Design and characterization of novel immunogens for AIDS vaccine development and evaluation of a sample inference method for NGS Illumina amplicon data

by

Heliang Shi

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY

Co-Majors: Toxicology; Statistics

Program of Study Committee:
Michael Cho, Co-Major Professor
Karin Dorman, Co-Major Professor
James Roth
Qijing Zhang
Chong Wang
Joan Cunnick

Iowa State University
Ames, Iowa
2017

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my advisors, Dr. Michael Cho and Dr. Karin Dorman, for giving me the opportunity to work on these challenging, but interesting topics. This dissertation would not have been finished without their encouragement and guidance. I would also like to thank Dr. James Roth, Dr. Qijing Zhang, Dr. Chong Wang, and Dr. Joan Cunnick for sharing their knowledge and expertise as members of my program of study committee.

Thanks to Dr.Saikat Banerjee, Dr. Marisa Banasik, Dr. Yali Qin, Aditi Agrawal, Andrew Harley, and Xiyu Peng for their insightful ideas and contributions to my thesis projects. I am also thankful to all current and past members of the Cho lab for their kind help. I would also express my special thanks to Dr. Marisa Banasik for her critical review of my thesis.

Last, but not the least, I would like to thank my family for their unconditional love and support. Most importantly, I am grateful to my wife Xinxin and daughter Olivia, who have enriched my life, for their tremendous ongoing support.

ABSTRACT

Since the beginning of the AIDS pandemic, an estimated 78 million people have become infected and 35 million people have died from AIDS-related illnesses. Despite the existence of effective antiretroviral therapy, 1.1 million people died of AIDS-related causes in 2015. A vaccine that could induce broadly neutralizing antibodies (bnAbs) is hypothesized to be the most efficient way to halt the AIDS pandemic. However, the majority of attempts to elicit bnAbs with HIV-1 vaccine candidates have failed due to the extensive variability and complex immune-evasion strategies of HIV-1. Recent advances in the isolation of bnAbs from HIV-1 infected individuals have revived interest in vaccine development. The membrane proximal external region (MPER) of gp41 and the CD4 binding site (CD4bs) of gp120 have become attractive targets for vaccine development because they contain highly conserved epitopes recognized by some of the broadest neutralizing antibodies. Here, we have designed and characterized multiple immunogens and vaccine strategies to induce bnAbs targeted to MEPR or CD4bs. Our findings indicate that 1) neighboring domains influence the immunogenicity of gp41 MPER, and 2) priming with a small gp41 or gp120 immunogen, then subsequently boosting with larger and more native immunogens, may have the potential to elicit antibodies towards the appropriate neutralizing epitopes.

Illumina amplicon sequencing is an important tool for the identification and quantification of species or variants in metagenomics studies, but sequencing errors make it challenging to correctly identify the authentic differences. Many denoising

algorithms have been developed, but most ignore the quality scores or compress that data. We developed *ampliclust*, an error modeling approach using uncompressed sequences and quality scores to infer samples in Illumina amplicon data. Our approach showed better accuracy than the popular denoising tool DADA2 when data are not well separated.

CHAPTER 1

GENERAL INTRODUCTION

1.1 Dissertation Organization

This dissertation is divided into seven chapters. Chapter 1 is the “General Introduction” describing the history of HIV-1 pandemic, the function of envelope glycoprotein, the development of envelope glycoprotein based HIV-1 vaccines. This is followed by a review on next generation sequencing techniques, the sequencing platforms, the error profiles, and error correction algorithms.

Chapter 2 is a published manuscript titled “Modulating immunogenic properties of HIV-1 gp41 membrane-proximal external region by destabilizing six-helix bundle structure”. This manuscript evaluates the immunogenic properties of four gp41 putative fusion intermediates. The contribution of each author is as follows: Heliang Shi produced HR1-Δ10-54K and HR1-Δ17-54K antigens and Saikat Banerjee produced HR1-AA-54Q and HR1-EE-54Q antigens. Heliang Shi and Saikat Banerjee performed all experiments together; Habtom Habte and Yali Qin provided help in the antigen design; Heliang Shi, Saikat Banerjee, and Michael Cho wrote and revised the manuscript.

Chapter 3 is a published manuscript titled “Evaluation of a novel multi-immunogen vaccine strategy for targeting 4E10/10E8 neutralizing epitopes on HIV-1 gp41 membrane proximal external region”. This manuscript evaluates a novel multi-immunogen vaccine strategy to induce 4E10/10E8-like broadly neutralizing antibodies. The contribution of each author is as follows: Heliang Shi generated 28x3 antigens,

immunized the rabbits, and analyzed the antibody responses; Saikat Banerjee immunized the multi-immunogen vaccine group and analyze the antibody responses; Saikat Banerjee, Marisa Banasik, and Hojin Moon generated the monoclonal antibodies; Andrew Harley helped with the epitope mapping analysis;William Lees and Adrian Shepherd performed the NGS analysis; Heliang Shi, Saikat Banerjee, Marisa Banasik, and Michael Cho wrote and revised the manuscript.

Chapter 4 is presented as a manuscript in preparation titled “Evaluation of a novel rapid heterologous prime-boost strategy for targeting CD4bs neutralizing epitopes on HIV-1 gp120”. This manuscript evaluates a heterologous prime-boost vaccine strategy using rapid immunization schedule to induce CD4bs directed broadly neutralizing antibodies. The contribution of each author is as follows: Heliang Shi generated the antigens, immunized the rabbits, and analyzed the cross-reactive antibodies; Heliang Shi and Saikat Banerjee performed the competitive ELISA; Heliang Shi and Aditi Agrawal performed the peptide ELISA; Heliang Shi and Michael Cho wrote and revised the manuscript.

Chapter 5 is presented as a manuscript in preparation titled “Evaluation of a multi-immunogen vaccine strategy for targeting CD4bs neutralizing epitopes on HIV-1 gp120”. This manuscript evaluates a vaccine strategy applying a sequential and phased mannered immunization approach with related but antigenically distinct immunogens to induce CD4bs directed broadly neutralizing antibodies. The contribution of each author is as follows: Heliang Shi produced the protein antigens, conducted the rabbit immunization, and performed the cross-reactivity analysis of antibodies; Heliang Shi

and Aditi Agrawal conducted peptide ELISA; Heliang Shi and Saikat Banerjee analyzed the competitive assay; Heliang Shi and Michael Cho wrote and revised the manuscript.

Chapter 6 is presented as a manuscript in preparation titled “A Novel Sample Inference Method For Illumina Amplicon Data”. This manuscript describes a model-based approach for the inference of Illumina amplicon data using uncompressed quality information. The contribution of each author is as follows: Heliang Shi and Karin Dorman developed the model; Heliang Shi and Xiyu Peng performed the programming; Heliang Shi evaluated the performance of the model; Heliang Shi and Karin Dorman wrote and revised the manuscript.

Chapter 7 summarizes the general conclusions along with future directions. The appendix contains a manuscript in preparation titled “HIV-1 gp41-HR1-HR2 six-helix bundle as a Novel Fusion Protein Partner for Efficient Recombinant Protein Expression”. This manuscript describes a method to improve the recombinant protein expression. Heliang Shi performed all the experiments; Andrew Harley helped with the purification of HR1-6H-HR2-TH-54 protein. Heliang Shi and Michael Cho wrote and revised the manuscript.

1.2 HIV-1 vaccine research

1.2.1 The HIV-1 pandemic

The first cases of Acquired Immune Deficiency Syndrome (AIDS) were identified in 1981 among gay men in the United States. The cause of this disease will officially be known as Human Immunodeficiency Virus Type 1 (HIV-1) originated in non-human

primates. (1, 2) HIV-2 was identified in 1986 among patients in Africa, although it was morphologically similar to HIV-1, they were proved to be antigenically distinct. (3, 4) Simian immunodeficiency viruses (SIVs) were found to be the source of HIV, but SIVs are non-pathogenic in their natural hosts. (5)The cross-species transmission of SIVs from non-human primates to humans is responsible for the emerging infections. Since HIV-2 is in general less virulent and largely restricted to western Africa, HIV-1 has been the main target for HIV research efforts.(6)

HIV-1 has been divided into four distinct clades, groups M (major), O (outlier), N (non-M, non-O), and P (pending the identification of further human cases). Group M was the fist discovered lineage, represents the pandemic form of HIV-1 and has infected millions of people worldwide. Group O (discovered in 1990) represents less than 1% of total HIV-1 infections and is mainly found in Cameroon and its neighboring countries. Group N (discovered in 1998) is less prevalent than group O and has been found only in Cameroon. Group P (discovered in 2009) has so far only been identified in a Cameroonian woman living in France. These phylogenic groups are further divided into subtypes. For instance, group M has 11 subtypes A to K and group O has 9 subtypes. (7-9)

According to UNAIDS, there were approximately 36.7 million HIV-1 infected individuals at the end of 2015. Since the beginning of the HIV pandemic, about 78 million people have become infected and 35 million people have died from AIDS-related illnesses (https://www.avert.org/global-hiv-and-aids-statistics#footnote4_l4cbfua). Although antiretroviral therapy has been effective in controlling HIV, it cannot cure or

eradicate the virus completely.(10, 11) A safe and effective vaccine remains the most efficient way to prevent HIV-1 infection.

