maryan Crum, Dr. Binh Vu, Dr. Gavin Garvey, Dr. Melisa Martinez Paniagua Grant, and Dr. Xingyue An, for looking after me when I was at the end of my wits. And special thanks to the folks in the Cirino lab for allowing me to use their equipment.

I would like to thank my Houston family and friends who've helped settle down here and made my stay here very memorable. To my best buds, Kristian and Milbert, thank you for your unwavering support and for putting up with the craziness. And to all the other people who've helped me in one way or another but I wasn't able to name specifically here, you know who you are and couldn't imagine making it without you.

Last but not least, to my parents — for my dad who couldn't be with us here and for my mom who endured a lot for us siblings — I hope I made you both proud. *Maraming maraming salamat po*.

To God be the greater glory....

QUANTIFYING THE KINETICS OF IMMUNE RESPONSES TO CANCER: APPLICATIONS IN SERUM PROFILING AND SINGLE-CELL BIOLOGY

An Abstract

of a

Dissertation

Presented to

the Faculty of the Department of Chemical and Biomolecular Engineering

University of Houston

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

in Chemical and Biomolecular Engineering

by

Jay R T. Adolacion

August 2019

Abstract

Immunotherapy that harnesses the body's immune system to fight cancer has revolutionized the treatment of the disease. Advances in utilizing checkpoint inhibitors that release the potential of the adaptive immune system, and in the design and manufacture of chimeric antigen receptor (CAR) T cells that utilize T cells as living drugs, has served to alter the landscape of cancer treatment. Profiling responses to immunotherapy requires the development of newer methods that can deal with the breadth and complexity of these responses. In this dissertation, I demonstrate the advancement of two approaches that answer central immunological questions: (1) measuring the breadth of humoral responses elicited upon the development of cancer and whether these have prognostic/therapeutic potential, and (2) modeling the kinetics of the single-cell killing mediated by T cells to identify mechanisms for the manufacture of more potent cells for immunotherapy.

In the first part, a combinatorial library of phage-displayed linear dodecapeptides was bio-panned against plasma samples from a cohort of AML patients undergoing checkpoint therapy. Subsequent to recovery of phage particles and high-throughput sequencing, we utilized a novel biodiversity-based analysis to identify candidate peptides. We validated this workflow by profiling humoral responses elicited upon seasonal vaccination against influenza. By utilizing this same methodology for the interrogation of the plasma of AML patients, we demonstrated the discovery and characterization of peptides derived from AML-specific oncoprotein fusions.

In the second part, we sought to understand the mechanistic basis of the failure of CAR T cells to kill antigen-positive target cells. We utilized single-cell timelapse imaging of CD19-targeting CAR T cells and their interactions with leukemia cells to distinguish killer and non-killer CAR T cells. We hypothesized that the nature in which effector-target associations transition from conjugation to killing/detachment occurs in a

gamma-distributed fashion, implying a sequential transition across intermediate states during contact. Our modeling results showed that kill events are kinetically homogeneous, characteristic of a single rate-limiting step, whereas no-kill events were heterogeneous mixtures of abortive and persistent contacts. The results of the model were validated by microscopy experiments that illustrated defects in lysosome polarization and degranulation within non-killer CAR T cells.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Serology and multiplexing

Serological testing form part of standard laboratory practices in diagnosing a number of different illnesses. However, conventional tests rely on knowledge of the antigen that triggered the immune response. As such, diagnosis based on these techniques are limited to antigen availability and are typically confirmatory for a suspected disease. On the other hand, the antibodies generated in the process provide amplified early means of detection as the body is capable to raising antibody titers against low antigen levels.

Antibody reactivity is dictated by specific regions on the antigen's surface called epitopes. Therefore, instead of presenting the actual antigen for antibody binding, probing the antibodies with epitope-mimicking peptide sequences called mimotopes can be done to virtually deduce the antigen. This approach enables multiple epitope discovery which can yield a list of biomarkers that can be traced to illnesses a person has been exposed to.

Epitope sequences can be 3 to 85 amino acids in length (Harinder Singh, Ansari, and Raghava, 2013); in contrast, combinatorial peptide libraries are typically 6 to 15 amino acids long (Gershoni et al., 2007). Consequently, several single mimotopes can span and produce multiple validation to the same relatively shorter epitope sequence but the rest can

only be completely mapped out by a collection of affinity-selected mimotopes (Ryvkin et al., 2012).

Surveying these mimotopes for unique epitope sequences of a disease or group of diseases creates a need for fast, affordable, accurate and easy-to-use sequencers. Currently, a number of next-generation sequencing systems available have characteristics favorable for clinical use (Desai and Jere, 2012).

Another aspect for consideration is the availability of databases that would contain epitopes that can uniquely identify a disease. As an antigen can have several epitopes that can mount an immune response, sequence matching is facilitated by searching for immunodominant epitopes to which majority of antibodies are directed against. With a growing number of reported epitopes available (Charoentong et al., 2012; Salimi et al., 2012; Vita, Vaughan, et al., 2006), a knowledge base of immunodominant epitopes as disease biomarkers can be consolidated. However, in the context of other diseases like cancer, such databases may be unavailable or unsatisfactory. On the other hand, the disease may be at a level that would be missed or difficult to assess by standard diagnostic methods. Thus, a method that can detect deviations from healthy patient samples would be beneficial.

Standard diagnostic methods are designed for the detection of a single analyte. However, to understand the complexity of a disease — the genetic and environmental factors that affect it — would require combining biological information from a number of disease markers. The ability to survey the breadth of immune responses across different individuals and profile for differences in serological signatures in an unbiased and comprehensive manner would facilitate a better understanding of the underlying biology. Thus, the need to simultaneously detect multiple analytes, that is, to multiplex.

1.2 Antibodies

Antibodies are Y-shaped proteins produced by the immune system in response to an immunological insult. A substance that elicits an immune response is called an antigen. As part of the adapted immune response, these proteins bind with high affinity to specific regions on the surface of an antigen known as an epitope. There can be multiple epitopes to the same antigen. Also known as immunoglobulins (Ig), these proteins originated from B-cell activation for its cognate antigen. Depending on the B cell they originated from, these antibodies come in variety shapes and affinities.

1.2.1 Structure and isotypes

Antibodies are composed of a variable region, which is the portion that binds to the antigen, and a constant region, which determines the antibody's isotype (Janeway Jr et al., 2001). As such, there are five different isotypes: IgM, IgD, IgG, IgE, and IgA. Subclasses are also found within certain isotypes. For example, IgGs can be further subdivided into IgG_1, IgG_2, IgG_3, and IgG_4, while IgAs can be IgA_1 or IgA_2. These subclasses differ in the length of their constant region and are conferred a variety of properties involving antigen affinity and effector functions (Vidarsson, Dekkers, and Rispens, 2014). In addition, certain antibody types are capable of producing multimeric forms. IgMs can form pentamers while IgAs form dimers which is the form found in mucosal secretions.

Majority of antibodies in normal sera are IgGs, comprising around 80 % of the total antibody concentration (Loh, Vale, and McLean-Tooke, 2013) and its quantitation forms the basis of many conventional antibody-based diagnostic tests.

1.2.2 Epitopes

Epitopes recognized by antibodies can be divided into two categories: linear or conformational. Linear epitopes are also called continuous epitopes as these consist of consecutive residues. It is the opposite case for conformational epitopes which contain residues which are in close proximity to each other in the native protein but are made up of discontinuous segments when the protein is unfolded. Most B-cell epitopes are conformational, with estimates numbering more than 90 % (Barlow, Edwards, and Thornton, 1986; Van Regenmortel, 2001). Despite this, most conformational epitopes contain short consecutive segments of four to seven in length (Berglund et al., 2008). Epitope lengths recognized by antibodies span about 12 to 15 residues; however, not all the residues participate in binding, with most of the binding estimated to be attributable to around five residues while the rest acts as scaffolding (Sykes, Legutki, and Stafford, 2013). Greer et al. (1997) observed that two known epitopes that differ in only one residue can induce different pathologies in mice, with only one of the two epitopes inducing encephalomyelitis. Thus, even a substitution of only one residue but at the correct position can result to new epitopes that map to different outcomes.

1.2.3 Immunodominance

Immune responses are initially directed to multiple epitopes. However, this response gradually focuses onto a few epitope targets favoring a select set of what are called immunodominant epitopes. Angeletti et al. (2017) observed this hierarchy of antibody response immunodominance towards the hemagglutinin globular domain of influenza A virus in mice. Interestingly, they showed that this hierarchy varied with vaccination in terms of antigen formulation and site of injection. In the view point of improving vaccine efficacy,

due to the high mutation rates in influenza, it is desirable to target conserved epitopes; however, these epitopes are frequently subdominant, eliciting a lesser immune response (Silva et al., 2017). By selective elimination of immunodominant B cells through the administration of the soluble antigen to active germinal centers, Silva et al. (2017) was able to create conditions favoring the expansion of subdominant B cells.

1.2.4 Autoantibodies

Under normal circumstances, the immune system's checks and balances allows it to detect foreign pathogens or damaged cells while sparring healthy tissue. However, in cases of cancer or autoimmune diseases, the boundary between self and nonself proteins becomes blurred and the body begins to produce antibodies to itself. As little to no expression of autoantigens that generate these autoantibodies is expected in healthy individuals, their detection is of interest as potential biomarkers (Pedersen and Wandall, 2011). Nevertheless, the presence of these autoantibodies does not necessitate a positive diagnosis as other factors may play a role toward disease progression (Q.-Z. Li et al., 2011; Slight-Webb et al., 2016).

In cancer, a number of mechanisms have been proposed that drives autoantibody production (Zaenker, Gray, and Ziman, 2016). The escape into circulation of lymphocytes that can recognize autoantigens from clonal deletion or regulatory T cell inactivation (Ding and Yan, 2007; H.-J. Kim et al., 2010), chronic inflammation (Nanda and Sercarz, 1995), abnormal protein expression (Alexandrov et al., 2013; Anderton, 2004; Hanash, 2003; Simpson et al., 2005; Watanabe et al., 2000), and release of cellular material from cell death (Bei et al., 2009; Doyle and Mamula, 2001) are theories that have been put forward as explanations to this phenomenon. All these categories highlight the combination of immune dysregulation and aberrant protein expression for inducing immunogenicity.

The same mechanisms are believed to be in play for autoimmune disease (Suurmond and Diamond, 2015).

1.3 Challenges in diagnostics

The expert committee convened by Institute of Medicine identified that there are more research allocations to treatment compared to diagnosis; however, there is a strong incentive to improve inaccuracies or delays in diagnosis due to its prevalence and to address the challenges it imposes on the current health care system (National Academies of Sciences and Medicine, 2016). The inherent complexities involved in the diagnostic process highlights the need for better measures in place to safeguard against diagnostic errors (figure 1).

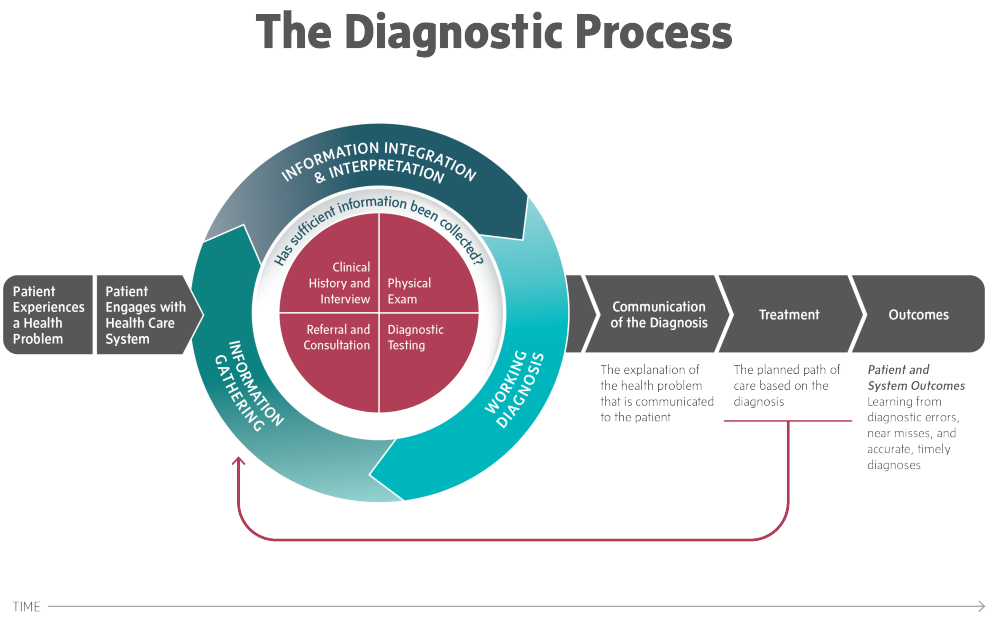


Figure 1: A conceptual model of the diagnostic process (National Academies of Sciences and Medicine, 2016).

In the US, Tehrani et al. (2013) surveyed the malpractice claims from 1986 to 2010 and found that more than a third were diagnosis-related (table 1). The highest number of disabilities or deaths were attributed to diagnostic errors and were roughly equal in frequency, leading the authors to call critical attention toward improving diagnostic safety.

Table 1: Attributable fraction of severe outcomes by malpractice allegation group (1986[†]-2010) (Tehrani et al., 2013)

Malpractice allegation group	Attributable fraction of disability (%)	Attributable fraction of death (%)	Attributable fraction of disability or death (%)
Diagnosis related	33.8	39.3	36.5
Surgery related	21.5	11.3	16.4
Treatment related	18.3	24.3	21.3
Obstetrics related	14.7	5.5	10.1
Medication related	3.6	7.1	5.4
Other (all categories with <5% of total each)	8.2	12.5	10.3
Total	100	100	100

[†] Data on severity of injury are available in the NPDB only for claims filed after 31 January 2004. This table does not include claims filed before 2004, so claims paid during the 1980s and 1990s are under-represented in these analyses.
NPDB, National Practitioner Data Bank.

Kostopoulou, Delaney, and Munro (2008) reviewed features common among misdiagnosed diseases and classified them into five categories (table 2). Classifications were based on some disease features being missing or unexpected (atypical presentation), features that are too common among diseases to be distinguishing (non-specific presentation), disease rarity that makes it unlikely to be diagnosed right away (very low prevalence), disease that is masked by the presence of one or more conditions (co-morbidity), or disease features that are easily missed or difficult to recognize (perceptual features). Of the diseases listed, many are notably types of cancer.

One avenue to reduce diagnostic errors would be improvements in diagnostic testing, and multiplexing biomarker signatures may be key to reducing diagnostic errors

Table 2: Conditions and associated features of difficulty (Kostopoulou, Delaney, and Munro, 2008).

Conditions[†]	Features of difficulty				
	Atypical presentation	Non-specific presentation	Very low prevalence	Co-morbidity	Perceptual features
Breast cancer	✗				
Testicular cancer	✗				
Oral cancer	✗				
Mycocardial infarction	✗				
Meningococcal disease	✗	✗			
Dementia (and depression)	✗			✗	
Asthma		✗			
Childhood cancers		✗	✗		
Upper gastrointestinal cancer		✗			
Tremor in the elderly		✗			
Metastatic spinal cord compression		✗	✗		
Iron deficiency anemia		✗			
Tongue cancer	✗		✗		
Retinoblastoma	✗	✗	✗		
Cancers (various)				✗	
HIV			✗		✗

[†] Conditions in bold suggest misattribution of symptoms to an obvious aetiology or readily available explanation.

(Borrebaeck, 2017). In human cancer, where single biomarkers have failed to achieve diagnostic performance comparable to well-established biomarkers in areas like heart disease (table 3), greater accuracy could be attained via biomarker panels (Polanski and Anderson, 2006). This observation serves an argument in establishing and improving multiplexing technologies.

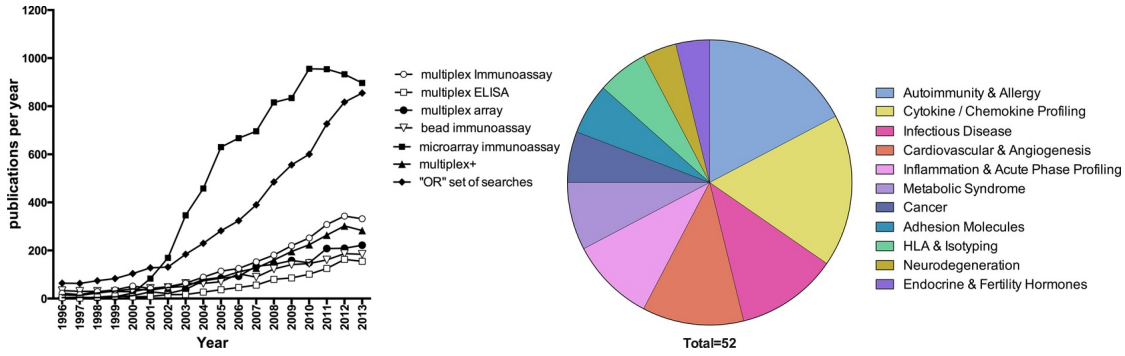
Table 3: Example sensitivities and specificities for nine FDA approved cancer biomarkers along with cancer biomarker panels and other biomarkers (Polanski and Anderson, 2006).

Marker	Disease	Cut Off	Sensitivity	Specificity	Reference
CEA	malignant pleural effusion	NA[1]	57.5 %	78.6 %	Li et. 2003
CEA	peritoneal cancer dissemination	0.5 ng/mL	75.8 %	90.8 %	Yamamoto et al. 2004
Her-2/neu	stage IV breast cancer	15 ng/mL	40 %	98 %[2]	Cook et al. 2001
Bladder Tumor Antigen	urothelial cell carcinoma	NA	52.8 %	70 %	Mian et al. 2000
Thyro-globulin	thyroid cancer metastasis	2.3 ng/mL[3]	74.5 %	95 %	Lima et al. 2002
Alpha-fetoprotein	hepatocellular carcinoma	20 ng/mL	50 %	70 %	De Masi et al. 2005
PSA	prostate cancer	4.0 ng/mL	46 %	91 %	Gann et al. 1995
CA 125	non-small cell lung cancer	95 IU/mL	84 %	80 %	Dabrowska et al. 2004
CA 19.9	pancreatic cancer	NA	75 %	80 %	Yamaguchi et al. 2004
CA 15.3	breast cancer	40 U/mL	58.2 %	96.0 %	Ciambellotti et al. 1993
leptin, prolactin, osteopontin, and IGF-II	ovarian cancer	NA	95 %	95 %	Mor et al. 2005
CD98, fascin, sPIgR[4], and 14-3-3 eta	lung cancer	NA	96 %	77 %	Xiao et al. 2005
Troponin I	myocardial infarction	0.1 µg/L	93 %	81 %	Eggers et al. 2004
B-type natriuretic peptide	congestive heart failure	8 pg/mL	98 %	92 %	Dao et al. 2001

[1] Not available.
[2] vs benign breast diseases
[3] vs 3rd week post surgery
[4] secreted chain of the polymeric immunoglobulin receptor

1.4 Multiplex assays

The need for better diagnostic tools is mirrored by the increasing number of research work on multiplex immunoassay formats — some of which has made it into the market (figure 2) — although most of these assays have been validated for research use only (Tighe et al., 2015). Tighe et al. (2015) reported clinical validity (the lack of characterization on clinical significance), reproducibility (inconsistent or insufficient expression) and nonspecificity (either also found in healthy tissue or common across several diseases) as reasons as to why only a few biomarkers are clinically qualified.



(a) (b)

Figure 2: Developments in multiplex immunoassays formats. (a) The number of publications published per year in PubMed, (b) validated commercially assays available based on clinical application (Tighe et al., 2015).

A growing consensus is to combine several biomarkers together to improve the diagnostic accuracy of a test. Irvine et al. (2016) built a model from seventeen analytes from sera that enhanced accuracy to an AUC of 0.938 compared to an AUC of 0.898 when using individual components for patients with advanced liver fibrosis. Goodison et al. (2016) developed a 10-biomarker assay based on urine proteins for bladder cancer with an AUC of 0.892 whereas individual components had AUCs ranging from 0.62 to 0.82. Augello et al. (2018) found 37 plasma proteins to be diagnostic and predictive biomarkers for stroke patients, achieving an AUC of 0.96 while the best AUC for a single biomarker is 0.81.

The tendency to achieve higher accuracies with multi-marker panels infers the potential of using the entire immune repertoire for more accurate diagnosis.

1.5 Phage display

One technique which can profile the diversity of the immune repertoire is phage display. Phage display involves the cloning of desired protein or peptide into the bacteriophage such that it is expressed on the bacteriophage's surface (Smith, 1985). By expressing a library of possible antigenic targets, one can probe the immune response by capturing antibody-antigen complexes that form. Given the genetic linkage to the expressed antigen, sequencing the region of the bacteriophage's DNA encoding for antigen allows identification of the antigenic target. This technology has been employed applications like epitope mapping, selection of peptide mimics, and drug discovery (Wu et al., 2016).

1.5.1 Display library formats

In order to interrogate the antibody profile, a suitable bait needs to be offered that the antibodies will have affinity to. The use of combinatorial peptide display libraries has facilitated this process as the displayed peptides are known to be able to mimic epitopes recognized by antibodies (Meloen, Puijk, and Slootstra, 2000). As such, these mimotopes have been displayed on phage in a variety of ways, with a number of commercially options available.

Phage display libraries have been made from using the coat proteins of various nonlytic and lytic bacteriophages (Bazan, Cakosiski, and Gamian, 2012). Peptides have been successfully made as fusions to the coat proteins of filamentous phages, the choice on which coat protein to display depends on factors like the maximum tolerable size, the num-

ber of copies displayed, as well as the structure of the peptide (Henry, Arbabi-Ghahroudi, and Scott, 2015). The way these type of phage propagate is by a process called chronic infection were the bacterial host remains viable during the release of phage virions (Hobbs and Abedon, 2016). A disadvantage of this system is the presence of propagation biases from fast-growing filamentous phage as the peptide displayed can affect its rate at which it is secreted (Derda et al., 2011). This problem is mitigated in lytic phage systems due to the due to the destruction of the host to release its virions and the lower demands on its host translational system (Krumpe et al., 2006).

Better epitope recognition has been observed when peptides are instead displayed as loops due to higher affinities and selectivities observed in this conformation (Deyle, Kong, and Heinis, 2017). This cyclization is achieved through a variety of chemistries like disulfide reduction, chemical linkers, and enzymes (Gang, D. Kim, and H.-S. Park, 2018). Even higher affinities were reported with the use of bicyclic peptides due to the additional constraint in conformation (Deyle, Kong, and Heinis, 2017).

1.5.2 Phage panning and high-throughput sequencing

Traditional phage panning involves adding the phage display library to the target to be interrogated, like antibodies (figure 3). These can be performed with the target already immobilized prior to addition or, alternatively, allow the peptide-antibody complex to form in solution prior to immobilization to a capture surface. This step is followed by washing off the unbound phage from the immobilized peptide-antibody complex, then an elution step is performed to release the captured phage. Plating the captured bacteriophage onto a plate of bacteria permits the picking of single bacteriophage plaques for repropagation in bacteria which can be used in another round of panning (Wu et al., 2016).

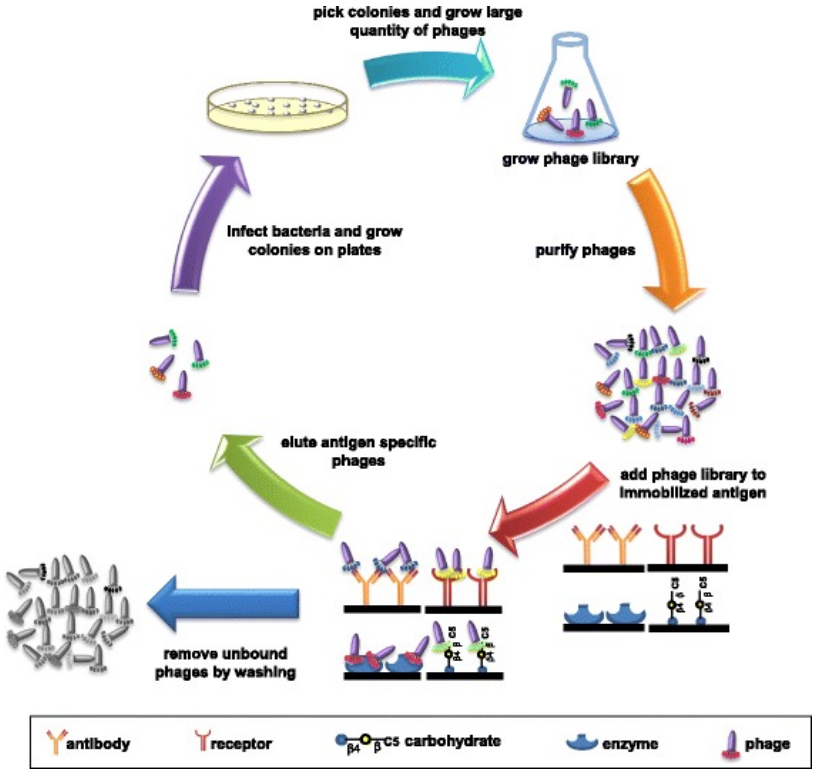


Figure 3: Steps during phage panning (Wu et al., 2016).

The peptides displayed by the picked bacteriophage plaques can be identified using Sanger sequencing. However, this method only allows the identification of small pool of captured bacteriophage. In order to analyze the peptides on capture bacteriophage in the millions, high-throughput sequencing is necessary (Gallo, 2019).

A number of high-throughput sequencing platforms has been successfully applied to phage display peptides after panning. Popular sequencing platforms include Illumina and Ion Torrent (Rouet et al., 2018). The Illumina platform is known to suffer from base substitution errors while deletions are the problem with Ion Torrent (E. J. Fox et al., 2014). Several groups found the performances of both platforms comparable (Lahens et al., 2017; Speranskaya et al., 2018). However, Ion Torrent can only generate single reads while Illumina is capable of performing paired-end reads (Lahens et al., 2017).

The utility of combining high-throughput sequencing with phage-displayed peptides for interrogating antibody repertoires has been demonstrated in a number of cases (Paull and Daugherty, 2018). The T7-Pep library which has been designed to bear peptides corresponding to the human proteome was applied in screening autoantibodies to diseases such as paraneoplastic syndromes, multiple sclerosis, type 1 diabetes, and rheumatoid arthritis (Larman, Laserson, et al., 2013; Larman, Zhao, et al., 2011). By the designing the same library to bear peptides to all viruses known to infect humans instead, virus-specific epitopes were identified across a cohort of 596 individuals (Xu et al., 2015). In this particular study, serological signatures were found across their cohort, which varied with age and geographical location. Two problems were discovered with this rationally-designed library: (1) a difficulty of locating the 5-10 core epitope residues in the 56-mer peptide display, and (2) the absence of signatures to common viruses. In Deep Panning, peptide motifs were identified using a random 7-mer NNK linear peptide library to probe HIV+ serum (Ryvkin et al., 2012). In another case, a random 7-mer library was also used to profile individuals with severe peanut allergies (Christiansen et al., 2015). Offering virus-like particles displaying random peptide loops to ovarian cancer patient plasma, Frietze et al. (2016) found a peptide that correlated to better survival. Similarly, Ikemoto et al. (2017) performed high-throughput sequencing of a panned cyclic random peptide phage display library to find a peptide that target peritoneal carcinomatosis.

Presented with two options to select a peptide library from, a design choice needs to be made regarding the appropriate library to use. In the context of a disease like cancer, it seems that the straightforward choice is a rationally-designed library towards the human proteome. Interestingly, Navalkar, Johnston, and Stafford (2015) found random sequences had better accuracy than epitope peptides. Paull and Daugherty (2018) suspects that affinities may be simply higher in random peptides and possibly associated with a higher peptide diversity in random peptide libraries. The challenge with random libraries though is figur-

ing out the antigen the random peptides correspond to. Even before that, there is a need to preprocess the data in order to separate signal from noise. Thus, a suit of computational tools is needed.

1.5.3 Informatics

With the availability of high-throughput sequencing data, computational resources have been developed to tackle problems that arise from analyzing millions of sequences. For examining library representativeness, PuLSE has been developed to evaluate randomness in naive phage display libraries (Shave et al., 2018). Nonspecific binding of phage-display peptides to panning components lead to the presence of target-unrelated peptides (TUPs) across samples (Bakhshinejad et al., 2016). The web service SAROTUP contains a suite of algorithms that allow the prediction of TUPs to assist in their identification and removal. Published panned sequences have been also made available via the Biopanning Data Bank database (BDB) (He et al., 2015) while the Immune Epitope Database and Analysis Resource (IEDB-AR) maintains a list of curated epitopes as well as tools for epitope prediction (Vita, Overton, et al., 2014).

Various computation pipelines have been built to be able to enrich for affinity-selected peptides. The program GuiTope was developed to make similarity matches between proteins and candidate peptides (Halperin et al., 2012). However, knowing which protein markers is necessary to carry out the alignment. Brinton et al. (2016) created the PHASTpep to be able to perform enrichment analysis based on the ratio of normalized reads taken from the average of positive and negative screens. A limitation of this method is that enrichment is confined to two samples at a time. It is argued that peptide screens become more meaningful when based on global patterns across a cohort. Thus, there is a definite advantage conferred by a method that shares information across several independent

samples. Ricotta (2017) proposed a new beta biodiversity metric, an information-theoretic measure, that represents a weighted measure of concentration which preserves the relative dispersion of abundances. This method is repurposed in the context of sequencing data as evidence of enrichment.

1.6 Challenges in multiplexing

As single markers have largely failed due to a lack of clinical relevance or poor specificity, an increase in multiplexing technologies has been seen to meet this need. One such technology that has been made more powerful by high-throughput sequencing is phage display. Phage display has been demonstrated to enable discovery of unbiased markers, facilitate the understanding of complex diseases, and reveal novel therapeutic targets. As health care advances toward the direction of personalized medicine, disease heterogeneity remains a challenge. With the power of high-throughput sequencing to accelerate the screening of markers from a person's antibody repertoire, phage display can be an orthogonal technique to support existing medical procedures.

To interrogate the antibody repertoire, we found that random peptides may have better antigen-affinity than a rationally-designed library. Upon making the choice to use a random peptide library, we observe that core epitope residues are reported to be 5-10 residues in length. Epitope motifs have been discovered using 7-mer libraries. However, 12-mer libraries are also available. Thus, a rational design choice would be to choose a 12-mer peptide that spans the length of such epitopes.

The main aim of this thesis is to be able to perform an unbiased and comprehensive sampling of humoral responses that can identify the complexity of the underlying diseases.

The specific aims are:

- To create a computational pipeline that can uncover true biological variations in the raw data.
- To assess immune responses in a well-characterized immune status like healthy individuals challenged with the influenza vaccine.
- To examine immune responses in an uncharacterized disease like AML.

Chapter 2

Biodiversity Analyses of High-Throughput Sequencing Results from a Peptide Phage-Display Library Reveal Candidate Peptide Profiles in Acute Myeloid Leukemia Patients Undergoing Nivolumab/Azacytidine Treatment

2.1 Introduction

Serum antibody levels rise and fall due to antigen-dependent and antigen-independent events that eventually lead to immunological memory (Traggiai, Puzone, and Lanzavecchia, 2003). In the presence of an antigen, antibody levels peak, plateau, then slowly decline due to the proliferation and death of antigen-driven short-lived and long-lived plasma cells. However, higher steady antibody levels are achieved eventually due to a larger memory B cell pool that results from addition of memory B cells specific to the antigen. The onset of a new systemic infection is met by an antibody response initially marked by the appearance of IgMs followed by elevated response in IgAs and IgGs within a few days, while reinfection results to primarily IgGs (Baron and Klimpel, 1996).

As mounted immune responses are recorded in memory B cell diversity, the antibody repertoire that is generated by it provides a rich information source of one's past

and present immune status (Weiss-Ottolenghi and Gershoni, 2014). Weiss-Ottolenghi and Gershoni (2014) reported that a direct analysis on antibodies or antibody-producing cells have been facilitated by the use of phage-displayed antibodies, single cell cloning, next-generation sequencing of antibody-related mRNA transcripts, as well as proteomic techniques. They also mentioned phage-displayed random peptides or gene fragments and deep-panning as examples of indirect analysis which target antibody specificities instead.

In deep-panning, serum sample is used to affinity-select a phage-displayed random peptide library and the entire peptide subset directly analyzed by next-generation sequencing with options for multiplexing (Ryvkin et al., 2012). This approach avoids biases in phage diversity introduced by further amplification in bacteria and panning (Derda et al., 2011; Matochko, S. C. Li, et al., 2014). Additionally, template enrichment can be achieved by the incorporation of emulsion PCR as part of sample preparation in sequencing protocols (Desai and Jere, 2012; Matochko, S. C. Li, et al., 2014; Rebollo et al., 2014).

While phage display is the oldest and most commonly applied molecular display technique, other scaffolds have been demonstrated to be potential peptide display platforms for affinity selection: bacteria, yeast, ribosome, and mRNA (Levin and Weiss, 2006). Strategies employing thermodynamic control (affinity-based selection) or kinetic control (off-rate-based selection) have been employed to achieve great affinity enhancements.

The wide array of affinities generated in random peptide libraries have been useful in uncovering epitope sequences (Bonnycastle et al., 1996; Gershoni et al., 2007; Ryvkin et al., 2012). These epitopes have been found to be either continuous (linear) or discontinuous (conformational), with discontinuous types comprising more than 90 % of B-cell epitopes (Cortese et al., 1994; Sun et al., 2014). However, only one or a few of these epitopes are highly immunogenic (Nowak, 1996). Thus, these immunodominant epitopes make good candidates as disease biomarkers. A comprehensive up-to-date collection of epitopes curated from published literature is available from the Immune Epitope Database

(IEDB) which has been used before in consolidating customized datasets and conducting meta-analysis of pathogens of interest (Vita, Overton, et al., 2014; Yasser and Honavar, 2010).

Several general diagnostic platforms have been demonstrated to be able to screen for multiple diseases from a single sample. In immunosignaturing, a combinatorial peptide microarrray captures a distinct reactivity pattern from antibodies in sera which is useful in characterizing an individual's immune status (Sykes, Legutki, and Stafford, 2013). This pattern is detected through the use of fluorescent or electrochemical probes. A different technology that leverages on PCR coupled with electronspray ionization time-of-flight mass spectrometry enables rapid diagnosis in patients with clinical suspicion of sepsis (Jordana-Lluch et al., 2013). Comparison of nucleotide base composition determined from clinical samples to a database have led to accurate detection of pathogens. Both technologies are actively researched to establish confidence in their clinical utility.

The concept of analyzing blood with molecular-based methods is emerging to be a promising technology in the field of diagnostics. Advancements in biomolecular devices combined with a growing knowledge on molecular signatures propels interest towards general diagnostic techniques. In one such approach, serum IgGs are probed with a phage display library corresponding to proteins from all viruses known to infect humans to create an epitope profile which can be traced back to viral pathogens or pathogenic material an individual was exposed (Xu et al., 2015). Under normal conditions, the immune system does not target self proteins. However, autoantibodies to various types of cancer has been reported in literature (Zaenker, Gray, and Ziman, 2016). Thus, profiling a patient's humoral response offers an alternate means of detecting and monitoring cancer.

Among acute leukemias in adults, acute myeloid leukemia (AML) is the most common, with a 2015 US estimate of over 20 000 new cases annually and is characterized by an accumulation of immature cells from the myeloid compartment of the immune sys-

tem (De Kouchkovsky and Abdul-Hay, 2016). From the review done by De Kouchkovsky and Abdul-Hay (2016), AML incidence is higher in the elderly, with 12.2 cases per 100 000 in patient over 65 years old versus 1.3 cases per 100 000 in patients less than 65 years old. They mentioned that despite the heterogeneity of this disease, chemotherapy with cytarabine and anthracycline remains the primary treatment with possible allogeneic stem cell transplantation in some patients. However, this chemotherapeutic regime is not well tolerated by elderly patients, and they reported that "as much as 70 % of patients 65 years or older will die of their disease within 1 year of diagnosis." The use of a milder chemotherapeutic like azacytidine in tandem with the monoclonal antibody nivolumab to suppress immune evasion is currently explored as a new treatment modality for elderly patients by N. Daver et al. (2019).

The current diagnosis for AML involves establishing the presence of abnormal myeoblasts in blood and bone marrow samples plus a host of other tests to examine morphology, immunophenotype, cytogenetic profile, and other disease markers (De Kouchkovsky and Abdul-Hay, 2016). On the research side, several groups have demonstrated the potential of profiling various components in patient sera for monitoring in AML such as serum proteins (Bai et al., 2013), metabolites (W.-L. Chen et al., 2014), and free fatty acids (Khalid et al., 2018). However, to the best of our knowledge, we have not yet seen a study that has investigated alterations in antibody profiles of AML patients.

In this work, the utility of profiling IgGs from human sera/plasma samples for immunodominant as well as subdominant epitope signatures in AML is explored. This approach is examined using a commercially available random peptide phage display library in combination with high-throughput sequencing. The method is validated using influenza-vaccinated samples and applied to cancer samples, specifically cases of acute myeloid leukemia that failed prior therapy and subsequently treated with nivolumab and azacytidine (N. Daver et al., 2019). Bioinformatic enrichment is carried out using meth-

ods in biodiversity analysis borrowed from mathematical ecology between test and control samples. Candidate peptides derived from this procedure are examined based on sequence alignment results to custom protein databases.

2.1.1 Hypothesis

Traditional phage display screens involves several round of panning to enrich for high affinity binders, biasing the search process towards immunodominant epitopes. However, antibodies of lesser affinity but possibly clinically relevant targeting subdominant epitopes would be lost to this process. Thus, it is desirable to reduce the number of pannings to probe a larger pool of epitope sequences. It is hypothesized that a single round of panning using a random dodecapeptide phage display library would be sufficient to uncover peptide sequences that characterize AML when paired with contrasting of samples by choosing an appropriate set of test and control cases.

2.1.2 Rationale

Stafford et al. (2014) demonstrated that the random sequence peptide microarrays are capable of assaying five different cancer types (GBM, PC, lung cancer, MM, BC) with 95 % classification accuracy with a blinded evaluation on a 100 cancer/20 noncancer cohort over the same set of cancers. Cross-validation of more than 1500 historical samples for 14 different diseases showed an average accuracy of greater than 98 %. Navalkar, Johnston, and Stafford (2015) reported performance benefits in using random sequence peptides over epitope peptides.

Combining high-throughput sequencing together with patient-control contrasting in panning with a Ph.D.7™ phage library, Christiansen et al. (2015) recovered a peptide cluster that generated a sequence motif towards major peanut allergen Ara h 1. The

control samples provided a bioinformatic means to segregate out target-unrelated peptide sequences and identify epitope candidates.

We speculate that utilization of a random phage display library would allow the exploration of a wider peptide sequence space that would otherwise be constrained by available space on a peptide microarray chip. High-throughput sequencing with appropriate patient controls would allow elucidation of peptide signatures specific for a particular disease.

2.2 Methods

Serum IgGs were affinity-captured along with the peptides displayed on phage they were bound. The protocol involves (1) incubation of serum with the phage library, (2) affinity capture of IgG and bound phage with protein G beads, (3) elution and amplification of captured phage — hence completing one round of panning, (4) enriching the phage library for binders through additional rounds of panning with serum, then (5) determination of peptide binders by high-throughput sequencing (figure 4). Afterwards, a bioinformatic analysis of these peptide sequences was carried out to yield a set of candidate sequences enriched in the test cohort relative to controls.

To circumvent mass transfer limitations associated with immobilization to a porous bead, serum was incubated with the phage library in solution prior to affinity capture with protein G beads. Protein G beads were chosen due to its strong affinity for all human IgG subclasses, with antibody recoveries of 60 to 80 % (Nath et al., 2015).

With the purpose of mitigating interferences due to serum albumin as well as promoting good mixing on a rotary mixer, serum and phage were placed in a 1.5 mL microcentrifuge tubes filled to roughly a third its volume of PBS. The resulting solution would

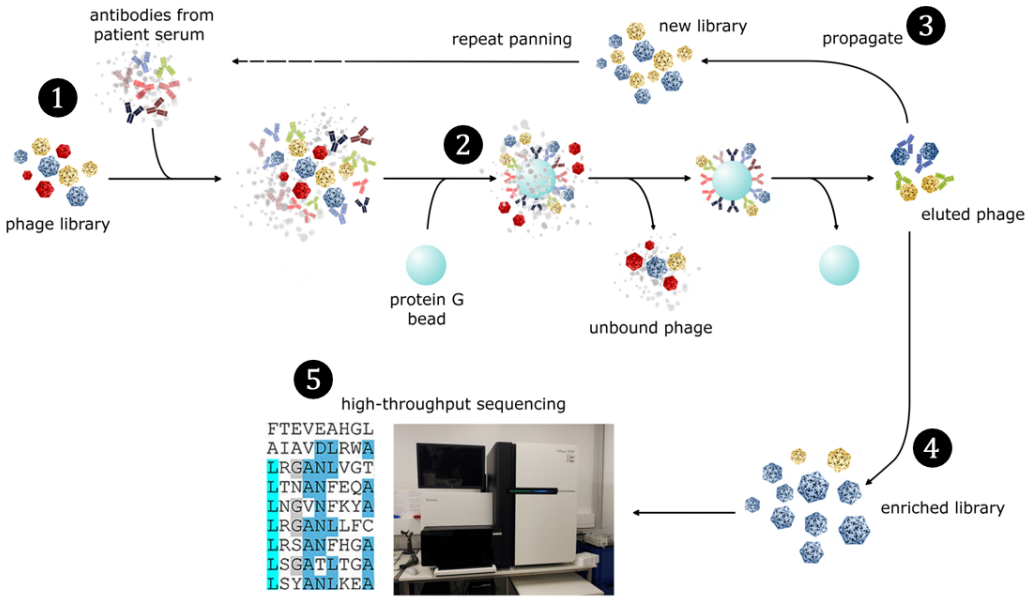


Figure 4: Schematic for antibody profiling by phage display.

be very dilute and nonspecific binding to the storage vessel can lead to denaturation and losses. To minimize these effects, incubation was carried out inside low-binding tubes with PBS containing 0.05 % Tween 20 (PBST).

As library preparation was to be carried out using long custom primers containing sequences required for the Illumina next-generation sequencing (NGS) platform, PCR-amplified library templates were validated using Sanger sequencing. Then, the libraries were prepooled and sent for NGS. Resulting reads were preprocessed to remove sequencing artifacts prior to analysis. Analyses consisted of examining read distributions and enriched sequences across the samples.

2.2.1 Samples

All work outlined in this study was performed according to protocols approved by the Institutional Review Boards at the University of Houston and the University of Texas M.D. Anderson Cancer Center. Plasma samples from flu-vaccinated individuals that were

recently immunized with the 2014-2015 influenza vaccine (*What You Should Know for the 2014-2015 Influenza Season* n.d.) and AML patients before and after the first round of combination therapy (azacytidine + nivolumab) (N. G. Daver et al., 2017) were collected and stored at −20 °C. Additionally, a sample one year after vaccination was obtained from one of flu-vaccinated cohort. A working stock of each sample was prepared by thawing the plasma sample at ambient temperature and diluting 1 µL of clear supernatant to 400 µL with 0.05 % PBST. Single-use aliquots were prepared from each stock and stored at −20 °C. Thawed aliquots were kept at 4 °C and used only up to a week . All dilutions were prepared and stored in low protein binding tubes.