1.2.2 HIV-1 envelope

The HIV-1 envelop glycoproteins (Env) consists of the external protein gp120 and the transmembrane protein gp41, which are generated by gp160 precursor through furin-like protease cleavage. (12, 13) Three gp120 subunits non-covalently associate with three gp41 subunits to form trimeric spikes on the virion surface. The Env plays a critical role in HIV-1 entry. To infect cells, gp120 binds to the primary cellular receptor (CD4), which initiates conformational changes in gp120 and exposes the binding domain of co-receptors (CCR5 or CXCR4). Binding of gp120 to a co-receptor would trigger the conformational changes in gp41 that allow the fusion peptide to insert into target membranes. The heptad repeat regions of gp41 fold into a six-helix bundle that drives the fusion of viral and cellular membranes.(14)

The elicitation of broadly neutralizing antibodies (bnAbs) is a major goal of HIV-1 vaccine development. The Env trimer on the surface of HIV-1 virions is the only target for bnAbs. A large number of potent bnAbs have been isolated from infected patients in recent years. They have been placed into five categories based on the location of their epitopes on the viral spike. These are the CD4 biding site (CD4bs)-directed antibodies, variable region 1 and variable region 2 (V1/V2)-directed antibodies, glycan variable region 3 (V3)-directed antibodies, membrane proximal external region (MPER)-directed antibodies, and the newly discovered glycan associated bridging region-directed antibodies.(15) The bnAbs could efficiently inhibit virus binding or preventing

conformational changes needed for virus entry. The great neutralizing capabilities of these bnAbs may be attributed to some of their unique features including high rate of somatic hypermutation and long heavy chain complementarity determining region 3 (HCDR3). (16-20) The bnAbs were induced only in a minority of HIV-1 infected individuals several years post infection. Although protection by passive transfer of bnAbs has been reported,(21-23) there is currently no effective immunization regimens could elicit bnAbs.

1.2.3 HIV-1 clinical trials

Although hundreds of vaccine trials have been conducted so far, (24) only five vaccines have advanced to Phase IIb and Phase III clinical trials. The VAX003 (conducted in Thailand) and VAX004 (conducted in North America and Europe) were the world's first phase III trials conducted between 2005 and 2007. They tested bivalent recombinant gp120 vaccines using subtypes B/E and B/B, respectively. However, they failed to elicit bnAbs and could not efficiently prevent HIV-1 acquisition in the vaccinated groups.(25, 26) The STEP (conducted in America and Australia) and Phambili (conducted in South Africa) Phase IIb trials were conducted between 2005 and 2007. They evaluated an adenovirus serotype 5 (Ad5) vector based vaccine expressing multiple HIV antigens comprising group specific antigen (gag), polymerase (pol), and negative regulation factor (nef) genes. Disappointingly, both trials failed to demonstrate any protection against HIV-1 infection.(27, 28) The HVTN505 was the most recent trial conducted between 2009 and 2013. It evaluated the efficacy of a prime-boost regimen consisting of DNA vector encoding HIV-1 clade B Gag, Pol and Nef proteins along with

Env proteins (subtypes A, B and C) followed by a recombinant Ad5 boost. (29) This trial failed to demonstrate any protective effect.

So far, the RV144 trial was the only candidate vaccine that displayed protection against HIV-1 infection. It was conducted between 2003 and 2006 in Thailand. This trial evaluated a prime-boost vaccine regimen consisting of canarypox vector expressing Env, Gag and protease (ALVAC-HIV [vCP1521]), and boosted by bivalent recombinant gp120 (AIDSVAX B/E).(30) This study demonstrated a modest vaccine efficacy of 31% but no bnAbs were induced. Further analysis have revealed that IgG antibodies induced against V1/V2 loops of gp120 correlated with the reduced risk of HIV-1 acquisition, IgA antibodies elicited against Env associated with the enhanced risk, and the antibody-dependent cell-mediated cytotoxicity (ADCC) may have contributed to the protection. (31, 32)

The findings of RV144 trial imply that an effective HIV-1 vaccine needs to induce both humoral and cellular immunity. The modest success of the RV144 trial suggests that a vaccine to prevent the establishment of HIV-1 infection is possible.

1.2.4 Gp41-based vaccine development

Elicitation of bnAbs is a major HIV-1 vaccine goal. As described earlier, the MPER of gp41 is the only neutralizing epitopes present on gp41. It is a very attractive vaccine target because it is a linear, unglycosylated, highly conserved, and tryptophan-rich region playing a crucial role in the membrane fusion step of HIV-1 virus entry.(33) Five anti-MPER antibodies have been discovered so far, including 2F5, Z13e1, 4E10, 10E8, and m66.6. They all target overlapping linear epitopes within MPER. 2F5 and

m66.6 recognize the N terminal of MEPR, while 4E10, Z13e1, and 10E8 target the C terminal. (34-38) Among all anti-MPER bnAbs, 10E8 is the broadest and most potent bnAb.

A variety of studies have been performed to induce anti-MPER bnAbs. Early immunogens tested were MPER peptide-based vaccines containing neutralizing epitopes alone or coupling them to carrier proteins. (39-41) They failed to show neutralizing activity, possibly due to lack of appropriate conformation of the MPER. Different approaches have been applied to display MPER epitopes, including chimeric viruses, engineered scaffolds, virus-like particles, and liposomes.(42-47) However, all of these studies failed to induce neutralizing antibodies, possibly due to either the presence of immunodominant regions outside the MPER or the improper MPER conformation.

The weak to modest neutralizing activities have been reported in a few studies. One prime-boost vaccine regimen consisting of rhinoviruses expressing the 2F5 epitopes and boosted by similar epitopes coupled to carrier proteins have demonstrated induction of modest neutralizing activities in immunized guinea pigs. (48) Similar results have been observed from another study using a rhinovirus expressing the 4E10 epitope (45). In one study, rabbits, which were primed with MPER 60-mer antigens and boosted with gp160 DNA, induced weak neutralizing activities against HIV-2/HIV-1 chimeric viruses. (49) One study using virus like particles displaying heamagglutinin mutation/gp41 chimeric proteins or DNA have induced weak neutralizing antibodies in immunized guinea pigs. (50) Thus far, the best MPER-specific neutralization activities have been observed from the guinea pigs immunized with gp41 fusion intermediate

displayed on liposomal. (51) Although weak to modest neutralizing activities have been induced, there is currently no vaccine could elicit anti-MPER bnAbs. Therefore, additional efforts are needed to develop effective gp41-based HIV-1 vaccines.

1.2.5 Gp120-based vaccine development

As described earlier, bnAbs have been induced against five distinct regions on the Env. Three of these neutralizing epitopes are located within gp120. Thus far, a variety of gp120-based vaccines have been evaluated to induce such bnAbs.

Early studies evaluating the V3 loop peptide-based vaccines induced neutralization activities but limited to Tier-1 viruses only, possibly due to the improper conformational epitopes. (52-55) Epitope scaffold immunogens transplanting epitopes, including V3 loop, CD4bs, and V3 glycan epitopes, onto heterologous protein scaffolds failed to induce an effective neutralization activities in animals. (41, 56, 57) The recombinant gp120 immunogens were evaluated in clinical trials, but met with limit success.

Since the immune responses against the inner domain (ID) of gp120 are generally non-neutralizing and immunodominant, some studies attempted to develop outer domain (OD) based immunogens.(58) However, most of the studies have failed to induce neutralization activities. (59-62) Recently, one study reported by our group has shown that both OD and ODx3 immunogens, based on an M group consensus sequence (MCON6), induced cross-reactive nAb responses against the clade B, C, and AE Tier-1 viruses, which was much better than previous studies.(63) Another study has designed engineered OD (eOD) immunoges to target bnAbs germline precursors. The

first optimized immunogen eOD-GT6, which was fused to Lumazine Synthase that self-assembles into 60mer nanoparticles, (64) failed to induce neutralizing responses according to our recent study. Another improved version eOD-GT8 has induced neutralizing activity against autologous viruses and the Tier-2 viruses lacking the glycan at position 276 in knock-in mice that carried the mature heavy chain of 3BNC60.(65)

The Env trimer immunogens have been designed to mimic the native spike. The stabilized soluble Evn trimers comprising of gp120 and gp41 ectodomain induced better neutralizing responses than gp120, but still failed to induce bnAbs,(66-69) probably due to the misfolding of trimer. Another approach has been used to stabilize the Env trimer by introducing a disulfide bond (SOS) between gp120 and gp41 and a Ile-to-Pro mutation (I559P). (70) Several SOSIP based trimeric immunogens were evaluated, BG505 SOSIP.664 trimers is one of the improved immuogens. (71-73) It could mimic the native spikes and induce neutralizing responses against Tier 1 viruses and autologous Tier 2 viruses, but not against heterologous Tier 2 viruses. (74)

Although a lot of efforts have been made to develop gp120-based vaccine, little success has been achieved. The main challenges are high sequence variability, extensive glycosylation, the presence of decoy immunodominant epitopes, and highly conformational nature of the bnAbs epitopes. Additional work is needed to overcome these challenges and develop an effective HIV-1 vaccine.

1.3 Denoising of Next-generation sequencing data

1.3.1 Next-generation sequencing

Deoxyribonucleic acid (DNA) and ribonucleic acid (RNA) are major biological macromolecules that are essential for all known forms of life. Sequencing refers to the process of determining the precise order of nucleotides within a DNA molecule in genome sequencing or RNA molecule in transcriptome sequencing; it has significantly accelerated biological research and discovery. Sanger sequencing was developed in 1970s by Frederick Sanger and has been widely used since then. The advent of Sanger sequencing gave a huge boost to sequencing field, it could provide high accurate sequencing results for long reads .(75) However, Sanger sequencing is labor intensive and high cost. The Human Genome Project took 13 years and cost 2.7 billion dollars. Its completion triggered the development of next-generation sequencing (NGS) technologies that are able to generate high-throughput sequences at a much lower cost.

Affordable high-throughput NGS has revolutionized genomics, making large scale of studies of genetic variation and the sequencing projects that involve large number of organisms feasible (76). A series of NGS technologies have been developed, the major platforms are 454 pyrosequencing, Illumina/Solexa, SOLiD, Ion Torrent Personal Genome Machine (PGM), and Pacific Biosciences (PacBio) single molecule real-time sequencing (SMRT). Compared to Sanger sequencing, the NGS technologies simultaneously monitor incorporations on millions of templates to enhance the sequencing throughput and apply imaging or semiconductor technologies to make rapid and highly automated determination of nucleotides. These NGS techniques utilize DNA polymerase or ligase to synthesize deoxynucleotides (dNTPs), a fraction of which is

labeled. The signals released by bases synthesis are captured by the sequencers and translated to reads. Errors could be introduced during the imperfected library preparation and sequencing processes.