2.2.2 Phage library

Plasma samples were panned against the Ph.D.™-12 Phage Display Peptide Library of New England Biolabs (NEB), a combinatorial library of random linear dodecapeptides fused to the N terminus of the minor coat protein pIII of M13 phage (*Datasheet for Ph.D.™-12 Library*, n.d.). Twenty microliters of the library was offered, giving slightly over a 100-fold representation of 1.7×10^9 unique clones for the first panning round while the amount of phage clones offered during the second panning round varied as specified in table 5.

2.2.3 Panning

A total of 23 runs were carried out: (1) eight panned AML samples from a cohort of four patients with one round of panning, (2) eleven panned flu-vaccinated samples from a cohort of four donors where in one donor the ratio of phage to IgG was varied as well as had one more round of panning, and (3) a repeat panning for one of the AML samples for which two rounds was carried out. Additionally, two unpanned reference libraries con-

sisting of the original and its repropagation were included (table 4). Fixing the volume of phage library used, the amount of IgGs were varied to satisfy the indicated phage-IgG ratios for the first round of panning; IgG titers were presumed to be at a nominal value of 100 mg mL^{-1}. For the second panning round, IgG quantities were fixed instead and the amount of phage library adjusted to account for variable amplified phage titers.

Table 4: Samples used in the NGS analysis of phage-panned AML and influenza-vaccinated cohorts.

ID	Sample name	Sample type	Donor
A	1p-I	AML baseline (responder)	1
B	2p-I	AML end-of-cycle (responder)	1
C	3p-I	AML baseline (responder)	2
D	4p-I	AML end-of-cycle (responder)	2
E	5p-I	AML baseline (nonresponder)	3
F	6p-I	AML end-of-cycle (nonresponder)	3
G	7p-I	AML baseline (nonresponder)	4
H	8p-I	AML end-of-cycle (nonresponder)	4
I	9p-I	flu-vaccinated (recent)	5
J	10p-I	flu-vaccinated (recent)	6
K	11p-I	flu-vaccinated (recent)	7
L	7-1-2017	AML baseline (repeat run of 7p-I, 1st panning)	4
M	7-2-2017	AML baseline (repeat run of 7p-I, 2nd panning)	4
N	R0-rnd-1	flu-vaccinated (recent; phage-IgG ratio varied)	8
O	R1-rnd-1	flu-vaccinated (recent; phage-IgG ratio varied)	8
P	R2-rnd-1	flu-vaccinated (recent; phage-IgG ratio varied)	8
Q	ctrl-rnd-1	flu-vaccinated (1-yr after)	8
R	R0-rnd-2	flu-vaccinated (recent; phage-IgG ratio varied)	8
S	R1-rnd-2	flu-vaccinated (recent; phage-IgG ratio varied)	8
T	R2-rnd-2	flu-vaccinated (recent; phage-IgG ratio varied)	8
U	ctrl-rnd-2	flu-vaccinated (1-yr after)	8
V	naïve	Ph.D.-12 library	
W	once-amp	once-amplified Ph.D.-12 library	

Runs were diluted to 500 µL with 0.05 % PBST in 1.5 µL low binding tubes and were incubated overnight on a rotary mixer inside a cold room. Then, 20 µL magnetic protein G beads were prepared as per manufacturer's instructions (Promega: *Antibody Purification Technical Manual*, 2015) and added to the runs for an additional 4 h of incubation.

Next, the rest of the antibody purification protocol was followed yielding a combined volume of 120 µL neutralized eluent. Amplified phage stock titers are measured to calculate the amount needed in carrying out the second round of panning (table 5). All AML samples were panned only once except 7p-I, which was repanned up to two rounds to examine reproducibility and as an additional sample to investigate repanning effects.

Table 5: Number ratios of phage incubated with plasma IgGs used in each panning round. The clonal representations of phage libraries in the first panning round is about 100-fold. Therefore, ratios can be thought of as the number of phage clones matched to IgGs.

Sample name	phage:IgG, in number of particles	
	1st panning	2nd panning
1p-I	100:10	NA
2p-I	100:10	NA
3p-I	100:10	NA
4p-I	100:10	NA
5p-I	100:10	NA
6p-I	100:10	NA
7p-I	100:10	NA
8p-I	100:10	NA
9p-I	100:10	NA
10p-I	100:10	NA
11p-I	100:10	NA
7-X-2017	100:100	100:10
R0-rnd-X	100:100	100:200
R1-rnd-X	100:10	100:20
R2-rnd-X	100:1	100:2
ctrl-rnd-X	100:10	100:20

X = 1 or 2, in reference to panning round

2.2.4 Sequencing

A protocol for high-throughput sequencing on Ion Torrent (Matochko and Derda, 2015) was modified and adapted for the Illumina platform. Briefly, phage library DNA from each run was isolated using a QIAprep spin M13 kit and the concentrations of the

resulting eluents were measured using a NanoDrop spectrophotometer. The isolated DNA were imaged on a gel to check for band consistency across runs. These phage library DNA samples were then PCR amplified using custom primers designed via Primer-BLAST (Ye et al., 2012) to flank the random peptide region within a read length of 75 bp and introduce the necessary Illumina adapter sequences (appendix B); an *NKKN* tetramer was inserted 5' of the forward primer to facilitate cluster identification (Matochko, Chu, et al., 2012). The resulting Illumina libraries were purified via band excision from preparative gels and sent for Sanger sequencing to examine for sequencing artifacts. Library sequences were composed of the 12-mer random peptide insert + Gly-Gly-Gly spacer, flanked on both sides by 15-mer peptide subsequences plus Illumina sequences were examined using the *R* packages *sangerseqR* (Hill et al., 2014) and *msa* (Bodenhofer et al., 2015). Then, the results were inspected for deviations from the consensus sequence in the flanking sequences and erroneous amino acid residues due to base miscalls in the random peptide insert region. Most of the sequences obtained appear to be consistent overall with the expected library template structure (appendix C). All 23 libraries were then prepooled and sent for Illumina sequencing.

2.2.5 NGS data analysis

Illumina sequencing results were preprocessed and analyzed using R (R Core Team, 2019). Briefly, reads obtained from Illumina sequencing underwent translation to peptide sequences with various filtering steps to remove sequencing artifacts. A number of R packages were used for reading, manipulating, and managing data, graphing and other data visualizations, various calculations, web scraping, and report generation (Analytics and Weston, 2018; Arora et al., 2018; Bache and Wickham, 2014; Bengtsson, 2018; Beygelzimer et al., 2019; Brown, 2012; Carlson, 2018; Corporation and Weston, 2018;

Csárdi and Chang, 2019; Dowle and Srinivasan, 2019; J. Fox and Weisberg, 2019; J. Fox, Weisberg, and Price, 2018; Genz et al., 2019; W. Huber et al., 2015; Huber et al., 2015; Michael Lawrence et al., 2013a; Michael Lawrence et al., 2013b; Michael Lawrence et al., 2013c; Maechler et al., 2019; Norm Matloff, 2016; Norm Matloff and Yingkang Xie, 2016; Norm Matloff, Yang, and Nguyen, 2017; Norman Matloff, 2016; Microsoft and Weston, 2017; Morgan, Anders, et al., 2009; Morgan, Obenchain, Jim Hester, et al., 2018; Morgan, Obenchain, Lang, et al., 2019; Morgan, Hervé Pagès, et al., 2018; Neuwirth, 2014; H. Pagès, P. Aboyoun, et al., 2019; H. Pagès, M. Lawrence, and P. Aboyoun, 2018; Hervé Pagès and Patrick Aboyoun, 2018; Hervé Pagès, Carlson, et al., 2018; Hervé Pagès and Peter Hickey, 2018; Ren and Russell, 2016; Rinker and Kurkiewicz, 2018; Rodriguez-Sanchez, 2018; Schloerke et al., 2018; Sievert, 2018; Solymos and Zawadzki, 2019; Wickham, 2016; Wickham, 2019; Wickham, James Hester, and Ooms, 2018; Wong, 2013; Yihui Xie, 2019; Yihui Xie, Cheng, and Tan, 2019; Zhu, 2019), supplemented by tools available in Linux. An adaptation of the beta diversity approach described by Ricotta (2017) was used for peptide enrichment. Peptide sequences were aligned to custom databases for influenza, human proteins, and human-associated microorganisms at a minimum bit score of 15 with BLOSUM62 scoring matrix using DIAMOND (Buchfink, C. Xie, and Huson, 2015). The custom protein databases for human, influenza, and human pathogens were generated from the non-redundant protein sequences (nr) database from NCBI via seqkit (Shen et al., 2016). The human protein sequence database was prepared by selecting for entries corresponding to the search term *Homo sapiens* and similarly using the search term *Influenza* for the influenza protein database. The protein sequence database for known human pathogens was made by filtering based on the names of organisms that has humans as hosts listed in the table of host-pathogen interactions from Wardeh et al. (2015). Web services GeneShot (Lachmann et al., 2019) and the GeneCards suite (Stelzer et al., 2016) were used to generate a list of genes associated with AML using "AML OR acute myeloid

leukemia" and "[all] (AML) OR [all] (acute AND myeloid AND leukemia)" as search terms respectively.

2.2.6 Software

All the codes used in the analyses are available upon request and can be found on `https://bitbucket.org/alcooj/phagelibanalysis`.

2.3 Results and discussion

2.3.1 Sequencing artifacts present were corrected prior to analysis

Sequencing results returned a total read count of 49 836 730, with 48 479 986 reads passing Illumina's filtering scheme. Of the reads that passed filtering, those which can be identified were only $(84.6410 \pm 0.4001)\,\%$. Majority of reads in each sample appeared to be of good quality overall ($>$Q30); however, there were plenty of samples with non-uniform coverage and dissimilar profiles between paired-end reads. In general, forward reads seemed to suffer more from non-uniform coverage than reverse reads (appendix D). Non-uniformity might be due to overrepresented sequences arising from sequencing biases.

Examination of the barcode sequences showed that around 97 % or more of the Illumina i5/i7 barcodes matched perfectly. Only a maximum of one mismatch was found for any barcode (appendix E). It may be that Illumina internally filters out reads with barcodes mismatching more than once.

To evaluate sequencing fidelity, the proportion of reads matching the sequences flanking the combinatorial library insert was measured relative to the maximum number of mismatches allowable. The proportion of reads in each sample library with the number of

mismatches less than or equal to some threshold is plotted for both the forward and reverse reads of the Illumina libraries (figure 5). All forward and reverse reads were first made uniform to a length of 76 bases by padding ends with *N*s such that the library insert will roughly lie between residues 41 to 76. Then, M13 phage sequences 36-nt long flanking both sides of the combinatorial library insert published by NEB (*Ph.D.™ Phage Display Libraries Instruction Manual* 2016) were matched against the NGS reads. Mismatches may have been due to sequencing error or possibly mutations in phage genome during propagation. As the stringency of matching is decreased, the risk of sequence matching becoming nonspecific increases, causing the flanking sequences to map to more than one region. Only reads with unique matches to the flanking sequences are counted, i.e. reads mapping more than once do not make it to the total count. Thus, the curves begin to descend when the stringency for mismatches is relaxed to beyond 15. Strangely, the forward reads of the R2 samples (sample IDs P and T) do not appear to match at all. Its forward reads had zero matches as a consequence of poor matching with the reference flanking sequences; however, more than 99 % of their reverse reads remained after filtering. Out of 46 samples, the curve peaks at a maximum mismatch of 12 in 23 samples, at 14 in 21 samples, and at 23 in two samples. From this observation, a maximum mismatch frequency of 12 out of 36 nt in the flanking sequence was chosen for all samples. To determine whether reads from the R2 samples are salvageable, only paired-end reads mapping more than once to a flanking sequence in either its forward or reverse read were removed, resulting to 40 965 648 after filtering.

Flanking sequences were trimmed off and paired-end combinatorial library sequences were examined for mismatches between its forward and reverse reads. Some reads became truncated after trimming off the flanking sequences. To make paired-end reads uniform in length, truncated reads were padded with *N*s to a length of 36 nt. In figure 6, the proportion of reads in each sample library with the number of mismatches less than or

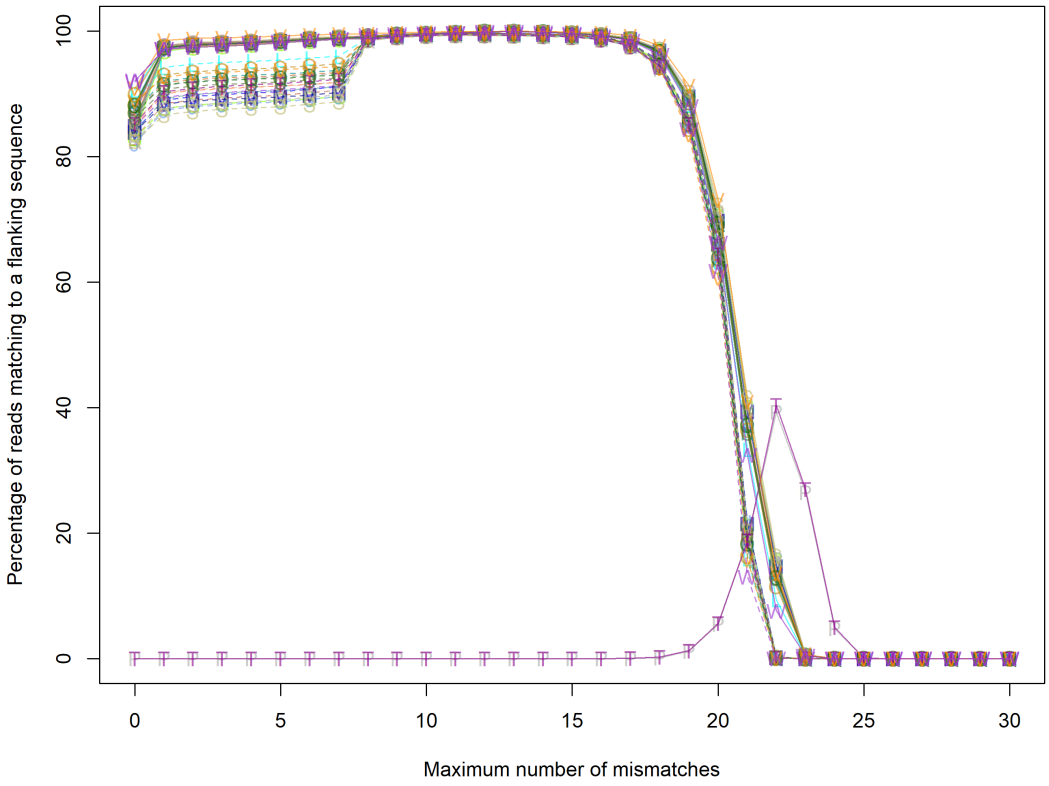


Figure 5: Mismatches in the sequence alignment of regions flanking the combinatorial library insert. Markers on each of the 46 curves correspond to one of the 23 sample IDs (forward and reverse reads of the 23 libraries).

equal to some threshold is plotted for both the aligned reads of the Illumina libraries. A maximum mismatch of three out of 36 residues was chosen for all samples except samples P and T (figure 6). At this stringency threshold except for samples P and T, about 85 % to 95 % is retained across the samples. Given that a substantial number of reads gets removed overall and therefore not be feasible to ignore filtering reads from samples P and T, analysis was carried out with the R2 samples excluded. Total reads passing filtering by paired-end overlap mismatch after excluding the R2 samples resulted to 32 415 186.

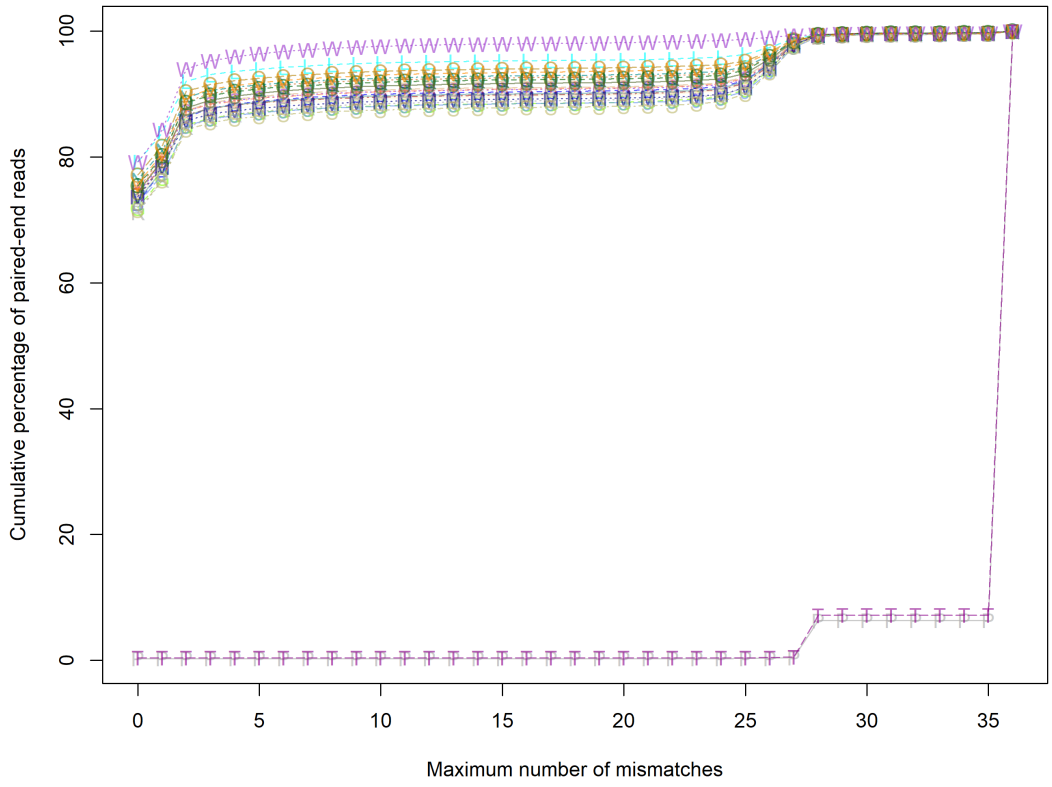


Figure 6: Mismatches in the paired-end reads of the combinatorial library insert. Markers on each of the 23 curves correspond to one of the 23 sample IDs.

The combinatorial library sequences are designed to contain 12 *NNK* codons and thus the sequences were validated whether there are deviations from the *NNK* motif. Validation was carried out on the filtered paired-end reads and the results are shown on figure 7.

The proportion of reads in each sample library with the number of mismatches less than or equal to some threshold is plotted for both the paired-end reads of the Illumina libraries. Mismatches arising from padded *N*s due truncation of the sequences were ignored. Since applying the maximum stringency would incur a loss of not more than 3 %, only reads with no mismatch from the random library inserts NNK motif were eventually carried over for further analysis, giving a total read count of 31 774 376. Retaining only paired-end sequences that are consistent with *NNK* motif on both reads, the number of reads goes down to a final count of 31 713 752.

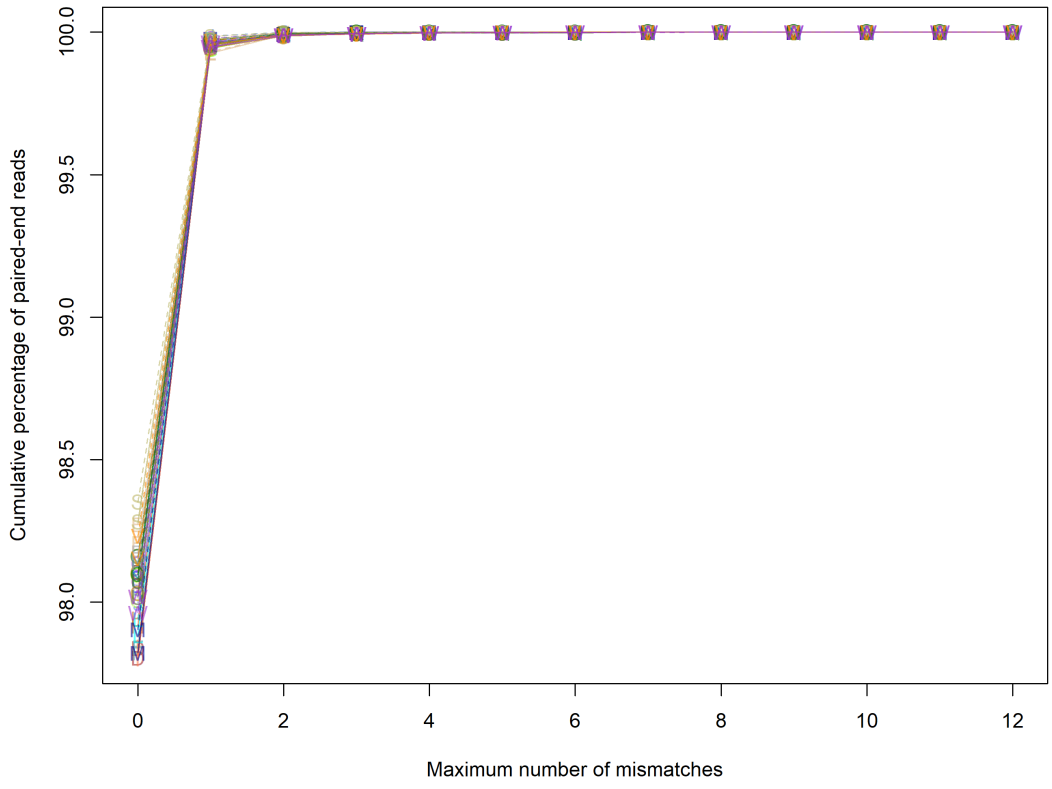


Figure 7: Mismatches from the NNK motif in the paired-end reads of the combinatorial library insert. Markers on each of the 23 curves correspond to one of the 23 sample IDs.

The filtered reads were translated, replacing all stop codons present with glutamine (*Ph.D.™ Phage Display Libraries Instruction Manual* 2016). Examination of

unique peptide sequences showed a fraction of the peptides containing ambiguous bases; ambiguous residues stem from *N*s from the padding of truncated sequences. The number of unique peptides containing ambiguous residues range from 6 % to 8 % of the total number of unique peptide sequences across samples. Residue ambiguity was resolved by substituting an ambiguous residue with the corresponding partner amino acid between paired-end reads. Reads that cannot be rescued were filtered out, yielding 31 713 494 peptide sequences for analysis.

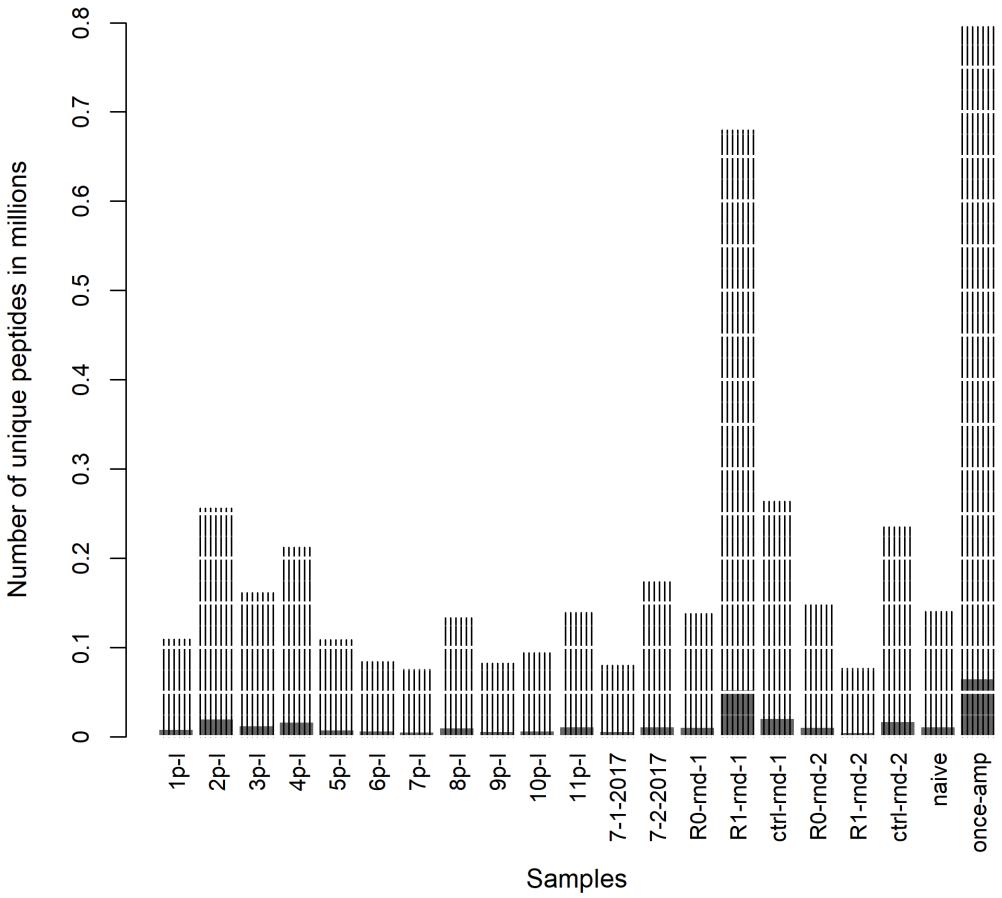


Figure 8: Proportion of peptides from the translated combinatorial library sequences containing ambiguous residues. Vertical black shading lines show the total number of unique reads. Black solid bars show unique reads with ambiguous residues.

2.3.2 Peptide diversity indicated varying degrees of enrichment across samples

The distributions in the total number and the number of unique sequences are shown in figure 9. From the figure, we see that the number of unique peptide sequences is relatively low with respect to the total number of sequences in all of the panned samples except in sample R1-rnd-1, indicating enrichment in a subset of peptides. On the other hand, the number of unique sequences in the unpanned reference libraries appear to be slightly more than half of the total sequence count. Since the forward and reverse reads of the paired-end sequences were kept separate as a consequence of retaining a fraction of mismatched pairs, the near 50 % unique read content in the naïve and once-amp libraries suggests a uniform representation of the peptides sampled. With next-generation sequencing yielding around 32 million sequences passing quality control, equal distribution across 21 samples would give each sample about 1.5 million reads on average. Figure 9 shows all except three samples (11p-I, 7-1-2017, naïve) are within 20 % (1.2 million or greater) from this average. One thing to note is that since paired-ends were not merged to accommodate mismatches, the maximum number of unique peptides that can be sampled should be half of the average — at most 0.75 million. Given that the first panning rounds were derived from the naïve library with a complexity of 1.7×10^9 (*Datasheet for Ph.D.™-12 Library*, n.d.), only a small fraction of the total peptide sequence space gets sampled, about 1/2300 if the peptide sequences sampled are the same across all 21 libraries to 1/100 if unique peptides sequences in the libraries are non-overlapping. Counting the total observed number of unique peptides obtained across all samples gives 3 428 327, which is about 1/500, implying a significant fraction of peptide sequences are uncommon and thus will show up as zero counts across the panned samples. For a table consisting of 3 428 327 observed peptide sequences by 21 libraries, around approximately 95 % of the values are zeroes. Thus, it is important to account for this sparsity when examining for peptide enrichment via test

and control case comparisons as it can confound the results. Simply taking differences in counts may lead to false positives due to failure of a peptide sequence getting sampled in the a control case.

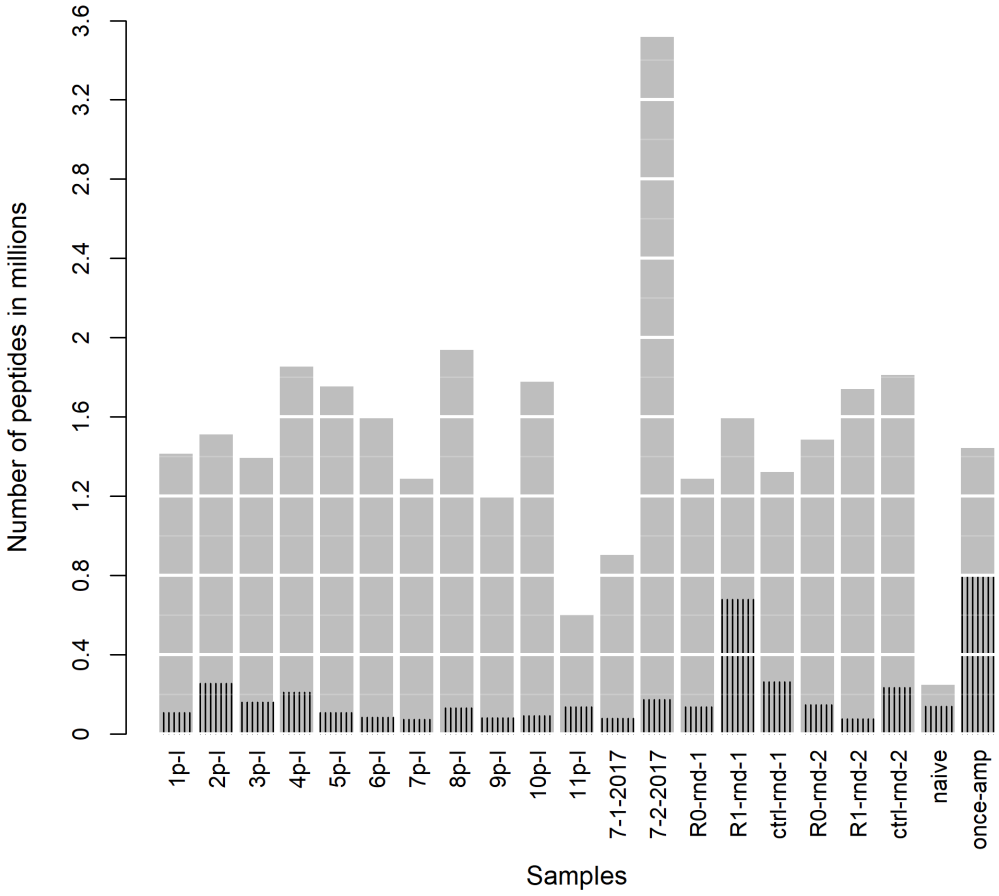


Figure 9: Peptide abundances of translated combinatorial library sequences after filtering out sequencing artifacts. Gray bars indicate the total number of reads. Black shading lines show the number of unique reads.

One source of concern is the very low number of peptides in the naïve library. As a way of checking how its library may have impacted its peptide diversity, an accounting of its amino acid composition was made relative to data published by the manufacturer (*Datasheet for Ph.D.™-12 Library*, n.d.). This approach was also applied to the once-amplified library to examine biases introduced by naïve library repropagation. In table 6,

the **Expected** column lists down theoretical frequencies expected from an *NNK* library. The **Observed** column contains the actual frequencies obtained from next-generation sequencing of the naïve library as reported by the manufacturer. The **naïve** and **once-amp** columns are experimentally determined values. Amino acid compositions of the reference libraries appear to be similar to what is reported by NEB; however, discrepancies are relatively larger in the naïve library (range: -1.1 % to 0.6 %) compared to the once-amplified library (range: -0.6 % to 0.7 %). These results suggest that one round of amplification does not introduce drastic changes to library complexity.

To further investigate the diversity in sequences across all the samples, diversity indices used in ecological studies are introduced here to demonstrate utility in describing peptide libraries. The Shannon index H_j of the j^{th} library defined as

$$H_j = -\sum_{i=1}^{N} p_{ij} \ln p_{ij} \tag{2.1}$$

is akin to the definition of entropy in statistical thermodynamics, where p_{ij} is the probability or relative frequency of the i^{th} peptide sequence in the j^{th} library containing N unique sequences. The Simpson index D_j of the j^{th} is given here as

$$D_j = \sum_{i=1}^{N} p_{ij}^2, \tag{2.2}$$

where squaring prior to summation puts more weight on dominant, more abundant sequences. When Simpson index is formulated as $-\ln D$, it shares a number of properties with the Shannon index:

- H and $-\ln D$ are both zero in the case of a library composed of only one sequence.
- H and $-\ln D$ are both identical and maximum with a value of $\ln N$ in the case of a library with N unique sequences when every sequence is equiprobable (equal frequencies in all unique sequences).

Table 6: Amino acid composition of reference libraries. Percent composition is highlighted in yellow at an intensity scaled relative to the value.

			% Composition			
1 letter abbreviation	Abbreviation	Amino acid name	Expected	Observed	naive	once-amp
A	Ala	Alanine	6.2	7.4	8.0	7.1
C	Cys	Cysteine	3.1	1.5	1.6	1.1
D	Asp	Aspartic acid	3.1	4.6	4.6	4.7
E	Glu	Glutamic acid	3.1	3.1	3.3	3.0
F	Phe	Phenylalanine	3.1	2.7	2.5	3.4
G	Gly	Glycine	6.2	5.8	6.4	6.5
H	His	Histidine	3.1	4.6	4.5	4.2
I	Ile	Isoleucine	3.1	3.4	2.6	3.4
K	Lys	Lysine	3.1	2.3	2.2	2.4
L	Leu	Leucine	9.4	8.9	8.3	8.3
M	Met	Methionine	3.1	3.1	3.6	3.3
N	Asn	Asparagine	3.1	4.5	3.4	4.7
P	Pro	Proline	6.2	8.1	7.8	7.5
Q	Gln	Glutamine	6.2	3.9	3.9	4.0
R	Arg	Arginine	9.4	5.7	6.3	5.7
S	Ser	Serine	9.4	11.2	10.7	11.0
T	Thr	Threonine	6.2	7.8	8.0	7.4
V	Val	Valine	6.2	6.1	6.6	6.2
W	Trp	Tryptophan	3.1	2.3	2.6	2.4
Y	Tyr	Tyrosine	3.1	3.6	3.0	3.7

The results of computing these indices for all 21 libraries are plotted on figure 10. The proximity of the naïve library to the equiprobable line indicates that the sequences it contains are more or less equally represented. On the other hand, the once-amplified library is located further away from this line, which is possibly a reflection of library bias brought about by repropagation. It is interesting to note that most of the libraries appear to fall on a line. It is hypothesized that libraries that originated from the same library introduce the linear behavior — in this case, first-panning-round libraries derived from the naïve library.

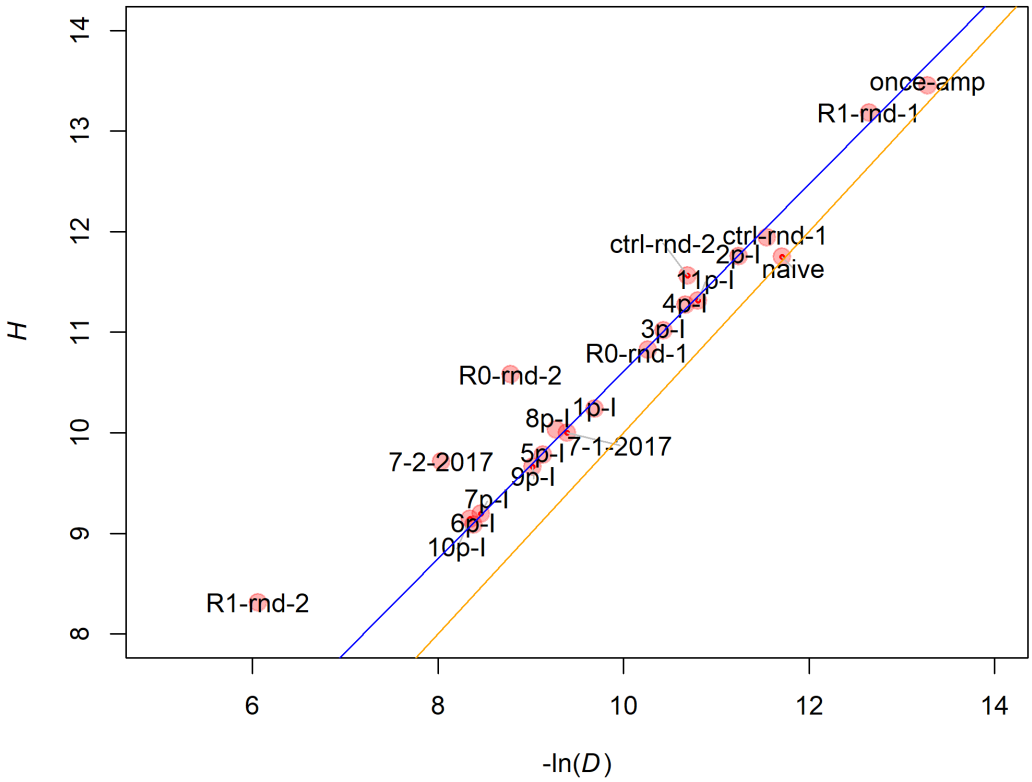


Figure 10: Library diversity as represented by the Shannon index H and the negative logarithm of the Simpson index D. Orange line demarcates libraries with equiprobable sequences. Libraries panned from the naïve library seem to follow a linear trend (blue line).

2.3.3 Peptide diversity is strongly influenced by the library it was panned from and phage-antibody ratio

Given the number of libraries that fell on the blue line in figure 10, it may be important to understand the origin this linearity — to determine under what conditions this behavior holds. Since the diversity indices are a function of the set of probabilities used as input and agnostic to the particular peptide sequences present, it is posited that the behavior is a function of the distributional pattern of the probabilities. Looking at simpler cases of discrete distributions of probabilities, it seems that libraries varying in the number of unique peptides but sharing the same distributional profile yield similarly-shaped curves when ordered by decreasing abundance (figure 11). It can be shown that a plot of diversity indices calculated from these curves akin to figure 10 would yield points that fall on a line with different slopes for each profile.

Carrying out the same analysis on first-panning round libraries yields similarly-shaped curves in support of similarity in distributional profiles as the basis of the observed linear behavior (figure 12). From the figure, R1-rnd-1 appears to be an outlier. It is interesting to note that it differs from the other libraries with respect to the phage-antibody ratio used during panning (see table 5). This result suggests that the linearity observed in figure 10 is most likely a function of the source library used in panning and the relative amount of phage offered to the plasma sample.

Looking at the other libraries, R1-rnd-1 is more similar to the once-amp library. This makes sense when we consider the similarity in their unique peptide abundance in figure 9. The distributional profiles of the second-panning-round libraries would look similar to the first-panning-round libraries if we excluded peptide abundances at their leading end, suggesting that the departures from the blue line in figure 10 is driven by the enrichment in a subset of peptides. It is intriguing that a tenth-fold lower amount of antibody offered

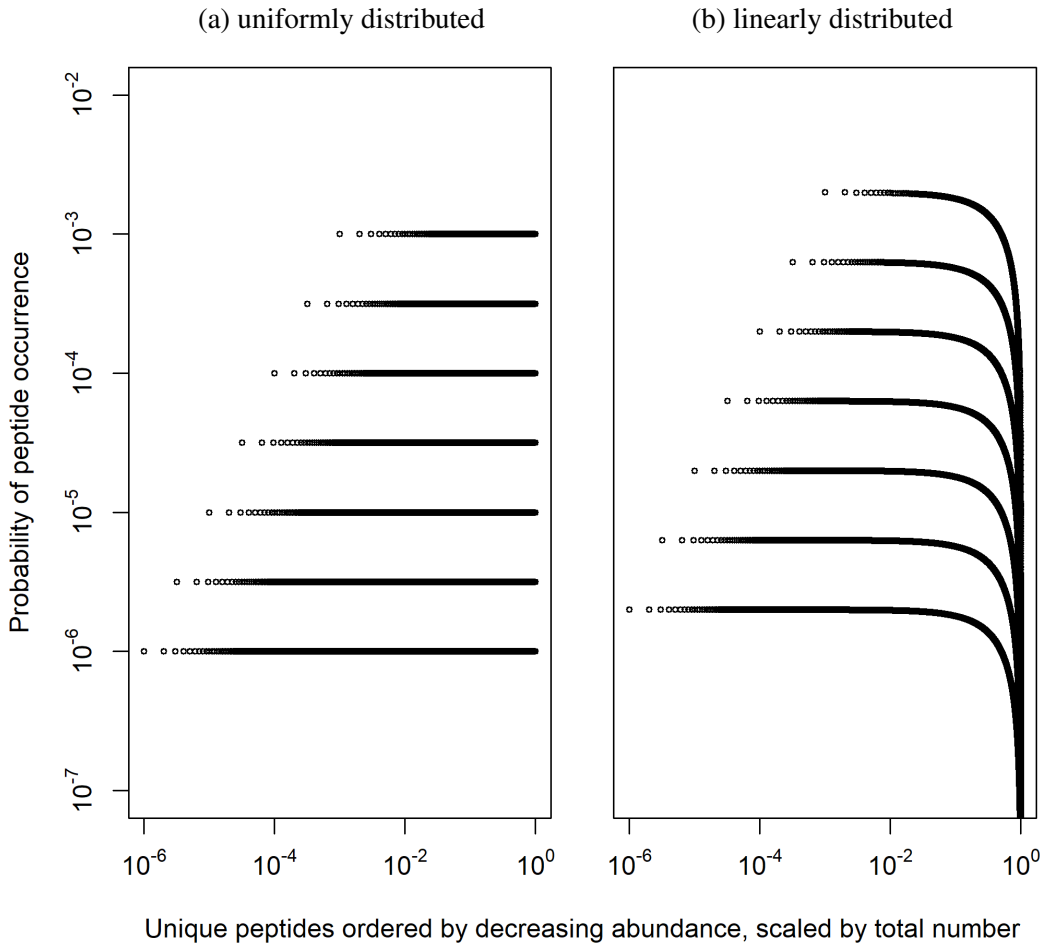


Figure 11: Distributional profiles of unique peptides ordered by decreasing abundance. Peptide order is scaled as a fraction between 0 to 1 based on the total number of unique peptides. Each dot represents a unique peptide sequence.

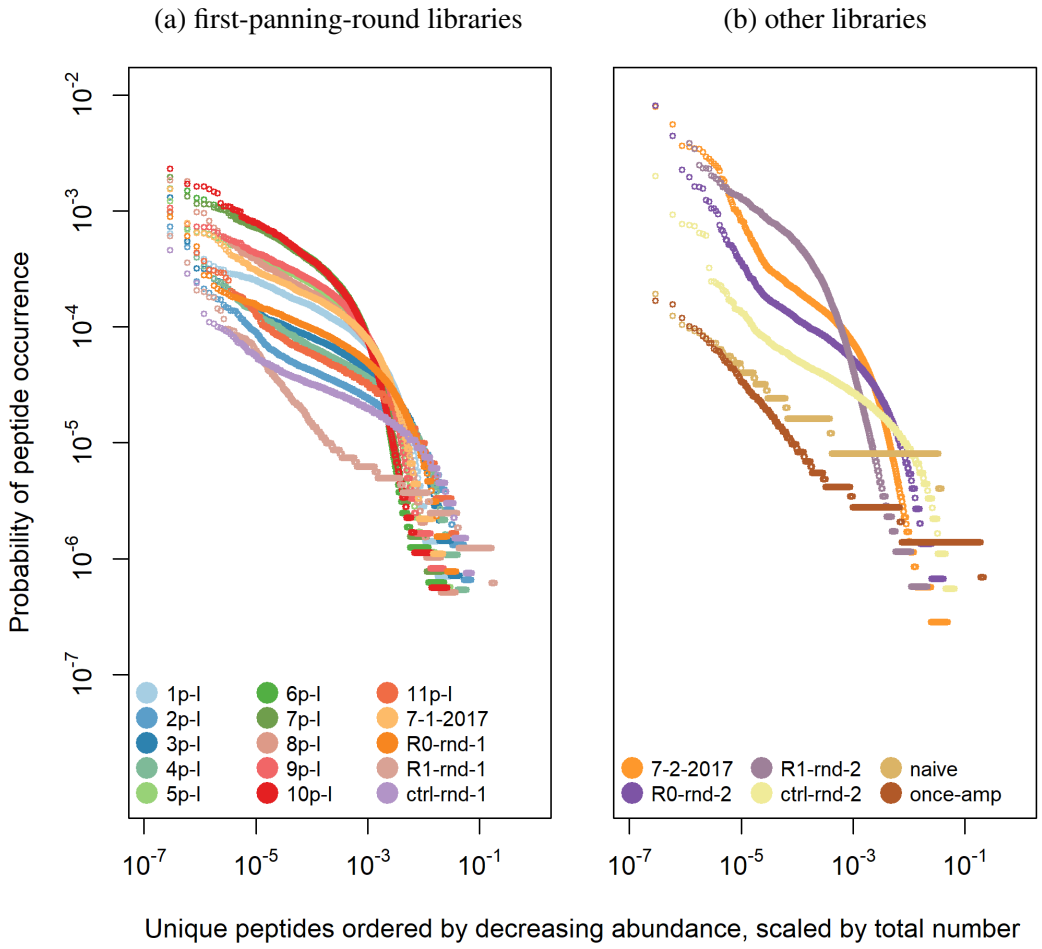


Figure 12: Distributional profiles of unique peptides ordered by decreasing abundance. Peptide order is scaled as a fraction between 0 to 1 based on the total number of unique peptides. Each dot represents a unique peptide sequence.

to the same amount of phage used in the other samples resulted to a profile similar to the once-amp library. A possible explanation could be that phage have some affinity to the protein G bead and tube wall surface during the capture step, albeit weak, with serum proteins readily displacing them. If true, then nonspecific binding contributes largely to the observed profile for R1-rnd-1, which is mitigated by the presence of sacrificial proteins.

2.3.4 Sequences with low overall read counts across all libraries obscure read count variations in sequences with higher read counts

Peptide sequences with a low number of reads would typically yield results of low confidence due to background noise arising from sequencing artifacts (Matochko and Derda, 2013; G. Park et al., 2017; Rabadan et al., 2018; Sha, Phan, and Wang, 2015). To determine a suitable cutoff to exclude these sequences, the total read count of each peptide across all libraries was calculated and a tally was made of how many peptides were of a total read count. Plotting these values reveals a peculiar dependence of abundance to whether the total read count of a peptide is odd or even in number for low counts (figure 13). This behavior may be a consequence of accommodating mismatched paired-end reads. As this oscillation diminishes with increasing total read count, a cutoff of 100 was chosen, resulting to the exclusion of around 98.5 % of the sequences. This leaves 52 447 sequences for analysis.

Examination of the distribution of nonzero counts across all the libraries before filtering reveal that majority of read counts are low, with 75 % or more of read counts having a value of ten or less (figure 14a). After filtering based on total read count, a shift in the median read count is observed in most libraries, suggesting an enrichment in a subset of peptides based on a comparison of panned to reference libraries naïve and once-amp (figure 14b). We again see similarity between the distributions of R1-rnd-1 and

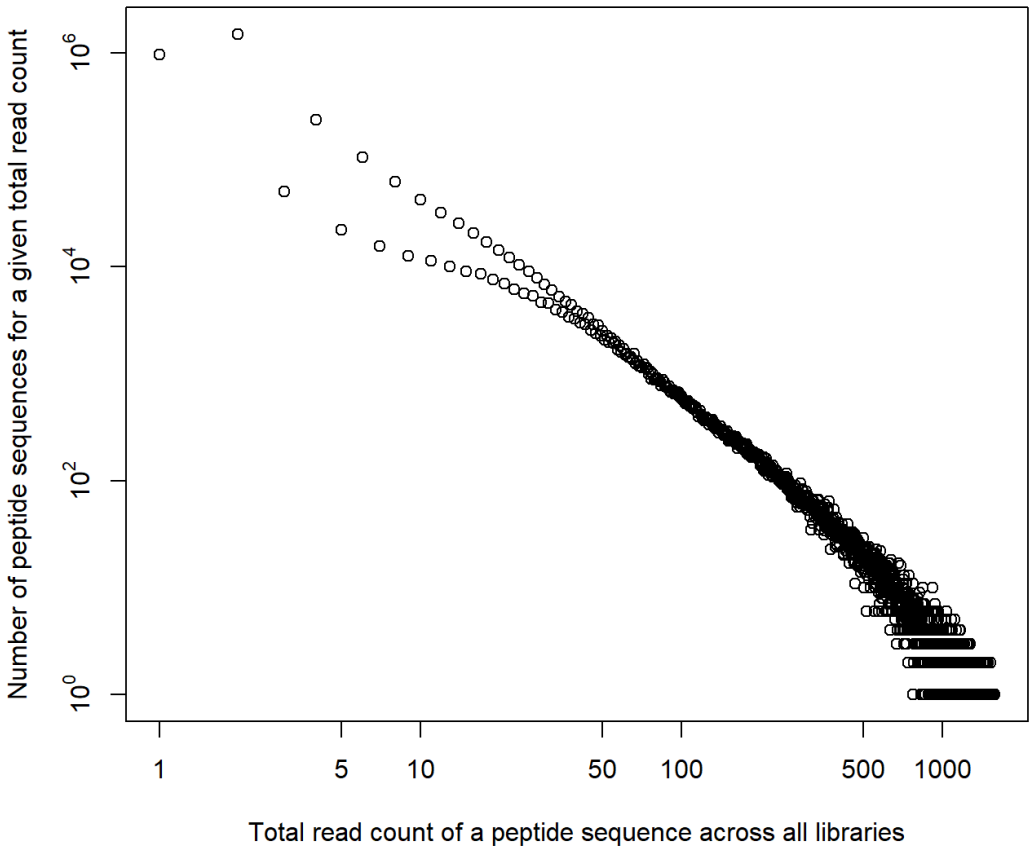


Figure 13: Total read count frequencies of peptide sequences across all libraries.

the once-amp library, supporting the statement made in the previous subsection regarding nonspecific binding when panning with very diluted serum.

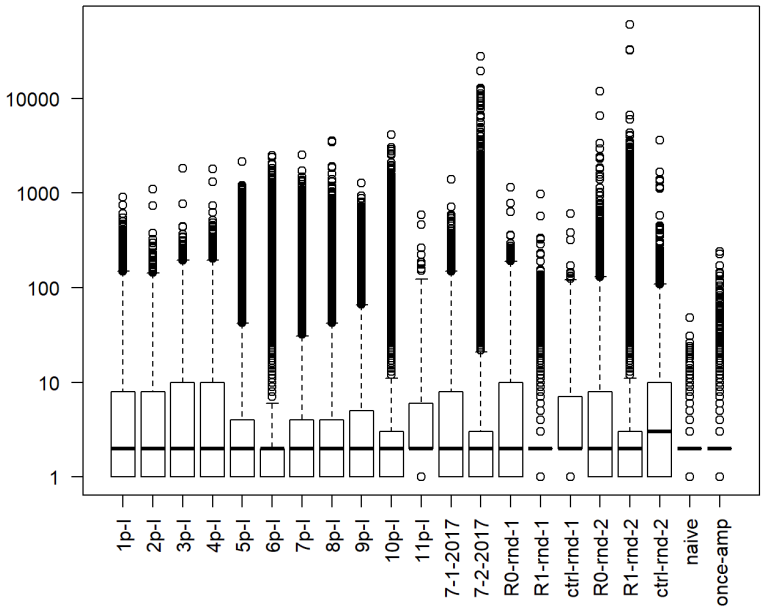
2.3.5 Unweighted beta diversity metric calculated peptide-wise across all libraries reveals target-unrelated peptide candidates

As suggested by Ricotta (2017), unweighted peptide-wise beta diversity β based on the Shannon index H were calculated across all libraries and profiled according to the number of libraries a peptide sequence had zero counts figure 15. Briefly, β is calculated as

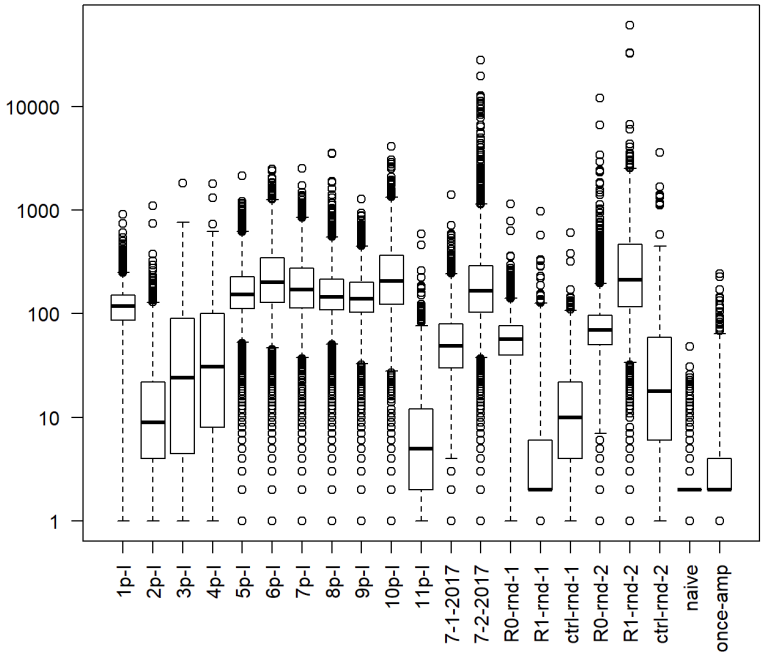
$$\begin{cases} H_i & = -\sum_{j=1}^{S} p_{ji} \ln p_{ji} \\ \beta_i & = 1 - \dfrac{H_{ji}}{\ln S} \\ \beta & = \dfrac{w_i \beta_i}{\sum_{i=1}^{N} w_i \beta_i}, \end{cases} \tag{2.3}$$

where p_{ji} is the probability or relative frequency of j^{th} library in the i^{th} peptide sequence for library samples 1 to S, where weights w_i of each peptide are set equal for the unweighted case. Given that the more even the distribution of counts across the libraries are the lower the value of the unweighted β will be, it is hypothesized that target-unrelated peptides (TUPs) will be located at the lower left corner of figure 15. To identify a window to gate these TUP candidates, highly probable TUPs were identified based on abundant, most frequently-occurring sequences across all the libraries.

To determine the most frequently occurring peptides, the top ten most abundant peptide sequences were determined for each library and tabulated according to how frequently they occurred across all 21 libraries. The results are given in table 7. From the table, sequences with the same frequency of occurrence starts to become more common



(a) Before peptide-wise total read count filtering



(b) After peptide-wise total read count filtering

Figure 14: Nonzero read count distributions of peptide sequences across all libraries.

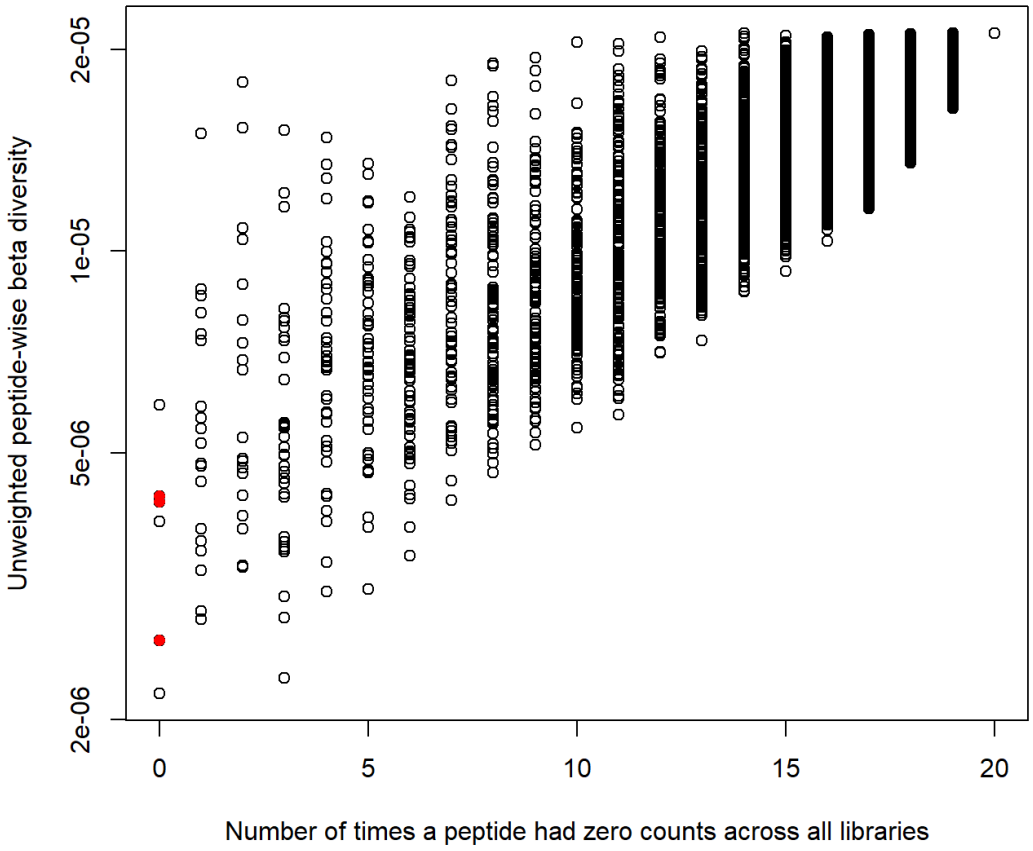


Figure 15: Relation between unweighted peptide-wise beta diversity and the number of zero counts across all libraries. Each dot represents a peptide sequence. Red solid dots mark the top three most frequently-occurring peptide sequences.

across all the libraries after the top five sequences. As sequences are more likely to be found in only a few libraries, these top three sequences were taken to be TUP candidates for further consideration.

From figure 15, the locations of the top three TUP candidates suggests a search region of all sequences with unweighted beta diversities of less than 5×10^{-6}. To facilitate identification, these sequences were run through SAROTUP website to check for polystyrene surface-binding peptides (PSBinder web service) and peptide appearing in multiple unrelated biopanning datasets (MimoSearch web service). The results are given in tables 8 and 9. Adding the locations of peptide sequences that have PSBinder probabilities greater than or equal to 0.90 or have been identified in three or more unrelated biopanning datasets are shown in figure 16. It is argued that these sequences may introduce false positives later on during enrichment analysis by virtue of association with the identified sequences. Thus, all sequences with unweighted beta diversities of less than 5×10^{-6} were excluded from further analysis, resulting in the removal of 71 sequences. It is interesting to note that two of the top three most frequently-occuring peptide sequences are predicted with high probability (SGVYKVAYDWQH, 0.91; GLHTSATNLYLH, 0.96) to be polystyrene surface-binding peptides although polystyrene is not known to be present in any of the panning components (polypropylene tubes, protein G-conjugated magnetic bead based on macroporous cellulose).

2.3.6 Weighted beta diversity metric calculated peptide-wise across various library subsets facilitates selection of enriched peptides

By taking weighted beta diversities across all peptide sequences for a given set of libraries, peptide distributions can be constructed based on this metric and stratified based on the number of times a peptide was zero in a library. The calculations of this metric

Table 7: Top ten most abundant sequences for each of the 21 libraries. The occurrence across the 21 libraries (Freq), the number of libraries with zero reads (# Zeroes), and the peptide-wise unweighted beta diversity (Unweighted Beta) is listed beside each of sequence.

	Sequence	Freq	# Zeroes	Unweighted Beta		Sequence	Freq	# Zeroes	Unweighted Beta
1	SGVYKVAYDWQH	19	0	4.22×10^{-6}	62	HPNTPAMQPVDG	1	7	1.63×10^{-5}
2	GLHTSATNLYLH	15	0	2.62×10^{-6}	63	HSPKDWRPHSFL	1	13	1.95×10^{-5}
3	STPIFAEATARS	10	0	4.31×10^{-6}	64	IITQAASKSNLV	1	5	7.39×10^{-6}
4	DDFRVWWPNFPR	7	1	3.33×10^{-6}	65	KASGSPSGFWPS	1	2	3.37×10^{-6}
5	QFDYMRPANDTH	5	1	4.82×10^{-6}	66	KCCFADLGPVTP	1	14	2.09×10^{-5}
6	GFAVGARDSLMF	4	1	4.53×10^{-6}	67	KCCFYNVPTSSA	1	19	2.11×10^{-5}
7	NHLSTPVWSITG	4	2	7.29×10^{-6}	68	KCCYSAPVETAM	1	8	1.89×10^{-5}
8	SALKGLFPADHH	4	0	2.18×10^{-6}	69	KCCYSFPAELLT	1	14	2.01×10^{-5}
9	SQPWDDSTNRRV	4	2	7.89×10^{-6}	70	KCCYVPDNAGMR	1	11	2.04×10^{-5}
10	TSIQISNAHPKS	4	1	7.35×10^{-6}	71	KFHTPSPIQVDL	1	19	1.64×10^{-5}
11	ALAKVWIVTSPP	3	7	1.44×10^{-5}	72	KIWDIYQGGHTY	1	2	1.79×10^{-5}
12	DWSSWVYRDPQT	3	8	1.18×10^{-5}	73	KLWQIPSGAIES	1	14	2.11×10^{-5}
13	GSAPLLTVDTSK	3	3	4.96×10^{-6}	74	KLWSLPTSTIDL	1	1	8.77×10^{-6}
14	SQDIRTWNGTRS	3	1	3.69×10^{-6}	75	KLWTIPYNTGTS	1	10	1.47×10^{-5}
15	VVSPDMNLLLTN	3	0	5.90×10^{-6}	76	KLWVLPDVTDVR	1	11	1.99×10^{-5}
16	AFHPRQMETQMY	2	1	8.59×10^{-6}	77	KVWMIGPHEPPV	1	5	1.03×10^{-5}
17	ASLTMAYNNPRF	2	13	1.63×10^{-5}	78	KVWQLYAGGDNI	1	1	1.50×10^{-5}
18	AWADQPVTAPNR	2	1	5.87×10^{-6}	79	MHPNAGHGSLMR	1	6	6.93×10^{-6}
19	GHGHKVWRVPPV	2	2	1.53×10^{-5}	80	MIQTNWDKLGLV	1	1	5.63×10^{-6}
20	HPHDYNDLTSPF	2	3	4.69×10^{-6}	81	MSPVMNERETER	1	16	2.08×10^{-5}
21	HTGLIGQDCWTC	2	4	6.85×10^{-6}	82	NASSFPTNSRWA	1	10	7.82×10^{-6}
22	IGLPHSANSTKP	2	3	7.43×10^{-6}	83	NISWPFATSNHW	1	8	1.12×10^{-5}
23	IPLGRDGGSYQR	2	2	4.32×10^{-6}	84	NYDGTRQSTPGW	1	11	1.83×10^{-5}
24	LPMHTNLPSGPL	2	6	1.08×10^{-5}	85	QGPGMGPGDQFK	1	19	2.06×10^{-5}
25	SMGPNTSYSLAH	2	3	3.05×10^{-6}	86	QSASYYHTLGKQ	1	20	2.12×10^{-5}
26	SPIADFGQPLNF	2	5	1.31×10^{-5}	87	QWNWPVRSVANV	1	8	1.62×10^{-5}
27	TENVSAELARSY	2	2	4.85×10^{-6}	88	RVFDPPWHVASM	1	12	2.09×10^{-5}
28	TFAANPLPVLLV	2	2	6.65×10^{-6}	89	SALTDILRSTAF	1	20	2.12×10^{-5}
29	THLPFSQNLADV	2	1	8.10×10^{-6}	90	SGRYVPDEHLYT	1	15	1.37×10^{-5}
30	VLKEASHLPYSG	2	2	3.85×10^{-6}	91	SLDGAGAALRTS	1	4	6.75×10^{-6}
31	ADAPGSMGWHKY	1	11	1.67×10^{-5}	92	SLFMQDPGVRIG	1	5	8.69×10^{-6}
32	AHTLTGTKTRDQ	1	13	8.96×10^{-6}	93	SLMKGLSGDQWI	1	10	1.66×10^{-5}
33	ALSPQHYTNLPD	1	6	8.07×10^{-6}	94	SLVPWPNSYEAG	1	7	1.54×10^{-5}
34	ANITLNHLPTLT	1	7	1.05×10^{-5}	95	SMSTNWTWWKEN	1	17	2.11×10^{-5}
35	APHSPYMKSLMS	1	17	2.09×10^{-5}	96	SNTQSERHPLSM	1	8	5.99×10^{-6}
36	AWFPSNAVTTLS	1	17	2.09×10^{-5}	97	SSNSYTPVSFGR	1	19	2.11×10^{-5}
37	AWRDGPTYSLHN	1	20	2.12×10^{-5}	98	STTSNFFGALVH	1	9	1.31×10^{-5}
38	DAYRAHAGPGQM	1	8	1.27×10^{-5}	99	SYGPNTLWVSEV	1	19	2.12×10^{-5}
39	DDLNSGTPPAWS	1	19	2.12×10^{-5}	100	SYPGHVGIFKIA	1	19	2.11×10^{-5}
40	DHAPSFLGTYNS	1	5	6.71×10^{-6}	101	TDGLKSGQGMSK	1	3	8.21×10^{-6}
41	DPILPKKLWIVK	1	7	1.11×10^{-5}	102	TGAPPRLDARPA	1	5	4.89×10^{-6}
42	DPVGLGGWWAKV	1	5	1.03×10^{-5}	103	TLPAILQSSGTR	1	5	8.06×10^{-6}
43	DRWVARDPASIF	1	2	4.90×10^{-6}	104	TLTSETPWSLNR	1	2	1.04×10^{-5}
44	DSGTKSHFKSMY	1	15	1.95×10^{-5}	105	TNVNSNLWQINR	1	19	2.12×10^{-5}
45	DSQFNKYSIATV	1	3	5.46×10^{-6}	106	TPAVHDSFRNPK	1	20	2.12×10^{-5}
46	DSSGMGPGDAIR	1	16	2.05×10^{-5}	107	TPQSFWQKGSLV	1	2	4.65×10^{-6}
47	DSSVRPGHSLNL	1	15	1.85×10^{-5}	108	TSLFPVSEHFSG	1	20	2.12×10^{-5}
48	FIPFDPMSMRWE	1	3	4.29×10^{-6}	109	TSNSPGPMWAGP	1	19	2.12×10^{-5}
49	FPNSISTVIRPV	1	20	2.12×10^{-5}	110	TSQTNAKVWQIY	1	7	1.03×10^{-5}
50	FSDPDMRAWALS	1	3	3.62×10^{-6}	111	TWAKCCYAGYAN	1	8	1.70×10^{-5}
51	GAVVNQLATVSF	1	18	1.68×10^{-5}	112	VAASPYYAPRVP	1	15	2.04×10^{-5}
52	GDLLTFQNFVMK	1	7	1.43×10^{-5}	113	VEAKCCFSMHKT	1	10	2.06×10^{-5}
53	GHGSGANPPDVR	1	19	1.64×10^{-5}	114	VGVVASEDKLYL	1	18	2.03×10^{-5}
54	GIGYELEHKAYI	1	19	1.99×10^{-5}	115	VHWDFRQWWQPS	1	1	7.52×10^{-6}
55	GKLDAVVLKTPT	1	20	2.12×10^{-5}	116	VIAKSSPVMDYH	1	16	2.09×10^{-5}
56	GLGDELKRDDWF	1	18	1.95×10^{-5}	117	VIVPPSGHQGAA	1	8	1.34×10^{-5}
57	GQSEHHMRVASF	1	7	1.25×10^{-5}	118	WSNNGASHTQIH	1	12	1.83×10^{-5}
58	GSTLGKSGALSQ	1	8	1.28×10^{-5}	119	WVSAEDSPPWIR	1	20	2.12×10^{-5}
59	GYTTENYKTTHP	1	19	2.03×10^{-5}	120	YDSDSKVAAPYR	1	20	2.12×10^{-5}
60	HKIVSWDWLSSR	1	15	1.49×10^{-5}	121	YHDPNRKCCYAA	1	7	1.52×10^{-5}
61	HLMGEGPSTAPR	1	17	2.08×10^{-5}	122	YPFSFSSGPQAI	1	20	2.12×10^{-5}

Table 8: Peptide sequences with unweighted beta diversities of less 5×10^{-6} and their probabilities of being polystyrene surface-binding peptides as identified using the PSBinder web service.

	Query Sequenceă	Lengthă	Probabilityă		Query Sequenceă	Lengthă	Probabilityă
1	GLHTSATNLYLH	12	0.96	37	SMRASYPMPTFI	12	0.44
2	QLAWQLSYSWPG	12	0.95	38	SQDIRTWNGTRS	12	0.43
3	GGVYKVAYDWQH	12	0.94	39	IPVKSWPIRPSS	12	0.42
4	SLHTSATNLYLH	12	0.93	40	KVFSIGDIQKHQ	12	0.42
5	KASGSPSGFWPS	12	0.91	41	DRWVARDPASIF	12	0.41
6	SGVYKVAYDWQH	12	0.91	42	ALVTSLMENEST	12	0.4
7	SSYYNSRWAFYP	12	0.9	43	VSGQRSVGTPLS	12	0.4
8	YSLRSDFLPFAT	12	0.9	44	YLQDRATRLSFG	12	0.38
9	HAQHMRVWGAVS	12	0.88	45	HSVRYDFTGLLE	12	0.37
10	SSLWSELYGGSM	12	0.88	46	VNNSKVVFPVTN	12	0.37
11	RGVYKVAYDWQH	12	0.85	47	YYRTDQVVNLRS	12	0.37
12	LKCYGSPLIDYL	12	0.81	48	SLHTTGRGIFHL	12	0.36
13	NDFRVWWPNFPR	12	0.78	49	EDLRKESSRLVD	12	0.35
14	TPQSFWQKGSLV	12	0.77	50	HPHDYNDLTSPF	12	0.35
15	VLKEASHLPYSG	12	0.77	51	KVFLLNMSDPNT	12	0.34
16	SMEEAVVSPTST	12	0.76	52	SLASEDTPNVLA	12	0.33
17	FIPFDPMSMRWE	12	0.75	53	SLHRDYPKLRSA	12	0.31
18	KIWFPMGNYQSN	12	0.75	54	IPLGRDGGSYQR	12	0.3
19	MHAIPGDHVVEN	12	0.75	55	HPLTWNLRSSPA	12	0.26
20	SVPMGSLASLES	12	0.73	56	DYANRLSGRGQV	12	0.24
21	QFDYMRPANDTH	12	0.66	57	NDRNLLPLSGNA	12	0.23
22	SALKGLFPADHH	12	0.66	58	TENVSAELARSY	12	0.22
23	YVKGQMPRSWFP	12	0.66	59	GHKVWMVPTVTR	12	0.2
24	YSLRHDAFWDVE	12	0.65	60	NAPIPSFSPLSK	12	0.19
25	DDFRVWWPNFPR	12	0.63	61	VDPTRDWQLLSS	12	0.18
26	SASYNMKRMSFV	12	0.61	62	KVPVGVLPLSHS	12	0.16
27	VSLSGVSSNSRV	12	0.61	63	SGASSDMLGMPN	12	0.16
28	MPYKIPSTFFNI	12	0.57	64	TGAPPRLDARPA	12	0.16
29	FSDPDMRAWALS	12	0.54	65	TVNPIFMVQLAE	12	0.16
30	GFAVGARDSLMF	12	0.53	66	MTARIFDPPLTV	12	0.15
31	SFAVGARDSLMF	12	0.53	67	GSAPLLTVDTSK	12	0.14
32	STPIFAEATARS	12	0.53	68	VAHSYRSDKTLI	12	0.13
33	SLLHTSMPSMIA	12	0.5	69	VAKVWQVQAPQE	12	0.08
34	MSWTDLHHQEYL	12	0.47	70	SKLTSYQSPTMQ	12	0.07
35	SMGPNTSYSLAH	12	0.46	71	LLVPQDPMAGAI	12	0.06
36	GEAKIWRMPQHP	12	0.44				

Table 9: Peptide sequences with unweighted beta diversities of less 5×10^{-6} that have been reported in at least one biopanning dataset submitted to the Biopanning Data Bank via the MimoSearch web service.

Query Sequence	BiopanningDataSet ID	Target Name
GLHTSATNLYLH	3261	Macrophage-stimulating protein receptor, MSP receptor
	2874	Cation-independent mannose-6-phosphate receptor
	3204	Particulate matter, PM
	3290	Extracellular domain of Fibroblast growth factor receptor 2 [1-377]
	3023	Human blood-brain barrier (BBB) cellular model
	3024	
	3386	BDNF/NT-3 growth factors receptor (EC:2.7.10.1)[1-430]
	3206	SECp43180
KASGSPSGFWPS	2874	Cation-independent mannose-6-phosphate receptor
	3023	Human blood-brain barrier (BBB) cellular model
	3057	Cellulose of paper
SGVYKVAYDWQH	3058	The printed toner of standard office laser printers
	3059	
	3310	Mouse heart
	3023	Human blood-brain barrier (BBB) cellular model
	3024	
	3206	SECp43180
	2874	Cation-independent mannose-6-phosphate receptor
	3086	Prominin-1
	3326	Folate receptor alpha
	3014	C-terminal half of aryl hydrocarbon receptor (AhR)
	3107	Anti-Dengue virus (DENV) polyclonal antibody
	3027	Deinagkistrodon acutus venom
	3386	BDNF/NT-3 growth factors receptor (EC:2.7.10.1)[1-430]
TPQSFWQKGSLV	3057	Cellulose of paper
	3261	Macrophage-stimulating protein receptor, MSP receptor
VLKEASHLPYSG	3023	Human blood-brain barrier (BBB) cellular model
FIPFDPMSMRWE	3024	Human blood-brain barrier (BBB) cellular model
	3226	Myeloperoxidase
QFDYMRPANDTH	3325	CD63 protein fragment containing the second extracellular loop (CD63 LEL)
SALKGLFPADHH	2954	Beta-lactamase 2 (EC:3.5.2.6)
STPIFAEATARS	3024	Human blood-brain barrier (BBB) cellular model
SLLHTSMPSMIA	3246	Cholera toxin subunit B, CTX-B
SMRASYPMPTFI	3023	Human blood-brain barrier (BBB) cellular model
SQDIRTWNGTRS	3325	CD63 protein fragment containing the second extracellular loop (CD63 LEL)
	3023	Human blood-brain barrier (BBB) cellular model
	3058	The printed toner of standard office laser printers
DRWVARDPASIF	3261	Macrophage-stimulating protein receptor, MSP receptor
	3262	
	3325	CD63 protein fragment containing the second extracellular loop (CD63 LEL)
	3309	MoS2
	3343	CD177 antigen
KVPVGVLPLSHS	3057	Cellulose of paper
GSAPLLTVDTSK	3335	Human gastric cancer cell line MKN-45

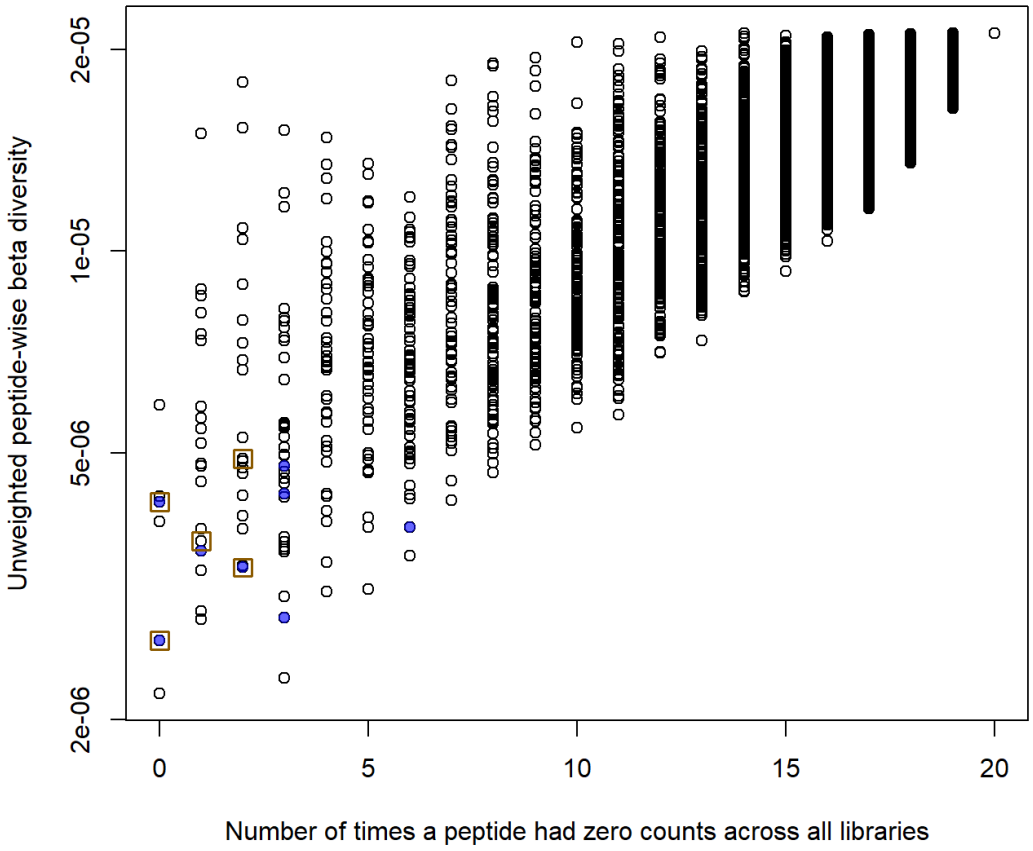


Figure 16: Relation between unweighted peptide-wise beta diversity and the number of zero counts across all libraries. Blue solid dots are highly suspected polystyrene surface-binding peptides. Orange boxes indicate presence in ≥ 3 unrelated Biopanning Data Bank datasets.

evaluates for peptide candidates based on the unevenness of distribution across the libraries and the total read counts. A minimum weighted beta diversity threshold is defined based on the lowest among the extremes of the upper whiskers of adjusted box plots of the zero-stratified data. Sequences passing this threshold are aligned to custom protein databases for influenza, human, and human-associated microorganisms. This approach is illustrated for the following cases.

2.3.6.1 Enrichment of influenza peptides from a single individual

Panned libraries R0-rnd-1, R1-rnd-1, ctrl-rnd-1 R0-rnd-2, R0-rnd-2, and ctrl-rnd-2 were derived from the same individual (donor 8) at varying phage-to-antibody ratios (R0 and R1) after a recent flu shot as well as post vaccination one year after (ctrl). To determine influenza sequences that were enriched in donor 8 (table 4) upon flu vaccination, all six donor 8-derived libraries were examined together with the two reference libraries (naïve, once-amp). Alignment with influenza sequences resulted to 153 sequences matching. Figure 17 shows the location of these influenza-matched peptides with respect to the other peptides present in the library set. Blue solid dots mark sequences that matched to influenza sequences passing the threshold marked by a blue dotted horizontal line. The threshold is based on the lowest upper-whisker extreme across adjusted box plots. To facilitate visualization, analysis was further restricted to high-scoring sequences with a bit score of 17 or better (figure 18), yielding 53 sequences. Inspection of figure 19 indicates enrichment for influenza based on read count differences between test (R0 and R1) and control (ctrl, naïve, once-amp) cases. An exception would be KVWSIEPVNSQH whose read count pattern (which includes relatively high read counts in post-vaccination sample ctrl-rnd-2) suggests a long-lasting humoral response elicited to the epitope mimicked by this peptide. Another observation to make would be the presence or absence of a peptide in R0 but not in

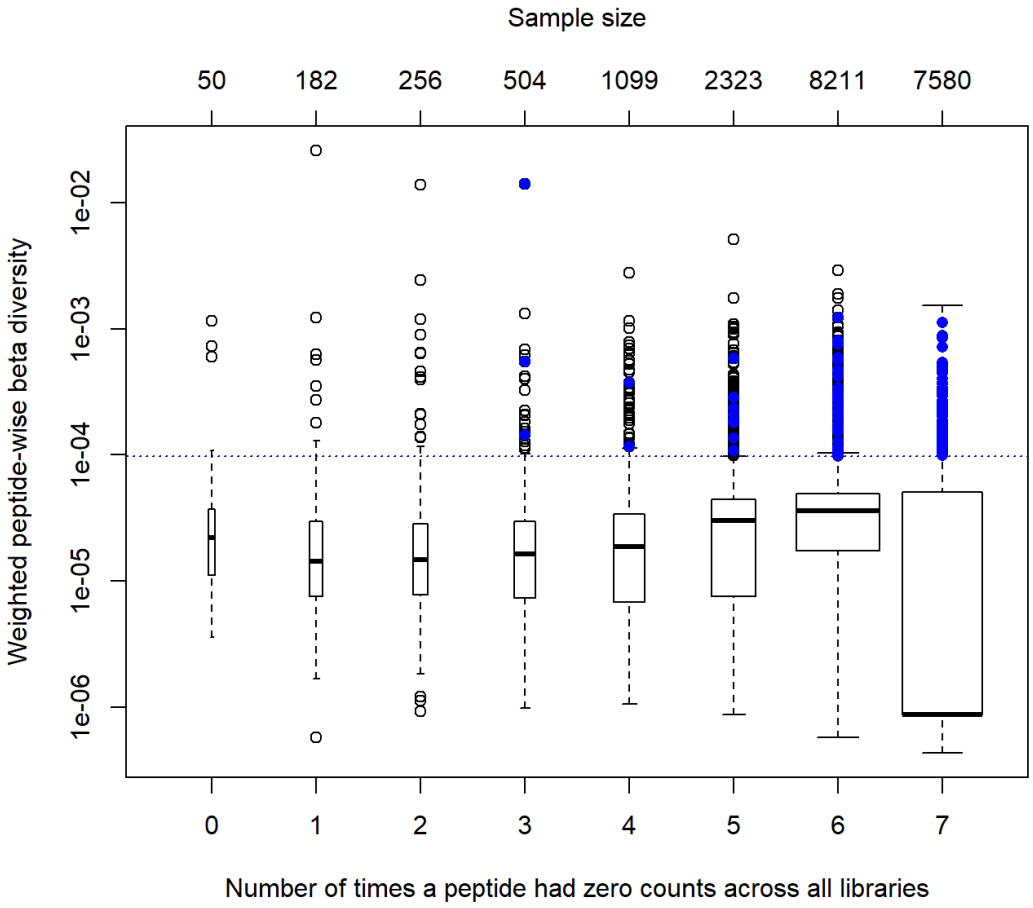


Figure 17: Relation between weighted peptide-wise beta diversity and the number of zero counts across six libraries derived from a single influenza-vaccinated donor and the two reference libraries. Outliers on the adjusted box plots are indicated by circles.

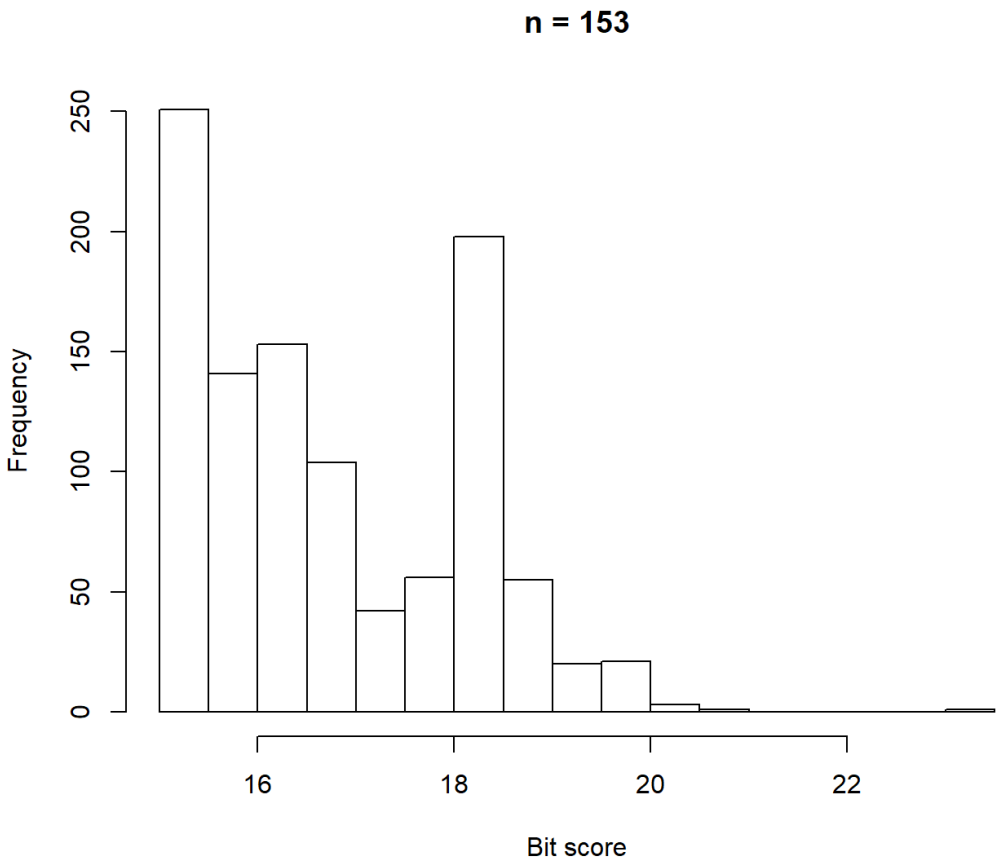


Figure 18: Bit score distribution of influenza-matching search hits that scored 15 or better, with the number of sequences matching indicated on top.

R1 and vice-versa which may be a consequence of sampling inadequately covering the entire sequence landscape. Looking at patterns across the 54 sequences selected (figure 20), the most frequently-occurring search hits (blue lines) correspond to sequences found in only one or two libraries (based on the number of zeroes). In figure 20, axes represent column headers returned by sequence alignment output merge with calculated diversity-related metrics: zeroes = number of libraries with zero read counts, weightBeta = weight beta diversity, mismatch = number of mismatches, qstart/qend = start/end of alignment in query, bitscore = bit score, evalue = expectation value, flu strain = influenza strains, protein = influenza viral proteins, seq = influenza-matching peptide sequences. Sequences corresponding to these search hits are characterized by relatively high weighted beta diversities and a moderate number of mismatches. It is unsurprising that the most frequently-occurring hits have relatively low bit scores and therefore correspondingly high E values. Hits match mostly to viral proteins from influenza A and B with some hits to influenza D which is not known to be pathogenic in humans (Su et al., 2017).