As new technologies emerge, new challenges will inevitably arise. NGS platforms produce sequences with shorter length and higher error rate compared to Sanger sequencing. The major hurdle to overcome for NGS is the high error rate, because long-read sequencing could overcome the length limitation, although it would be more expensive and lower throughput than other platforms(77). The presence of sequencing errors could severely influence the downstream analysis of variant calling (78), genome assembly(79), transcript quantification and so forth. Thus, there is a compelling need for developing new algorithms for correcting errors in sequencing data.

1.3.2 Denoising of NGS amplicon data

Each of the platforms tends to have its own error profiles that differ slightly from one another due to the differing chemistries. Possibly due to the dominance of Illumina sequencers in the market, most error correction methods so far have been designed for substitution errors. This dissertation mainly focuses on the development of statistical method to denoise Illumina amplicon data. In the remainder of this chapter, we will briefly discuss some relevant prior knowledge.

The amplicon-specific error-correction methods were initially developed for pyrosequencing platform (80, 81). More recently, a number of denoising algorithms have been developed for Illumina including UNOISE(82), MED(83), IPED(84), UNOISE2 (85), and DADA2 (86). The IPED utilizes an artificial intelligent classifier to

detect and correct the erroneous positions. The training data is required for this machine learning based method. MED divides amplicon reads into partitions and iteratively partitions dataset using only the informative nucleotide positions; it has been used widely to identify fine-scale variation. UNOISE exploits unique sequence abundances to conduct error correction. Among all the available algorithm,s DADA2 has been claimed as the best peer-reviewed methods for denoising Illumina amplicon data although UNOISE2 has shown its competitive performance. Both DADA2 and UNOISE2 group all amplicon reads with the same sequence into unique sequences with an associated abundance, but their partitioning algorithms are different. UNOISE2 applies a one-pass clustering algorithm ignoring quality scores and has only two parameters. DADA2 utilizes a Poisson model-based iterative divisive partitioning clustering algorithm using quality score and has hundreds of parameters.

Although many advances have been made for denoising Illumina amplicon datasets, there is still ample room for improvement. Most Illumina denoiser algorithms often discard quality scores completely or turn to compression. Quality scores are related to error rates and informative for error correction, underutilizing them may influence the downstream analysis and result to undesirable consequences. Some algorithms assign all reads of the same sequence to the same cluster, which may influence the clustering accuracy, because some sites within the reads are misreads of the true amplicon nucleotides which is possibly due to the low quality score of those sites.

Inspired by DADA2, we proposed *ampliclust*, an error modeling approach using uncompressed data to correct errors and infer samples in Illumina amplicon data. We

model the reads as independent components from the model. We compare the error correction performance of our method against DADA2 on the Mock, real HIV, and simulated data sets.

1.4 References

1. **Greene WC**. 2007. A history of AIDS: looking back to see ahead. Eur J Immunol **37 Suppl 1**:S94–102.

2. **Coffin J**, **Haase A**, **Levy JA**, **Montagnier L**, **Oroszlan S**, **Teich N**, **Temin H**, **Toyoshima K**, **Varmus H**, **Vogt P**. 1986. What to call the AIDS virus? Nature **321**:10.

3. **Clavel F**, **Guétard D**, **Brun-Vézinet F**, **Chamaret S**, **Rey MA**, **Santos-Ferreira MO**, **Laurent AG**, **Dauguet C**, **Katlama C**, **Rouzioux C**. 1986. Isolation of a new human retrovirus from West African patients with AIDS. Science **233**:343–346.

4. **Guyader M**, **Emerman M**, **Sonigo P**, **Clavel F**, **Montagnier L**, **Alizon M**. 1987. Genome organization and transactivation of the human immunodeficiency virus type 2. Nature **326**:662–669.

5. **Huet T**, **Cheynier R**, **Meyerhans A**, **Roelants G**, **Wain-Hobson S**. 1990. Genetic organization of a chimpanzee lentivirus related to HIV-1. Nature **345**:356–359.

6. **Hahn BH**, **Shaw GM**, **De Cock KM**, **Sharp PM**. 2000. AIDS as a zoonosis: scientific and public health implications. Science **287**:607–614.

7. **Buonaguro L**, **Tornesello ML**, **Buonaguro FM**. 2007. Human immunodeficiency virus type 1 subtype distribution in the worldwide epidemic: pathogenetic and therapeutic implications. journal of virology **81**:10209–10219.

8. **Hemelaar J**, **Gouws E**, **Ghys PD**, **Osmanov S**, **WHO-UNAIDS Network for HIV Isolation and Characterisation**. 2011. Global trends in molecular epidemiology of HIV-1 during 2000-2007. AIDS **25**:679–689.

9. **Hemelaar J**. 2012. The origin and diversity of the HIV-1 pandemic. Trends Mol Med **18**:182–192.

10. **Arts EJ**, **Hazuda DJ**. 2012. HIV-1 antiretroviral drug therapy. Cold Spring Harb Perspect Med **2**:a007161.

11. **Spinner CD**, **Boesecke C**, **Zink A**, **Jessen H**, **Stellbrink H-J**, **Rockstroh JK**, **Esser S**. 2016. HIV pre-exposure prophylaxis (PrEP): a review of current knowledge of oral systemic HIV PrEP in humans. Infection **44**:151–158.

12. **Fennie C**, **Lasky LA**. 1989. Model for intracellular folding of the human immunodeficiency virus type 1 gp120. journal of virology **63**:639–646.

13. **Moulard M**, **Decroly E**. 2000. Maturation of HIV envelope glycoprotein precursors by cellular endoproteases. Biochim Biophys Acta **1469**:121–132.

14. **Melikyan GB**. 2008. Common principles and intermediates of viral protein-mediated fusion: the HIV-1 paradigm. Retrovirology **5**:111.

15. **Mouquet H**. 2014. Antibody B cell responses in HIV-1 infection. Trends Immunol **35**:549–561.

16. **Corti D**, **Langedijk JPM**, **Hinz A**, **Seaman MS**, **Vanzetta F**, **Fernandez-Rodriguez BM**, **Silacci C**, **Pinna D**, **Jarrossay D**, **Balla-Jhagjhoorsingh S**, **Willems B**, **Zekveld MJ**, **Dreja H**, **O'Sullivan E**, **Pade C**, **Orkin C**, **Jeffs SA**, **Montefiori DC**, **Davis D**, **Weissenhorn W**, **McKnight A**, **Heeney JL**, **Sallusto F**, **Sattentau QJ**, **Weiss RA**, **Lanzavecchia A**. 2010. Analysis of memory B cell responses and isolation of novel monoclonal antibodies with neutralizing breadth from HIV-1-infected individuals. PLoS ONE **5**:e8805.

17. **Liao H-X**, **Lynch R**, **Zhou T**, **Gao F**, **Alam SM**, **Boyd SD**, **Fire AZ**, **Roskin KM**, **Schramm CA**, **Zhang Z**, **Zhu J**, **Shapiro L**, **NISC Comparative Sequencing Program**, **Mullikin JC**, **Gnanakaran S**, **Hraber P**, **Wiehe K**, **Kelsoe G**, **Yang G**, **Xia S-M**, **Montefiori DC**, **Parks R**, **Lloyd KE**, **Scearce RM**, **Soderberg KA**, **Cohen M**, **Kamanga G**, **Louder MK**, **Tran LM**, **Chen Y**, **Cai F**, **Chen S**, **Moquin S**, **Du X**, **Joyce MG**, **Srivatsan S**, **Zhang B**, **Zheng A**, **Shaw GM**, **Hahn BH**, **Kepler TB**, **Korber BTM**, **Kwong PD**, **Mascola JR**, **Haynes BF**. 2013. Co-evolution of a broadly neutralizing HIV-1 antibody and founder virus. Nature **496**:469–476.

18. **Walker LM**, **Huber M**, **Doores KJ**, **Falkowska E**, **Pejchal R**, **Julien J-P**, **Wang S-K**, **Ramos A**, **Chan-Hui P-Y**, **Moyle M**, **Mitcham JL**, **Hammond PW**, **Olsen OA**, **Phung P**, **Fling S**, **Wong C-H**, **Phogat S**, **Wrin T**, **Simek MD**, **Protocol G Principal Investigators**, **Koff WC**, **Wilson IA**, **Burton DR**, **Poignard P**. 2011. Broad neutralization coverage of HIV by multiple highly potent antibodies. Nature **477**:466–470.

19. **Kwong PD**, **Mascola JR**. 2012. Human antibodies that neutralize HIV-1: identification, structures, and B cell ontogenies. Immunity **37**:412–425.

20. **Mouquet H**, **Nussenzweig MC**. 2013. HIV: Roadmaps to a vaccine. Nature **496**:441–442.

21. **Balazs AB**, **Chen J**, **Hong CM**, **Rao DS**, **Yang L**, **Baltimore D**. 2011. Antibody-based protection against HIV infection by vectored immunoprophylaxis. Nature **481**:81–84.

22. **Hessell AJ**, **Rakasz EG**, **Poignard P**, **Hangartner L**, **Landucci G**, **Forthal DN**, **Koff WC**, **Watkins DI**, **Burton DR**. 2009. Broadly neutralizing human anti-HIV antibody 2G12 is effective in protection against mucosal SHIV challenge even at low serum neutralizing titers. PLoS Pathog **5**:e1000433.

23. **Mascola JR**, **Stiegler G**, **VanCott TC**, **Katinger H**, **Carpenter CB**, **Hanson CE**, **Beary H**, **Hayes D**, **Frankel SS**, **Birx DL**, **Lewis MG**. 2000. Protection of

macaques against vaginal transmission of a pathogenic HIV-1/SIV chimeric virus by passive infusion of neutralizing antibodies. Nat Med **6**:207–210.

24. **Esparza J**. 2013. A brief history of the global effort to develop a preventive HIV vaccine. Vaccines **31**:3502–3518.

25. **Flynn NM**, **Forthal DN**, **Harro CD**, **Judson FN**, **Mayer KH**, **Para MF**, **rgp120 HIV Vaccine Study Group**. 2005. Placebo-controlled phase 3 trial of a recombinant glycoprotein 120 vaccine to prevent HIV-1 infection. J Infect Dis **191**:654–665.