2.3.6.2 Enrichment of influenza peptides across several individuals

Panned libraries 9p-I, 10p-I, and 11p-I are single-round pannings from different donors who have taken a flu shot. To explore enrichment in these libraries, first-round libraries from donor 8 and the two reference libraries are included in the set. It is argued that the sequence selection process across different donors will tend to favor those from libraries with higher mean read counts. However, it is desirable to enrich for peptide sequences from all donors instead of only a few especially when there are no common sequences enriched across libraries. A workaround would be to express read counts as proportions to make the total count of each library the same. Nevertheless, doing so will inflate the read counts in libraries with a lesser number of sequences whereas all first-round libraries

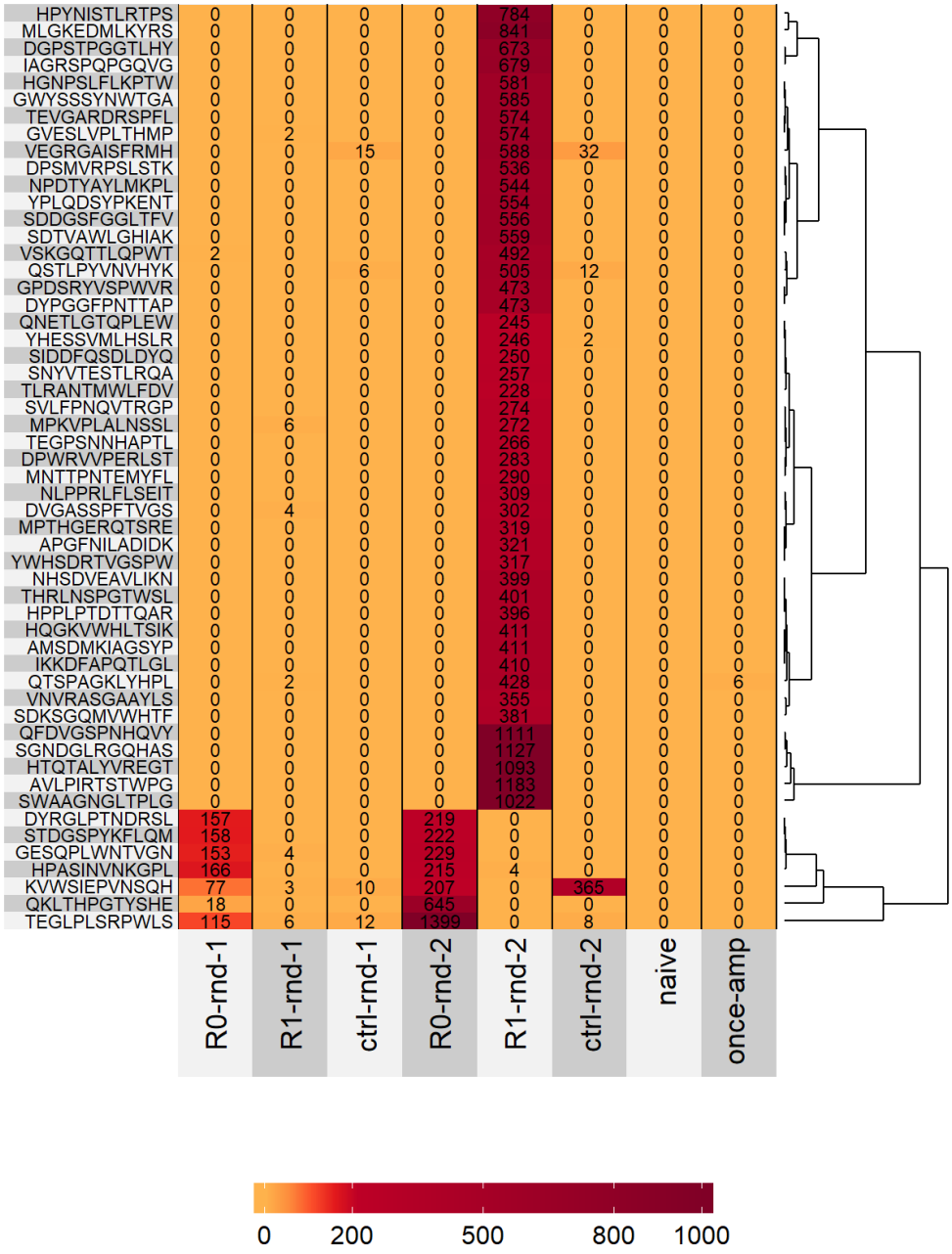


Figure 19: Heatmap of 54 influenza-matched peptide sequences derived from a single donor. Color intensity is relative to the number of read counts. Rows and columns are reordered using hierarchical clustering via Ward's linkage method.

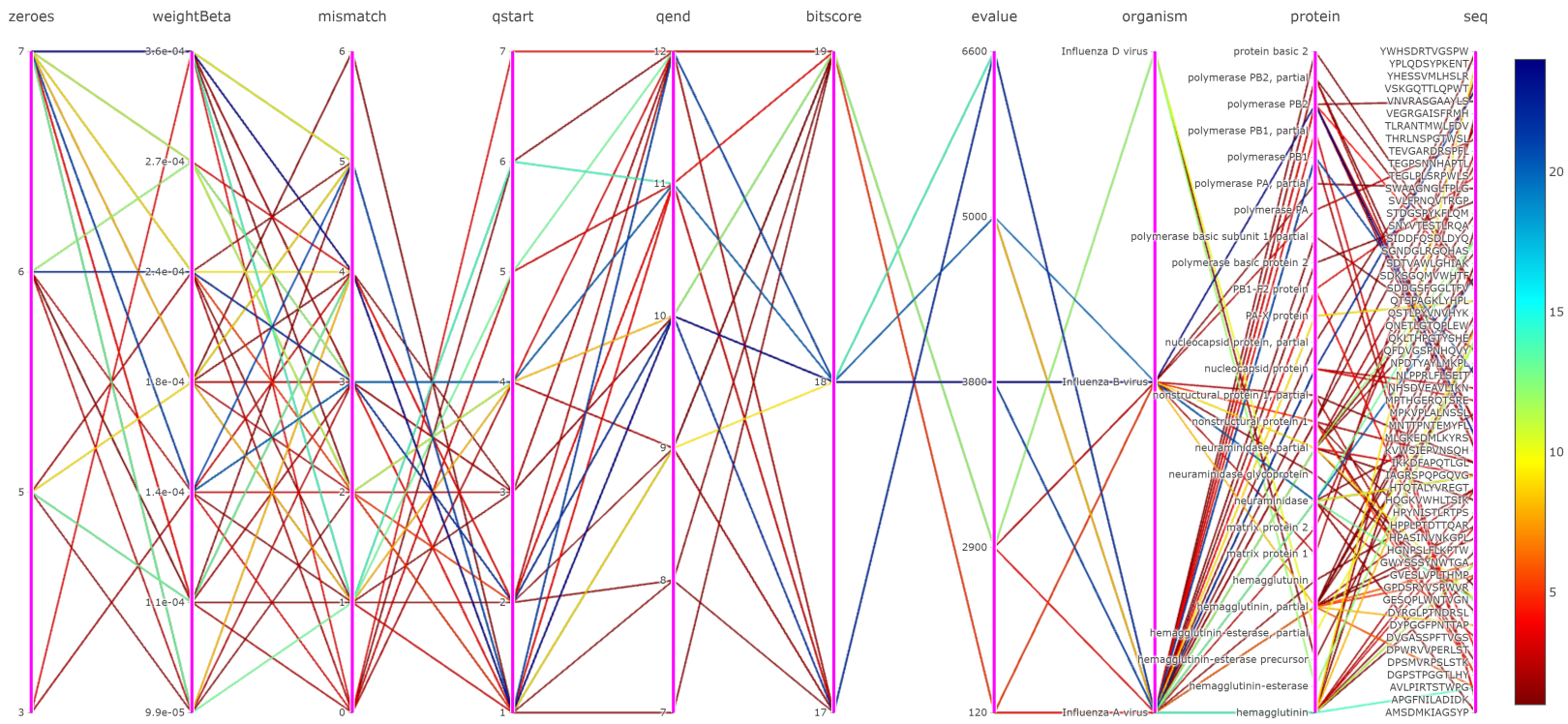
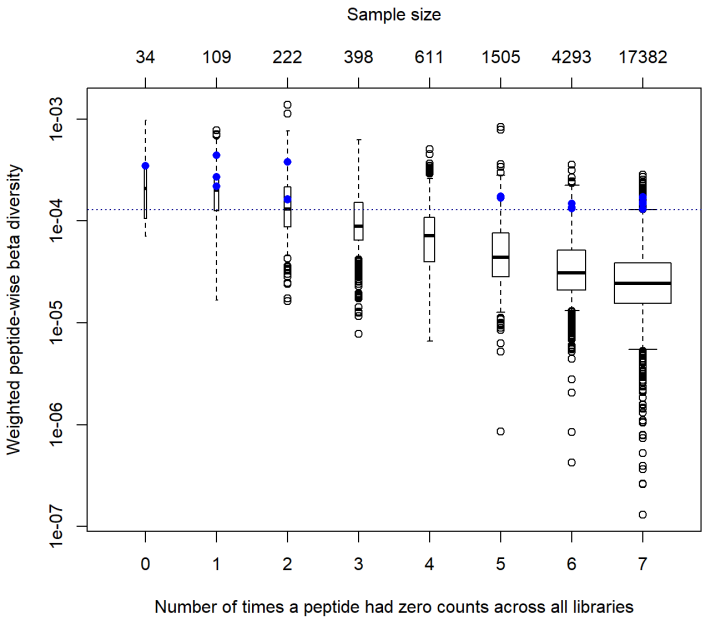


Figure 20: Parallel coordinates plot of 54 influenza-matching sequences with a bit score of 17 or better from the tabular output of the sequence aligner DIAMOND. Lines are color-coded to indicate frequency of occurrence.

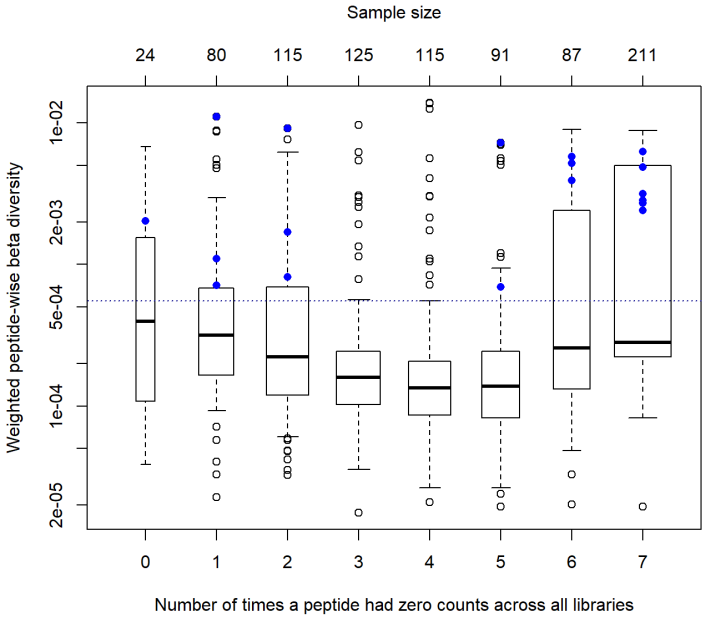
were derived from the same starting naïve library. Thus, it is proposed to perform an initial enrichment based on read proportions follow by a second enrichment based on the read counts of sequences passing the initial enrichment. This approach is illustrated in figure 21. Blue solid dots mark sequences that matched to influenza sequences passing the threshold marked by a blue dotted horizontal line. The threshold is based on the lowest upper-whisker extreme across adjusted box plots.

A preliminary screening for enriched peptide candidates is carried out using read proportions, resulting to 848 sequences making it pass the threshold (figure 21a). These sequences are then subsetted from the read count table from which enrichment on their read counts instead is performed, resulting to 229 sequences above the threshold (figure 21b). The same collection of sequences matching to influenza (blue dots) is shown on both subfigures. An examination of all the sequences obtained thru this two-step process in comparison to one-step direct enrichment using read counts directly (which yields 1048 sequences) is shown in figure 22. One-step enrichment is based on directly using read counts. Two-step enrichment perform a initial screen based on read proportions followed by enrichment using read counts. From figure 22a, one-step enrichment appears to emphasize heavily sequences derived from the libraries with higher mean read counts (10p-I: 25.95, 9p-I: 12.86, R0-rnd-1: 9.24, ctrl-rnd-1: 1.14, R1-rnd-1: 0.72, once-amp: 0.44, 11p-I: 0.43, naive: 0.08). On the other hand, two-step enrichment allows inclusion of sequences takes into consideration information contained in the other libraries used in the comparison (figure 22b).

An analysis of bit scores of influenza-matching sequences from the two-step process results in only 18 sequences at a bit score threshold of 15 (figure 23). Two of the entries were found to correspond to antibody complexes with influenza proteins. Thus, excluding those leaves 16 sequences for generating a heatmap.

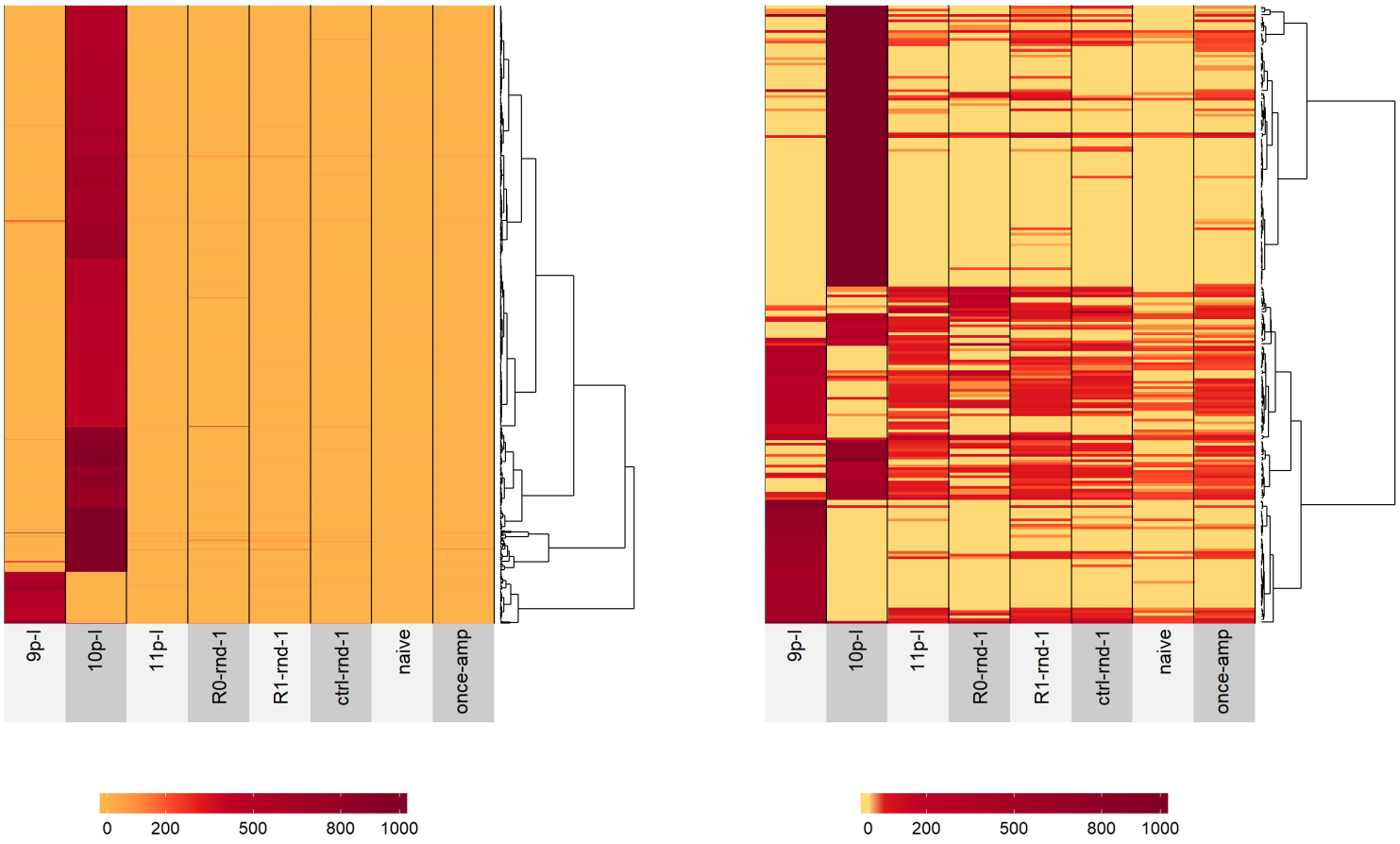


(a) First enrichment based on read proportions



(b) Second enrichment based on read counts of sequences passing first enrichment

Figure 21: Relation between weighted peptide-wise beta diversity and the number of zero counts across first-round panning libraries from influenza-vaccinated donors with the two reference libraries. Outliers on the adjusted box plots are indicated by circles.



(a) One-step enrichment

(b) Two-step enrichment

Figure 22: Heatmaps of read counts across first-round panning libraries from influenza-vaccinated donors with the two reference libraries. Color intensity is relative to the number of read counts. Rows and columns are reordered using hierarchical clustering via Ward's linkage method.

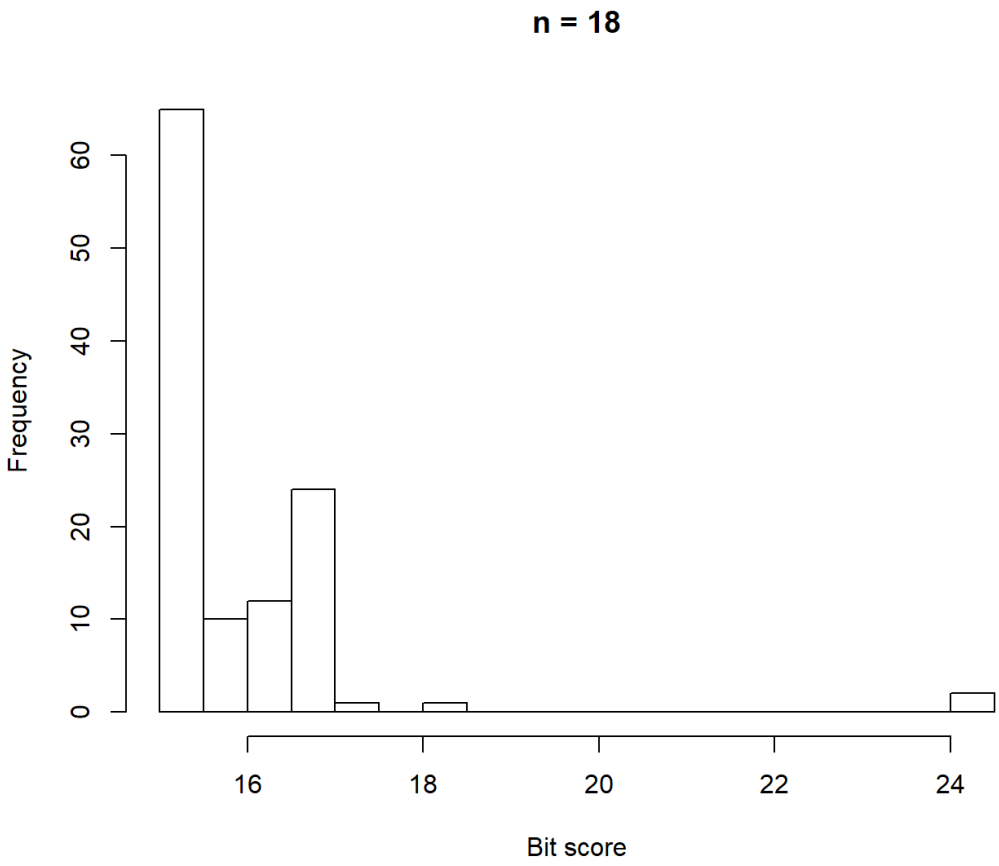


Figure 23: Bit score distribution of influenza-matching search hits that scored 15 or better, with the number of sequences matching indicated on top.

From figure 24, we find sequences enriched in the libraries 9p-I and 10p-I. We also find a new peptide sequence NLQYTLVHRDPY, which did show up in the previous analysis. The exclusion of the second-panning-round library R0-rnd-2 from the set resulted in peptides from the previous analysis failing to make it pass the initial screening of the two-step enrichment process. We also see sequences that appear to be enriched in more than one library like TDGLKSGQGMSK, THLPFSQNLADV, and TLQLPSSAGTTR. The first two sequences correspond to hemagglutinin which is a viral coat protein while the third sequence matches to nonstructural protein 1 which is a glycoprotein involved in the inhibition of the host innate immune response, with both proteins mapping to influenza A.

Looking at patterns across the 16 sequences selected (figure 25), the search hits that most frequently occur (blue lines) appear to be associated with more than one libraries (based on the number of zeroes). In figure 25, axes represent column headers returned by sequence alignment output merge with calculated diversity-related metrics: zeroes = number of libraries with zero read counts, weightBeta = weight beta diversity, mismatch = number of mismatches, qstart/qend = start/end of alignment in query, bitscore = bit score, evalue = expectation value, flu strain = influenza strains, protein = influenza viral proteins, seq = influenza-matching peptide sequences. Sequences corresponding to these search hits are characterized by moderate to high weighted beta diversities and number of mismatches. The most frequently-occurring hits both come from low and high bit scores corresponding to high E values. Hits match to a number of viral proteins from influenza A and B only.

2.3.6.3 Enrichment of AML-associated peptides across several patients

We repeat the two-step process in the previous section on the libraries from AML patients with the libraries from influenza-vaccinated individual as controls, including the two reference libraries, taking sequences passing the threshold (figure 26). Blue solid dots

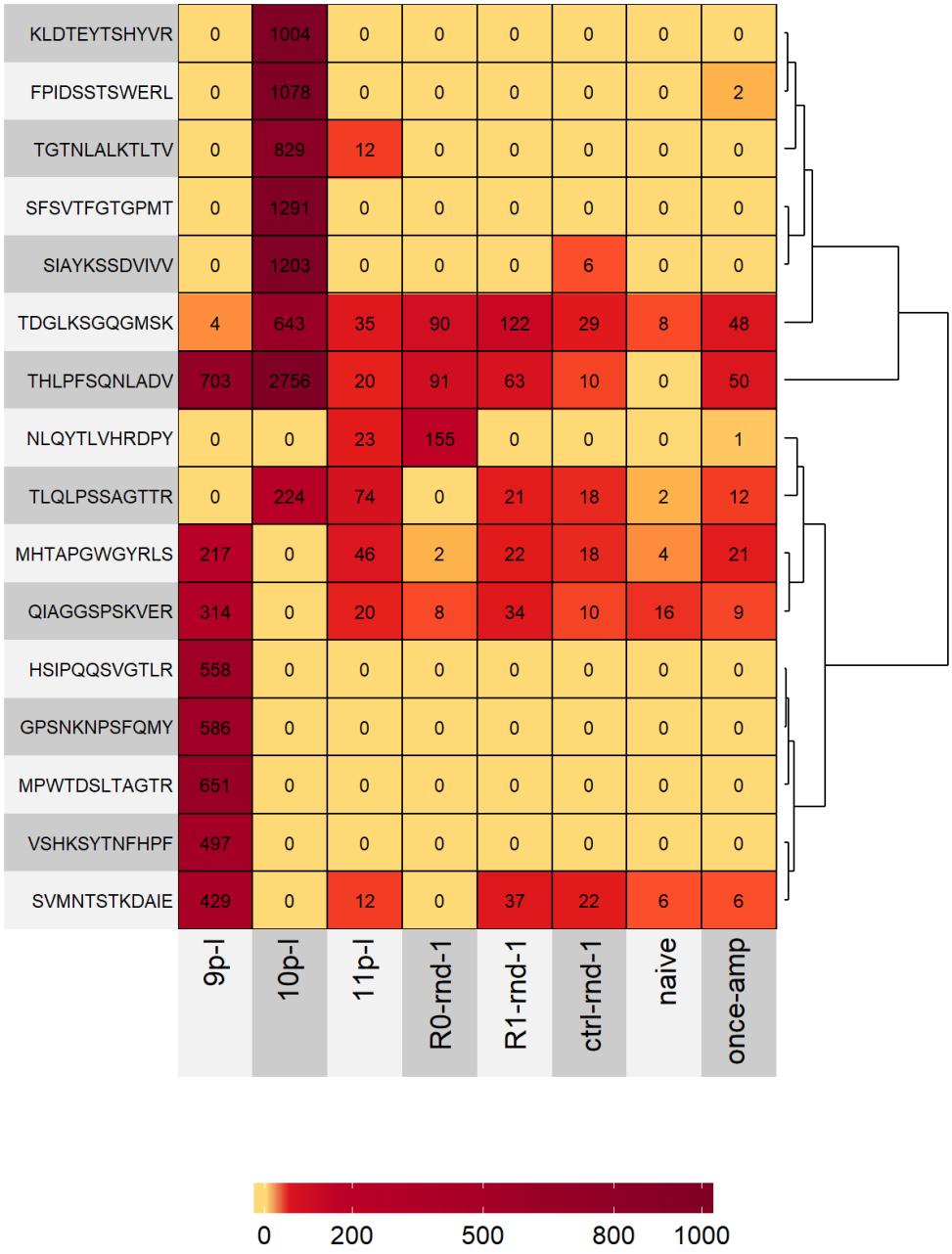


Figure 24: Heatmap of influenza-matched peptide sequences derived across several donors. Color intensity is relative to the number of read counts. Rows and columns are reordered using hierarchical clustering via Ward's linkage method.

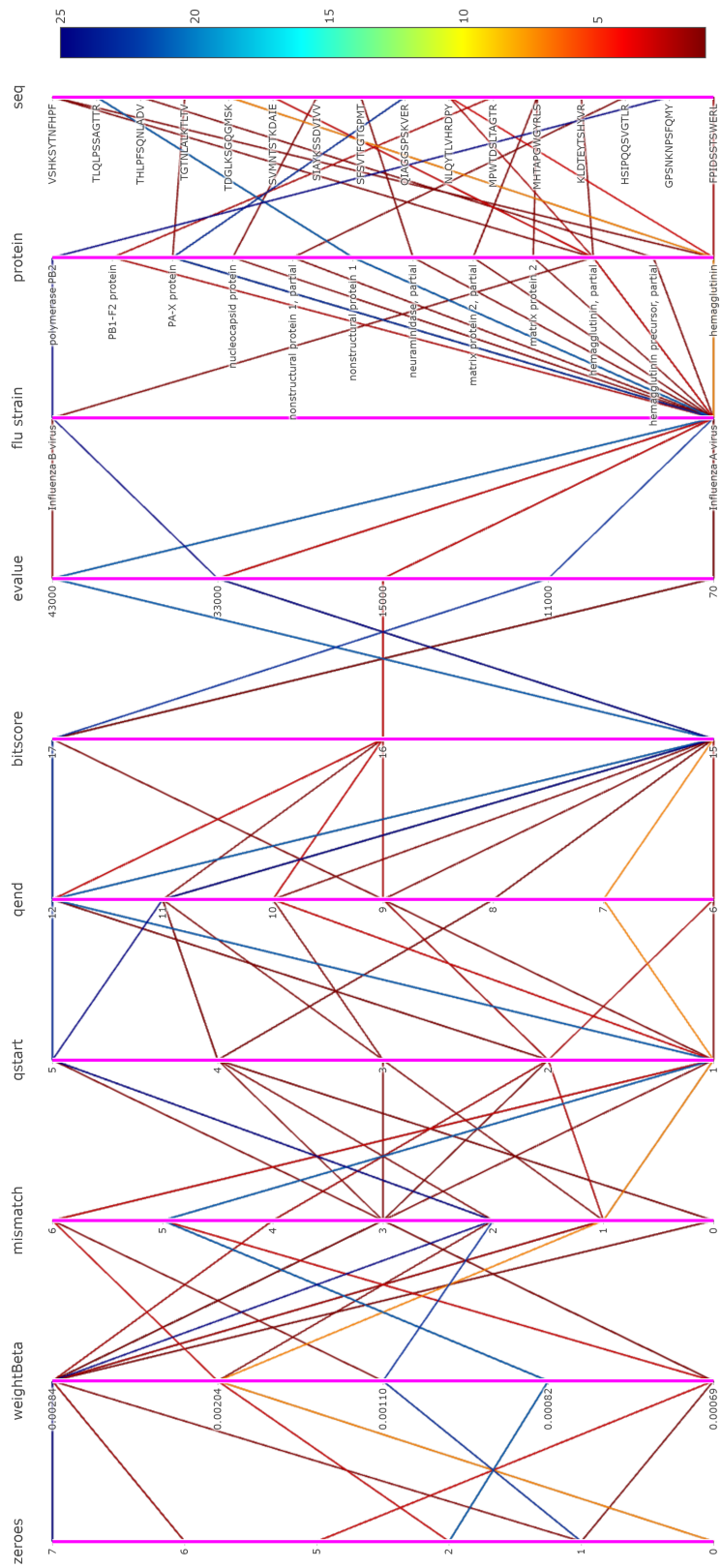
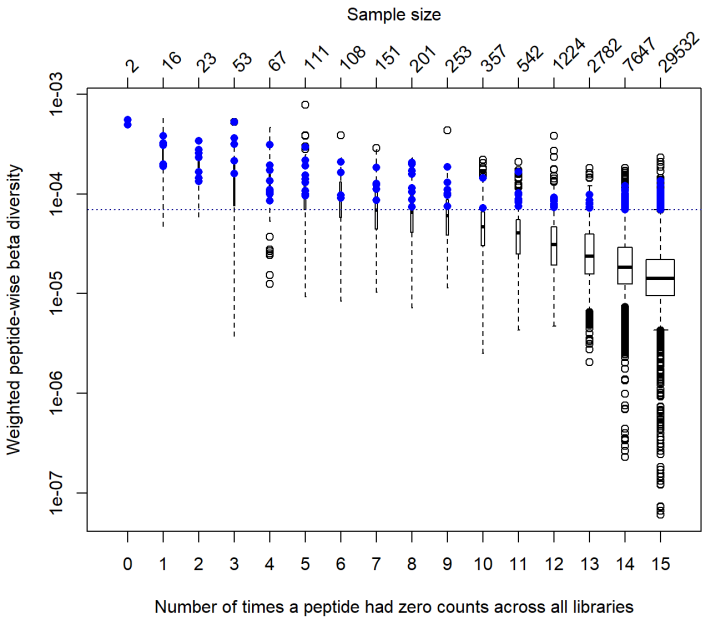


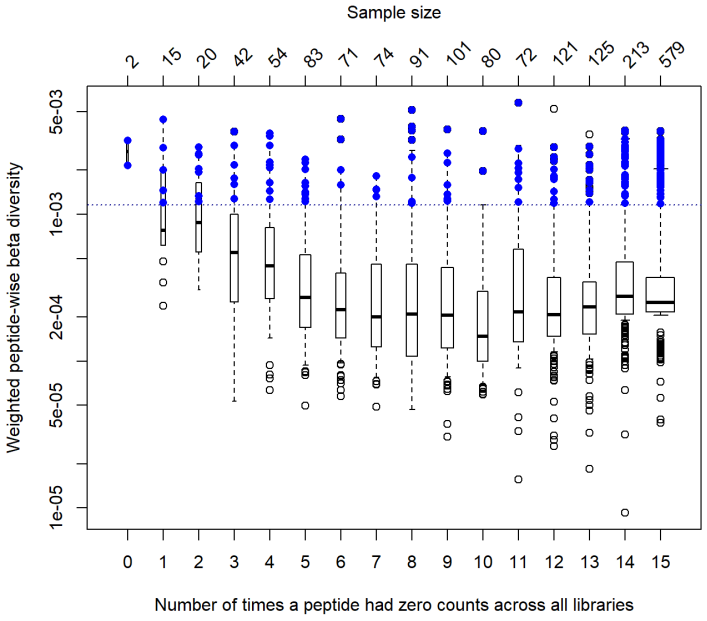
Figure 25: Parallel coordinates plot of 16 influenza-matching sequences with a bit score of 15 or better from the tabular output of the sequence aligner DIAMOND. Lines are color coded to indicate frequency of occurrence.

mark sequences that matched to human protein sequences passing the threshold marked by a blue dotted horizontal line. The threshold is based on the lowest upper-whisker extreme across adjusted box plots. An analysis of bit scores of human-protein-matching sequences from the two-step process results in only 233 sequences at a bit score threshold of 15 (figure 27). All these sequences were carried over for generating a heatmap. Plotting the read counts of these sequences on a heatmap reveal bands of peptide sequences that are possibly enriched across the AML samples, specially in the nonresponder libraries 5p-I, 6p-I, 7p-I, and 8p-I (figure 28). Interestingly, we also see a strong band with the influenza-vaccinated library 10p-I. Subsetting for the top 50 frequently-occurring combination of column values (ex. zeroes, weightBeta, mismatch) yield patterns for 43 sequences (figure 29). In figure 29, axes represent column headers returned by sequence alignment output merge with calculated diversity-related metrics: zeroes = number of libraries with zero read counts, weightBeta = weight beta diversity, mismatch = number of mismatches, qstart/qend = start/end of alignment in query, bitscore = bit score, evalue = expectation value, organism = source organism, protein = proteins from source organism, seq = peptide sequences. The selected sequences indicate that the most frequently-occurring search hits (lines colored green to blue from the color scale) appear to be associated with one to three libraries (based on the number of zeroes). These sequences span the full range of weighted beta diversities and number of mismatches. The same may be said for their bit scores and corresponding E values. Some proteins in the search hits has been reported in literature to be AML-associated like circ-ANAPC7 which has been recently suggested as a potential biomarker and novel drug target (H. Chen et al., 2018), and AF-6 which is a component of the fusion oncoprotein MLL-AF6 (Manara et al., 2014).

By mapping the 233 human-protein-matching sequences to the protein sequence databases for influenza and known human pathogens, we find sequences that have corresponding matches to nonhuman proteins (figure 30) as well and may serve as the expla-



(a) First enrichment based on read proportions



(b) Second enrichment based on read counts of sequences passing first enrichment

Figure 26: Relation between weighted peptide-wise beta diversity and the number of zero counts across first-round panning libraries from AML patients and influenza-vaccinated donors with the two reference libraries. Outliers are indicated by circles.

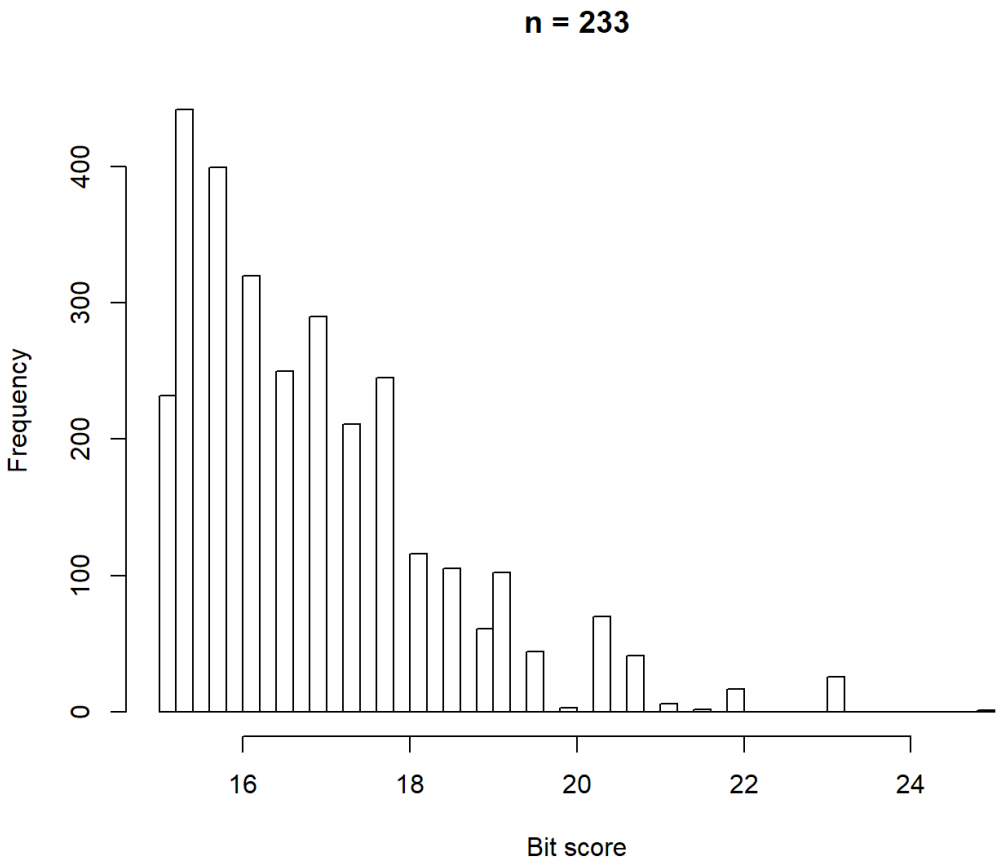


Figure 27: Bit score distribution of human-protein-matching search hits that scored 15 or better, with the number of sequences matching indicated on top.

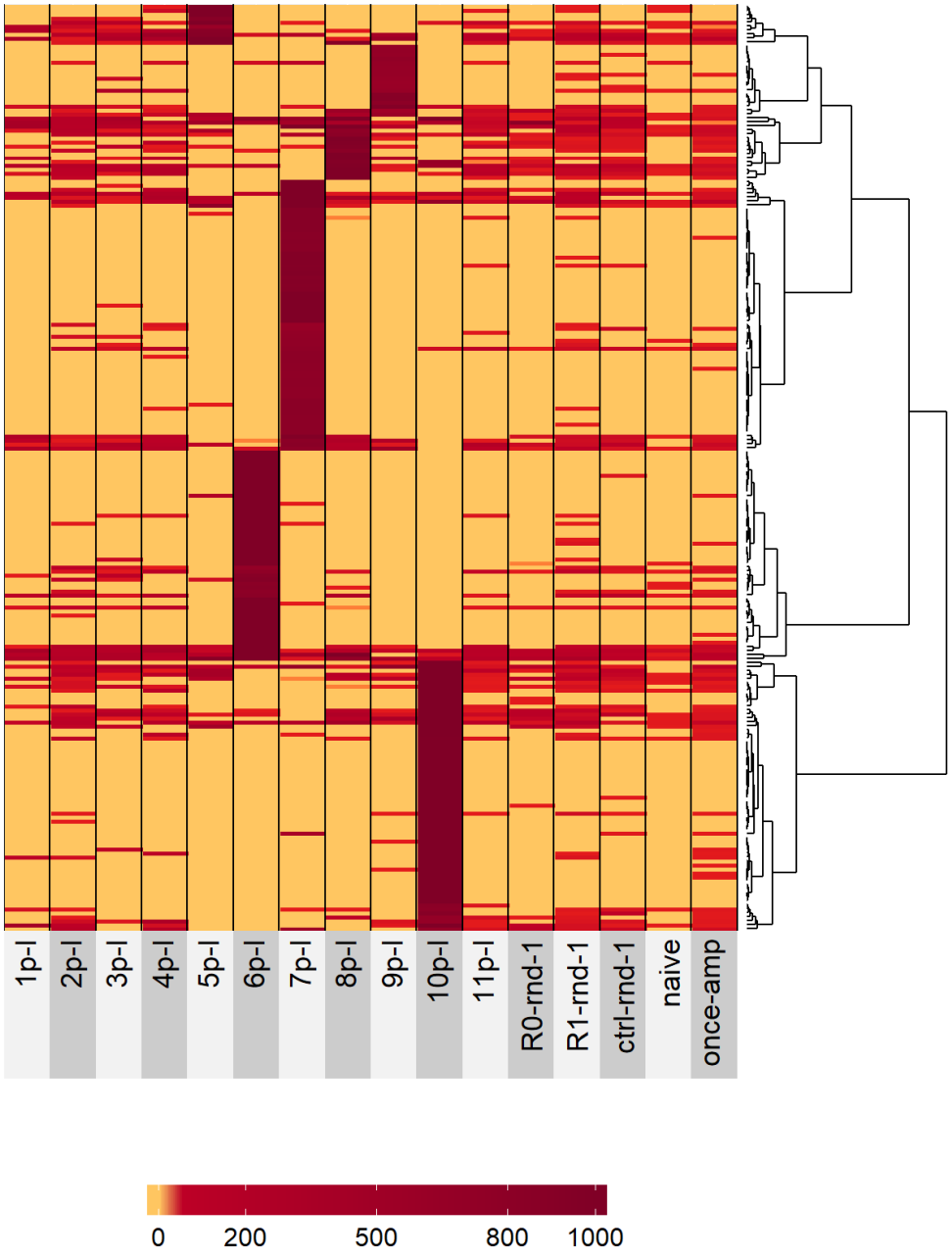


Figure 28: Heatmap of 233 peptide sequences matching human proteins. Color intensity is relative to the number of read counts. Rows are peptide sequences while columns are libraries. Reordering is via hierarchical clustering via Ward's linkage method.

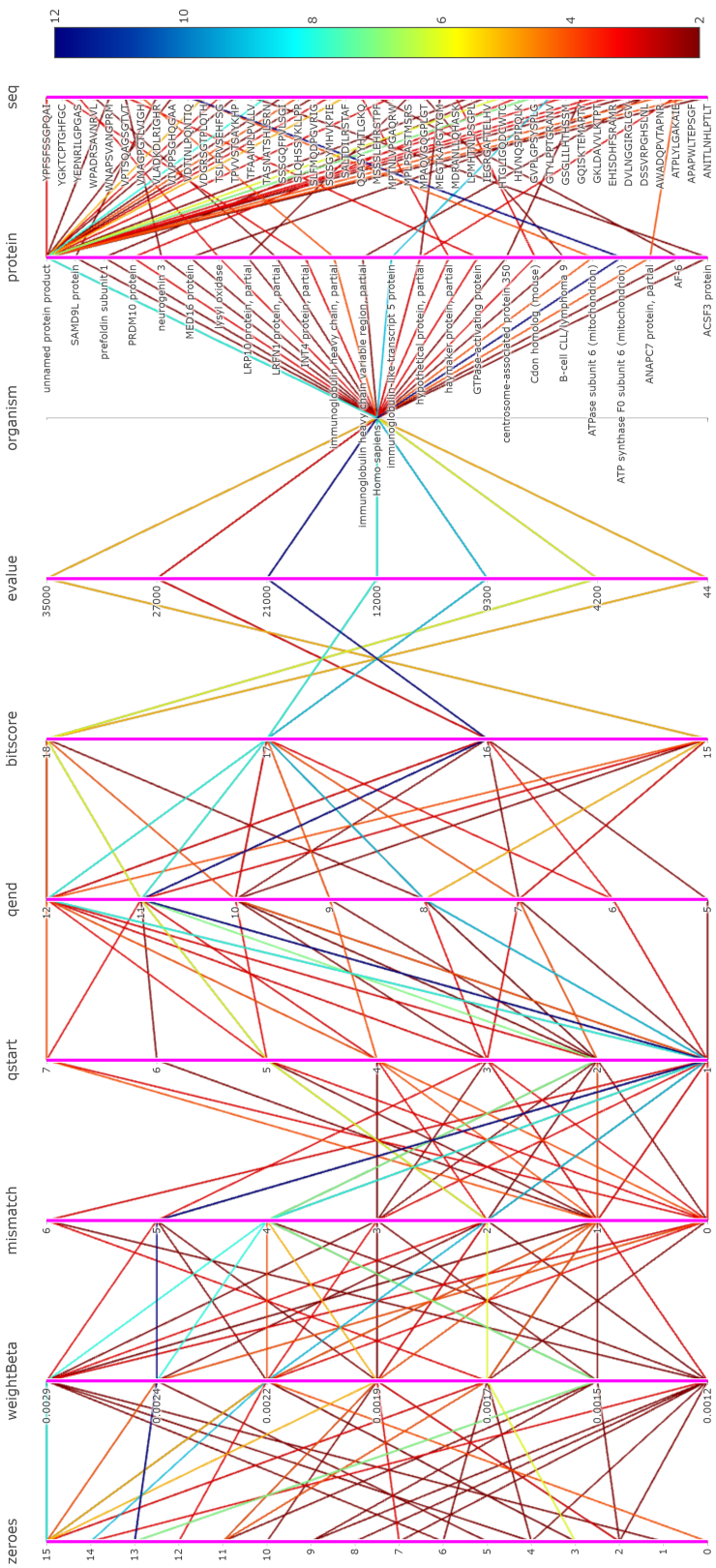


Figure 29: Parallel coordinates plot of 43 sequences matching human proteins with a bit score of 15 or better from the tabular output of the sequence aligner DIAMOND. Lines are color coded to indicate frequency of occurrence.

nation to the strong band observed in the influenza-vaccinated library 10p-I in figure 28. In figure 30, lines correspond to the top 500 column-value combinations based on occurrence after merging the tabular results from the three protein databases. Lines are color coded to indicate frequency of occurrence. This result suggests the presence of mimotopes that can bind a diverse selection of antibodies that have different antigenic targets. A heatmap of the five peptide sequences from figure 30 show read counts across multiple libraries except for STIAVVTYSGLS and GVVYDYSFLPKP (figure 31). From figure 30, these two sequences have matches to influenza A proteins but have zero reads in the influenza-vaccinated libraries (9p-I, 10p-I, 11p-I, R0-rnd-1, R1-rnd-1, and ctrl-rnd-1), which may be one or more of the following: failure to capture/amplify during library preparation/sequencing, and the peptides may have negligible affinity for the influenza serotype the antibodies were raised against in the influenza-vaccinated libraries; the same may be argued for DVLNGGIRGLGV. Sequences MSSSLEHRSTPF and LPMHTNLPSGPL may have affinities for antibodies against human and influenza proteins based on figure 31. In figure 31, rows and columns are reordered using hierarchical clustering via Ward's linkage method. There are also peptide sequences present that were not found to map to influenza sequences (figure 32). In figure 32, lines correspond to the top 50 column-value combinations based on occurrence after merging the tabular results from the three protein databases. Sequences that have no database matches are marked with NA [NA]. Among the 27 sequences, the peptide VAASPYYAPRVP only matched to a human pathogenic protein, specifically the TonB-dependent receptor of the gram-negative bacteria *Moraxella catarrhalis*. On a heatmap, we find that the nonresponder libraries (5p-I, 6p-I, 7p-I, and 8p-I) and the influenza-vaccinated library 10p-I are characterized by relatively high read counts (figure 33). In figure 33, rows and columns are reordered using hierarchical clustering via Ward's linkage method. It is intriguing to find some of the peptides with high read counts in 10p-I (TASNATSHLSRN, DNAGIRLPSYTL, NSLVQSCGILCS, AEAN-

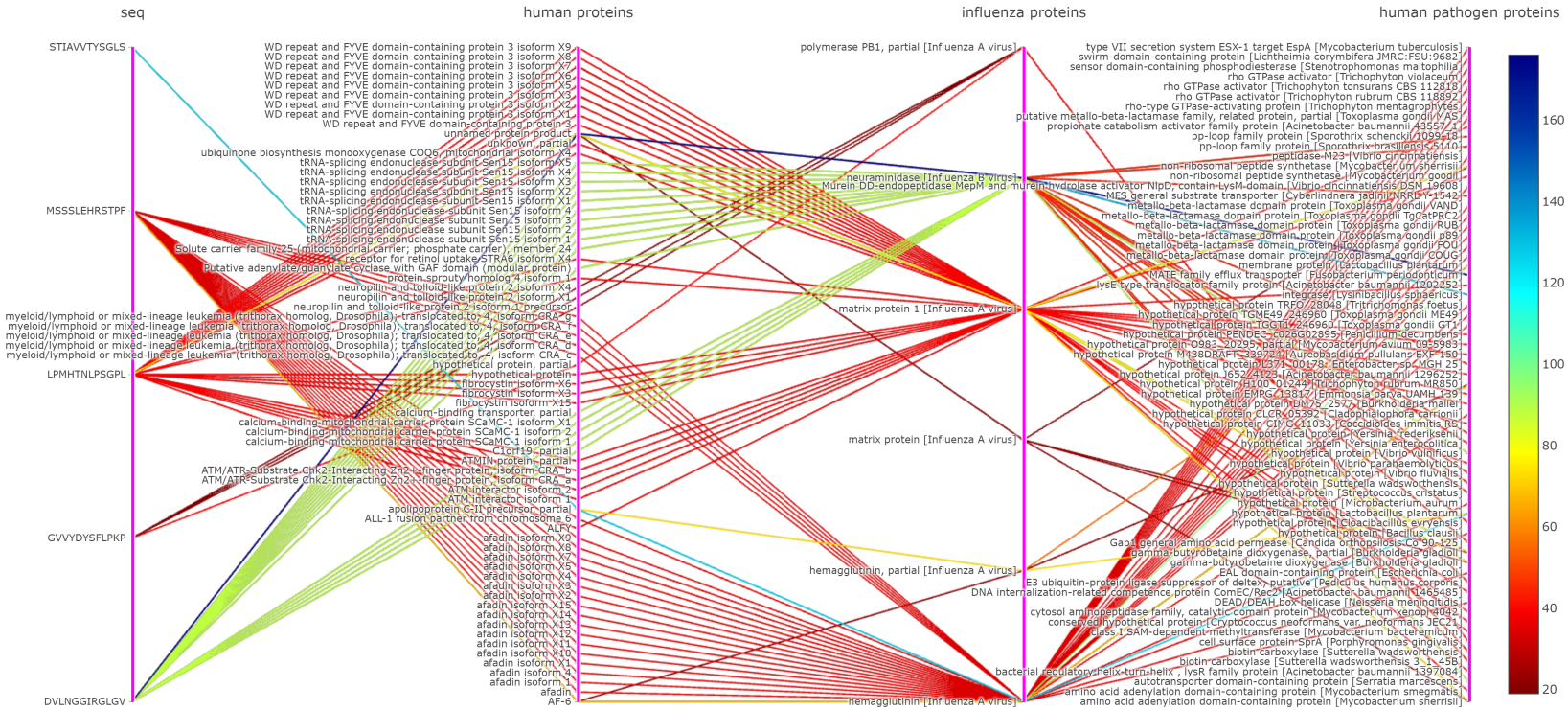


Figure 30: Parallel coordinates plot of five sequences matching to protein sequences in human, influenza, and human pathogens with bit scores of 15 or better from all three protein sequence databases as returned in the tabular output of the sequence aligner DIAMOND.

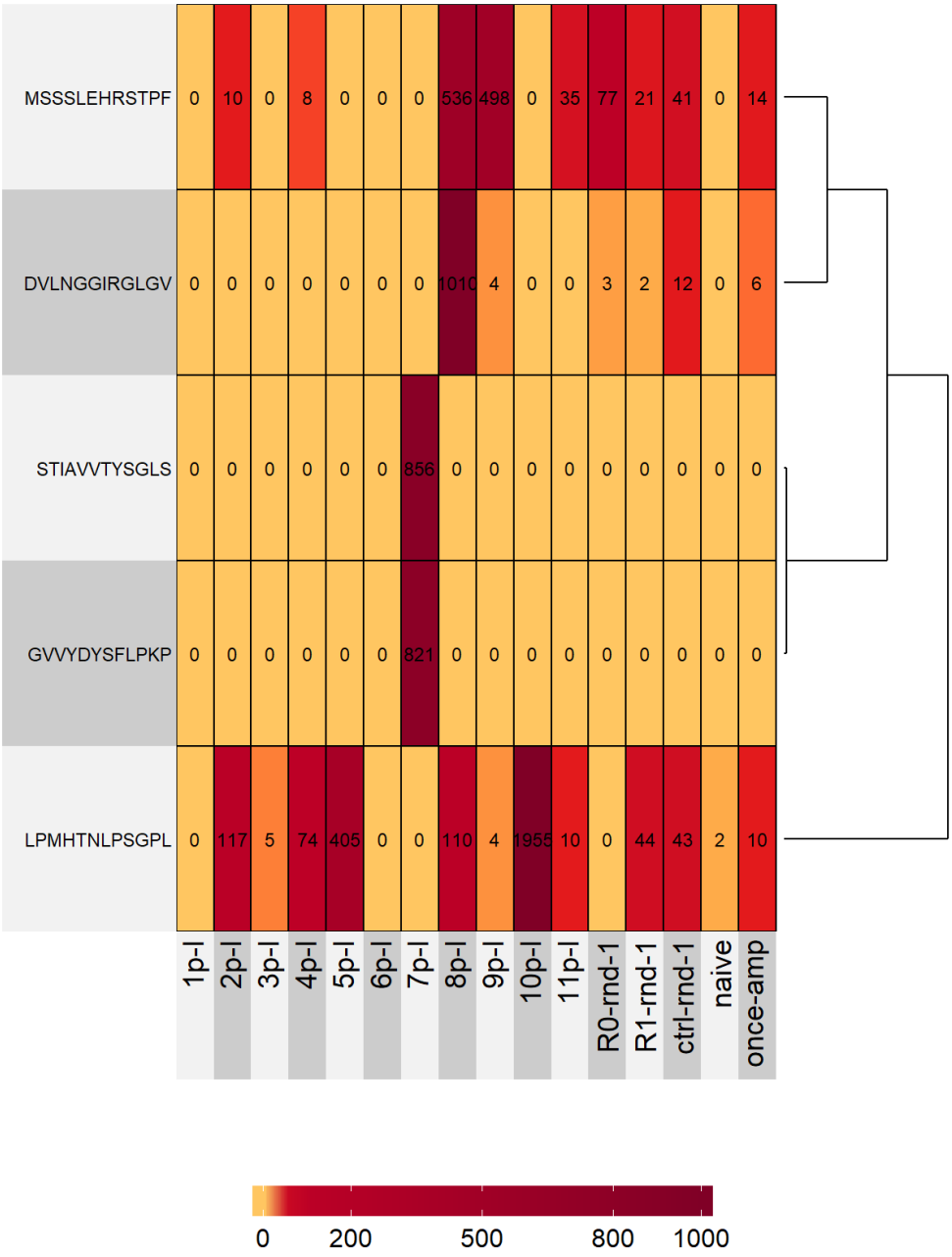


Figure 31: Heatmap of five peptide sequences matching human proteins derived across several donors that also match to nonhuman proteins. Color intensity is relative to the number of read counts.

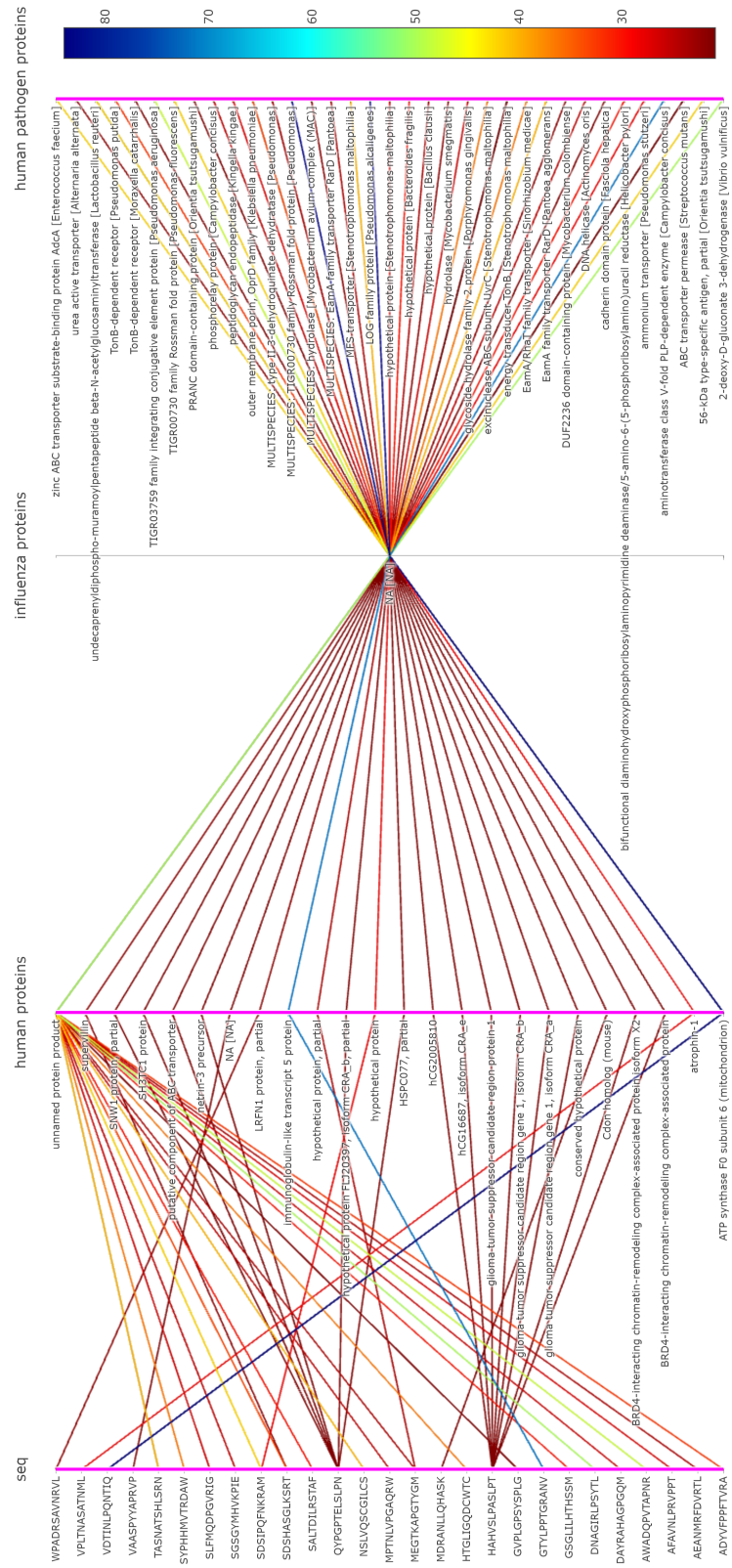


Figure 32: Parallel coordinates plot of 27 sequences matching to protein sequences in either human, influenza, and human pathogens but not in all three with bit scores of 15 or better from all three protein sequence databases as returned in the tabular output of the sequence aligner DIAMOND. Lines are color coded to indicate frequency of occurrence.

MRFDVRTL) mapping to an unnamed human protein products (figure 32), but given the interconnectedness to nonhuman protein sequences leads one to argue that these peptide sequences are more likely to be pathogenic in origin.

To determine which among the 233 peptide sequences can be traced to AML pathology, the GenBank accession number of the peptide sequences' human protein matches were converted to gene symbols and matched with the a combined list of 7402 AML-associated genes from GeneShot (1167 genes) and GeneCards (7311 genes). Out of the 233 sequences, 131 peptides matched to AML-associated genes. Some of these peptides are shown in figures 34 and 35, which belong to the subset of peptides having the most number of matches to influenza and other human pathogen proteins. In figure 35, lines correspond to the top 100 column-value combinations based on occurrence after merging the tabular results from the three protein databases for human, influenza, and human pathogens. Looking at their library read counts, we find preferential enrichment in either the AML libraries or in the influenza-vaccinated libraries. Further analysis would be required to validate whether the enriched peptides in the AML libraries evince pathological features of AML or a possibly traces of previous influenza vaccinations of the AML patients.

2.3.6.4 Enrichment of AML-associated peptides from a single individual

From the baseline sample of AML nonresponder donor 4, two libraries, 7p-I and 7-1-2017, were panned at different phage-to-antibody ratios (table 5). Also, a second panning round is available, 7-2-2017, which was panned from 7-1-2017. These were compared with first panning round libraries from influenza-vaccinated donors (9p-I, 10p-I, 11p-I, R0-rnd-1, R1-rnd-1, ctrl-rnd-1) and the two reference libraries (naïve and once-amp). Alignment with human protein sequences resulted to 1285 sequences above the threshold

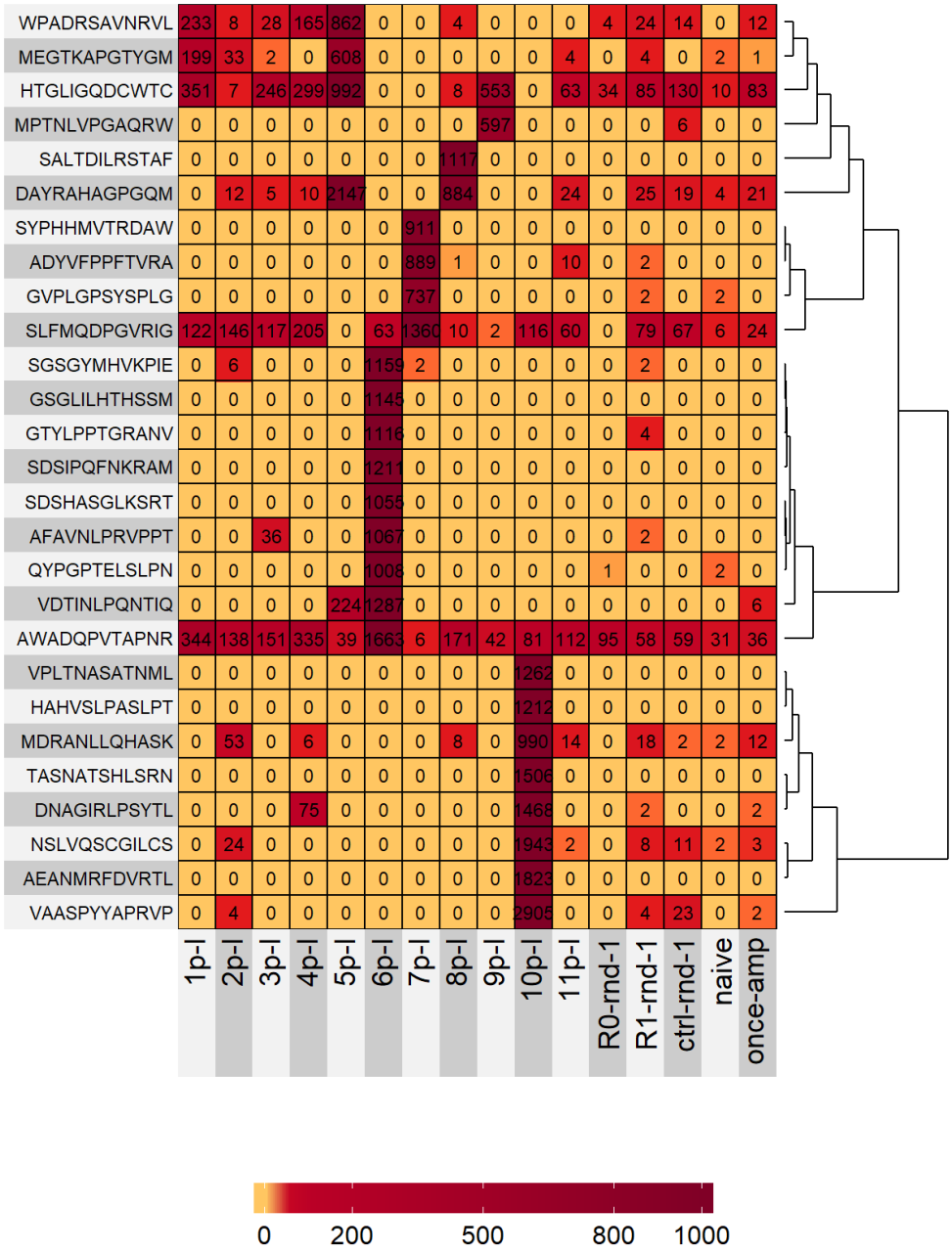


Figure 33: Heatmap of 27 peptide sequences matching to protein sequences in either human, influenza, and human pathogens but not in all three. Color intensity is relative to the number of read counts.

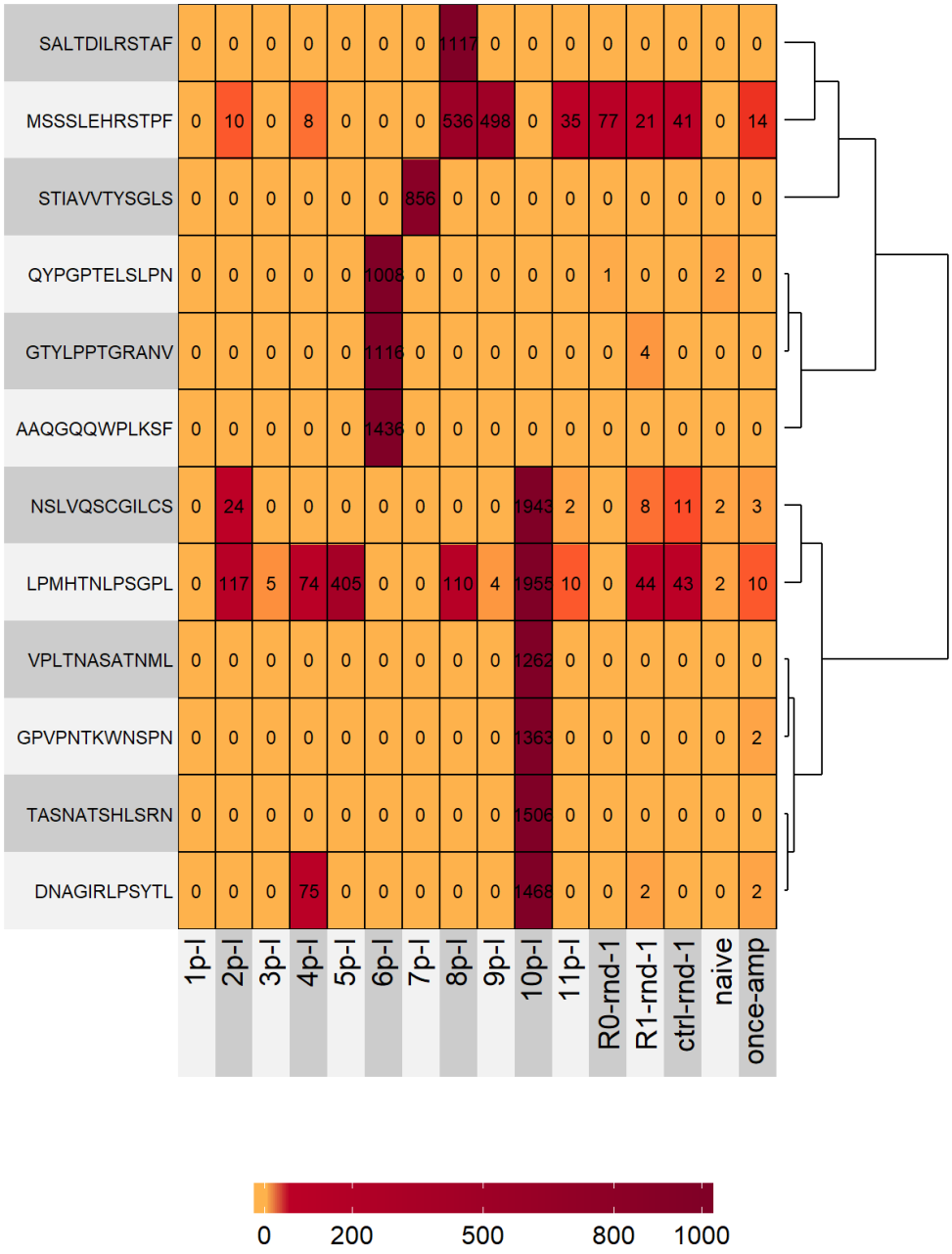


Figure 34: Heatmap of 12 peptide sequences matching to human protein sequences that are coded by genes associated with acute myeloid leukemia. Color intensity is relative to the number of read counts.

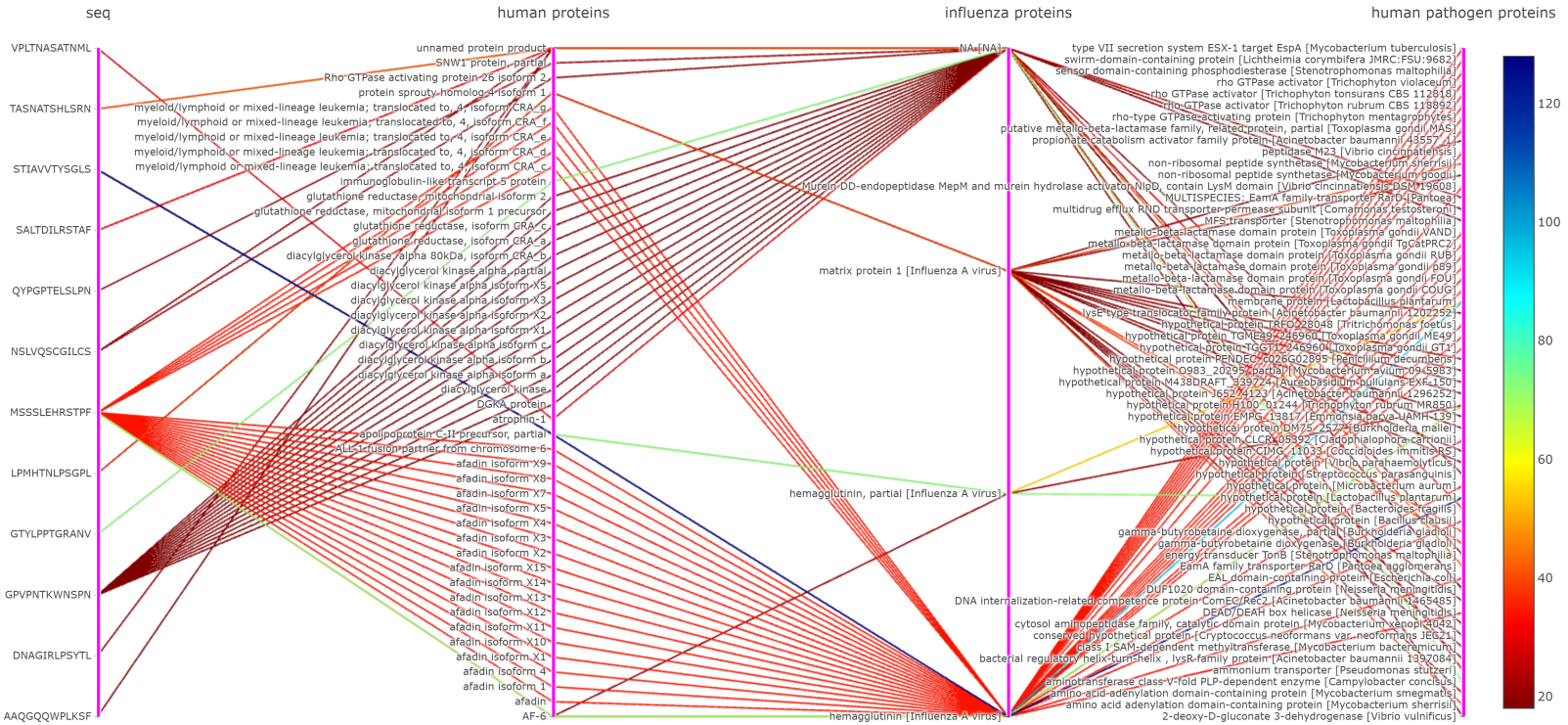


Figure 35: Parallel coordinates plot of 12 sequences matching to human protein sequences from the tabular output of the sequence aligner DIAMOND that are coded by genes associated with acute myeloid leukemia. Lines are color coded to indicate frequency of occurrence.

(figure 36), with most search hits having a bit score of around 16.5 (figure 37). In figure 36, each dot represents an outlier peptide sequence. Blue solid dots mark sequences that matched to human protein sequences passing the threshold marked by a blue dotted horizontal line. The threshold is based on the lowest upper-whisker extreme across adjusted box plots. Removal of search hits that map to organisms aside from human were removed resulting to 1281 sequences. These sequences map unto 2512 genes of which 931 are within the combined list of AML-associated genes from GeneShot and GeneCards.

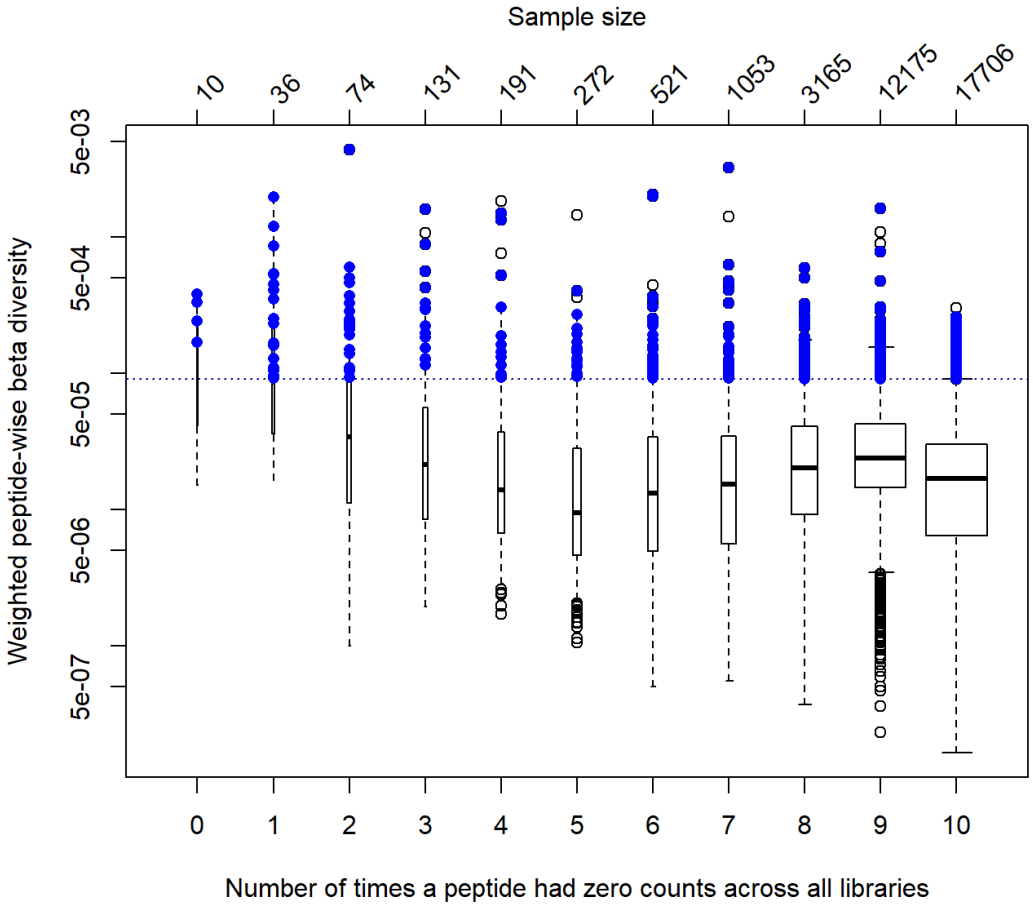


Figure 36: Relation between weighted peptide-wise beta diversity and the number of zero counts across three libraries derived from a single AML patient with the first-panning-round influenza-vaccinated libraries and the two reference libraries as controls.

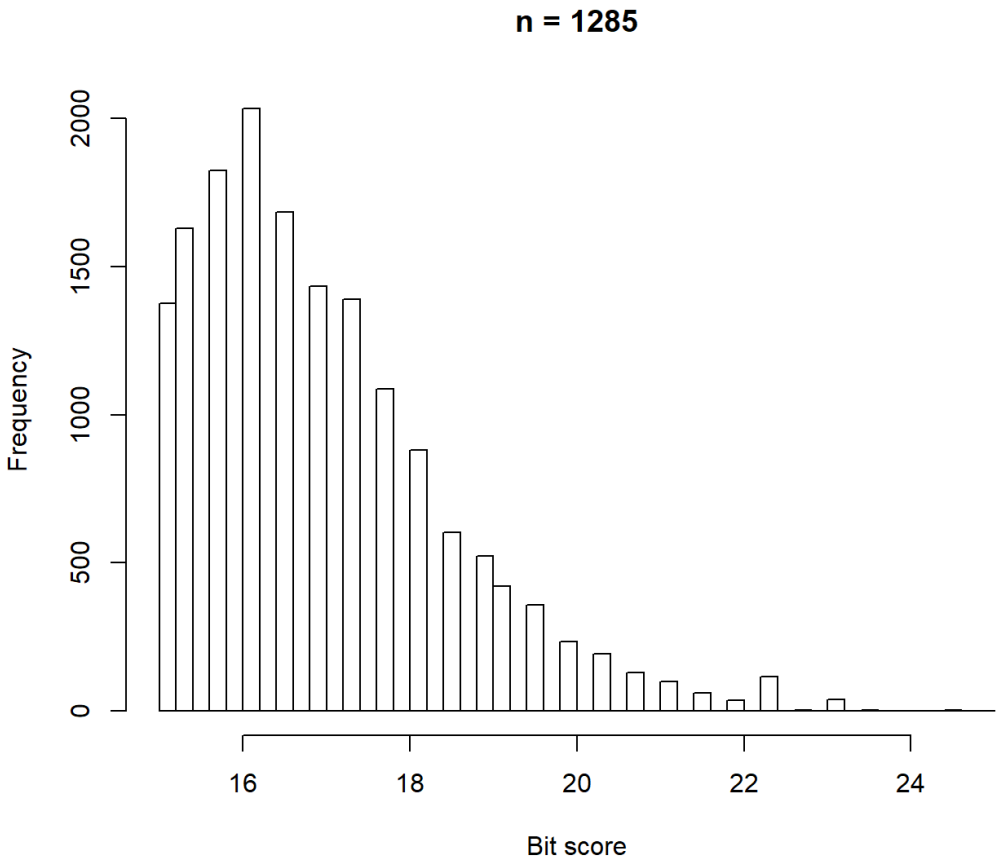


Figure 37: Bit score distribution of search hits matching to human proteins that scored 15 or better, with the number of sequences matching indicated on top.

A heatmap of the read counts of the peptide sequences from all the search hits reveal that there is very little overlap in terms of sequences common to 7p-I and 7-1-2015/7-2-2015 (figure 38). In figure 38, rows are reordered using hierarchical clustering using complete linkage method with Manhattan distances. Sequences sharing similar read count patterns are divided into three major clusters A, B, and C; sequences that do not belong to any of these clusters are grouped together into cluster D. Cluster A contains sequences enriched in libraries 7-1-2015/7-2-2015, cluster B from 10p-I, and cluster C from 7p-I. Most overlap between the AML libraries is found in cluster D. This pattern of mutual exclusion in peptide sequences panned from the same plasma sample but at different phage-to-antibody ratios was also seen with the influenza libraries (figure 17), reinforcing the observation that phage-to-antibody ratios strongly influences the peptides captured during panning.

To test whether the uniqueness in peptide sequences between clusters A and C collapses when mapped to their corresponding genes, the Genbank accession numbers of the sequences were converted to gene symbols when available. However, inspection of figure 39 suggests that there is little to no overlap between the genes corresponding to these two enriched-AML-library clusters. In figure 39, lines correspond to the top 50 column-value combinations based on occurrence after merging the tabular results for human proteins with their clustering results based on their read count pattern. Axes represent column headers returned by sequence alignment output merge with calculated diversity-related metrics: zeroes = number of libraries with zero read counts, weightBeta = weight beta diversity, mismatch = number of mismatches, qstart/qend = start/end of alignment in query, bitscore = bit score, evalue = expectation value, protein = human proteins, seq = peptide sequences, cluster = grouping based on read-count clustering, symbol = human gene symbol. Sequences that have no gene matches have their gene symbols marked with [NA]. Checking across all gene matches, we find only 40 out of 2512 or 1.6 % to be com-

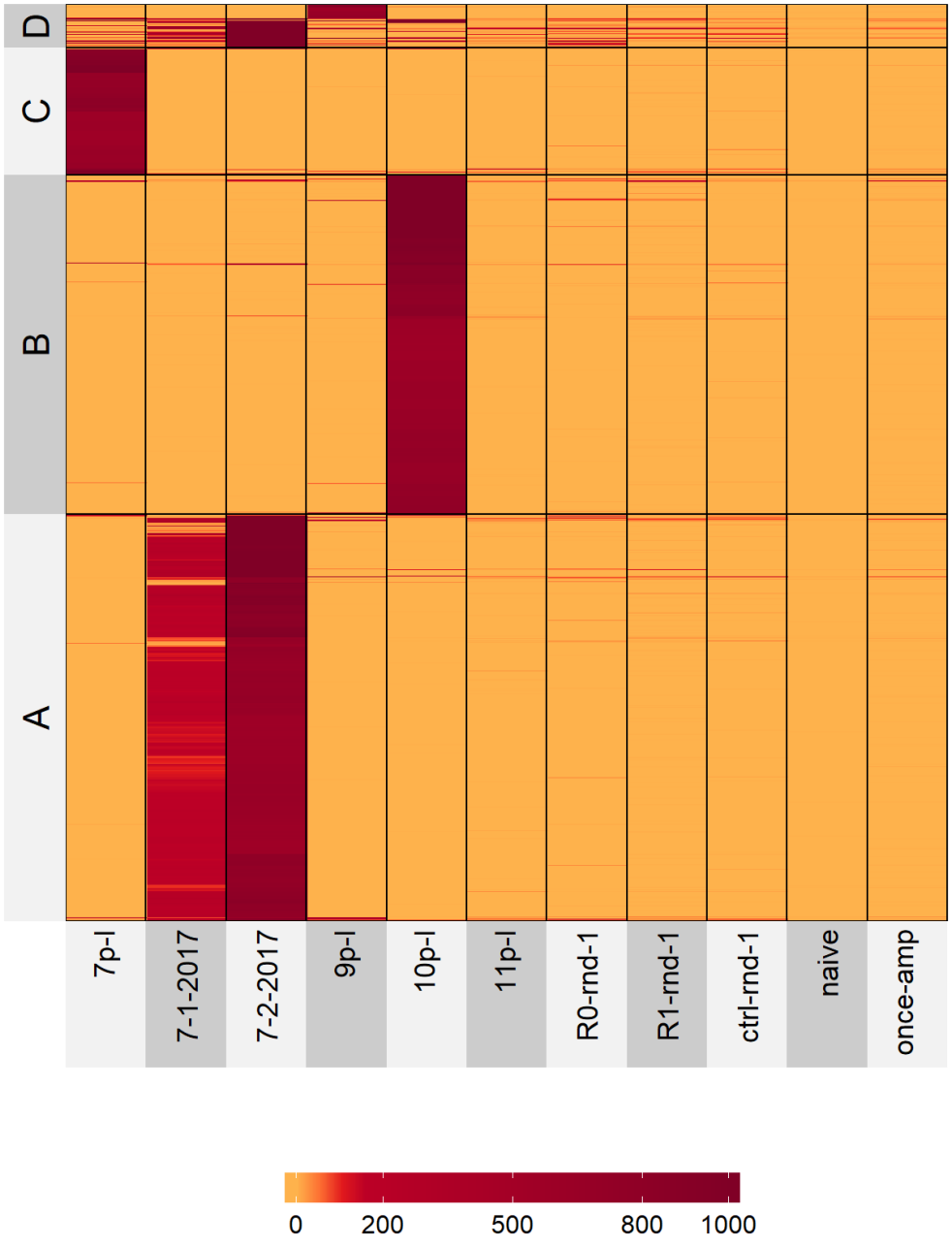


Figure 38: Heatmap of 1281 peptide sequences matching to protein sequences in either human, influenza, and human pathogens but not in all three. Color intensity is relative to the number of read counts. Rows represent peptide sequences while columns are libraries.

mon between these two clusters. Recalling that the change in phage-to-antibody ratio is achieved by diluting sera while keeping the amount of the phage library fixed, these results suggest that different peptides are probed by almost mutually exclusive sets of antibodies. Unlike for the case of influenza-vaccinated library R1-rnd-1 where the peptide distribution profile suggests nonspecificity of binding, the library with the higher phage-to-antibody ratio 7p-I is probably capturing a subset of peptides whose presence are otherwise masked by other peptides found in the libraries with a lower phage-to-antibody ratio 7-1-2017 and 7-2-2017, compounded further by limitations in sequencing coverage. It is argued that a greater sequencing coverage (higher number of reads per sample) would result in a higher overlap in the peptide sequences present between these donor 4-derived AML libraries.

2.4 Conclusions and future directions

We have demonstrated in this study the feasibility of enriching for peptides associated with AML from a single round of panning. Several processing steps were undertaken in order to mitigate the presence of sequencing artifacts in the analysis. Most libraries were found to be enriched relative to the naive library. However, one library, R1-rnd-1, was found to have a distributional profile akin to the once-amp library suggesting the possibility of nonspecific binding at high dilutions of the plasma samples. Relative to this, we find that the phage-to-antibody ratio strongly influences the peptides panned from donor samples. Peptide sequences with low read counts make up the vast majority of the peptide sequences across libraries, which was filtered out to reveal diversity in the sequences with higher read counts. Borrowing concepts from mathematical ecology, an observation was made on the linear trend with respect to calculated Shannon and Simpson diversity indices across the first-panning-round libraries, which may be a characteristic of libraries

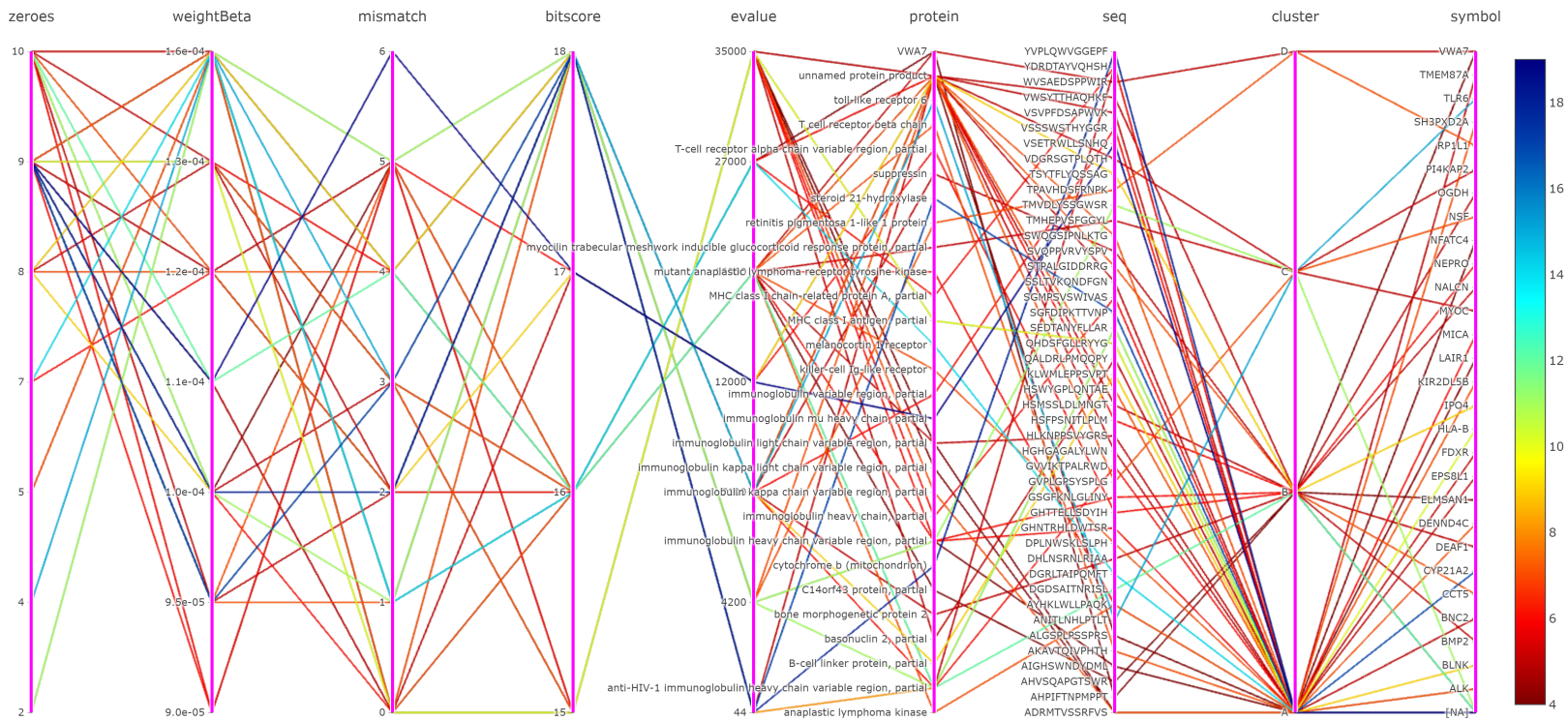


Figure 39: Parallel coordinates plot of 46 sequences matching to human protein sequences from the tabular output of the sequence aligner DIAMOND that are coded by genes associated with acute myeloid leukemia. Lines are color coded to indicate frequency of occurrence.

derived from the same source library. Similarly, the concept of beta diversity was utilized in uncovering candidate peptide sequences that are either target-unrelated peptides or enriched in a particular cohort. TUPs were selected on the basis of using an unweighted beta diversity metric across all the libraries. On the other hand, enrichment was carried out using a weighted version of the beta diversity metric together with the contrasting of test and control cases. This approach permitted the simultaneous examination of several libraries, deriving candidate peptide sequences by incorporating information from all the libraries comprising the subset under consideration to determine enrichment. Combined with sequence alignment to further refine the candidate peptide selection, this enrichment procedure yielded peptide sequences that map to known AML-associated proteins. The discovery of a peptide that mapped to ANAPC7 which was only recently associated with AML suggests the potential of this approach for protein discovery. Likewise, other candidate sequences that did not match to known AML-associated genes may be useful as biomarkers for diagnosis, prognosis, treatment-response prediction, or targeted therapy in AML patients undergoing nivolumab/azacytidine treatment.