26. **Pitisuttithum P**, **Gilbert P**, **Gurwith M**, **Heyward W**, **Martin M**, **van Griensven F**, **Hu D**, **Tappero JW**, **Choopanya K**, **Bangkok Vaccine Evaluation Group**. 2006. Randomized, double-blind, placebo-controlled efficacy trial of a bivalent recombinant glycoprotein 120 HIV-1 vaccine among injection drug users in Bangkok, Thailand. J Infect Dis **194**:1661–1671.

27. **Buchbinder SP**, **Mehrotra DV**, **Duerr A**, **Fitzgerald DW**, **Mogg R**, **Li D**, **Gilbert PB**, **Lama JR**, **Marmor M**, **Del Rio C**, **McElrath MJ**, **Casimiro DR**, **Gottesdiener KM**, **Chodakewitz JA**, **Corey L**, **Robertson MN**, **Step Study Protocol Team**. 2008. Efficacy assessment of a cell-mediated immunity HIV-1 vaccine (the Step Study): a double-blind, randomised, placebo-controlled, test-of-concept trial. Lancet **372**:1881–1893.

28. **Gray G**, **Buchbinder S**, **Duerr A**. 2010. Overview of STEP and Phambili trial results: two phase IIb test-of-concept studies investigating the efficacy of MRK adenovirus type 5 gag/pol/nef subtype B HIV vaccine. Curr Opin HIV AIDS **5**:357–361.

29. **Hammer SM**, **Sobieszczyk ME**, **Janes H**, **Karuna ST**, **Mulligan MJ**, **Grove D**, **Koblin BA**, **Buchbinder SP**, **Keefer MC**, **Tomaras GD**, **Frahm N**, **Hural J**, **Anude C**, **Graham BS**, **Enama ME**, **Adams E**, **DeJesus E**, **Novak RM**, **Frank I**, **Bentley C**, **Ramirez S**, **Fu R**, **Koup RA**, **Mascola JR**, **Nabel GJ**, **Montefiori DC**, **Kublin J**, **McElrath MJ**, **Corey L**, **Gilbert PB**, **HVTN 505 Study Team**. 2013. Efficacy trial of a DNA/rAd5 HIV-1 preventive vaccine. N Engl J Med **369**:2083–2092.

30. **Rerks-Ngarm S**, **Pitisuttithum P**, **Nitayaphan S**, **Kaewkungwal J**, **Chiu J**, **Paris R**, **Premsri N**, **Namwat C**, **de Souza M**, **Adams E**, **Benenson M**, **Gurunathan S**, **Tartaglia J**, **McNeil JG**, **Francis DP**, **Stablein D**, **Birx DL**, **Chunsuttiwat S**, **Khamboonruang C**, **Thongcharoen P**, **Robb ML**, **Michael NL**, **Kunasol P**, **Kim JH**, **MOPH-TAVEG Investigators**. 2009. Vaccination with ALVAC and AIDSVAX to prevent HIV-1 infection in Thailand. N Engl J Med **361**:2209–2220.

31. **Haynes BF**, **Gilbert PB**, **McElrath MJ**, **Zolla-Pazner S**, **Tomaras GD**, **Alam SM**, **Evans DT**, **Montefiori DC**, **Karnasuta C**, **Sutthent R**, **Liao H-X**, **DeVico AL**, **Lewis GK**, **Williams C**, **Pinter A**, **Fong Y**, **Janes H**, **deCamp A**, **Huang Y**, **Rao M**, **Billings E**, **Karasavvas N**, **Robb ML**, **Ngauy V**, **de Souza MS**, **Paris R**,

Ferrari G, **Bailer RT**, **Soderberg KA**, **Andrews C**, **Berman PW**, **Frahm N**, **De Rosa SC**, **Alpert MD**, **Yates NL**, **Shen X**, **Koup RA**, **Pitisuttithum P**, **Kaewkungwal J**, **Nitayaphan S**, **Rerks-Ngarm S**, **Michael NL**, **Kim JH**. 2012. Immune-correlates analysis of an HIV-1 vaccine efficacy trial. N Engl J Med **366**:1275–1286.

32. **Kim JH**, **Excler J-L**, **Michael NL**. 2015. Lessons from the RV144 Thai phase III HIV-1 vaccine trial and the search for correlates of protection. Annu Rev Med **66**:423–437.

33. **Montero M**, **van Houten NE**, **Wang X**, **Scott JK**. 2008. The membrane-proximal external region of the human immunodeficiency virus type 1 envelope: dominant site of antibody neutralization and target for vaccine design. Microbiol Mol Biol Rev **72**:54–84– table of contents.

34. **Purtscher M**, **Trkola A**, **Gruber G**, **Buchacher A**, **Predl R**, **Steindl F**, **Tauer C**, **Berger R**, **Barrett N**, **Jungbauer A**. 1994. A broadly neutralizing human monoclonal antibody against gp41 of human immunodeficiency virus type 1. AIDS Res Hum Retroviruses **10**:1651–1658.

35. **Stiegler G**, **Kunert R**, **Purtscher M**, **Wolbank S**, **Voglauer R**, **Steindl F**, **Katinger H**. 2001. A potent cross-clade neutralizing human monoclonal antibody against a novel epitope on gp41 of human immunodeficiency virus type 1. AIDS Res Hum Retroviruses **17**:1757–1765.

36. **Nelson JD**, **Brunel FM**, **Jensen R**, **Crooks ET**, **Cardoso RMF**, **Wang M**, **Hessell A**, **Wilson IA**, **Binley JM**, **Dawson PE**, **Burton DR**, **Zwick MB**. 2007. An affinity-enhanced neutralizing antibody against the membrane-proximal external region of human immunodeficiency virus type 1 gp41 recognizes an epitope between those of 2F5 and 4E10. journal of virology **81**:4033–4043.

37. **Zhu Z**, **Qin HR**, **Chen W**, **Zhao Q**, **Shen X**, **Schutte R**, **Wang Y**, **Ofek G**, **Streaker E**, **Prabakaran P**, **Fouda GG**, **Liao H-X**, **Owens J**, **Louder M**, **Yang Y**, **Klaric K-A**, **Moody MA**, **Mascola JR**, **Scott JK**, **Kwong PD**, **Montefiori D**, **Haynes BF**, **Tomaras GD**, **Dimitrov DS**. 2011. Cross-reactive HIV-1-neutralizing human monoclonal antibodies identified from a patient with 2F5-like antibodies. journal of virology **85**:11401–11408.

38. **Huang J**, **Ofek G**, **Laub L**, **Louder MK**, **Doria-Rose NA**, **Longo NS**, **Imamichi H**, **Bailer RT**, **Chakrabarti B**, **Sharma SK**, **Alam SM**, **Wang T**, **Yang Y**, **Zhang B**, **Migueles SA**, **Wyatt R**, **Haynes BF**, **Kwong PD**, **Mascola JR**, **Connors M**. 2012. Broad and potent neutralization of HIV-1 by a gp41-specific human antibody. Nature **491**:406–412.

39. **Ni J**, **Powell R**, **Baskakov IV**, **DeVico A**, **Lewis GK**, **Wang L-X**. 2004. Synthesis, conformation, and immunogenicity of monosaccharide-centered multivalent HIV-1 gp41 peptides containing the sequence of DP178. Bioorg Med Chem **12**:3141–3148.

40. **Decroix N**, **Hocini H**, **Quan CP**, **Bellon B**, **Kazatchkine MD**, **Bouvet JP**. 2001. Induction in mucosa of IgG and IgA antibodies against parenterally administered soluble immunogens. Scand J Immunol **53**:401–409.

41. **Joyce JG**, **Krauss IJ**, **Song HC**, **Opalka DW**, **Grimm KM**, **Nahas DD**, **Esser MT**, **Hrin R**, **Feng M**, **Dudkin VY**, **Chastain M**, **Shiver JW**, **Danishefsky SJ**. 2008. An oligosaccharide-based HIV-1 2G12 mimotope vaccine induces carbohydrate-specific antibodies that fail to neutralize HIV-1 virions. Proc Natl Acad Sci USA **105**:15684–15689.

42. **Eckhart L**, **Raffelsberger W**, **Ferko B**, **Klima A**, **Purtscher M**, **Katinger H**, **Rüker F**. 1996. Immunogenic presentation of a conserved gp41 epitope of human immunodeficiency virus type 1 on recombinant surface antigen of hepatitis B virus. J Gen Virol **77 (Pt 9)**:2001–2008.

43. **Kusov YY**, **Zamjatina NA**, **Poleschuk VF**, **Michailov MI**, **Morace G**, **Eberle J**, **Gauss-Müller V**. 2007. Immunogenicity of a chimeric hepatitis A virus (HAV) carrying the HIV gp41 epitope 2F5. Antiviral Research **73**:101–111.

44. **Marusic C**, **Rizza P**, **Lattanzi L**, **Mancini C**, **Spada M**, **Belardelli F**, **Benvenuto E**, **Capone I**. 2001. Chimeric plant virus particles as immunogens for inducing murine and human immune responses against human immunodeficiency virus type 1. journal of virology **75**:8434–8439.

45. **Yi G**, **Lapelosa M**, **Bradley R**, **Mariano TM**, **Dietz DE**, **Hughes S**, **Wrin T**, **Petropoulos C**, **Gallicchio E**, **Levy RM**, **Arnold E**, **Arnold GF**. 2013. Chimeric rhinoviruses displaying MPER epitopes elicit anti-HIV neutralizing responses. PLoS ONE **8**:e72205.

46. **Correia BE**, **Ban Y-EA**, **Holmes MA**, **Xu H**, **Ellingson K**, **Kraft Z**, **Carrico C**, **Boni E**, **Sather DN**, **Zenobia C**, **Burke KY**, **Bradley-Hewitt T**, **Bruhn-Johannsen JF**, **Kalyuzhniy O**, **Baker D**, **Strong RK**, **Stamatatos L**, **Schief WR**. 2010. Computational design of epitope-scaffolds allows induction of antibodies specific for a poorly immunogenic HIV vaccine epitope. Structure **18**:1116–1126.

47. **Ofek G**, **Guenaga FJ**, **Schief WR**, **Skinner J**, **Baker D**, **Wyatt R**, **Kwong PD**. 2010. Elicitation of structure-specific antibodies by epitope scaffolds. Proc Natl Acad Sci USA **107**:17880–17887.

48. **Arnold GF**, **Velasco PK**, **Holmes AK**, **Wrin T**, **Geisler SC**, **Phung P**, **Tian Y**, **Resnick DA**, **Ma X**, **Mariano TM**, **Petropoulos CJ**, **Taylor JW**, **Katinger H**, **Arnold E**. 2009. Broad neutralization of human immunodeficiency virus type 1 (HIV-1) elicited from human rhinoviruses that display the HIV-1 gp41 ELDKWA epitope. journal of virology **83**:5087–5100.

49. **Krebs SJ**, **McBurney SP**, **Kovarik DN**, **Waddell CD**, **Jaworski JP**, **Sutton WF**, **Gomes MM**, **Trovato M**, **Waagmeester G**, **Barnett SJ**, **DeBerardinis P**, **Haigwood NL**. 2014. Multimeric scaffolds displaying the HIV-1 envelope MPER

induce MPER-specific antibodies and cross-neutralizing antibodies when co-immunized with gp160 DNA. PLoS ONE **9**:e113463.

50. **Ye L**, **Wen Z**, **Dong K**, **Wang X**, **Bu Z**, **Zhang H**, **Compans RW**, **Yang C**. 2011. Induction of HIV neutralizing antibodies against the MPER of the HIV envelope protein by HA/gp41 chimeric protein-based DNA and VLP vaccines. PLoS ONE **6**:e14813.

51. **Lai RPJ**, **Hock M**, **Radzimanowski J**, **Tonks P**, **Hulsik DL**, **Effantin G**, **Seilly DJ**, **Dreja H**, **Kliche A**, **Wagner R**, **Barnett SW**, **Tumba N**, **Morris L**, **LaBranche CC**, **Montefiori DC**, **Seaman MS**, **Heeney JL**, **Weissenhorn W**. 2014. A fusion intermediate gp41 immunogen elicits neutralizing antibodies to HIV-1. J Biol Chem **289**:29912–29926.

52. **Javaherian K**, **Langlois AJ**, **LaRosa GJ**, **Profy AT**, **Bolognesi DP**, **Herlihy WC**, **Putney SD**, **Matthews TJ**. 1990. Broadly neutralizing antibodies elicited by the hypervariable neutralizing determinant of HIV-1. Science **250**:1590–1593.

53. **Hart MK**, **Palker TJ**, **Matthews TJ**, **Langlois AJ**, **Lerche NW**, **Martin ME**, **Scearce RM**, **McDanal C**, **Bolognesi DP**, **Haynes BF**. 1990. Synthetic peptides containing T and B cell epitopes from human immunodeficiency virus envelope gp120 induce anti-HIV proliferative responses and high titers of neutralizing antibodies in rhesus monkeys. J Immunol **145**:2677–2685.

54. **Letvin NL**. 1998. Progress in the development of an HIV-1 vaccine. Science **280**:1875–1880.

55. **Moseri A**, **Tantry S**, **Sagi Y**, **Arshava B**, **Naider F**, **Anglister J**. 2010. An optimally constrained V3 peptide is a better immunogen than its linear homolog or HIV-1 gp120. Virology **401**:293–304.

56. **Totrov M**, **Jiang X**, **Kong X-P**, **Cohen S**, **Krachmarov C**, **Salomon A**, **Williams C**, **Seaman MS**, **Abagyan R**, **Cardozo T**, **Gorny MK**, **Wang S**, **Lu S**, **Pinter A**, **Zolla-Pazner S**. 2010. Structure-guided design and immunological characterization of immunogens presenting the HIV-1 gp120 V3 loop on a CTB scaffold. Virology **405**:513–523.

57. **Chakraborty K**, **Durani V**, **Miranda ER**, **Citron M**, **Liang X**, **Schleif W**, **Joyce JG**, **Varadarajan R**. 2006. Design of immunogens that present the crown of the HIV-1 V3 loop in a conformation competent to generate 447-52D-like antibodies. Biochem J **399**:483–491.

58. **Nabel GJ**, **Kwong PD**, **Mascola JR**. 2011. Progress in the rational design of an AIDS vaccine. Philos Trans R Soc Lond, B, Biol Sci **366**:2759–2765.

59. **Yang X**, **Tomov V**, **Kurteva S**, **Wang L**, **Ren X**, **Gorny MK**, **Zolla-Pazner S**, **Sodroski J**. 2004. Characterization of the outer domain of the gp120 glycoprotein from human immunodeficiency virus type 1. journal of virology **78**:12975–12986.

60. **Chen H**, **Xu X**, **Jones IM**. 2007. Immunogenicity of the outer domain of a HIV-1 clade C gp120. Retrovirology **4**:33.

61. **Wu L**, **Zhou T**, **Yang Z-Y**, **Svehla K**, **O'Dell S**, **Louder MK**, **Xu L**, **Mascola JR**, **Burton DR**, **Hoxie JA**, **Doms RW**, **Kwong PD**, **Nabel GJ**. 2009. Enhanced exposure of the CD4-binding site to neutralizing antibodies by structural design of a membrane-anchored human immunodeficiency virus type 1 gp120 domain. journal of virology **83**:5077–5086.

62. **Joyce MG**, **Kanekiyo M**, **Xu L**, **Biertümpfel C**, **Boyington JC**, **Moquin S**, **Shi W**, **Wu X**, **Yang Y**, **Yang Z-Y**, **Zhang B**, **Zheng A**, **Zhou T**, **Zhu J**, **Mascola JR**, **Kwong PD**, **Nabel GJ**. 2013. Outer domain of HIV-1 gp120: antigenic optimization, structural malleability, and crystal structure with antibody VRC-PG04. journal of virology **87**:2294–2306.

63. **Qin Y**, **Banasik M**, **Kim S**, **Penn-Nicholson A**, **Habte HH**, **LaBranche C**, **Montefiori DC**, **Wang C**, **Cho MW**. 2014. Eliciting neutralizing antibodies with gp120 outer domain constructs based on M-group consensus sequence. Virology **462-463**:363–376.

64. **Jardine J**, **Julien J-P**, **Menis S**, **Ota T**, **Kalyuzhniy O**, **McGuire A**, **Sok D**, **Huang P-S**, **MacPherson S**, **Jones M**, **Nieusma T**, **Mathison J**, **Baker D**, **Ward AB**, **Burton DR**, **Stamatatos L**, **Nemazee D**, **Wilson IA**, **Schief WR**. 2013. Rational HIV immunogen design to target specific germline B cell receptors. Science **340**:711–716.

65. **Dosenovic P**, **Boehmer von L**, **Escolano A**, **Jardine J**, **Freund NT**, **Gitlin AD**, **McGuire AT**, **Kulp DW**, **Oliveira T**, **Scharf L**, **Pietzsch J**, **Gray MD**, **Cupo A**, **van Gils MJ**, **Yao K-H**, **Liu C**, **Gazumyan A**, **Seaman MS**, **Björkman PJ**, **Sanders RW**, **Moore JP**, **Stamatatos L**, **Schief WR**, **Nussenzweig MC**. 2015. Immunization for HIV-1 Broadly Neutralizing Antibodies in Human Ig Knockin Mice. Cell **161**:1505–1515.

66. **Yang X**, **Wyatt R**, **Sodroski J**. 2001. Improved elicitation of neutralizing antibodies against primary human immunodeficiency viruses by soluble stabilized envelope glycoprotein trimers. journal of virology **75**:1165–1171.

67. **Forsell MNE**, **Schief WR**, **Wyatt RT**. 2009. Immunogenicity of HIV-1 envelope glycoprotein oligomers. Curr Opin HIV AIDS **4**:380–387.

68. **Kim M**, **Qiao Z-S**, **Montefiori DC**, **Haynes BF**, **Reinherz EL**, **Liao H-X**. 2005. Comparison of HIV Type 1 ADA gp120 monomers versus gp140 trimers as immunogens for the induction of neutralizing antibodies. AIDS Res Hum Retroviruses **21**:58–67.

69. **Li Y**, **Svehla K**, **Mathy NL**, **Voss G**, **Mascola JR**, **Wyatt R**. 2006. Characterization of antibody responses elicited by human immunodeficiency virus

type 1 primary isolate trimeric and monomeric envelope glycoproteins in selected adjuvants. journal of virology **80**:1414–1426.

70. **Sanders RW**, **Vesanen M**, **Schuelke N**, **Master A**, **Schiffner L**, **Kalyanaraman R**, **Paluch M**, **Berkhout B**, **Maddon PJ**, **Olson WC**, **Lu M**, **Moore JP**. 2002. Stabilization of the soluble, cleaved, trimeric form of the envelope glycoprotein complex of human immunodeficiency virus type 1. journal of virology **76**:8875–8889.

71. **Klasse PJ**, **Depetris RS**, **Pejchal R**, **Julien J-P**, **Khayat R**, **Lee JH**, **Marozsan AJ**, **Cupo A**, **Cocco N**, **Korzun J**, **Yasmeen A**, **Ward AB**, **Wilson IA**, **Sanders RW**, **Moore JP**. 2013. Influences on trimerization and aggregation of soluble, cleaved HIV-1 SOSIP envelope glycoprotein. journal of virology **87**:9873–9885.

72. **Hoffenberg S**, **Powell R**, **Carpov A**, **Wagner D**, **Wilson A**, **Kosakovsky Pond S**, **Lindsay R**, **Arendt H**, **Destefano J**, **Phogat S**, **Poignard P**, **Fling SP**, **Simek M**, **LaBranche C**, **Montefiori D**, **Wrin T**, **Phung P**, **Burton D**, **Koff W**, **King CR**, **Parks CL**, **Caulfield MJ**. 2013. Identification of an HIV-1 clade A envelope that exhibits broad antigenicity and neutralization sensitivity and elicits antibodies targeting three distinct epitopes. journal of virology **87**:5372–5383.