The results presented here are preliminary and serve to demonstrate proof of concept. Future work would involved performing additional rounds of panning on all the panned libraries to validate patterns of enrichment seen in some of the library samples that had two rounds of panning. This study involved only a small cohort of patients and thus larger cohorts are needed to improve confidence in the results. Candidate peptide sequences require validation with an immunoassay to confirm the presence of antibodies that recognize these sequences. Given that majority of peptide sequences were discarded due to low read counts, it would be desirable to incorporate read counts from these sequences in the analysis by combining sequences based on similarity. Since molecular recognition of epitopes are mostly from a few peptide residues with the rest acting as scaffold, the idea of combining read counts from peptides that are recognized by the same antibody paratope

would be interesting to pursue. The computational framework laid out by Caoili (2016) looks promising but lacks a computational pipeline for implementation on the NGS data. Thus, the development of such a pipeline would be another area for research. In addition, the performance of the biodiversity metric used in this study has not been evaluated for its accuracy so far. Aside from the Ph.D.-12 library which is based on linear peptides displayed on filamentous phage, the consequences of using other phage display library formats where the shape of the phage, the shape and length of the peptide displayed, and where these peptides are displayed on the surface of the phage have not yet been explored with respect to peptide enrichment. This work can also be extended to other diseases like allergies and autoimmunity that are capable of generating immunological signatures.

References

Alexandrov, Ludmil B, Serena Nik-Zainal, David C Wedge, Samuel AJR Aparicio, Sam Behjati, Andrew V Biankin, Graham R Bignell, Niccolo Bolli, Ake Borg, and Anne-Lise Børresen-Dale (2013). "Signatures of mutational processes in human cancer." In: *Nature* 500.7463, p. 415.

An, Xingyue, Victor G Sendra, Ivan Liadi, Balakrishnan Ramesh, Gabrielle Romain, Cara Haymaker, Melisa Martinez-Paniagua, Yanbin Lu, Laszlo G Radvanyi, and Badrinath Roysam (2017). "Single-cell profiling of dynamic cytokine secretion and the phenotype of immune cells." In: *PloS one* 12.8, e0181904.

Analytics, Revolution and Steve Weston (2018). *iterators: Provides Iterator Construct for R*. R package version 1.0.10.

Anderton, Stephen M (2004). "Post-translational modifications of self antigens: implications for autoimmunity." In: *Current opinion in immunology* 16.6, pp. 753–758.

Angeletti, Davide, James S Gibbs, Matthew Angel, Ivan Kosik, Heather D Hickman, Gregory M Frank, Suman R Das, Adam K Wheatley, Madhu Prabhakaran, and David J Leggat (2017). "Defining B cell immunodominance to viruses." In: *Nature immunology* 18.4, p. 456.

Arora, Sonali, Martin Morgan, Marc Carlson, and H. Pagès (2018). *GenomeInfoDb: Utilities for manipulating chromosome and other 'seqname' identifiers*. R package version 1.18.1.

Augello, Catherine J, Jessica M Noll, Timothy J Distel, Jolita D Wainright, Charles E Stout, and Byron D Ford (2018). "Identification of novel blood biomarker panels to detect ischemic stroke in patients and their responsiveness to therapeutic intervention." In: *Brain research* 1698, pp. 161–169.

Bache, Stefan Milton and Hadley Wickham (2014). *magrittr: A Forward-Pipe Operator for R*. R package version 1.5.

Bai, Ju, Aili He, Wanggang Zhang, Chen Huang, Juan Yang, Yun Yang, Jianli Wang, and Yang Zhang (2013). "Potential biomarkers for adult acute myeloid leukemia minimal residual disease assessment searched by serum peptidome profiling." In: *Proteome science* 11.1, p. 39.

Bailey, R Clifton, GS Eadie, and Frederick H Schmidt (1974). "Estimation procedures for consecutive first order irreversible reactions." In: *Biometrics*, pp. 67–75.

Bakhshinejad, Babak, Hesam Motaleb Zade, Hosna Sadat Zahed Shekarabi, and Sara Neman (2016). "Phage display biopanning and isolation of target-unrelated peptides: in search of nonspecific binders hidden in a combinatorial library." In: *Amino acids* 48.12, pp. 2699–2716.

Barlow, DJ, MS Edwards, and JM Thornton (1986). "Continuous and discontinuous protein antigenic determinants." In: *Nature* 322.6081, p. 747.

Baron, Samuel and Gary R Klimpel (1996). "Immune Defenses." In: *Medical Microbiology*. 4th. University of Texas Medical Branch at Galveston.

Bazan, Justyna, Ireneusz Cakosiski, and Andrzej Gamian (2012). "Phage displayA powerful technique for immunotherapy: 1. Introduction and potential of therapeutic applications." In: *Human vaccines & immunotherapeutics* 8.12, pp. 1817–1828.

Bei, R, L Masuelli, C Palumbo, M Modesti, and A Modesti (2009). "A common repertoire of autoantibodies is shared by cancer and autoimmune disease patients: Inflammation in their induction and impact on tumor growth." In: *Cancer letters* 281.1, pp. 8–23.

Bengtsson, Henrik (2018). *matrixStats: Functions that Apply to Rows and Columns of Matrices (and to Vectors)*. R package version 0.54.0.

Berglund, Lisa, Jorge Andrade, Jacob Odeberg, and Mathias Uhlén (2008). "The epitope space of the human proteome." In: *Protein Science* 17.4, pp. 606–613.

Beygelzimer, Alina, Sham Kakadet, John Langford, Sunil Arya, David Mount, and Shengqiao Li (2019). *FNN: Fast Nearest Neighbor Search Algorithms and Applications*. R package version 1.1.3.

Bodenhofer, Ulrich, Enrico Bonatesta, Christoph Horejs-Kainrath, and Sepp Hochreiter (2015). "msa: an R package for multiple sequence alignment." In: *Bioinformatics* 31.24, pp. 3997–3999. DOI: 10.1093/bioinformatics/btv494.

Bonnycastle, LLC, JS Mehroke, M Rashed, X Gong, and JK Scott (1996). "Probing the basis of antibody reactivity with a panel of constrained peptide libraries displayed by filamentous phage." In: *Journal of molecular biology* 258.5, pp. 747–762.

Borchers, Hans W. (2018). *pracma: Practical Numerical Math Functions*. R package version 2.2.2.

Borrebaeck, Carl AK (2017). "Precision diagnostics: moving towards protein biomarker signatures of clinical utility in cancer." In: *Nature Reviews Cancer* 17.3, p. 199.

Brinton, Lindsey T, Dustin K Bauknight, Siva Sai Krishna Dasa, and Kimberly A Kelly (2016). "PHASTpep: analysis software for discovery of cell-selective peptides via phage display and next-generation sequencing." In: *PloS one* 11.5, e0155244.

Brown, Christopher (2012). *dummies: Create dummy/indicator variables flexibly and efficiently*. R package version 1.5.6.

Buchfink, Benjamin, Chao Xie, and Daniel H Huson (2015). "Fast and sensitive protein alignment using DIAMOND." In: *Nature methods* 12.1, p. 59.

Budhu, Sadna, John D Loike, Ashley Pandolfi, Soo Han, Geoffrey Catalano, Andrei Constantinescu, Raphael Clynes, and Samuel C Silverstein (2010). "CD8+ T cell concentration determines their efficiency in killing cognate antigen–expressing

syngeneic mammalian cells in vitro and in mouse tissues." In: *Journal of Experimental Medicine* 207.1, pp. 223–235.

Caoili, Salvador Eugenio C (2016). "Expressing Redundancy among Linear-Epitope Sequence Data Based on Residue-Level Physicochemical Similarity in the Context of Antigenic Cross-Reaction." In: *Advances in bioinformatics* 2016.

Carlson, Marc (2018). *org.Hs.eg.db: Genome wide annotation for Human*. R package version 3.7.0.

Charoentong, Pornpimol, Mihaela Angelova, Mirjana Efremova, Ralf Gallasch, Hubert Hackl, Jerome Galon, and Zlatko Trajanoski (2012). "Bioinformatics for cancer immunology and immunotherapy." In: *Cancer Immunology, Immunotherapy* 61.11, pp. 1885–1903.

Chen, Hongli, Tian Liu, Jing Liu, Yuandong Feng, Baiyan Wang, Jianli Wang, Ju Bai, Wanhong Zhao, Ying Shen, and Xiaman Wang (2018). "Circ-ANAPC7 is upregulated in acute myeloid leukemia and appears to target the MiR-181 family." In: *Cellular Physiology and Biochemistry* 47.5, pp. 1998–2007.

Chen, Wen-Lian, Jing-Han Wang, Ai-Hua Zhao, Xin Xu, Yi-Huang Wang, Tian-Lu Chen, Jun-Min Li, Jian-Qing Mi, Yong-Mei Zhu, and Yuan-Fang Liu (2014). "A distinct glucose metabolism signature of acute myeloid leukemia with prognostic value." In: *Blood* 124.10, pp. 1645–1654.

Christiansen, Anders, Jens V Kringelum, Christian S Hansen, Katrine L Bøgh, Eric Sullivan, Jigar Patel, Neil M Rigby, Thomas Eiwegger, Zsolt Szépfalusi, Federico De Masi, Morten Nielsen, Ole Lund, and Martin Dufva (2015). "High-throughput sequencing enhanced phage display enables the identification of patient-specific epitope motifs in serum." In: *Scientific reports* 5.

Corporation, Microsoft and Steve Weston (2018). *doParallel: Foreach Parallel Adaptor for the 'parallel' Package*. R package version 1.0.14.

Cortese, Riccardo, Franco Felici, Giovanni Galfre, Alessandra Luzzago, Paolo Monaci, and Alfredo Nicosia (1994). "Epitope discovery using peptide libraries displayed on phage." In: *Trends in biotechnology* 12.7, pp. 262–267.

Csárdi, Gábor and Winston Chang (2019). *processx: Execute and Control System Processes*. R package version 3.3.1.

Datasheet for Ph.D.™-12 Library, NEB: (n.d.). *Datasheet for Ph.D.™-12 Phage Display Peptide Library (E8111; Lot 0131408)*. New England Biolabs.

Daver, Naval Guastad, Sreyashi Basu, Guillermo Garcia-Manero, Jorge E Cortes, Farhad Ravandi, Elias Jabbour, Stephany Hendrickson, Mark Brandt, Sherry Pierce, and Tauna Gordon (2017). *Phase IB/II study of nivolumab with azacytidine (AZA) in patients (pts) with relapsed AML.*

Daver, Naval, Guillermo Garcia-Manero, Sreyashi Basu, Prajwal C Boddu, Mansour Alfayez, Jorge E Cortes, Marina Konopleva, Farhad Ravandi-Kashani, Elias Jabbour, and Tapan Kadia (2019). "Efficacy, Safety, and Biomarkers of Response to Azacitidine and Nivolumab in Relapsed/Refractory Acute Myeloid Leukemia: A Nonrandomized, Open-Label, Phase II Study." In: *Cancer discovery* 9.3, pp. 370–383.

De Kouchkovsky, I and M Abdul-Hay (2016). "Acute myeloid leukemia: a comprehensive review and 2016 update." In: *Blood cancer journal* 6.7, e441.

Delignette-Muller, Marie Laure and Christophe Dutang (2015). "fitdistrplus: An R Package for Fitting Distributions." In: *Journal of Statistical Software* 64.4, pp. 1–34.

Derda, Ratmir, Sindy Tang, S Cory Li, Simon Ng, Wadim Matochko, and Mohammad Jafari (2011). "Diversity of phage-displayed libraries of peptides during panning and amplification." In: *Molecules* 16.2, pp. 1776–1803.

Desai, Aarti N and Abhay Jere (2012). "Next-generation sequencing: ready for the clinics?" In: *Clinical genetics* 81.6, pp. 503–510.

Deyle, Kaycie, Xu-Dong Kong, and Christian Heinis (2017). "Phage selection of cyclic peptides for application in research and drug development." In: *Accounts of chemical research* 50.8, pp. 1866–1874.

Ding, Chuanlin and Jun Yan (2007). "Regulation of autoreactive B cells: checkpoints and activation." In: *Archivum immunologiae et therapiae experimentalis* 55.2, p. 83.

Dowle, Matt and Arun Srinivasan (2019). *data.table: Extension of 'data.frame'*. R package version 1.12.2.

Doyle, Hester A and Mark J Mamula (2001). "Post-translational protein modifications in antigen recognition and autoimmunity." In: *Trends in immunology* 22.8, pp. 443–449.

Floyd, Daniel L, Stephen C Harrison, and Antoine M Van Oijen (2010). "Analysis of kinetic intermediates in single-particle dwell-time distributions." In: *Biophysical journal* 99.2, pp. 360–366.

Fox, Edward J, Kate S Reid-Bayliss, Mary J Emond, and Lawrence A Loeb (2014). "Accuracy of next generation sequencing platforms." In: *Next generation, sequencing & applications* 1.

Fox, John and Sanford Weisberg (2019). *An R Companion to Applied Regression*. Third. Thousand Oaks CA: Sage.

Fox, John, Sanford Weisberg, and Brad Price (2018). *carData: Companion to Applied Regression Data Sets*. R package version 3.0-2.

Frietze, Kathryn M, Richard BS Roden, Ji-Hyun Lee, Yang Shi, David S Peabody, and Bryce Chackerian (2016). "Identification of anti-CA125 antibody responses in ovarian cancer patients by a novel deep sequence–coupled biopanning platform." In: *Cancer immunology research* 4.2, pp. 157–164.

Gadhamsetty, Saikrishna, Athanasius FM Marée, Joost B Beltman, and Rob J de Boer (2017). "A sigmoid functional response emerges when cytotoxic T lymphocytes start killing fresh target cells." In: *Biophysical journal* 112.6, pp. 1221–1235.

Gallo, Eugenio (2019). "A High-Throughput Platform for the Generation of Synthetic Ab Clones by Single-Strand Site-Directed Mutagenesis." In: *Molecular biotechnology* 61.6, pp. 410–420.

Gang, Donghyeok, Do Kim, and Hee-Sung Park (2018). "Cyclic peptides: Promising scaffolds for biopharmaceuticals." In: *Genes* 9.11, p. 557.

Ganusov, Vitaly V, Daniel L Barber, and Rob J De Boer (2011). "Killing of targets by CD8+ T cells in the mouse spleen follows the law of mass action." In: *PloS one* 6.1, e15959.

Genz, Alan, Frank Bretz, Tetsuhisa Miwa, Xuefei Mi, Friedrich Leisch, Fabian Scheipl, and Torsten Hothorn (2019). *mvtnorm: Multivariate Normal and t Distributions*. R package version 1.0-10.

Gershoni, Jonathan M, Anna Roitburd-Berman, Dror D Siman-Tov, Natalia Tarnovitski Freund, and Yael Weiss (2007). "Epitope Mapping." In: *BioDrugs* 21.3, pp. 145–156.

Goodison, Steve, Osamu Ogawa, Yoshiyuki Matsui, Takashi Kobayashi, Makito Miyake, Sayuri Ohnishi, Kiyohide Fujimoto, Yunfeng Dai, Yoshiko Shimizu, and Kazue Tsukikawa (2016). "A multiplex urinary immunoassay for bladder cancer detection: analysis of a Japanese cohort." In: *Journal of translational medicine* 14.1, p. 287.

Greer, Judith M, Christine Klinguer, Elisabeth Trifilieff, Raymond A Sobel, and Marjorie B Lees (1997). "Encephalitogenicity of murine, but not bovine, DM20 in SJL mice is due to a single amino acid difference in the immunodominant encephalitogenic epitope." In: *Neurochemical research* 22.4, pp. 541–547.

Halperin, Rebecca F, Phillip Stafford, Jack S Emery, Krupa Arun Navalkar, and Stephen Albert Johnston (2012). "GuiTope: an application for mapping random-sequence peptides to protein sequences." In: *BMC bioinformatics* 13.1, p. 1.

Hanash, Sam (2003). *Harnessing immunity for cancer marker discovery*.

He, Bifang, Guoshi Chai, Yaocong Duan, Zhiqiang Yan, Liuyang Qiu, Huixiong Zhang, Zechun Liu, Qiang He, Ke Han, and Beibei Ru (2015). "BDB: biopanning data bank." In: *Nucleic acids research* 44.D1, pp. D1127–D1132.

Henry, Kevin A, Mehdi Arbabi-Ghahroudi, and Jamie K Scott (2015). "Beyond phage display: non-traditional applications of the filamentous bacteriophage as a vaccine carrier, therapeutic biologic, and bioconjugation scaffold." In: *Frontiers in microbiology* 6, p. 755.

Hill, Jonathon T, Bradley L Demarest, Brent W Bisgrove, Yi-chu Su, Megan Smith, and H. Joseph Yost (2014). "Poly peak parser: Method and software for identification of unknown indels using sanger sequencing of polymerase chain reaction products." In: *Developmental Dynamics*. DOI: 10.1002/dvdy.24183..

Hobbs, Zack and Stephen T Abedon (2016). "Diversity of phage infection types and associated terminology: the problem with Lytic or lysogenic." In: *FEMS microbiology letters* 363.7.

Huber, W., V. J. Carey, R. Gentleman, S. Anders, M. Carlson, B. S. Carvalho, H. C. Bravo, S. Davis, L. Gatto, T. Girke, R. Gottardo, F. Hahne, K. D. Hansen, R. A. Irizarry, M. Lawrence, M. I. Love, J. MacDonald, V. Obenchain, A. K. Ole's, H. Pag'es, A. Reyes, P. Shannon, G. K. Smyth, D. Tenenbaum, L. Waldron, and M. Morgan (2015). "Orchestrating high-throughput genomic analysis with Bioconductor." In: *Nature Methods* 12.2, pp. 115–121.

Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., Bravo, H. C., Davis, S., Gatto, L., Girke, T., Gottardo, R., Hahne, F., Hansen, K. D., Irizarry,

R. A., Lawrence, M., Love, M. I., MacDonald, J., Obenchain, V., Ole's, A. K., Pag'es, H., Reyes, A., Shannon, P., Smyth, G. K., Tenenbaum, D., Waldron, L., Morgan, and M. (2015). "Orchestrating high-throughput genomic analysis with Bioconductor." In: *Nature Methods* 12.2, pp. 115–121.

Ikemoto, Hideki, Prakash Lingasamy, Anne-Mari Anton Willmore, Hedi Hunt, Kaarel Kurm, Olav Tammik, Pablo Scodeller, Lorena Simón-Gracia, Venkata Ramana Kotamraju, and Andrew M Lowy (2017). "Hyaluronan-binding peptide for targeting peritoneal carcinomatosis." In: *Tumor Biology* 39.5, p. 1010428317701628.

Irvine, Katharine M, Leesa F Wockner, Isabell Hoffmann, Leigh U Horsfall, Kevin J Fagan, Veonice Bijin, Bernett Lee, Andrew D Clouston, Guy Lampe, and John E Connolly (2016). "Multiplex serum protein analysis identifies novel biomarkers of advanced fibrosis in patients with chronic liver disease with the potential to improve diagnostic accuracy of established biomarkers." In: *PloS one* 11.11, e0167001.

Janeway Jr, Charles A, Paul Travers, Mark Walport, and Mark J Shlomchik (2001). "The structure of a typical antibody molecule." In: *Immunobiology: The Immune System in Health and Disease. 5th edition*. Garland Science.

Jordana-Lluch, Elena, Heather E Carolan, Montserrat Giménez, Rangarajan Sampath, David J Ecker, M Dolores Quesada, Josep M Mòdol, Fernando Arméstar, Lawrence B Blyn, and Lendell L Cummins (2013). "Rapid diagnosis of bloodstream infections with PCR followed by mass spectrometry." In: *PloS one* 8.4, e62108.

Khalid, Ayesha, Amna Jabbar Siddiqui, Jian-Hua Huang, Tahir Shamsi, and Syed Ghulam Musharraf (2018). "Alteration of Serum Free Fatty Acids are Indicators for Progression of Pre-leukaemia Diseases to Leukaemia." In: *Scientific reports* 8.1, p. 14883.

Kim, Hye-Jung, Bert Verbinnen, Xiaolei Tang, Linrong Lu, and Harvey Cantor (2010). "Inhibition of follicular T-helper cells by CD8+ regulatory T cells is essential for self tolerance." In: *Nature* 467.7313, p. 328.

Kostopoulou, Olga, Brendan C Delaney, and Craig W Munro (2008). "Diagnostic difficulty and error in primary carea systematic review." In: *Family practice* 25.6, pp. 400–413.

Krumpe, Lauren RH, Andrew J Atkinson, Gary W Smythers, Andrea Kandel, Kathryn M Schumacher, James B McMahon, Lee Makowski, and Toshiyuki Mori (2006). "T7 lytic phage-displayed peptide libraries exhibit less sequence bias than M13 filamentous phage-displayed peptide libraries." In: *Proteomics* 6.15, pp. 4210–4222.

Lachmann, Alexander, Brian M Schilder, Megan L Wojciechowicz, Denis Torre, Maxim V Kuleshov, Alexandra B Keenan, and Avi Maayan (2019). "Geneshot: search engine for ranking genes from arbitrary text queries." In: *Nucleic acids research*.

Lahens, Nicholas F, Emanuela Ricciotti, Olga Smirnova, Erik Toorens, Eun Ji Kim, Giacomo Baruzzo, Katharina E Hayer, Tapan Ganguly, Jonathan Schug, and Gregory R Grant (2017). "A comparison of Illumina and Ion Torrent sequencing platforms in the context of differential gene expression." In: *BMC genomics* 18.1, p. 602.

Larman, H Benjamin, Uri Laserson, Luis Querol, Katrijn Verhaeghen, Nicole L Solimini, George Jing Xu, Paul L Klarenbeek, George M Church, David A Hafler, and Robert M Plenge (2013). "PhIP-Seq characterization of autoantibodies from patients with multiple sclerosis, type 1 diabetes and rheumatoid arthritis." In: *Journal of autoimmunity* 43, pp. 1–9.

Larman, H Benjamin, Zhenming Zhao, Uri Laserson, Mamie Z Li, Alberto Ciccia, M Angelica Martinez Gakidis, George M Church, Santosh Kesari, Emily M LeProust,

and Nicole L Solimini (2011). “Autoantigen discovery with a synthetic human peptidome.” In: *Nature biotechnology* 29.6, p. 535.

Lawrence, Michael, Wolfgang Huber, Hervé Pagès, Patrick Aboyoun, Marc Carlson, Robert Gentleman, Martin Morgan, and Vincent Carey (2013a). “Software for Computing and Annotating Genomic Ranges.” In: *PLoS Computational Biology* 9 (8). DOI: 10.1371/journal.pcbi.1003118.

Lawrence, Michael, Wolfgang Huber, Hervé Pagès, Patrick Aboyoun, Marc Carlson, Robert Gentleman, Martin Morgan, and Vincent Carey (2013b). “Software for Computing and Annotating Genomic Ranges.” In: *PLoS Computational Biology* 9 (8). DOI: 10.1371/journal.pcbi.1003118.

Lawrence, Michael, Wolfgang Huber, Hervé Pagès, Patrick Aboyoun, Marc Carlson, Robert Gentleman, Martin Morgan, and Vincent Carey (2013c). “Software for Computing and Annotating Genomic Ranges.” In: *PLoS Computational Biology* 9 (8). DOI: 10.1371/journal.pcbi.1003118.

Levin, AM and GA Weiss (2006). “Optimizing the affinity and specificity of proteins with molecular display.” In: *Molecular Biosystems* 2.1, pp. 49–57.

Li, Quan-Zhen, David R Karp, Jiexia Quan, Valerie K Branch, Jinchun Zhou, Yun Lian, Benjamin F Chong, Edward K Wakeland, and Nancy J Olsen (2011). “Risk factors for ANA positivity in healthy persons.” In: *Arthritis research & therapy* 13.2, R38.

Liadi, Ivan, Jason Roszik, Gabrielle Romain, Laurence JN Cooper, and Navin Varadarajan (2013). “Quantitative high-throughput single-cell cytotoxicity assay for T cells.” In: *JoVE (Journal of Visualized Experiments)* 72, e50058.

Liadi, Ivan, Harjeet Singh, Gabrielle Romain, Badrinath Roysam, Laurence JN Cooper, and Navin Varadarajan (2018). “Defining potency of CAR+ T cells: Fast and furious or slow and steady.” In: *OncoImmunology*, e1051298.

Loh, Richard KS, Sandra Vale, and Andrew McLean-Tooke (2013). "Quantitative serum immunoglobulin tests." In: *Australian family physician* 42.4, p. 195.

Maechler, Martin, Peter Rousseeuw, Christophe Croux, Valentin Todorov, Andreas Ruckstuhl, Matias Salibian-Barrera, Tobias Verbeke, Manuel Koller, Eduardo L. T. Conceicao, and Maria Anna di Palma (2019). *robustbase: Basic Robust Statistics*. R package version 0.93-5.

Manara, Elena, Emma Baron, Claudia Tregnago, Sanja Aveic, Valeria Bisio, Silvia Bresolin, Riccardo Masetti, Franco Locatelli, Giuseppe Basso, and Martina Pigazzi (2014). "MLL-AF6 fusion oncogene sequesters AF6 into the nucleus to trigger RAS activation in myeloid leukemia." In: *Blood* 124.2, pp. 263–272.

Matloff, Norm (2016). *regtools: Regression Tools*. R package version 1.0.1.

Matloff, Norm and Yingkang Xie (2016). *freqparcoord: Novel Methods for Parallel Coordinates*. R package version 1.0.1.

Matloff, Norm, Vincent Yang, and Harrison Nguyen (2017). *cdparcoord: Top Frequency-Based Parallel Coordinates*. R package version 1.0.0.

Matloff, Norman (2016). "Software Alchemy: Turning Complex Statistical Computations into Embarrassingly-Parallel Ones." In: *Journal of Statistical Software* 71.4, pp. 1–15. DOI: 10.18637/jss.v071.i04.

Matochko, Wadim L, Kiki Chu, Bingjie Jin, Sam W Lee, George M Whitesides, and Ratmir Derda (2012). "Deep sequencing analysis of phage libraries using Illumina platform." In: *Methods* 58.1, pp. 47–55.

Matochko, Wadim L and Ratmir Derda (2013). "Error analysis of deep sequencing of phage libraries: peptides censored in sequencing." In: *Computational and mathematical methods in medicine* 2013.

Matochko, Wadim L and Ratmir Derda (2015). "Next-Generation Sequencing of Phage-Displayed Peptide Libraries." In: *Peptide Libraries: Methods and Protocols*, pp. 249–266.

Matochko, Wadim L, S Cory Li, Sindy KY Tang, and Ratmir Derda (2014). "Prospective identification of parasitic sequences in phage display screens." In: *Nucleic acids research* 42.3, pp. 1784–1798.

Meloen, RH, WC Puijk, and JW Slootstra (2000). "Mimotopes: realization of an unlikely concept." In: *Journal of molecular recognition* 13.6, pp. 352–359.

Merouane, Amine, Nicolas Rey-Villamizar, Yanbin Lu, Ivan Liadi, Gabrielle Romain, Jennifer Lu, Harjeet Singh, Laurence JN Cooper, Navin Varadarajan, and Badrinath Roysam (2015). "Automated profiling of individual cell–cell interactions from high-throughput time-lapse imaging microscopy in nanowell grids (TIMING)." In: *Bioinformatics* 31.19, pp. 3189–3197.

Microsoft and Steve Weston (2017). *foreach: Provides Foreach Looping Construct for R*. R package version 1.4.4.

Minka, Thomas P. (2002). *Estimating a Gamma distribution*.

Mohammadi, Reza (2018-07-5). *bmixture: Bayesian Estimation for Finite Mixture of Distributions*. R package version 1.2.

Morgan, Martin, Simon Anders, Michael Lawrence, Patrick Aboyoun, Hervé Pagès, and Robert Gentleman (2009). "ShortRead: a Bioconductor package for input, quality assessment and exploration of high-throughput sequence data." In: *Bioinformatics* 25, pp. 2607–2608. DOI: 10.1093/bioinformatics/btp450.

Morgan, Martin, Valerie Obenchain, Jim Hester, and Hervé Pagès (2018). *SummarizedExperiment: SummarizedExperiment container*. R package version 1.12.0.

Morgan, Martin, Valerie Obenchain, Michel Lang, Ryan Thompson, and Nitesh Turaga (2019). *BiocParallel: Bioconductor facilities for parallel evaluation*. R package version 1.16.5.

Morgan, Martin, Hervé Pagès, Valerie Obenchain, and Nathaniel Hayden (2018). *Rsamtools: Binary alignment (BAM), FASTA, variant call (BCF), and tabix file import*. R package version 1.34.0.

Mukherjee, Malini, Emily M Mace, Alexandre F Carisey, Nabil Ahmed, and Jordan S Orange (2017). "Quantitative imaging approaches to study the CAR immunological synapse." In: *Molecular Therapy* 25.8, pp. 1757–1768.

Nanda, Navreet K and Eli E Sercarz (1995). "Induction of anti-self-immunity to cure cancer." In: *Cell* 82.1, pp. 13–17.

Nath, Nidhi, Becky Godat, Hélène Benink, and Marjeta Urh (2015). "On-bead antibody-small molecule conjugation using high-capacity magnetic beads." In: *Journal of immunological methods* 426, pp. 95–103.

National Academies of Sciences, Engineering and Medicine (2016). *Improving Diagnosis in Health Care*. Washington, DC: The National Academies Press.

Navalkar, Krupa Arun, Stephan Albert Johnston, and Phillip Stafford (2015). "Peptide based diagnostics: Are random-sequence peptides more useful than tiling proteome sequences?" In: *Journal of immunological methods* 417, pp. 10–21.

Neuwirth, Erich (2014). *RColorBrewer: ColorBrewer Palettes*. R package version 1.1-2.

Nowak, Martin A (1996). "Immune responses against multiple epitopes: a theory for immunodominance and antigenic variation." In: *seminars in VIROLOGY*. Vol. 7. 1. Elsevier, pp. 83–92.

Pagès, H., P. Aboyoun, R. Gentleman, and S. DebRoy (2019). *Biostrings: Efficient manipulation of biological strings*. R package version 2.50.2.

Pagès, H., M. Lawrence, and P. Aboyoun (2018). *S4Vectors: S4 implementation of vector-like and list-like objects*. R package version 0.20.1.

Pagès, Hervé and Patrick Aboyoun (2018). *XVector: Representation and manipulation of external sequences*. R package version 0.22.0.

Pagès, Hervé, Marc Carlson, Seth Falcon, and Nianhua Li (2018). *AnnotationDbi: Annotation Database Interface*. R package version 1.44.0.

Pagès, Hervé and with contributions from Peter Hickey (2018). *DelayedArray: Delayed operations on array-like objects*. R package version 0.8.0.

Park, Gahee, Joo Kyung Park, Seung-Ho Shin, Hyo-Jeong Jeon, Nayoung KD Kim, Yeon Jeong Kim, Hyun-Tae Shin, Eunjin Lee, Kwang Hyuck Lee, and Dae-Soon Son (2017). "Characterization of background noise in capture-based targeted sequencing data." In: *Genome biology* 18.1, p. 136.

Paull, Michael L and Patrick S Daugherty (2018). "Mapping serum antibody repertoires using peptide libraries." In: *Current opinion in chemical engineering* 19, pp. 21–26.

Pedersen, Johannes W and Hans H Wandall (2011). "Autoantibodies as biomarkers in cancer." In: *Laboratory Medicine* 42.10, pp. 623–628.

Pennisi, Marzio (2012). "A mathematical model of immune-system-melanoma competition." In: *Computational and mathematical methods in medicine* 2012.

Ph.D.™ Phage Display Libraries Instruction Manual (July 2016). Version 2.1. New England Biolabs, pp. 19–22.

Pillis, Lisette G de, Ami E Radunskaya, and Charles L Wiseman (2005). "A validated mathematical model of cell-mediated immune response to tumor growth." In: *Cancer research* 65.17, pp. 7950–7958.

Polanski, Malu and N Leigh Anderson (2006). "A list of candidate cancer biomarkers for targeted proteomics." In: *Biomarker insights* 1, p. 117727190600100001.

Promega: *Antibody Purification Technical Manual*, TM371 2015 (July 2015). *Magne Protein A Beads and Magne Protein G Beads for Antibody Purification Technical Manual, TM371*. Promega, p. 8.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

Rabadan, Raul, Gyan Bhanot, Sonia Marsilio, Nicholas Chiorazzi, Laura Pasqualucci, and Hossein Khiabanian (2018). "On statistical modeling of sequencing noise in high depth data to assess tumor evolution." In: *Journal of statistical physics* 172.1, pp. 143–155.

Rebollo, Inmaculada Rentero, Michal Sabisz, Vanessa Baeriswyl, and Christian Heinis (2014). "Identification of target-binding peptide motifs by high-throughput sequencing of phage-selected peptides." In: *Nucleic acids research* 42.22, e169–e169.

Regoes, Roland R, Daniel L Barber, Rafi Ahmed, and Rustom Antia (2007). "Estimation of the rate of killing by cytotoxic T lymphocytes in vivo." In: *Proceedings of the National Academy of Sciences* 104.5, pp. 1599–1603.

Ren, Kun and Kenton Russell (2016). *formattable: Create 'Formattable' Data Structures*. R package version 0.2.0.1.

Ricotta, Carlo (2017). "Of beta diversity, variance, evenness, and dissimilarity." In: *Ecology and evolution* 7.13, pp. 4835–4843.

Rinker, Tyler W. and Dason Kurkiewicz (2018). *pacman: Package Management for R*. version 0.5.0. Buffalo, New York.

Rodriguez-Sanchez, Francisco (2018). *grateful: Facilitate Citation of R Packages*. R package version 0.0.2.

Romain, Gabrielle, Harjeet Singh, Ivan Liadi, Jay R Adolacion, Badrinath Roysam, Laurence Cooper, and Navin Varadarajan (2015). "Single cell metrics of the efficacy of CAR T cells." In: *Journal for immunotherapy of cancer* 3.S2, P324.

Rouet, Romain, Katherine JL Jackson, David B Langley, and Daniel Christ (2018). "Next-generation sequencing of antibody display repertoires." In: *Frontiers in immunology* 9, p. 118.

Ryvkin, Arie, Haim Ashkenazy, Larisa Smelyanski, Gilad Kaplan, Osnat Penn, Yael Weiss-Ottolenghi, Eyal Privman, Peter B Ngam, James E Woodward, Gregory D May, Callum Bell, Tal Pupko, and Jonathan M. Gershoni (2012). "Deep panning: steps towards probing the IgOme." In: *PloS one* 7.8, e41469.

Salimi, Nima, Ward Fleri, Bjoern Peters, and Alessandro Sette (2012). "The immune epitope database: a historical retrospective of the first decade." In: *Immunology* 137.2, pp. 117–123.

Schloerke, Barret, Jason Crowley, Di Cook, Francois Briatte, Moritz Marbach, Edwin Thoen, Amos Elberg, and Joseph Larmarange (2018). *GGally: Extension to 'ggplot2'*. R package version 1.4.0.

Sha, Ying, John H Phan, and May D Wang (2015). "Effect of low-expression gene filtering on detection of differentially expressed genes in RNA-seq data." In: *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, pp. 6461–6464.

Shave, Steven, Stefan Mann, Joanna Koszela, Alastair Kerr, and Manfred Auer (2018). "PuLSE: Quality control and quantification of peptide sequences explored by phage display libraries." In: *PloS one* 13.2, e0193332.

Shen, Wei, Shuai Le, Yan Li, and Fuquan Hu (2016). "SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation." In: *PLoS One* 11.10, e0163962.

Sievert, Carson (2018). *plotly for R*.

Silva, Murillo, Thao H Nguyen, Phaethon Philbrook, Matthew Chu, Olivia Sears, Stephen Hatfield, Robert K Abbott, Garnett Kelsoe, and Michail V Sitkovsky (2017). "Targeted elimination of immunodominant B cells drives the germinal center reaction toward subdominant epitopes." In: *Cell reports* 21.13, pp. 3672–3680.

Simpson, Andrew JG, Otavia L Caballero, Achim Jungbluth, Yao-Tseng Chen, and Lloyd J Old (2005). "Cancer/testis antigens, gametogenesis and cancer." In: *Nature Reviews Cancer* 5.8, p. 615.

Singh, Harinder, Hifzur Rahman Ansari, and Gajendra PS Raghava (2013). "Improved method for linear B-cell epitope prediction using Antigens primary sequence." In: *PloS one* 8.5, e62216.

Slight-Webb, Samantha, Rufei Lu, Lauren L Ritterhouse, Melissa E Munroe, Holden T Maecker, Charles G Fathman, Paul J Utz, Joan T Merrill, Joel M Guthridge, and Judith A James (2016). "Autoantibody-positive healthy individuals display unique immune profiles that may regulate autoimmunity." In: *Arthritis & rheumatology* 68.10, pp. 2492–2502.

Smith, George P (1985). "Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface." In: *Science* 228.4705, pp. 1315–1317.

Solymos, Peter and Zygmunt Zawadzki (2019). *pbapply: Adding Progress Bar to '*apply' Functions*. R package version 1.4-0.

Speranskaya, Anna S, Kamil Khafizov, Andrey A Ayginin, Anastasia A Krinitsina, Denis O Omelchenko, Maya V Nilova, Elena E Severova, Evgeniya N Samokhina, German A Shipulin, and Maria D Logacheva (2018). "Comparative analysis of Illumina and Ion Torrent high-throughput sequencing platforms for identification of plant components in herbal teas." In: *Food control* 93, pp. 315–324.

Stafford, Phillip, Zbigniew Cichacz, Neal W Woodbury, and Stephen Albert Johnston (2014). "Immunosignature system for diagnosis of cancer." In: *Proceedings of the National Academy of Sciences* 111.30, E3072–E3080.

Stelzer, Gil, Naomi Rosen, Inbar Plaschkes, Shahar Zimmerman, Michal Twik, Simon Fishilevich, Tsippi Iny Stein, Ron Nudel, Iris Lieder, and Yaron Mazor (2016). "The GeneCards suite: from gene data mining to disease genome sequence analyses." In: *Current protocols in bioinformatics* 54.1, pp. 1–30.

Su, Shuo, Xinliang Fu, Gairu Li, Fiona Kerlin, and Michael Veit (2017). "Novel Influenza D virus: Epidemiology, pathology, evolution and biological characteristics." In: *Virulence* 8.8, pp. 1580–1591.

Sun, Pingping, Haixu Ju, Baowen Zhang, Yu Gu, Bo Liu, Yanxin Huang, Huijie Zhang, and Yuxin Li (2014). "Conformational B-Cell Epitope Prediction Method Based on Antigen Preprocessing and Mimotopes Analysis." In: *BioMed Research International* 2014.

Suurmond, Jolien and Betty Diamond (2015). "Autoantibodies in systemic autoimmune diseases: specificity and pathogenicity." In: *The Journal of clinical investigation* 125.6, pp. 2194–2202.

Sykes, Kathryn F, Joseph B Legutki, and Phillip Stafford (2013). "Immunosignaturing: a critical review." In: *Trends in biotechnology* 31.1, pp. 45–51.

Tehrani, Ali S Saber, HeeWon Lee, Simon C Mathews, Andrew Shore, Martin A Makary, Peter J Pronovost, and David E Newman-Toker (2013). "25-Year summary of US malpractice claims for diagnostic errors 1986–2010: an analysis from the National Practitioner Data Bank." In: *BMJ Qual Saf* 22.8, pp. 672–680.

Tighe, Patrick J, Richard R Ryder, Ian Todd, and Lucy C Fairclough (2015). "ELISA in the multiplex era: potentials and pitfalls." In: *PROTEOMICS–Clinical Applications* 9.3-4, pp. 406–422.

Traggiai, Elisabetta, Roberto Puzone, and Antonio Lanzavecchia (2003). "Antigen dependent and independent mechanisms that sustain serum antibody levels." In: *Vaccine* 21, S35–S37.

Van Regenmortel, MHV (2001). "Antigenicity and immunogenicity of synthetic peptides." In: *Biologicals* 29.3-4, pp. 209–213.

Varadarajan, Navin, Ivan Liadi, Gabrielle Romain, Harjeet Singh, and Laurence Cooper (2014). "Functional and molecular characterization of single serial killer CAR+ T cells demonstrates adaptive modification of behavior and fate based on tumor cell density." In: *Journal for immunotherapy of cancer* 2.S3, P145.

Vidarsson, Gestur, Gillian Dekkers, and Theo Rispens (2014). "IgG subclasses and allotypes: from structure to effector functions." In: *Frontiers in immunology* 5, p. 520.

Vita, Randi, James A Overton, Jason A Greenbaum, Julia Ponomarenko, Jason D Clark, Jason R Cantrell, Daniel K Wheeler, Joseph L Gabbard, Deborah Hix, and Alessandro Sette (2014). "The immune epitope database (IEDB) 3.0." In: *Nucleic acids research*, gku938.