73. **Sanders RW**, **Derking R**, **Cupo A**, **Julien J-P**, **Yasmeen A**, **de Val N**, **Kim HJ**, **Blattner C**, **la Peña de AT**, **Korzun J**, **Golabek M**, **de Los Reyes K**, **Ketas TJ**, **van Gils MJ**, **King CR**, **Wilson IA**, **Ward AB**, **Klasse PJ**, **Moore JP**. 2013. A next-generation cleaved, soluble HIV-1 Env trimer, BG505 SOSIP.664 gp140, expresses multiple epitopes for broadly neutralizing but not non-neutralizing antibodies. PLoS Pathog **9**:e1003618.

74. **Sanders RW**, **van Gils MJ**, **Derking R**, **Sok D**, **Ketas TJ**, **Burger JA**, **Ozorowski G**, **Cupo A**, **Simonich C**, **Goo L**, **Arendt H**, **Kim HJ**, **Lee JH**, **Pugach P**, **Williams M**, **Debnath G**, **Moldt B**, **van Breemen MJ**, **Isik G**, **Medina-Ramírez M**, **Back JW**, **Koff WC**, **Julien J-P**, **Rakasz EG**, **Seaman MS**, **Guttman M**, **Lee KK**, **Klasse PJ**, **LaBranche C**, **Schief WR**, **Wilson IA**, **Overbaugh J**, **Burton DR**, **Ward AB**, **Montefiori DC**, **Dean H**, **Moore JP**. 2015. HIV-1 VACCINES. HIV-1 neutralizing antibodies induced by native-like envelope trimers. Science **349**:aac4223.

75. **Sanger F**, **Nicklen S**. 1977. DNA sequencing with chain-terminating inhibitors. *In*.

76. **Stratton M**. 2008. Genome resequencing and genetic variation. Nature biotechnology.

77. **Goodwin S**, **McPherson JD**, **McCombie WR**. 2016. Coming of age: ten years of next-generation sequencing technologies. Nature Reviews Genetics.

78. **Kinde I**, **Wu J**, **Papadopoulos N**. 2011. Detection and quantification of rare mutations with massively parallel sequencing. *In*.

79. **Salzberg SL**, **Phillippy AM**, **Zimin A**, **Puiu D**, **Magoc T**, **Koren S**, **Treangen TJ**, **Schatz MC**, **Delcher AL**, **Roberts M**, **Marçais G**, **Pop M**, **Yorke JA**. 2012. GAGE: A critical evaluation of genome assemblies and assembly algorithms. Genome Res **22**:557–567.

80. **Quince C**, **Lanzen A**, **Davenport RJ**, **Turnbaugh PJ**. 2011. Removing noise from pyrosequenced amplicons. BMC Bioinformatics **12**:38.

81. **Reeder J**, **Knight R**. 2010. Rapid denoising of pyrosequencing amplicon data: exploiting the rank-abundance distribution. Nat Methods.

82. **Edgar RC**, **Flyvbjerg H**. 2015. Error filtering, pair assembly and error correction for next-generation sequencing reads. Bioinformatics **31**:3476–3482.

83. **Eren AM**, **Morrison HG**, **Lescault PJ**, **Reveillaud J**, **Vineis JH**, **Sogin ML**. 2015. Minimum entropy decomposition: unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. ISME J **9**:968–979.

84. **Mysara M**, **Leys N**, **Raes J**. 2016. IPED: a highly efficient denoising tool for Illumina MiSeq Paired-end 16S rRNA gene amplicon sequencing data. Bioinformatics.

85. **Edgar RC**. 2016. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. bioRxiv.

86. **Callahan BJ**, **McMurdie PJ**, **Rosen MJ**, **Han AW**, **Johnson AJA**, **Holmes SP**. 2016. DADA2: High-resolution sample inference from Illumina amplicon data. Nat Methods **13**:581–583.

CHAPTER 2

MODULATING IMMUNOGENIC PROPERTIES OF HIV-1 GP41 MEMBRANE-PROXIMAL EXTERNAL REGION BY DESTABILIZING SIX-HELIX BUNDLE STRUCTURE

Saikat Banerjee[†], Heliang Shi[†], Habtom H. Habte, Yali Qin, Michael W. Cho

[†]These authors contributed equally.

Abstract

The C-terminal a-helix of gp41 membrane-proximal external region (MPER; ^{671}NWFDITNWLWYIK^{683}) encompassing 4E10/10E8 epitopes is an attractive target for HIV-1 vaccine development. We previously reported that gp41-HR1-54Q, a trimeric protein comprised of the MPER in the context of a stable six-helix bundle (6HB), induced strong immune responses against the helix, but antibodies were directed primarily against the non-neutralizing face of the helix. To better target 4E10/10E8 epitopes, we generated four putative fusion intermediates by introducing double point mutations or deletions in the heptad repeat region 1 (HR1) that destabilize 6HB in varying degrees. One variant, HR1-Δ10-54K, elicited antibodies in rabbits that targeted W672, I675 and L679, which are critical for 4E10/10E8 recognition. Overall, the results demonstrated that altering structural parameters of 6HB can influence immunogenic properties of the MPER and antibody targeting. Further exploration of this strategy could allow development of immunogens that could lead to induction of 4E10/10E8-like antibodies.

2.1 Introduction

It is widely hypothesized that a successful AIDS vaccine should induce antibodies that can neutralize a large number of HIV-1 isolates from multiple clades. However, such broadly neutralizing antibodies (bnAbs) have been observed only in a small fraction of HIV-1 infected patients (1, 2), suggesting that the generation of these bnAbs is a complex, difficult process. Nevertheless, efforts to develop immunogens and/or vaccine strategies that can elicit bnAbs must continue.

Recent isolation and characterization of potent bnAbs from patients has helped the field of vaccine research immensely by providing better understanding of both the epitopes targeted and the unique features of antibodies that contribute to their broad neutralizing ability. Most of the bnAbs that target gp120 recognize highly conformational, non-linear epitopes(3). In contrast, those that target gp41 recognize linear epitopes that are structurally simpler and reside in a highly conserved, ~22 amino acid long domain called the membrane proximal external region (MPER) (4). These bnAbs are thought to inhibit conformational changes that are critical for fusion between viral and cellular membranes. A more recent discovery of the highly potent and broadly neutralizing 10E8 mAb (5), along with previously characterized 2F5, 4E10 and Z13e1 (6-8), has renewed interests in designing MPER-based immunogens.

To date, eliciting anti-MPER bnAbs through vaccination has been elusive. Several approaches have been examined, including (1) immunization with short MPER peptides either alone or coupled to carrier proteins, (2) neutralizing epitopes presented on scaffolds, (3) MPER peptides delivered on liposomes, (4) MPER containing hybrid/fusion proteins, and (5) chimeric viruses or virus like particles (see (9) and

references therein). Although a handful of studies have shown induction of modest levels of neutralizing activity with limited breadth against tier 1 HIV-1 isolates (10-14), neither the identity of antibodies responsible for neutralization nor the mechanistic nature of inhibition have been further demonstrated.

Multiple crystal structures of short MPER peptides in complex with bnAbs have been solved (15-17). Despite simpler epitope structures, the difficulty in designing immunogens that can induce similar antibodies lies partly on the fact that the MPER structure in the context of a native trimeric envelope spike on virus particles remains unknown. In this regard, it is possible that antibody-bound MPER structures do not accurately represent the MPER conformation on native trimers. Furthermore, gp41 undergoes large structural changes during the fusion process (18), and it is likely that the MPER also assumes several different conformations. Thus, studies that characterize the structural and immunological properties of MPER in context of larger gp41-based proteins are much needed.

As an initial effort, we generated a soluble gp41 construct named gp41-HR1-54Q consisting of heptad repeat regions 1 and 2 (HR1and HR2, respectively) and the MPER (19). While the HR1 and HR2 domains formed a stable six-helix bundle (6HB), much of the MPER domain remained quite flexible and free from association with the 6HB (19). Surprisingly, this protein induced strong antibody responses against a peptide that encompass 4E10/10E8 epitopes (^{671}NWFDITNWLW^{680}; (9). Further analyses indicated that these antibodies targeted the non-neutralizing face of the C-terminal a-helix, but partly overlapping with 4E10/10E8 epitopes. One possible reason for the preferential targeting of the non-neutralizing face of the helix could be its

orientation when the MPER is presented in the context of a stable 6HB structure, which represents a post-fusion conformation. It had been suggested that MPER might be more exposed during the fusion process as gp41 undergoes conformational changes (20-25).

Recent crystal structures of BG505 SOSIP gp140 provided partial information on the pre-fusion state of gp41(18, 26). However, fusion intermediate structures are completely unknown. In particular, there is no information on how the MPER is oriented relative to the rest of the protein, and when and whether it make any contact with the rest of gp41 or gp120. The only certainty is that HR1 and HR2 are in the process of coming together to form 6HB. As such, we took an empirical approach of generating four variants of gp41-HR1-54Q that might represent different stages of fusion process by disrupting 6HB formation in varying degrees. Biochemical, antigenic, and immunogenic properties of these putative fusion intermediates (PFIs) were characterized. Although we did not succeed in inducing bnAbs against the MPER in rabbits, the results from the study should facilitate development of improved MPER immunogens.

2.2 Results

2.2.1 Designing gp41-HR1-54Q variants with destabilized 6HB.

The trimeric structure of gp41-HR1-54Q is stabilized by both inter- and intramolecular interactions between HR1 and HR2 (19). The exact order of molecular interactions between HR1 and HR2 that leads to 6HB formation is not known, although a leading working model suggests a zipping process along HR1-HR2 that begins at the

C- and N-termini of respective domains in an anti-parallel fashion, with a central trimeric HR1 core(27). As such, we hypothesized that it might be possible to generate partially opened hairpin loop structures that might simulate fusion intermediates if HR1-HR1 or HR1-HR2 interactions were destabilized.

Four variants were generated by either introducing point mutations or deletions (Fig. 1A). All mutations were restricted to the HR1 only, so as to avoid altering the native conformation of the HR2 or MPER domains. The C-terminal half of HR1 was mutated by introducing two point mutations (HR1-AA-54Q and HR1-EE-54Q), whereas the N-terminal half of HR1 was mutated more drastically by deletions (HR1Δ10-54K and HR1Δ17-54K). The HR1-AA-54Q variant contained L565A and L568A mutations with an intent of weakening hydrophobic interactions with I635 and Y638 residues on HR2 (Fig. 1B). In contrast, HR1-EE-54Q contained L568E and K574E mutations designed destabilize 6HB formation by introducing intra- and inter-molecular charge-charge repulsions with E634 and E632 residues on HR2, respectively (Fig. 1C). Deleting the N-terminal 10 or 17 amino acids of HR1 is designed to allow initiation of 6HB formation, but halt the zipping process in the middle to generate structures that might resemble fusion intermediates (Figs. 1D and 1E, respectively). For these constructs with deletions, the terminal 683Q residue was reverted back to the wild type lysine as it was later reported to be critical for 10E8 binding (5). Although we do not know whether any of these constructs would mimic true fusion intermediates, they will be referred herein as putative fusion intermediates (PFIs) for simplicity.

2.2.2 Biochemical and antigenic properties of PFIs.

As we previously reported, gp41-HR1-54Q is highly resistant to trypsin digestion (9), presumably due to its rigid structure. To determine how mutations might affect the protein, trypsin sensitivity of the PFIs was assessed. As shown in Fig. 2A, gp41-HR1-54Q was completely resistant to trypsin digestion even after one hour. In contrast, PFIs exhibited varying degrees of trypsin sensitivity. Not surprisingly, HR1-AA-54Q was least sensitive. Unexpectedly, however, HR1-EE-54Q was most sensitive and that HR1-Δ10-54K was more sensitive than HR1-Δ17-54K. The differences in trypsin sensitivity among the PFIs suggested that they likely have folded into structures different from each other, and certainly different from gp41-HR1-54Q.

Next, the PFIs were probed with NC-1, a mouse monoclonal antibody (mAb) that recognizes post-fusion 6HB structure (28). NC-1 recognizes amino acid residues from 643 to 655 within HR2 (29), which is present in all four PFIs. Thus, any changes in NC-1 binding should be the result of conformational changes induced by the mutations. As shown in Fig. 2B (left panel), NC-1 binding was completely abolished for HR1-EE-54Q, HR1-Δ10-54K and HR1-Δ17-54K. Although HR1-AA-54Q could be recognized, the binding was substantially weaker than gp41-HR1-54Q. These results indicate that introduced mutations were able to disrupt formation of the post-fusion 6HB conformation.

Next, the PFIs were probed with 126-7, a human mAb (an IgG2 version of 126-6) that only recognizes a trimeric conformation of gp41 shared between both pre- and post-fusion state (29-33). It recognizes residues from 641 to 648 in the cluster II of gp41. As shown in Fig. 2B (right panel), gp41-HR1-54Q, HR1-AA-54Q and HR1-EE-

54Q bound nearly equally to 126-7 suggesting that the trimeric conformations of these proteins were similar (at least at the 126-7 epitope). Not surprisingly, 126-7 failed to recognize both HR1-Δ10-54K and HR1-Δ17-54K. The elimination of three and five helical turns in HR1, respectively, likely prevented formation of stable trimeric HR1 core.

To examine whether epitopes recognized by MPER bnAbs remained conformationally intact and accessible on PFIs, they were probed with 2F5, Z13e1, 4E10 and 10E8 by ELISA (Fig. 2C). For 2F5, there were only minor differences between HR1-54Q and PFIs. The results were similar for Z13e1, although HR1-EE-54Q showed slightly weaker binding. The reduced binding to HR1-EE-54Q was more pronounced with 4E10 and 10E8. To a lesser extent, binding was also reduced for HR1-AA-54Q. In general, antibody binding to HR1-Δ10-54K and HR1-Δ17-54K was quite similar to HR1-54Q, except for 10E8, for which there was significantly better binding. However, this enhanced binding is most likely due to reverting back to lysine at position 683, rather than deletions themselves, since K683 is one of the critical residues that 10E8 recognizes. Taken together, these results suggest that destabilization of 6HB, depending on the approach taken, could potentially affect MPER structure, which could in turn alter conformation or accessibility of epitopes targeted by bnAbs.

2.2.3 Immunogenic properties of PFIs

Immunogenic properties of PFIs were evaluated in rabbits as we have done for gp41-HR1-54Q (9). For this initial study, two animals were used for each of the four PFIs. Serum samples were collected two weeks after each immunization (on weeks 0, 4, 9 and 15). Antibody titers against autologous antigens were assessed by ELISA (Fig.

3). Overall, antibody responses against the PFIs were weaker than gp41-HR1-54Q, which induced end point antibody titers greater than $5x10^6$ even after a single immunization. PFIs with point mutations were more immunogenic than the ones with deletions, which was more noticeable after the first immunization. The reduced antibody responses could be due in part to elimination of helper T cell epitopes, especially for the deletion mutants, in addition to altered conformations or loss of epitopes.

To better understand how mutations on PFIs altered immune responses, immunogenic linear epitopes were mapped by ELISA using overlapping biotinylated peptides as we previously described (9). Since mutations and deletions were in the HR1 domain, we focused on antibody responses directed against the HR2 and the MPER. Notwithstanding some animal-to-animal variations, the immunogenic epitope profile of HR1-AA-54Q was somewhat similar to that of HR1-54Q, with 671 peptide (^{671}NWFDITNWLW^{680}) being highly immunogenic in both animals. Interestingly, antibody responses against HR1-EE-54Q were directed towards N-terminus of HR2 and the C-terminus of MPER with little to no antibodies against peptides spanning the cluster II region (^{644}RLIEESQNQQEKNEQELLAL^{663}) that typically elicits non-neutralizing antibodies (34-36). Compared to HR1-54Q, one major difference in immune responses against HR1-Δ10-54K is strong antibody responses against the N-terminal end of HR2. Although the 671 peptide remained immunogenic, peptides 629 (^{629}MEWEREISNY^{638}), 632 (^{632}EREISNYTDI^{641}) and 635 (^{635}ISNYTDIIYR^{634}) were clearly immunodominant. By far, the most striking change in immunogenic epitope

profile was with HR1-Δ17-54K. Virtually all of the peptides, except for peptides 665, 668 and 674, were highly immunogenic in one or both of the rabbits.

Despite strong antibody responses against the 671 peptide (^{671}NWFDITNWLW^{680}) that contained all or most of 4E10 and 10E8 epitopes, none of the rabbit sera exhibited neutralizing activity when tested against HIV-1 pseudoviruses SF162 (tier 1A, clade B), MW965.26 (tier 1A, clade C), and MN.3 (tier 1A, clade B) in a standard TZM-bl assay.

2.2.4 Detailed analyses of antibodies targeting near 4E10/10E8 epitopes.

Although we did not succeed in inducing bnAbs against gp41 MPER, better characterization of antibody responses near 4E10/10E8 epitopes could facilitate improving immunogens. In particular, we were curious to see whether and how epitope targeting was altered for PFIs compared to HR1-54Q. Towards this goal, we conducted fine epitope mapping analyses using alanine-scanning mutants of a 13-mer peptide (^{671}NWFDITNWLWYIK^{683}), which we previously used to characterize antibody responses against HR1-54Q (4, 9). Initially, antibody reactivity against the wild-type 13-mer 671 peptide was measured (Fig. 5). Interestingly, some of the antisera reacted poorly to the 13-mer peptide when compared to their reactivity against the 10-mer 671 peptide (Fig. 4). This was particularly severe with rabbit #2 immunized with HR1-Δ17-54K, and, to a lesser degree, rabbit #2 immunized with HR1-AA-54Q. Since ELISA using 10-mer peptides (Fig. 4) was done with a mixture of both N- and C-terminally biotinylated peptides, and that the 13-mer is biotinylated at the C-terminal K683 residue, it was possible that the orientation of the peptide attachment to an ELISA plate, could

have affected antibody binding. However, this might not be the case since antibodies reacted strongly to the 10-mer peptide that was biotinylated at the C-terminus (data not shown). Thus, the reason for the discrepancy is unknown at the present time.

With the caveat that we would be evaluating only a subset of antibodies targeting near 4E10/10E8 epitopes, we proceeded to characterize antibodies using the panel of 13-mer mutant peptides. All sera, except from rabbit #2 immunized with HR1-Δ17-54K, were analyzed. In doing so, serum samples were first normalized to yield comparable signals when bound to the wild-type peptide (Fig. 6A). For all rabbits tested, D674 residue was critical for antibody binding, which is likely due to a critical role it plays in maintaining a helical conformation of the peptide (5, 37). For HR1-AA-54Q, the two animals exhibited different antibody epitope profiles (Fig. 6B); such animal-to-animal variations have been observed with HR1-54Q also (6-9). Interestingly, similar patterns were also observed in animals immunized with HR1-EE-54Q (Fig. 6C). This might suggest structural similarity of the C-terminal MPER for HR1-AA-54Q and HR1-EE-54Q. For the both groups, the profile shown on the top panels resembled one of the patterns observed from animals immunized with HR1-54Q. The profile shown on the bottom panels (critical residues being D674, W678, L679, and I675 for HR1-AA-54Q) was not observed in any of the six animals immunized with HR1-54Q. Thus, the latter profile could be specific to antibody responses against PFIs. Coincidently, a similar profile was observed for a rabbit immunized with HR1-Δ17-54K (Fig. 6E). The same four residues were also critical for antibodies induced by HR1-Δ10-54K (Fig. 6D). In addition, antibodies induced by HR1-Δ10-54K also targeted W672, which may be highly significant since this residue was never targeted by antibodies induced with HR1-54Q or

with any of the other PFIs. More importantly, this is one of the critical residues recognized by both 4E10 and 10E8.

To better compare antibodies induced by HR1-54Q and HR1-Δ10-54K, amino acid residues critical for binding were visualized on a structure of a peptide co-crystallized with 4E10 (Fig. 7; Cardoso et al., 2007). This analysis revealed that antibodies induced by HR1-Δ10-54K targeted nearly the opposite face of the helix compared to those induced by HR1-54Q with W678 being targeted by both. More importantly, HR1-Δ10-54K-induced antibodies targeted three of the five most critical residues for 4E10 binding (**W672**, F673, **I675**, T676 and **L679**). So, although we were not able to induce nAbs, the results of our study demonstrate that it is possible to alter immunogenicity of epitopes simply by changing structural context of an immunogen.