Vita, Randi, Kerrie Vaughan, Laura Zarebski, Nima Salimi, Ward Fleri, Howard Grey, Muthu Sathiamurthy, John Mokili, Huynh-Hoa Bui, Philip E Bourne, Julia Ponomarenko, Romulo Rde Castro Jr, Russell K Chan, John Sidney, Stephen S Wilson, Scott Stewart, Scott Way, Bjoern Peters, and Alessandro Sette (2006). "Curation of complex, context-dependent immunological data." In: *BMC bioinformatics* 7.1, p. 341.

Wardeh, Maya, Claire Risley, Marie Kirsty McIntyre, Christian Setzkorn, and Matthew Baylis (2015). "Database of host-pathogen and related species interactions, and their global distribution." In: *Scientific data* 2, p. 150049.

Watanabe, Norihiko, Hisashi Arase, Makoto Onodera, Pamela S Ohashi, and Takashi Saito (2000). "The quantity of TCR signal determines positive selection and lineage commitment of T cells." In: *The Journal of Immunology* 165.11, pp. 6252–6261.

Weiss-Ottolenghi, Yael and Jonathan M Gershoni (2014). "Profiling the IgOme: Meeting the challenge." In: *FEBS letters* 588.2, pp. 318–325.

What You Should Know for the 2014-2015 Influenza Season (n.d.). *What flu viruses did the 2014-2015 flu vaccines protect against?* CDC: *2014-2015 Influenza Season* accessed on 11/30/2016. Centers for Disease Control and Prevention. URL: `http://www.cdc.gov/flu/pastseasons/1415season.htm` (visited on 11/30/2016).

Wickham, Hadley (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wickham, Hadley (2019). *rvest: Easily Harvest (Scrape) Web Pages*. R package version 0.3.4.

Wickham, Hadley and Jennifer Bryan (2018). *readxl: Read Excel Files*. R package version 1.2.0.

Wickham, Hadley, James Hester, and Jeroen Ooms (2018). *xml2: Parse XML*. R package version 1.2.0.

Wong, Jeffrey (2013). *pdist: Partitioned Distance Function*. R package version 1.2.

Wu, Chien-Hsun, I-Ju Liu, Ruei-Min Lu, and Han-Chung Wu (2016). "Advancement and applications of peptide phage display technology in biomedical science." In: *Journal of biomedical science* 23.1, p. 8.

Xie, Yihui (2019). *knitr: A General-Purpose Package for Dynamic Report Generation in R*. R package version 1.23.

Xie, Yihui, Joe Cheng, and Xianying Tan (2019). *DT: A Wrapper of the JavaScript Library 'DataTables'*. R package version 0.6.

Xu, George J, Tomasz Kula, Qikai Xu, Mamie Z Li, Suzanne D Vernon, Thumbi Ndungu, Kiat Ruxrungtham, Jorge Sanchez, Christian Brander, and Raymond T Chung (2015). "Comprehensive serological profiling of human populations using a synthetic human virome." In: *Science* 348.6239, aaa0698.

Yasser, EL-Manzalawy and Vasant Honavar (2010). "Recent advances in B-cell epitope prediction methods." In: *Immunome research* 6.Suppl 2, S2.

Yates, Andrew, Frederik Graw, Daniel L Barber, Rafi Ahmed, Roland R Regoes, and Rustom Antia (2007). "Revisiting estimates of CTL killing rates in vivo." In: *PloS one* 2.12, e1301.

Ye, Jian, George Coulouris, Irena Zaretskaya, Ioana Cutcutache, Steve Rozen, and Thomas L Madden (2012). "Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction." In: *BMC bioinformatics* 13.1, p. 134.

Zaenker, Pauline, Elin Solomonovna Gray, and Melanie Ruth Ziman (2016). "Autoantibody production in cancerthe humoral immune response toward autologous antigens in cancer patients." In: *Autoimmunity reviews* 15.5, pp. 477–483.

Zhu, Hao (2019). *kableExtra: Construct Complex Table with 'kable' and Pipe Syntax*. R package version 1.1.0.

Appendix A Modeling of Effector-Target Contact Times in T Cells Uncovers Kinetic Homogeneity and Heterogeneity in Kill and No-kill Events

A.1 Introduction

The physical contact of an effector T cell with a target tumor cell it can recognize should in principle lead to a kill event where the effector causes the target to undergo programmed cell death. However, in these same effector cells, no-kill events can also be observed where the encounter had no noticeable effect on the target. As the fraction of effectors that can kill plays a role toward T cell efficacy, a fundamental understanding of the underlying processes can lead to avenues of improving T cell-based therapies.

Various models have been published studying the dynamics of T cell-mediated killing. A number of investigators have applied systems of equations to model killing *in vivo* (Pillis, Radunskaya, and Wiseman, 2005; Regoes et al., 2007; Yates et al., 2007). Budhu et al. (2010) established a model that points to a critical concentration of T cells determining the efficiency in killing target melanoma cells that varies depending on the tumor environment. Ganusov, Barber, and De Boer (2011) showed a proportionality between target death rate and number of effector T cells. Pennisi (2012) applied compartmental modeling to quantify the synergy in combined T cell-monoclonal antibody treatment. Gadhamsetty et al. (2017) used systems of ordinary differential equations and 2D cellular Potts model simulations to qualitatively explore the implications of transient and multistage killing. However, these models are various combinations of empirical and semi-empirical equations describing killing dynamics at the populations level. To the best of the author's knowledge, no model has been proposed yet that recapitulates mechanistically

changes that occur within the effector T cell upon conjugation to a target tumor cell. Nevertheless, its feasibility lies in having the capability to take measurements of such events at the single-cell level.

Timelapse Imaging Microscopy in Nanowell Grids (TIMING) is a technique that allows the collection of temporal and spatial features from live-cell video microscopy in a high-throughput manner (Merouane et al., 2015). This method has been used to study *in vitro* multi-killing as a function of tumor cell density (Varadarajan et al., 2014), multi-killing in CD4 and CD8 subsets of CAR T cells (Liadi, Harjeet Singh, et al., 2018), as well as phenotypic biomarkers of CAR T efficacy (Romain et al., 2015). This study takes advantage of this technology in order to obtain a distribution of durations of physical contact between effector-target pairs, which from here on we simply name contact times, and stratify the data based on the order of encounter an effector makes up to three different targets, i.e. first, second, and third encounters specially in multi-killing T cells.

A.1.1 Hypothesis

Kinetic models have been used in describing effector-target interactions at the population level, but there seems to be no kinetic models at the single-cell level that makes a fundamental account of the cellular mechanisms involved. It is hypothesized that no-kill events arise from defects in one or more intermediate transitional states occurring during contact and can be differentiated from kill events via a gamma-distributional analysis.

A.1.2 Rationale

Floyd, Harrison, and Van Oijen (2010) frames the problem of elucidating reaction trajectories via analysis of waiting-time distributions. The phenomenon is viewed to be stochastic, following a Poisson process. As a consequence, waiting times toward the

occurrence of the first event follows an exponential distribution. Furthermore, the waiting times until the occurrence of the n^{th} event is a gamma distribution. A gamma distribution describes the kinetics of intermediates in the following multistep reaction:

$$A \xrightarrow{k} X_1 \xrightarrow{k} X_2 \longrightarrow \xrightarrow{k} X_{N-1} \xrightarrow{k} B \quad \text{and} \tag{A.1}$$

$$p_{A\to B}(\tau) = \frac{k^N \tau^{N-1}}{\Gamma(N)} e^{-k\tau}, \tag{A.2}$$

where $p_{A\to B}(\tau)$ is the transition probability from initial state A to final state B, passing through a number of intermediate states $X_1, X_2, ..., X_{N-1}$ as a function of time τ and is characterized by a rate constant k across N reaction steps. Γ and e correspond to the gamma and exponential functions respectively.

By the same token, one effector/one target associations $(E:T)$ are conceived to transition through several states sequentially, i.e.,

$$(E:T)_1 \xrightarrow{k} (E:T)_2 \xrightarrow{k} (E:T)_3 \longrightarrow \xrightarrow{k} (E:T)_{N-1} \xrightarrow{k} (E:T)_N. \tag{A.3}$$

These states correspond to a number of steps involved succeeding association, a depiction of which is given in the following schematic by Mukherjee et al. (2017) (figure 40):

The result above assumes identical rate constants. Unequal rate constants, such as in the presence of a rate-determining step, will mask the true number of steps N. As demonstrated by Floyd, Harrison, and Van Oijen (2010), the effect of a single rate-determining step in a multistep process of otherwise identical rate constants is to give an apparent rate constant k_{app} and apparent number of steps N_{app} that is lower than if all steps were identical. Nevertheless, it is still possible to model multistep processes with heterogeneous rate constants by a gamma distribution analysis (Floyd, Harrison, and Van Oijen, 2010).

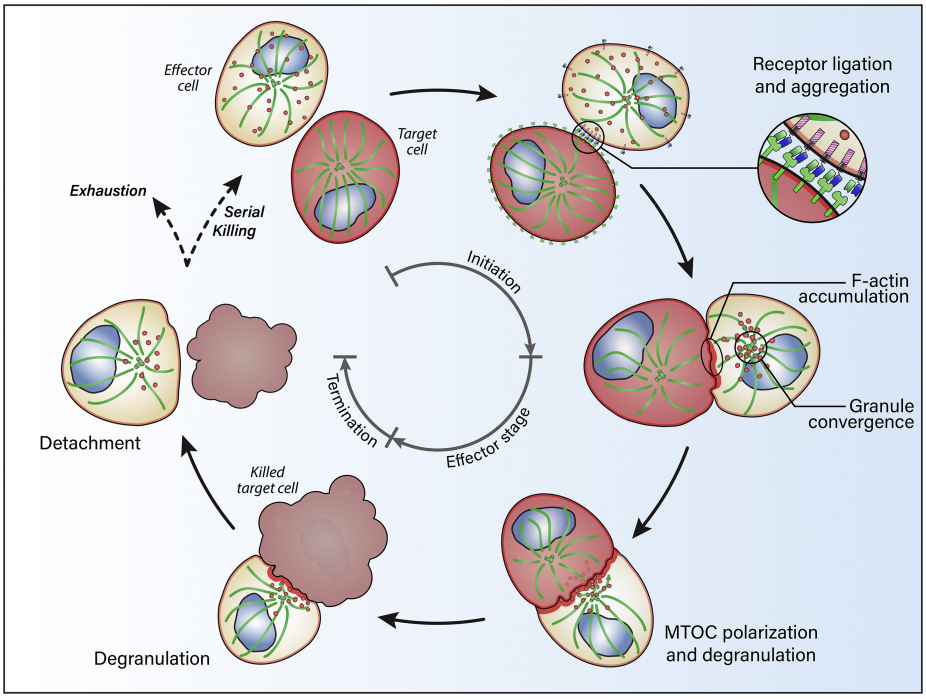


Figure 40: Schematic representation of the critical steps of the killing of a sensitive target by a CAR-expressing effector cell (Mukherjee et al., 2017).

A.2 Methods

All work outlined in this study was performed according to protocols approved by the Institutional Review Boards at the University of Houston and the University of Texas M.D. Anderson Cancer Center.

Second generation CAR containing a CD28 and CD3-ζ endodomain were expressed in healthy donor pan-T cells by electroporation with DNA plasmids from the Sleeping Beauty (SB) transposon/transposase system as described previously (Harinder Singh, Ansari, and Raghava, 2013). Cells were used between 10 days and 28 days after transfection. Where indicated, CAR+ T cells were co-transduced with an additional transposon vector encoding for CD137L (4-1BB-L). Mouse EL-4 cells (ATCC) were engineered for stable expression of human CD19. Human pre-B cell line NALM-6 (ATCC) was used as CD19+ target cells.

We previously described a method called TIMING that allows high throughput, timelapse, and single cell level imaging of thousands of nanowells, each containing 1-4 cells (Liadi, Roszik, et al., 2013). We improved the assay by utilizing thin nanowell arrays and glass bottom petri dishes as described previously (An et al., 2017). Effector (T cells) and target cells (mouse EL4 cells stably expressing human CD19), were labeled respectively with 1 μM PKH67 and PKH26 fluorescent dyes (Sigma-Aldrich) according to the manufacturers protocol and loaded on the array. Cell apoptosis was detected by immersing the array in complete, phenol red-free cell-culture media containing a dilution of 1:50 Annexin V - Alexa Fluor 647 (AF647) (Life Technologies). Arrays were imaged for 6hr at interval of 5 min using an Axio fluorescent microscope (Carl Zeiss) utilizing a 20x 0.8 NA objective, a scientific CMOS camera (Orca Flash 4.0), a humidity / CO_2 controlled chamber, and the tile function of the Zen software. Image analysis and cell segmentation/tracking were performed as described by us before (Merouane et al., 2015).

TIMING data collected from CAR T cells stratified from a transwell migration assay were analyzed using R (R Core Team, 2019). These two datasets involving CD19-targeting CAR T cells with CD-19 expressing leukemia cells were combined into a single dataset. Analysis involved probability distribution fitting by maximum likelihood carried out using fitdistrplus 1.0-11 (Delignette-Muller and Dutang, 2015) in conjunction with a number of helper functions (Borchers, 2018; Mohammadi, 2018-07-5; Rinker and Kurkiewicz, 2018; Wickham and Bryan, 2018; Yihui Xie, 2019). Only data on the contact times of kill and no-kill events were considered, with ambiguous cases excluded (e.g. simultaneous contact of an effector to two or more targets). For an effector with multiple targets, data arising these cases were sorted based on the chronological sequence the effector made contact with a target. Interval and right censoring of the data were accounted for in the fitting, where an uncertainty of 10 min was attributed to interval-censored data and contact times at the maximum experimental time of 350 min were right censored. An

additional 25 min was included for cases where the effector was already in contact with a target at the start of image acquisition, with the condition that the target was still observed to be viable.

A.3 Results and discussion

Waiting times governed by a gamma distribution underlie a multistep process occurring in series (Floyd, Harrison, and Van Oijen, 2010). This serial multistep process seems to coincide with our current understanding of the cellular changes occurring in CAR T cell-mediated killing (Mukherjee et al., 2017). We thus hypothesize that the nature in which effector-target associations transition from conjugation to detachment occurs in a gamma-distributed fashion. This assumption implies that effector-target contact times characterized by this type of distribution are mechanistically driven by a consecutive irreversible first-order kinetic system, with an average transit time across this system given by $\sum(1/k_i)$, where k_i is the first-order rate constant of the i^{th} step; for equal k_i's, $\sum(1/k_i) = N/k$ (Bailey, Eadie, and Schmidt, 1974). As a consequence, the apparent number of steps for such a system will always be bounded between 1 and N, corresponding to the extreme cases of a system dominated by a single rate-determining step and a system with all rate constants equal respectively (Bailey, Eadie, and Schmidt, 1974; Floyd, Harrison, and Van Oijen, 2010).

To illustrate, a hypothetical reaction series consisting of five steps with identical rate constants will have gamma-distributed transitions of its initial and intermediates states (figure 41, left panel). On the top left, waiting-time distributions across all five states A to F of an $N = 5$ multistep process with equal rate constants $k = 1$ showing the relative abundance of each state as a function of time. Waiting times for states A to E can be shown to be gamma-distributed. On the bottom left, cumulative distributions of states A to F showing the total fraction of a given state that has transitioned to the next state. Note that the

waiting-time distribution of the final state F is always identical to the cumulative distribution of the state before it, state E, when the kinetic steps occur only in series. However, in the presence of a rate-determining step, transitions from this slowest step and onwards will have that delay reflected in their temporal profiles (figure 41, middle panel). On the top middle, a similar $N = 5$ multistep process but with the last step occurring relative slower at a rate constant $k' = 0.25$. States are written as $E' \rightarrow F'$ to distinguish it from the case $E \rightarrow F$ with identical rate constants $k = 1$, which is plotted alongside for comparison. The slow down at $E' \rightarrow F'$ results in the broadening of the waiting-time distribution and slower decay of state E' towards zero. The true kinetics for state E' can approximately be captured by a gamma distribution fitting, yielding apparent N and k values. On the bottom middle, comparison of the cumulative distributions of E and E'. Since the waiting distribution of the final state is identical to the cumulative distribution of the previous state, we see a slower increase in the abundance of the final state for $E' \rightarrow F'$. In general, it is likely for the steps to have unequal rate constants and we would thus expect departures from this idealized system. Additionally, the inability to observe intermediate states confines the analysis to only data on the initial and final states. Fortunately, a best fit can still be afforded by performing a gamma-distribution analysis under these conditions (Bailey, Eadie, and Schmidt, 1974; Floyd, Harrison, and Van Oijen, 2010). For cases exhibiting multimodal patterns, we anticipate that such a system can be modeled using a finite mixture of gamma distributions (figure 41, right panel). On the top right, the case of a system composed of a mixture of three multistep processes (in gray) and the overall waiting-time distribution for state E. This system may be governed by the same underlying multistep process but differences in rates in one or more steps in only some events can result to two or more distinct multistep processes of varying apparent N and k. Dotted curve shows the gamma distribution fit of the system. On the bottom right, the overall cumulative distribution of E and the com-

ponent cumulative distributions that comprise it. Mixtures of multistep processes exhibit strong departures from its gamma distribution fit.

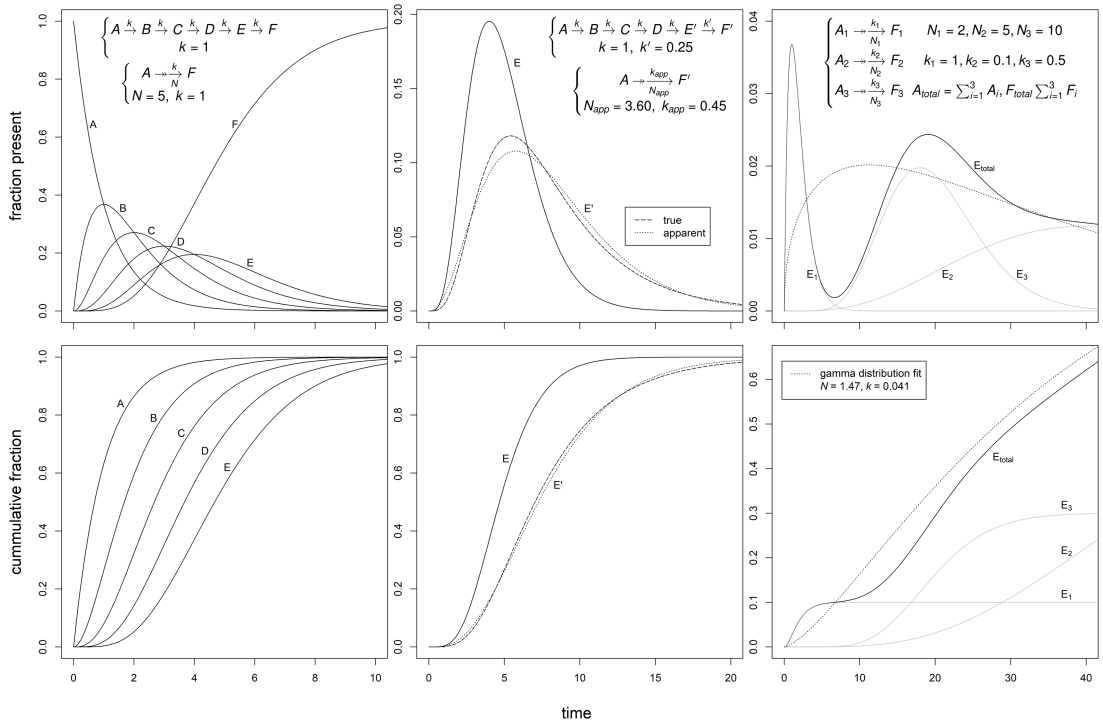


Figure 41: Temporal profiles of a multistep process consisting of consecutive first order irreversible kinetic step, illustrating the transition from an initial state A to final state F.

When applied to different cases of effector-target contact events, a gamma distribution analysis yielded fitting parameters that distinguish between kill and no-kill events (figure 42). In figure 42, contact time is defined to be the duration of effector-target association conditioned on the target cell remaining viable. CAR T cell contact times are stratified based on whether it is the time the T cell made contact with a target, a different target each time (1E_1T, 1E_2T, and 1E_3T), and whether it resulted to target death (kill vs. no-kill). The number of datapoints present in each case is indicated on the right. Fitting parameters are given as estimate $\pm$ standard deviation. Histograms show the distribution of contact times without accounting for censoring. Censoring is considered in the gamma

distribution fits drawn as a solid red curve and in the CDF, Q-Q, and P-P plots. The gamma distribution appear to fit kill events better over their no-kill counterparts. Values of N_{app} and k_{app} from no-kill events were observed to be about half and roughly a quarter of the fitted values from kill events respectively. Recalling that the average transit time is given by N/k, it follows that mean contact times for no-kill events are expected to relatively be twice of kill events. We find that this statement to hold for the fitted values but tends to overestimate when the calculations were based directly on the data (table 10). This departure stems from lower mean values when directly computed from no-kill contact events, which is a consequence of the right-censoring in the data being more pronounced in the no-kill cases. For kill events, we find estimated values of $N_{\text{app}} \geq 1$, suggesting that the kinetics for kill events is dominated by a rate-determining step. In contrast, no-kill events have $N_{\text{app}} < 1$, which is physically impossible for the kinetic system described by a gamma distribution. We propose that these phenomena may be the result of the presence of more than one multistep process akin to what is depicted on the right panel of figure 41. It can be shown that $N_{\text{app}} < 1$ occurs when the ratio of the geometric mean to the arithmetic mean becomes approximately less than 0.56 (appendix F). This statement is corroborated by calculations made on the data (table 11). In table 11, contact times are segregated based on whether the outcome is a kill or no-kill event and whether it is the first, second, or third encounter of an effector with different targets.

Table 10: Comparison of mean contact times calculated from fitting parameters vs from data. Contact times are segregated based on whether the outcome is a kill or no-kill event and whether it is the first, second, or third encounter of an effector with different targets.

	N_{app}	k_{app}, min^{-1}	mean (from fit), min	mean (from data), min	difference, min
kill.1E_1T	1.43	0.0157	90.7	80.1	10.7
no-kill.1E_1T	0.67	0.0043	157.3	123.4	33.9
kill.1E_2T	1.15	0.0153	74.8	65.5	9.4
no-kill.1E_2T	0.72	0.0049	147.1	123.0	24.2
kill.1E_3T	1.21	0.0142	84.9	76.0	8.8
no-kill.1E_3T	0.69	0.0035	193.6	145.6	48.0

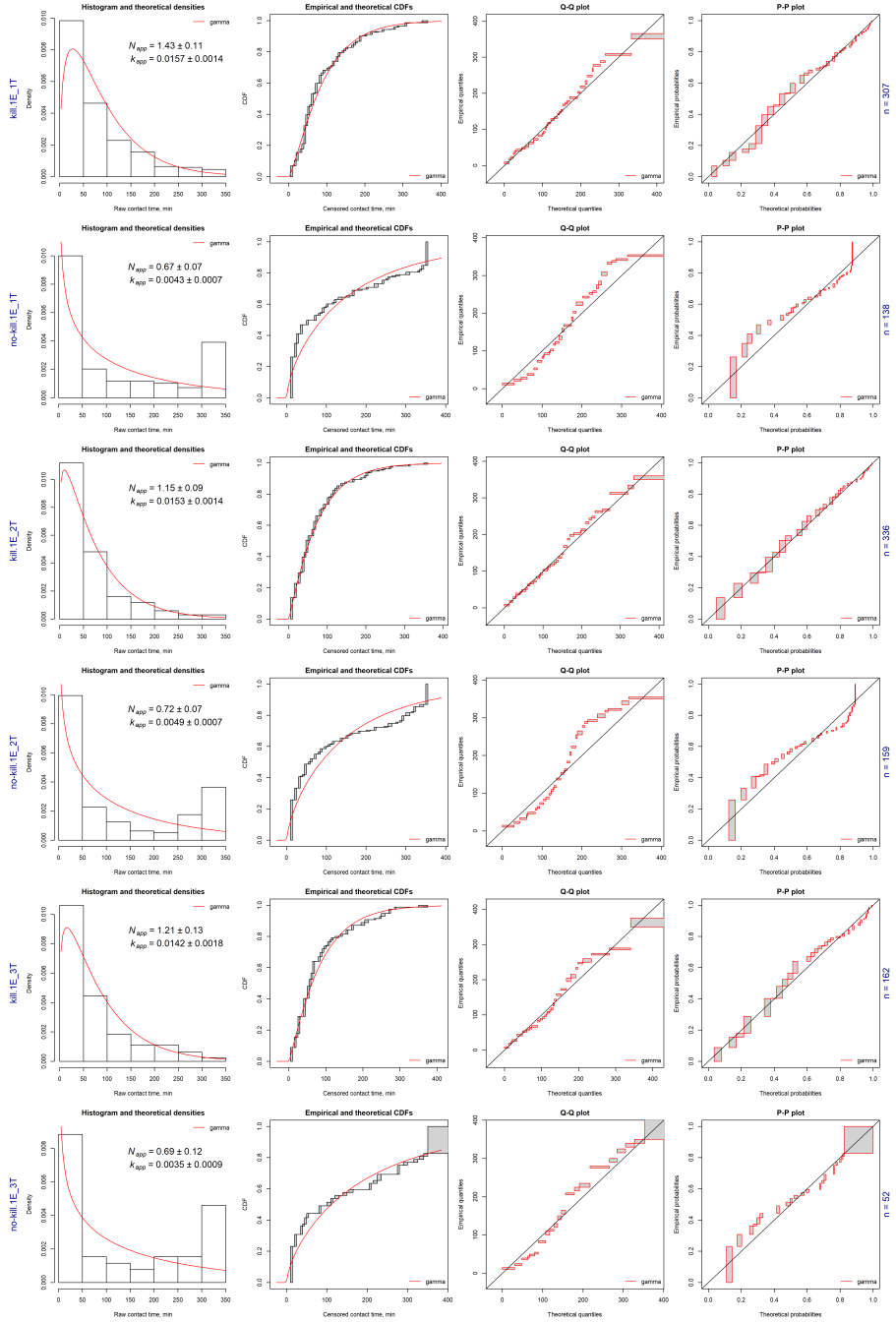


Figure 42: Comparison of the goodness-of-fit with a gamma distribution for different effector-target contact events.

Table 11: Comparison of the ratio of geometric mean $\bar{x}_g$ to the arithmetic mean $\bar{x}_a$ directly calculated from raw contact times between kill and no-kill cases. Zeroes were dropped in the calculation of the means.

	$\bar{x}_g$, min	$\bar{x}_a$, min	$\bar{x}_g/\bar{x}_a$
kill.1E_1T	54.5	84.2	0.65
no-kill.1E_1T	48.5	123.4	0.39
kill.1E_2T	42.3	71.0	0.60
no-kill.1E_2T	50.0	123.0	0.41
kill.1E_3T	52.7	81.6	0.65
no-kill.1E_3T	61.7	145.6	0.42

Bootstrapped estimates of the fitting parameters reinforces the observation of kill and no-kill events splitting characteristically at $N = 1$ (figure 43). In figure 43, CAR T cell contact times are stratified based on whether it is the 1st, 2nd, or 3rd time the T cell made contact with a target, a different target each time (1E_1T, 1E_2T, and 1E_3T), and whether it resulted to target death (kill vs. no-kill). Estimates of N_{app} and k_{app} values from the gamma distribution fit were generated from around 10 000 bootstrapped samples (black for kill, blue for no-kill); values are reported as median [95 % CI] together with the sample size n of the empirical CDF. Shaded regions indicate 95 % pointwise confidence intervals. Kill events are modeled better by a gamma distribution compared to no-kill events. The no-kill CDF crossing over the kill CDF suggests the presence of abortive contacts while the greater number of right-censored data indicates persistent contacts in no-kill events. Overlapping confidence intervals suggest N_{app} and k_{app} remain invariant across contacts. Another interesting observation to make is the crossover between the CDFs of the kill and no-kill events, where we see more effector-target contacts terminating earlier in no-kill events compared to kill events. We speculate that stalling occurs somewhere in the series of transitions the effector undergoes while in contact with its target. Plausibly, this stalling happens at a kinetic step that does not fully commit the effector to undergo all the transitions that occur in kill events. By a similar argument, persistent contacts are probably

the result of stalling at a kinetic step where the effector is required to clear a certain number of kinetic steps before detachment becomes permissible.

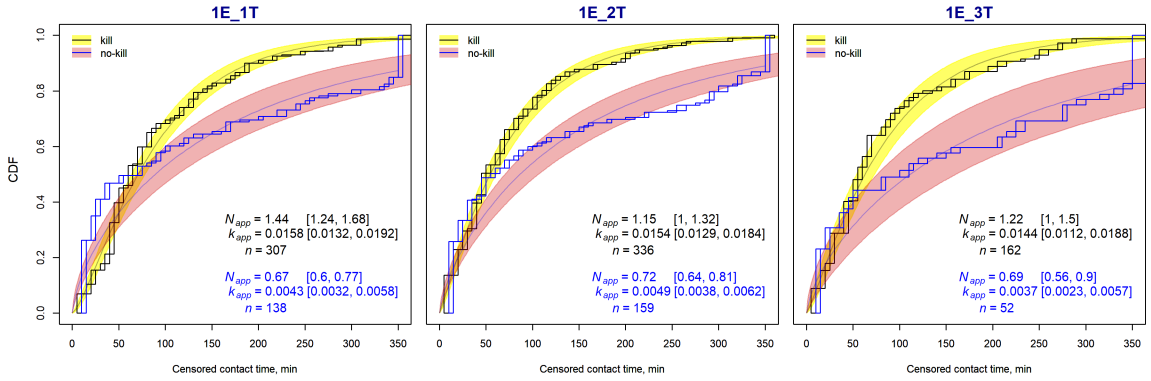


Figure 43: Nonparametric bootstrap simulations of uncertainty for censored contact times.

In order to investigate the implications of the model experimentally, and to understand the mechanistic basis of the failure, we performed TIMING assays by tracking lysosomes. As expected, 100 % of killer T cells polarized the lytic granules to the immunological synapse and showed sustained polarization (figure 44A). This result is consistent with previous data that demonstrated that granule exocytosis is essential for the killing mediated by T cells. By contrast, non-killer T cells behavior could be classified into two categories: (1) 80 % of the T cells showed only transient but not sustained polarization of the lysosomes towards the synapse (figure 44B), and (2) 20 % of the T cells showed sustained polarization similar to the killer T cells yet failed to kill (figure 44C).

A.4 Conclusions and future directions

In summary, the gamma-distribution analysis provides a framework in obtaining insights to the kinetics of effector-target contact events. Gamma-distributed contact times support the idea of an effector-target complex formation-dissociation as a sequential process, transitioning across a series of first-order intermediate states. Based on this model, the theoretical number of transitions cannot be lower than one, which we see in kill events

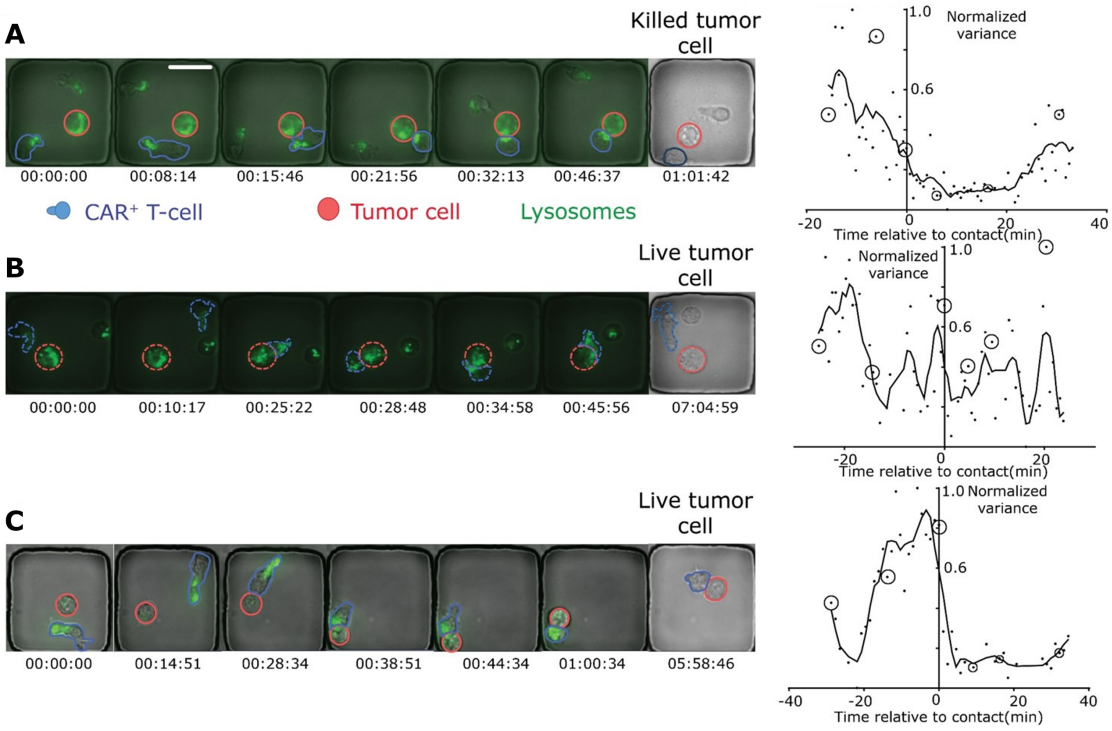


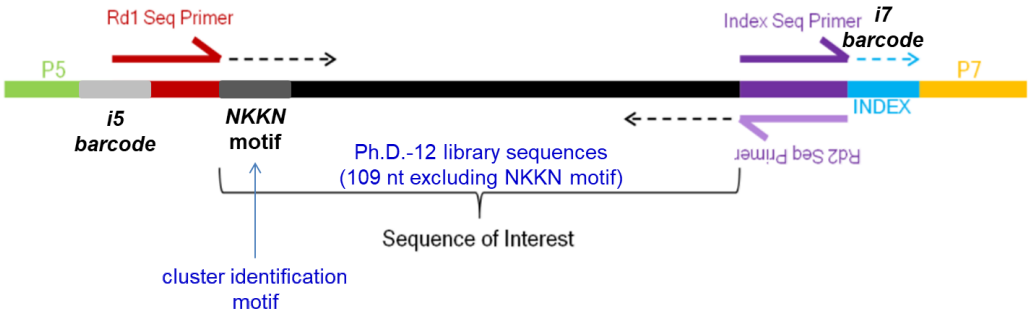
Figure 44: Tracking of lysosomes using TIMING assays.

but becomes anomalous in no-kill events. We explain this observation as a consequence of the kinetic heterogeneity in no-kill events, suggested by the presence of abortive and persistent contacts. Kill events appear to be kinetically homogeneous while no-kill events seem to stratify into kinetically distinct subpopulations.

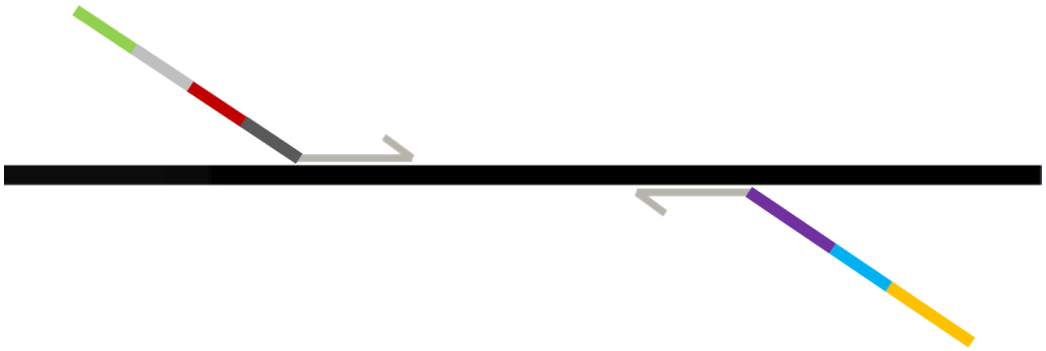
One aspect of the study that has not been touched is the modeling of no-kill events using a finite mixture of gamma distributions. Though there are algorithms available that allow elucidation of the gamma distribution components making up a mixture distribution, the presence of censoring complicates the analysis and prevents the application of these methods. Thus, an area to work on would writing a code that carries out the deconvolution for censored data. Additionally, this work should be evaluated with respect to other effector-target cell-type combinations to determine the robustness of the model in kill events and examine a wider array of heterogeneous kinetics in no-kill events.

Appendix B Illumina custom primer design for Ph.D.-12 libraries

B.1 Illumina library template structure



Source: http://nextgen.mgh.harvard.edu/CustomPrimer.html



Illumina libraries are generated by PCR amplification of the region on the phage genome encoding the combinatorial 12-mer sequence using primers tailed by Illumina sequences.

B.2 Library template sequences

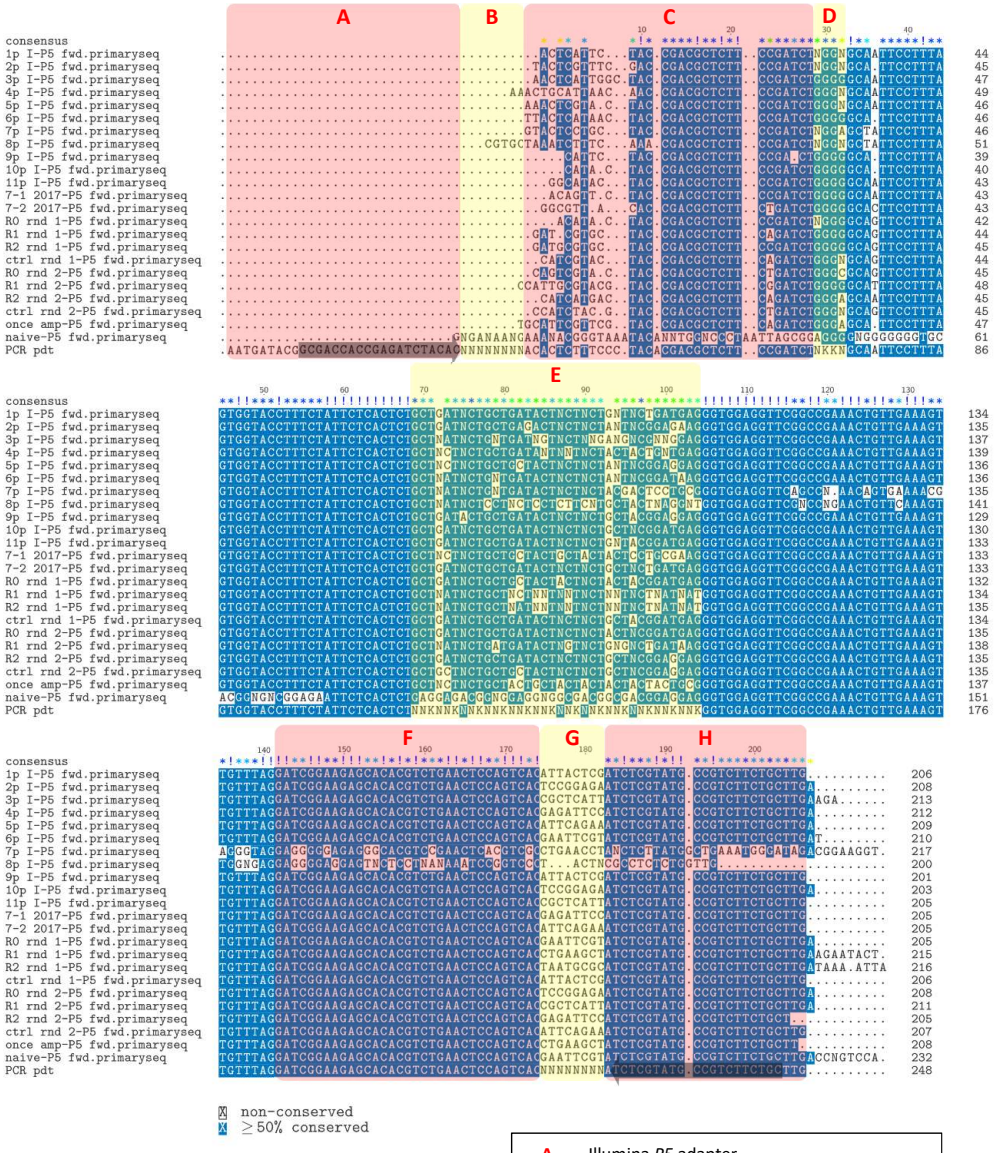
Component	Sequence	Source
P5 (29 nt)	AAT GAT ACG GCG ACC ACC GA GAT CTA CAC	http://nextgen.mgh.harvard.edu/CustomPrimer.html
i5 (8 nt)	Illumina TruSeq HT D505-D508: D505 AGG CGA AG D506 TAA TCT TA D507 CAG GAC GT D508 GTA CTG AC	page 7 of SMARTer Stranded Total RNA-Seq Kit - Pico Input Mammalian User Manual https://www.takarabio.com/assets/documents/User%20Manual/SMARTer%20Stranded%20Total%20RNA-Seq%20Kit%20-%20Pico%20Input%20Mammalian%20User%20Manual_112216.pdf
read 1 primer (33 nt)	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT	page 18 of Illumina Adapter Sequences https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/experiment-design/illumina-adapter-sequences-1000000002694-09.pdf
random 4-mer (for proper clustering)	NKKN	Scheme S4 in Supplementary Data 1 of Deep sequencing analysis of phage libraries using Illumina platform https://www.sciencedirect.com/science/article/pii/S1046202312001612
phage library insert (109 nt)	LFS — $(\text{NKK})_{12}$ — RFS	page 24 of Ph.D. Phage Display Libraries Instruction Manual https://www.neb.com/-/media/catalog/Datacards%20or%20Manuals/manualE8102.pdf
read 2 primer (33 nt)	GAT CGG AAG AGC ACA CGT CTG AAC TCC AGT CAC	page 18 of Illumina Adapter Sequences https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/experiment-design/illumina-adapter-sequences-1000000002694-09.pdf
i7 (8 nt)	Illumina TruSeq HT D701-D707: D701 ATT ACT CG D702 TCC GGA GA D703 CGC TCA TT D704 GAG ATT CC D705 ATT CAG AA D706 GAA TTC GT D707 CTG AAG CT	page 7 of SMARTer Stranded Total RNA-Seq Kit - Pico Input Mammalian User Manual https://www.takarabio.com/assets/documents/User%20Manual/SMARTer%20Stranded%20Total%20RNA-Seq%20Kit%20-%20Pico%20Input%20Mammalian%20User%20Manual_112216.pdf
P7 (24 nt)	ATC T CGT ATG CCG TCT TCT GCT TG	http://nextgen.mgh.harvard.edu/CustomPrimer.html

left flanking sequence (LFS)	G CAA TTC CTT TAG TGG TAC CTT TCT ATT CTC ACT CT
right flanking sequence (RFS)	G GTG GAG GTT CGG CCG AAA CTG TTG AAA GTT GTT TAG
tailed forward primer	GCA ATT CCT TTA GTG GTA CCT TTC T
tailed reverse primer	CTA AAC AAC TTT CAA CAG TTT CGG C

Appendix C Sanger sequencing of Illumina libraries

Illumina libraries were prepared by PCR amplification using 25-nt primers tailed by Illumina sequences. Location of Sanger sequencing primers are marked by arrows. The expected library sequence is shown on the `PCR pdt` row.

all forward reads



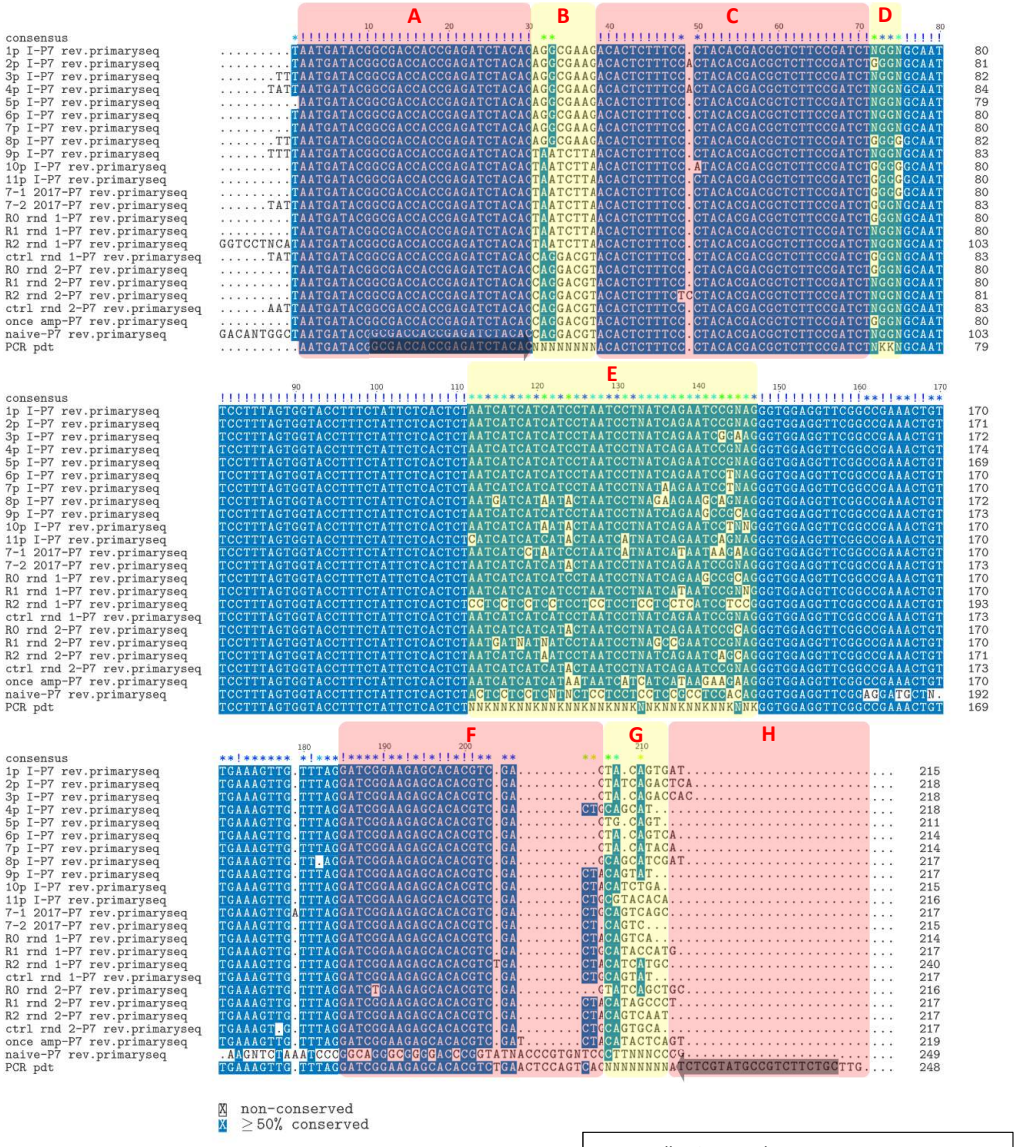
i7 Index (Tube Label)	Barcode Sequence
R1	ATTACTCG
R2	TCCGGAGA
R3	CGCTCATT
R4	GAGATTCC
R5	ATTCAGAA
R6	GAATTCGT
R7	CTGAAGCT

A	Illumina *P5* adapter
B	Illumina TruSeq HT *i5* barcode
C	paired-end read 1 sequencing primer
D	random 4-mer
E	combinatorial Ph.D.-12 library sequence
F	paired-end read 2 sequencing primer
G	Illumina TruSeq HT *i7* barcode
H	Illumina *P7* adapter

Sanger sequencing primers

- forward primer: GCG ACC ACC GAG ATC TAC AC
- reverse primer: GCA GAA GAC GGC ATA CGA GA

all reverse reads



i5 Index (Tube Label)	Barcode Sequence
F5	AGGCGAAG
F6	TAATCTTA
F7	CAGGACGT
F8	GTACTGAC

A	Illumina *P5* adapter
B	Illumina TruSeq HT *i5* barcode
C	paired-end read 1 sequencing primer
D	random 4-mer
E	combinatorial Ph.D.-12 library sequence
F	paired-end read 2 sequencing primer
G	Illumina TruSeq HT *i7* barcode
H	Illumina *P7* adapter

Appendix D Quality assessment report of Illumina libraries

ShortRead Quality Assessment

Overview

This document provides a quality assessment of Genome Analyzer results. The assessment is meant to complement, rather than replace, quality assessment available from the Genome Analyzer and its documentation. The narrative interpretation is based on experience of the package maintainer. It is applicable to results from the 'Genome Analyzer' hardware single-end module, configured to scan 300 tiles per lane. The 'control' results refered to below are from analysis of PhiX-174 sequence provided by Illumina.

Run Summary

Subsequent sections of the report use the following to identify figures and other information.

	Key
1p-I_S1_L001_R1_001	1
1p-I_S1_L001_R2_001	2
2p-I_S2_L001_R1_001	3
2p-I_S2_L001_R2_001	4
3p-I_S3_L001_R1_001	5
3p-I_S3_L001_R2_001	6
4p-I_S4_L001_R1_001	7
4p-I_S4_L001_R2_001	8
5p-I_S5_L001_R1_001	9
5p-I_S5_L001_R2_001	10
6p-I_S6_L001_R1_001	11
6p-I_S6_L001_R2_001	12
7p-I_S7_L001_R1_001	13
7p-I_S7_L001_R2_001	14
8p-I_S8_L001_R1_001	15
8p-I_S8_L001_R2_001	16
9p-I_S9_L001_R1_001	17
9p-I_S9_L001_R2_001	18
10p-I_S10_L001_R1_001	19
10p-I_S10_L001_R2_001	20
11p-I_S11_L001_R1_001	21
11p-I_S11_L001_R2_001	22
7-1-2017_S12_L001_R1_001	23
7-1-2017_S12_L001_R2_001	24
7-2-2017_S13_L001_R1_001	25
7-2-2017_S13_L001_R2_001	26

```
R0-rnd-1_S14_L001_R1_001     27
R0-rnd-1_S14_L001_R2_001     28
R1-rnd-1_S15_L001_R1_001     29
R1-rnd-1_S15_L001_R2_001     30
R2-rnd-1_S16_L001_R1_001     31
R2-rnd-1_S16_L001_R2_001     32
ctrl-rnd-1_S17_L001_R1_001   33
ctrl-rnd-1_S17_L001_R2_001   34
R0-rnd-2_S18_L001_R1_001     35
R0-rnd-2_S18_L001_R2_001     36
R1-rnd-2_S19_L001_R1_001     37
R1-rnd-2_S19_L001_R2_001     38
R2-rnd-2_S20_L001_R1_001     39
R2-rnd-2_S20_L001_R2_001     40
ctrl-rnd-2_S21_L001_R1_001   41
ctrl-rnd-2_S21_L001_R2_001   42
naive_S22_L001_R1_001        43
naive_S22_L001_R2_001        44
once-amp_S23_L001_R1_001     45
once-amp_S23_L001_R2_001     46
```

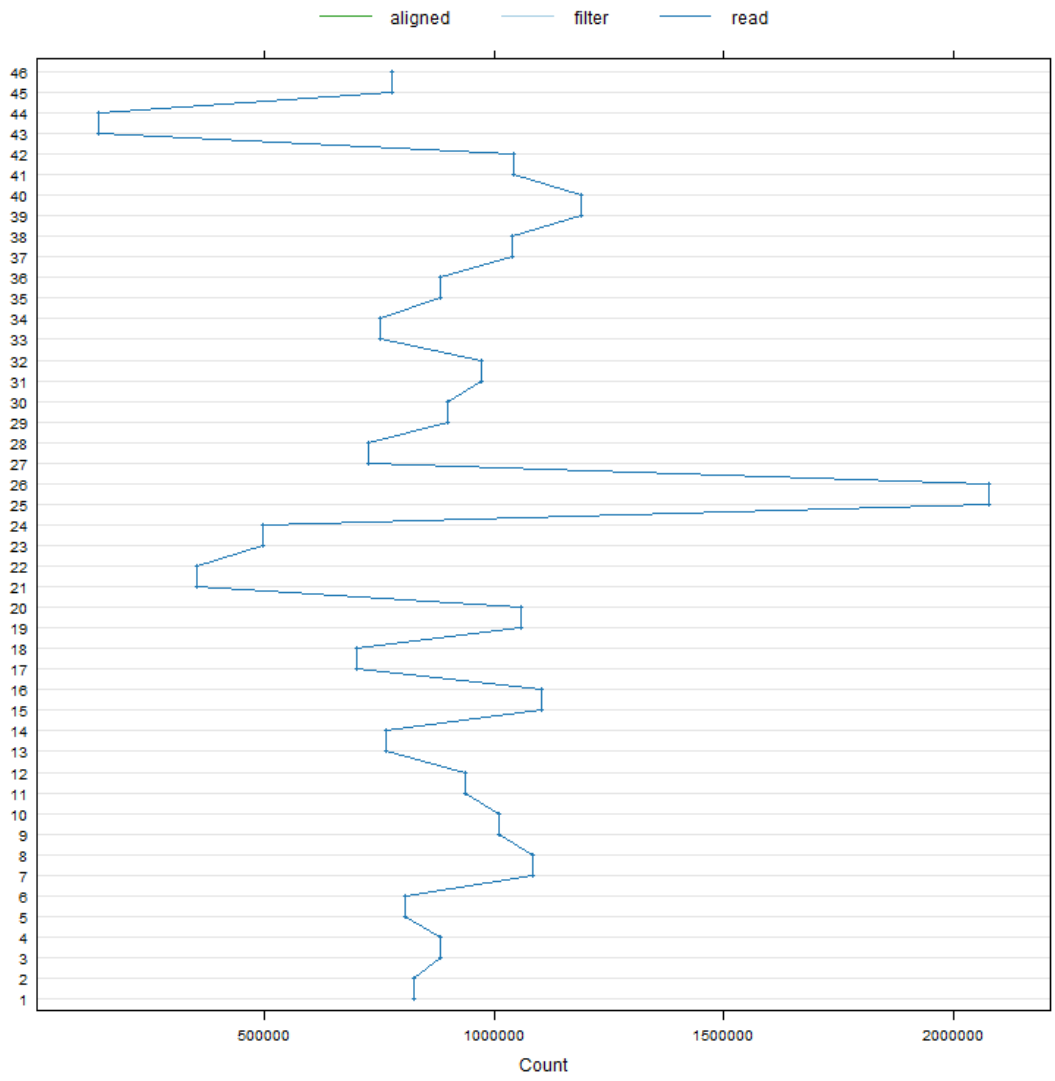
Read counts. Filtered and aligned read counts are reported relative to the total number of reads (clusters; if only filtered or aligned reads are available, total read count is reported). Consult Genome Analyzer documentation for official guidelines. From experience, very good runs of the Genome Analyzer 'control' lane result in 25-30 million reads, with up to 95% passing pre-defined filters.

```
ShortRead:::.ppnCount(qa[["readCounts"]])
```

```
     read      filter   aligned
1    825207
2    825207
3    881544
4    881544
5    805357
6    805357
7    1084328
8    1084328
9    1010639
10   1010639
11   938148
12   938148
13   764111
14   764111
15   1102561
16   1102561
17   702260
18   702260
19   1058171
20   1058171
21   352722
22   352722
23   497719
24   497719
25   2076298
26   2076298
27   727218
28   727218
```

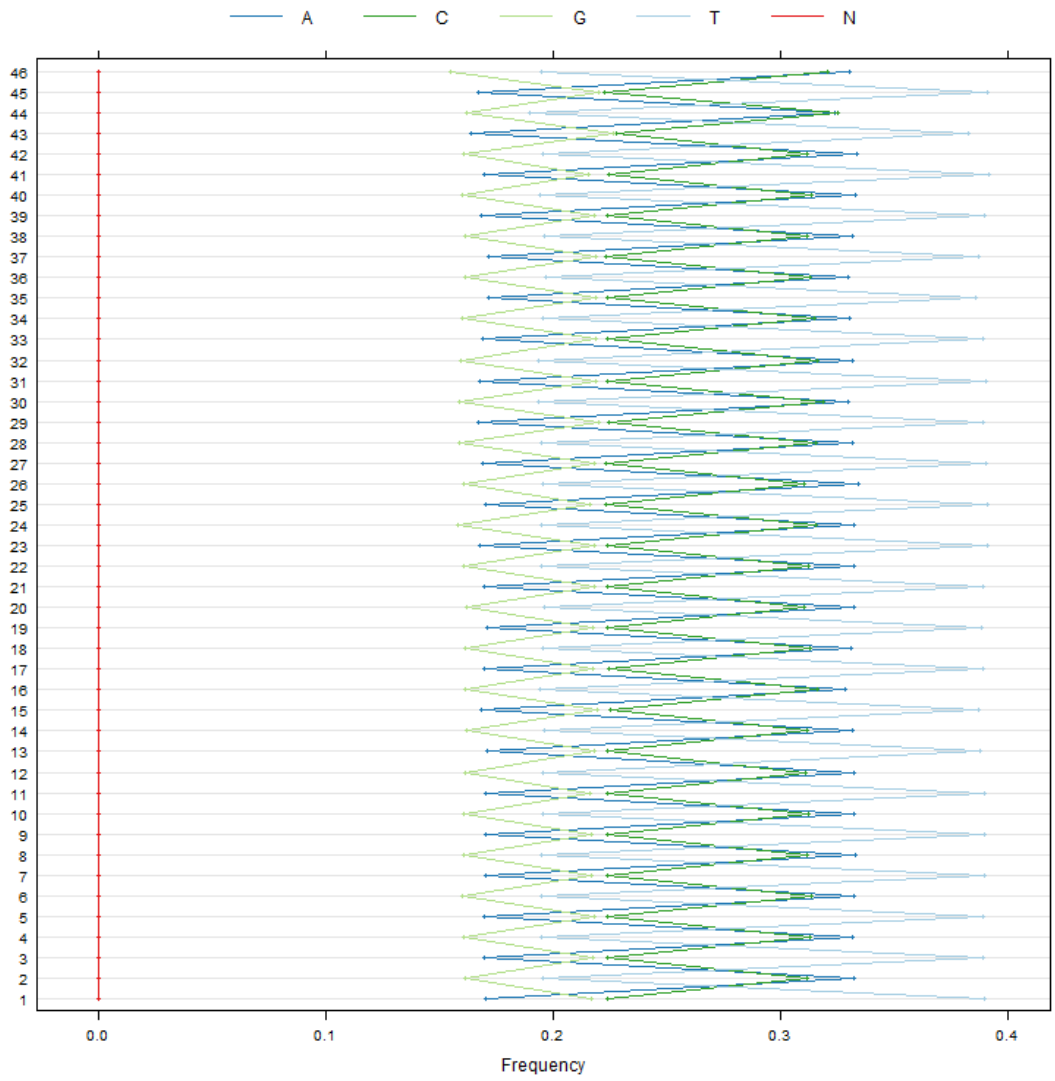
```
29   898403
30   898403
31   972837
32   972837
33   752815
34   752815
35   882101
36   882101
37   1039766
38   1039766
39   1188227
40   1188227
41   1040993
42   1040993
43   139586
44   139586
45   775960
46   775960
```

```
ShortRead:::.plotReadCount(qa)
```



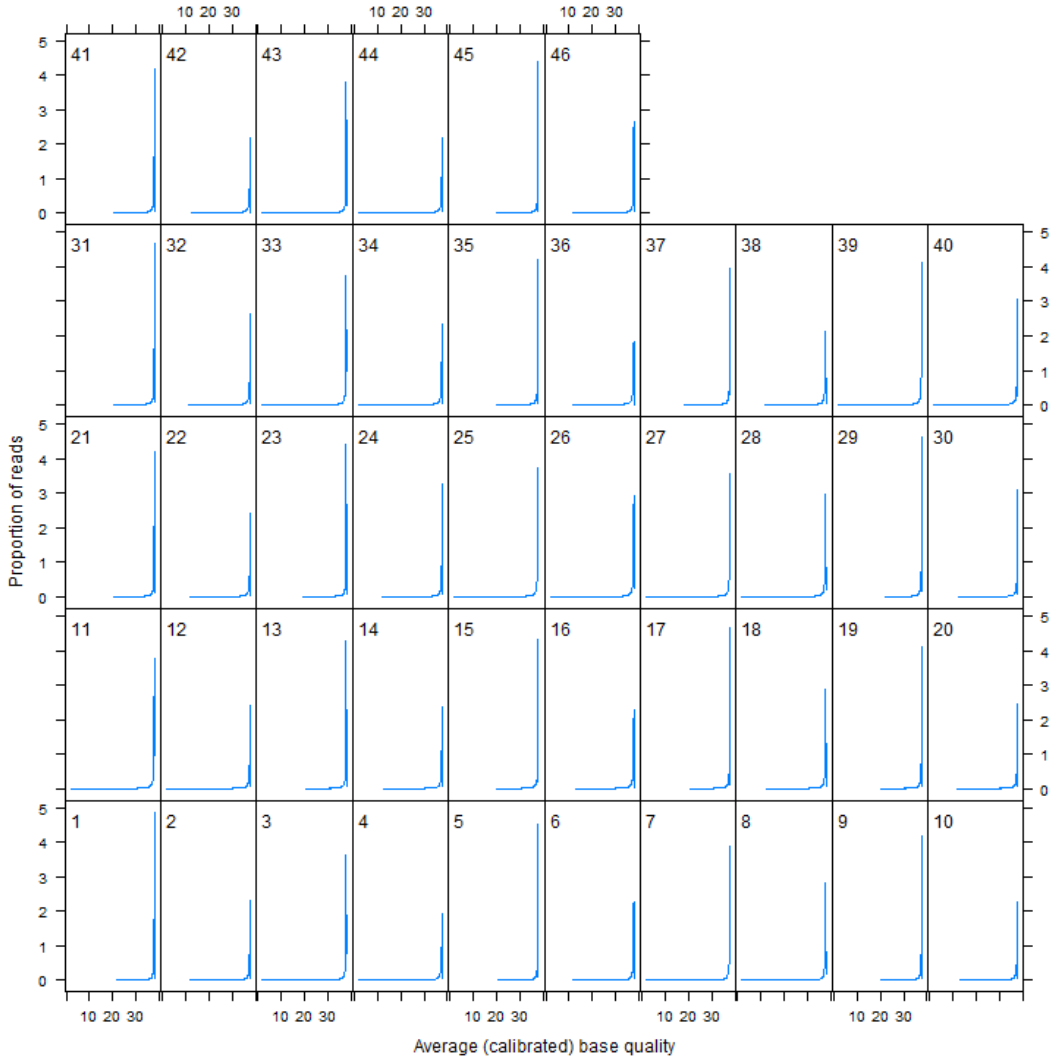
Base call frequency over all reads. Base frequencies should accurately reflect the frequencies of the regions sequenced.

```
ShortRead:::.plotNucleotideCount(qa)
```



Overall read quality. Lanes with consistently good quality reads have strong peaks at the right of the panel.

```
df <- qa[["readQualityScore"]]
ShortRead:::.plotReadQuality(df[df$type=="read",])
```



Read Distribution

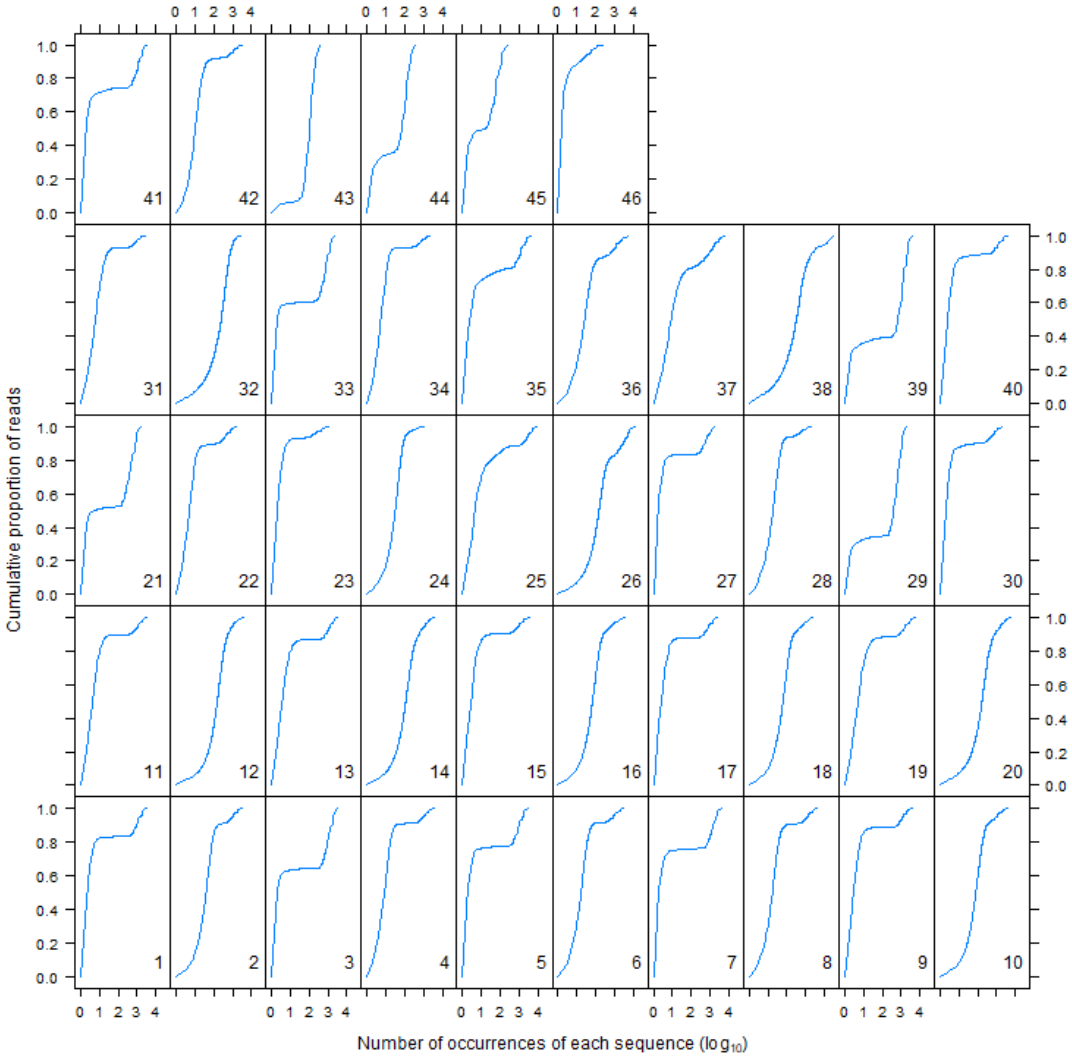
These curves show how coverage is distributed amongst reads. Ideally, the cumulative proportion of reads will transition sharply from low to high.

Portions to the left of the transition might correspond roughly to sequencing or sample processing errors, and correspond to reads that are represented relatively infrequently. 10-15%; of reads in a typical Genome Analyzer 'control' lane fall in this category.

Portions to the right of the transition represent reads that are over-represented compared to expectation. These might include inadvertently sequenced primer or adapter sequences, sequencing or base calling artifacts (e.g., poly-A reads), or features of the sample DNA (highly repeated regions) not adequately removed during sample preparation. About 5% of Genome Analyzer 'control' lane reads fall in this category.

Broad transitions from low to high cumulative proportion of reads may reflect sequencing bias or (perhaps intentional) features of sample preparation resulting in non-uniform coverage. the transition is about 5 times as wide as expected from uniform sampling across the Genome Analyzer 'control' lane.

```
df <- qa[["sequenceDistribution"]]
ShortRead:::.plotReadOccurrences(df[df$type=="read",], cex=.5)
```



Common duplicate reads might provide clues to the source of over-represented sequences. Some of these reads are filtered by the alignment algorithms; other duplicate reads might point to sample preparation issues.

```
ShortRead:::.freqSequences(qa, "read")
```

sequence	count	lane
TAAACAACTTTCAACAGTTTCGGCCGAACCTCCACCCATAGAAGCCACATGCCACGGAGGATCAAAAACCCTAG	28331	38
TAAACAACTTTCAACAGTTTCGGCCGAACCTCCACCCGTCTTATGCATCGAAAAACAACACTTAGCCTCCACAG	15557	38
TAAACAACTTTCAACAGTTTCGGCCGAACCTCCACCCGACTCAATAGCCCCACTAGGAATCTACCAAAGCTTAG	15224	38
TAAACAACTTTCAACAGTTTCGGCCGAACCTCCACCATACGTATGACCACCCTGATAAATATCCCAAATCTTAG	13433	26
TAAACAACTTTCAACAGTTTCGGCCGAGTGAGAATAGAAAGGTACCACTAAAGGAATTGCCCCC	9365	26
TAAACAACTTTCAACAGTTTCGGCCGAACCTCCACCAGGAGTAACAGGCCCCAAATCAGCAAAACAACACTTAG	9337	26

```
  GGGGGCAATTCCTTTAGTGGTACCTTTCTATTCTCACTCGGCCGAAACTGTTGAAAGTTGTTTA              9162    25
  TAAACAACTTTCAACAGTTTCGGCCGAGTGAGAATAGAAAGGTACCACTAAAGGAATTGCCCAC              6371    26
  GTGGGCAATTCCTTTAGTGGTACCTTTCTATTCTCACTCGGCCGAAACTGTTGAAAGTTGTTTA              6237    25
  TAAACAACTTTCAACAGTTTCGGCCGAACCTCCACCCACCGGCGGAACCCTCCAAACCTTATGCCCATGACCAG   6130    26
  TAAACAACTTTCAACAGTTTCGGCCGAACCTCCACCCCTCACATCCGTCACATCAGGCAGCACCCACAGCTTAG   6048    26
  TAAACAACTTTCAACAGTTTCGGCCGAGTGAGAATAGAAAGGTACCACTAAAGGAATTGCCCCC             5883    20
  TAAACAACTTTCAACAGTTTCGGCCGAACCTCCACCCCGCATACCAGCATTATCCGGAACATAACAACACTTAG   5866    26
  TAAACAACTTTCAACAGTTTCGGCCGAGTGAGAATAGAAAGGTACCACTAAAGGAATTGCCCCC             5810    38
  TAAACAACTTTCAACAGTTTCGGCCGAACCTCCACCATGCTGCCAATCATACGCAACCTTATACACACCACTAG   5799    26
  GGGGGCAATTCCTTTAGTGGTACCTTTCTATTCTCACTCGGCCGAAACTGTTGAAAGTTGTTTA              5761    19
  GGGGGCAATTCCTTTAGTGGTACCTTTCTATTCTCACTCGGCCGAAACTGTTGAAAGTTGTTTA              5728    37
  TAAACAACTTTCAACAGTTTCGGCCGAGTGAGAATAGAAAGGTACCACTAAAGGAATTGCCCCA              5490    26
  TAAACAACTTTCAACAGTTTCGGCCGAGTGAGAATAGAAAGGTACCACTAAAGGAATTGCCACC              5475    26
  TAAACAACTTTCAACAGTTTCGGCCGAGTGAGAATAGAAAGGTACCACTAAAGGAATTGCGCCC              5399    26
```

Common duplicate reads after filtering

```
  ShortRead:::.freqSequences(qa, "filtered")
```

```
NA
```

Common aligned duplicate reads are

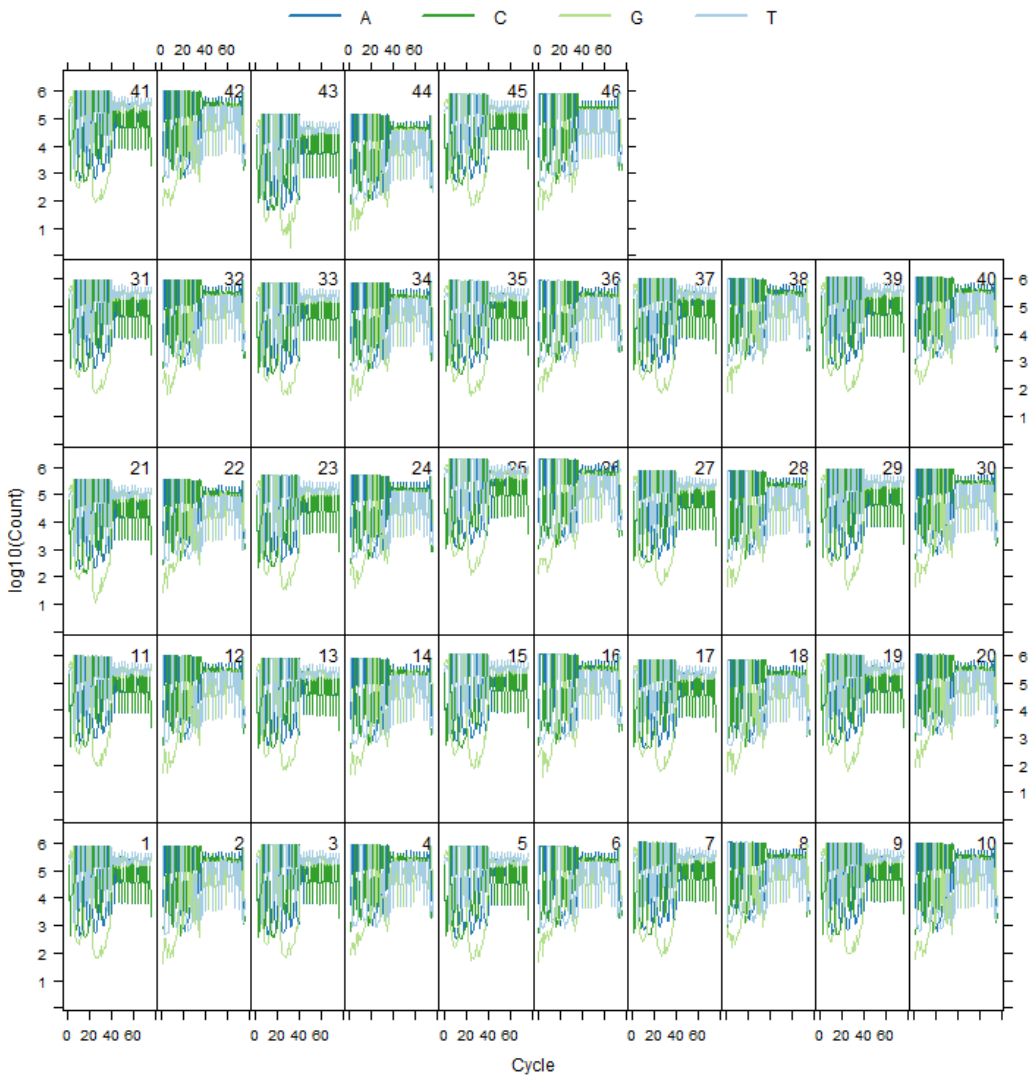
```
  ShortRead:::.freqSequences(qa, "aligned")
```

```
NA
```

Cycle-Specific Base Calls and Read Quality

Per-cycle base call should usually be approximately uniform across cycles. Genome Analyzer `control' lane results often show a deline in A and increase in T as cycles progress. This is likely an artifact of the underlying technology.

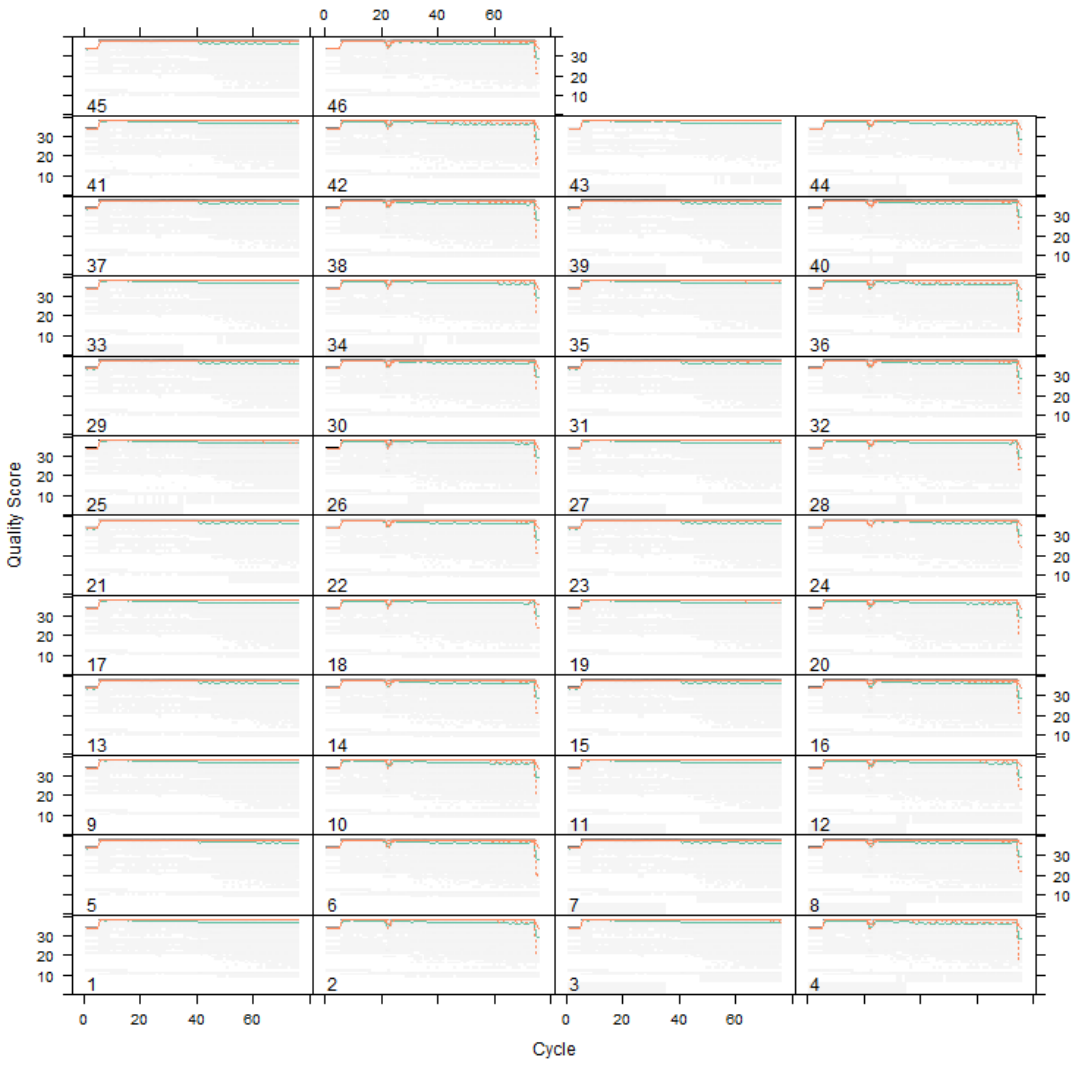
```
  perCycle <- qa[["perCycle"]]
  ShortRead:::.plotCycleBaseCall(perCycle$baseCall)
```



Per-cycle quality score. Reported quality scores are `calibrated', i.e., incorporating phred-like adjustments following sequence alignment. These typically decline with cycle, in an accelerating manner. Abrupt transitions in quality between cycles toward the end of the read might result when only some of the cycles are used for alignment: the cycles included in the alignment are calibrated more effectively than the reads excluded from the alignment.

The reddish lines are quartiles (solid: median, dotted: 25, 75), the green line is the mean. Shading is proportional to number of reads.

```
perCycle <- qa[["perCycle"]]
ShortRead:::.plotCycleQuality(perCycle$quality)
```



Adapter Contamination

Adapter contamination is defined here as non-genetic sequences attached at either or both ends of the reads. The 'contamination' measure is the number of reads with a right or left match to the adapter sequence over the total number of reads. Mismatch rates are 10% on the left and 20% on the right with a minimum overlap of 10 nt.

```
ShortRead:::.ppnCount(qa[["adapterContamination"]])
```

```
  contamination
1    NA
2    NA
```

```
3     NA
4     NA
5     NA
6     NA
7     NA
8     NA
9     NA
10    NA
11    NA
12    NA
13    NA
14    NA
15    NA
16    NA
17    NA
18    NA
19    NA
20    NA
21    NA
22    NA
23    NA
24    NA
25    NA
26    NA
27    NA
28    NA
29    NA
30    NA
31    NA
32    NA
33    NA
34    NA
35    NA
36    NA
37    NA
38    NA
39    NA
40    NA
41    NA
42    NA
43    NA
44    NA
45    NA
46    NA
```

Thu Apr 12 11:42:49 2018; ShortRead v. 1.36.0
Report template: Martin Morgan

Appendix E Barcode validation of Illumina libraries

	with 0 mismatch	with 1 mismatch
1p-I_S1_L001_R1_001	802,629 (97.3%)	22,578 (2.7%)
1p-I_S1_L001_R2_001	802,629 (97.3%)	22,578 (2.7%)
2p-I_S2_L001_R1_001	865,911 (98.2%)	15,633 (1.8%)
2p-I_S2_L001_R2_001	865,911 (98.2%)	15,633 (1.8%)
3p-I_S3_L001_R1_001	788,147 (97.9%)	17,210 (2.1%)
3p-I_S3_L001_R2_001	788,147 (97.9%)	17,210 (2.1%)
4p-I_S4_L001_R1_001	1,060,519 (97.8%)	23,809 (2.2%)
4p-I_S4_L001_R2_001	1,060,519 (97.8%)	23,809 (2.2%)
5p-I_S5_L001_R1_001	996,571 (98.6%)	14,068 (1.4%)
5p-I_S5_L001_R2_001	996,571 (98.6%)	14,068 (1.4%)
6p-I_S6_L001_R1_001	921,042 (98.2%)	17,106 (1.8%)
6p-I_S6_L001_R2_001	921,042 (98.2%)	17,106 (1.8%)
7p-I_S7_L001_R1_001	751,958 (98.4%)	12,153 (1.6%)
7p-I_S7_L001_R2_001	751,958 (98.4%)	12,153 (1.6%)
8p-I_S8_L001_R1_001	1,086,837 (98.6%)	15,724 (1.4%)
8p-I_S8_L001_R2_001	1,086,837 (98.6%)	15,724 (1.4%)
9p-I_S9_L001_R1_001	680,744 (96.9%)	21,516 (3.1%)
9p-I_S9_L001_R2_001	680,744 (96.9%)	21,516 (3.1%)
10p-I_S10_L001_R1_001	1,036,905 (98%)	21,266 (2%)
10p-I_S10_L001_R2_001	1,036,905 (98%)	21,266 (2%)
11p-I_S11_L001_R1_001	344,002 (97.5%)	8,720 (2.5%)
11p-I_S11_L001_R2_001	344,002 (97.5%)	8,720 (2.5%)
7-1-2017_S12_L001_R1_001	487,185 (97.9%)	10,534 (2.1%)
7-1-2017_S12_L001_R2_001	487,185 (97.9%)	10,534 (2.1%)
7-2-2017_S13_L001_R1_001	2,044,619 (98.5%)	31,679 (1.5%)
7-2-2017_S13_L001_R2_001	2,044,619 (98.5%)	31,679 (1.5%)
R0-rnd-1_S14_L001_R1_001	714,107 (98.2%)	13,111 (1.8%)
R0-rnd-1_S14_L001_R2_001	714,107 (98.2%)	13,111 (1.8%)
R1-rnd-1_S15_L001_R1_001	884,398 (98.4%)	14,005 (1.6%)
R1-rnd-1_S15_L001_R2_001	884,398 (98.4%)	14,005 (1.6%)
R2-rnd-1_S16_L001_R1_001	958,586 (98.5%)	14,251 (1.5%)
R2-rnd-1_S16_L001_R2_001	958,586 (98.5%)	14,251 (1.5%)
ctrl-rnd-1_S17_L001_R1_001	730,413 (97%)	22,402 (3%)
ctrl-rnd-1_S17_L001_R2_001	730,413 (97%)	22,402 (3%)
R0-rnd-2_S18_L001_R1_001	864,253 (98%)	17,848 (2%)
R0-rnd-2_S18_L001_R2_001	864,253 (98%)	17,848 (2%)
R1-rnd-2_S19_L001_R1_001	1,015,784 (97.7%)	23,982 (2.3%)
R1-rnd-2_S19_L001_R2_001	1,015,784 (97.7%)	23,982 (2.3%)
R2-rnd-2_S20_L001_R1_001	1,163,382 (97.9%)	24,845 (2.1%)
R2-rnd-2_S20_L001_R2_001	1,163,382 (97.9%)	24,845 (2.1%)
ctrl-rnd-2_S21_L001_R1_001	1,025,779 (98.5%)	15,214 (1.5%)
ctrl-rnd-2_S21_L001_R2_001	1,025,779 (98.5%)	15,214 (1.5%)
naive_S22_L001_R1_001	136,524 (97.8%)	3,062 (2.2%)
naive_S22_L001_R2_001	136,524 (97.8%)	3,062 (2.2%)
once-amp_S23_L001_R1_001	762,778 (98.3%)	13,182 (1.7%)
once-amp_S23_L001_R2_001	762,778 (98.3%)	13,182 (1.7%)

Appendix F Derivations

Given the gamma distribution

$$p(x|a,b) = \frac{x^{a-1}}{\Gamma(a)b^a}\exp\left(-\frac{x}{b}\right), \tag{F.1}$$

Minka (2002) demonstrated that the log-likelihood of (F.1) can be written as

$$\ln p(D|a,\hat{b}) = n(a-1)\overline{\ln x} - n\ln\Gamma(a) - na\ln\bar{x} + na\ln a - na. \tag{F.2}$$

Taking the derivative of (F.2) and solving for the maximum value of a yields

$$\frac{d}{da}\ln\Gamma(a) - \ln a = \Psi(a) - \ln a = \overline{\ln x} - \ln\bar{x} = \ln\bar{x}_g - \ln\bar{x}_a, \tag{F.3}$$

where $\Psi(a)$ is the digamma function and $\bar{x}_g, \bar{x}_a$ are the geometric and arithmetic means respectively. From (F.3), we obtain the relation

$$\frac{\bar{x}_g}{\bar{x}_a} = \exp(\Psi(a) - \ln a). \tag{F.4}$$

For $a < 1$, it follows that

$$\frac{\bar{x}_g}{\bar{x}_a} < \exp(\Psi(1) - \ln 1) = e^{\gamma}, \tag{F.5}$$

where γ is the Euler-Mascheroni constant. Thus,

$$\frac{\bar{x}_g}{\bar{x}_a} \lessapprox 0.56. \tag{F.6}$$