2.3 Discussion

In previous studies, we described structural and immunological properties of gp41-HR1-54Q, which likely represents a near-post-fusion conformation of gp41 (9, 19). Although antibodies elicited in rabbits by this antigen bound epitopes that partially overlap with those targeted by 4E10 and 10E8, they were largely directed against the non-neutralizing face of the helix and failed to exhibit neutralizing activity. It had been speculated that anti-MPER bnAbs primarily target fusion intermediate forms of gp41 (10-14, 22, 23, 38). However, their structures are not yet known, except for those of short peptides bound to the antibodies. As such, we decided to take an empirical approach of generating fusion intermediates with a simple assumption that they would have minimal or partial HR1-HR2 pairing and 6HB formation. We further assumed that

the MPER would likely exist in a conformation that is different from the one observed on gp41-HR1-54Q. With these assumptions, four putative fusion intermediates (PFIs) were generated by introducing double point mutations or deletions into the HR1 of gp41-HR1-54Q to destabilize HR1-HR1 and HR1-HR2 interactions, and their biochemical and immunological properties were evaluated in this study.

Although we do not have structural evidence, the increased sensitivity of the PFIs to trypsin digestion and their altered reactivity to NC-1 and/or 126-7 mAbs demonstrated that the structure of 6HB on PFIs has been disrupted in varying degrees. Not surprisingly, the four PFIs revealed different immunological profiles with respect to the overall immunogenicity (*i.e.* total antibody titers induced; Fig. 3), immunodominance of epitopes across the HR2 and MPER (Fig. 4), and specific amino acid residues targeted by antibodies directed against the C-terminal region containing 4E10/10E8 epitopes (^{671}NWFDITNWLWYIK^{683}; Fig. 6). Together, these results indicate that immunogenicity of HR2 and MPER domains are highly dependent on the structural context in which they are presented.

While we did not succeed in inducing bnAbs, detailed epitope mapping analyses revealed a few important findings about antibodies induced by PFIs, HR1-Δ10-54K in particular. First, antibodies induced in both rabbits immunized with HR1-Δ10-54K targeted W672. Targeting of this residue was never observed in any of the six rabbits immunized with gp41-HR1-54Q (9, 15-17, 19) or in any of the animals immunized with other PFIs. Replacement of this residue with an alanine has been reported to reduce 4E10 binding by over 1000-fold, highlighting its overarching importance (19, 37). W672 is also critical for 10E8 binding (5, 19). Second, animals immunized with HR1-Δ10-54K

and HR1-Δ17-54K induced antibodies that bound strongly to I675 and L679, both of which line up with W672 along the same side of the helix and contribute significantly to 4E10 and 10E8 binding (9, 37, 39). Targeting these three residues highlights a remarkable shift from the binding pattern of antibodies elicited by gp41-HR1-54Q. Third, while gaining recognition of these three residues, antibodies induced by HR1-Δ10-54K seemed to have lost recognition of F673 and T676, which were quite well recognized by antibodies induced with gp41-HR1-54Q and targeted by both 4E10 and 10E8.

One cautionary note for interpreting the results of our study is that the epitope mapping analyses shown in Fig. 6 were conducted with polyclonal antibodies. Thus, the phenotypic changes we observed are average of all antibodies that bind the peptide. Accordingly, when there are many antibodies that bind to different epitopes, we might not see significant reduction in antibody binding for mutations at any given position; this might be the case for rabbit #1 of HR1-AA-54Q and HR1-EE-54Q. Significant reduction in binding would be seen only when the antibody response is homogeneous or when all or most antibodies target same residues. In this regard, characterizing antibodies at the monoclonal level would provide a more accurate assessment.

In the absence of a crystal structure of HR1-Δ10-54K, it is hard to speculate how deleting ten residues from the N-terminal end of HR1 (with a potential contribution from K683) influenced the overall MPER conformation to promote such a major shift in antibody response. Nevertheless, our results clearly demonstrated that changes in HR1, which in turn affect stability of 6HB, significantly influence how antibodies target the MPER. Considering the difficulties in crystallizing proteins that contain the

hydrophobic C-terminal ectodomain of gp41, MPER-based vaccine development may have to rely on reiterative, empirical approaches. Based on the results from this study, HR1-Δ10-54K would be an excellent starting point.

To improve the immunogen, one factor that could be adjusted is the length of HR1. In this study, we deleted 10 or 17 residues, which account for roughly 3 and 5 helical turns, respectively. As shown in Fig. 4, the antibodies induced by HR1-Δ17-54K were drastically different from all others, which we believe is due to complete disruption of 6HB, thereby rendering the protein highly flexible and inducing greater diversity of antibodies. By adding two helical turns (HR1-Δ10-54K), which likely increased the stability of 6HB, antibody responses against HR2 was largely restricted to the N-terminal end and W672 could be targeted. This raises a question as to what would happen to the antibody response if less than 10 residues were deleted. Would any of them allow targeting of F673 and T676 (as did gp41-HR1-54Q), while also targeting W672, I675 and L679?

Another factor that could be considered for improving antibody responses is to minimize immunogenicity of W678. As shown in Fig. 6, W678 was targeted on all of the PFIs as well as on gp41-HR1-54Q (9, 20-25), suggesting its dominant role in determining antibody responses. Most likely, W678 is not exposed during the natural course of the fusion process, rendering these antibodies useless. Thus, preventing antibody responses against W678 could improve the chance of inducing bnAbs. Perhaps substituting W678 with a less immunogenic residue (*e.g.* glycine or alanine) might redirect the immune system to shift the focus away from W678 towards F673 and T676.

In all of our immunization studies, we used Zn-chitosan not only as an adjuvant, but also as an antigen delivery platform. Although all of our PFIs, as well as gp41-HR1-54Q, are soluble proteins, they all have 6xHis tag at the C-terminus, which was used to affix the proteins to Zn-chitosan. Thus, the flexibility of the C-terminal end of the MPER was most likely limited. However, neither the spatial distribution nor the orientation of the MPER when it is bound to Zn-chitosan are known. More importantly, how it would affect MPER immunogenicity is completely unknown. In this regard, it would be worthwhile to evaluate immunogenicity of some of the PFIs in the context of lipid membranes (*e.g.* delivered on liposomes or expressed directly on the cell surface), which would better resemble the MPER structure and the microenvironment of virus particles.

2.4 Materials and Methods

2.4.1 Cloning, Expression and Purification of PFIs

To generate PFI constructs with point mutations, the QuikChange® XL Site directed mutagenesis kit was used as per the manufacturer's instructions using the original gp41-HR1-54Q plasmid as the template (19). For HR1-AA-54Q, the mutations L565A and L568A were introduced using the sense primer 5'-GAGGCCCAGCAGCAC**GCC**CTGCAG**GCC**ACCGTGTGGGGCATC-3' and the antisense primer 5'-GATGCCCCACACGGT**GGC**CTGCAG**GGC**GTGCTGCTGGGCCTC-3'. For HR1-EE-54Q, the mutations L568E and K574E were introduced using the sense primer 5'-GCACCTGCTGCAG**GA**GACCGTGTGGGGCATC**G**AGCAGGGAGGAGG-3' and the

antisense primer 5'-CCTCCTCCCTGCT**C**GATGCCCCACACGGTC**TC**CTGCAGCAGGTGC-3'.

For the deletion variants, 10 and 17 residues were deleted from the N terminus end of the HR1 domain as shown in Fig 1A. Both constructs were synthesized from IDT (Integrated DNA Technology) in the pUC57 backbone with flanking restriction sites for BamHI and EcoRI at the 5' and 3' ends of the constructs, respectively. The sequence was also altered to encode terminal 683K residue instead of the 683Q as in gp41-HR1-54Q. These constructs were cloned into the pET-21a vector (Novagen; cat#69740-3) using BamHI and EcoRI. All constructs were expressed and purified similar to gp41-HR1-54Q (5, 19). The final proteins were dialyzed into 1x PBS (pH 8.0) and stored at -80 degrees.

2.4.2 Trypsin sensitivity assay

All PFIs were incubated with trypsin at 1:100 (enzyme:protein) mass ratio for one hour at 37 °C. 3 µg of untreated and trypsin treated samples were then run on a Novex® 10-20% tricine gel (Thermo Fisher Scientific; cat# EC6625BOX)

2.4.3 Rabbit immunization

Eight New Zealand white female rabbits (2.5 to 3 kg) were purchased from Charles River (USA), housed under specific pathogen free environments and used in compliance with the animal protocol approved by IACUC of Iowa State University. Two animals were immunized with each antigen using Zn-chitosan as an adjuvant. The immunization protocol including the adjuvant preparation, antigen/adjuvant dosage, the

immunization and bleeding schedule were all exactly the same as that previously described for gp41-HR1-54Q (9).

2.4.4 Enzyme-linked immunosorbent assay (ELISA)

All ELISAs were performed using the standard protocol described for gp41-HR1-54Q (9, 28) except for the use of an alternate blocking buffer consisting of PBS (pH 7.5) with 2.5% skim milk and 5% calf sera. For ELISAs testing the binding of antibodies NC-1, 126-7, 2F5, 4E10, Z13e1 and 10E8, coating antigen amounts for all other antigens equimolar to 30 ng/well of gp41-HR1-54Q using the same coating conditions as described for gp41-HR1-54Q. In order to determine end point titers, all antigens were coated at 30 ng/well. The end-point ELISA titers were defined as serum dilution factor that gave readings of average + 2xSD (standard deviation) of the background as described previously (29, 40). Coating for linear epitope mapping using 10-mer biotinylated peptides and 13-mer alanine scanning was also performed as previously described (9, 29-33).

2.4.5 Neutralization assays

Neutralization assays were performed in TZM-bl cells as previously described (9, 40-42). Viruses tested included SF162 (tier 1A, clade B), MW965.26 (tier 1A, clade C), and MN.3 (tier 1A, clade B). Murine leukemia virus Env-pseudotyped virus was used as a negative control.

2.5 Acknowledgments

The following reagents were obtained through the NIH AIDS Reagent Program, Division of AIDS, NIAID, NIH: HIV-1 anti-gp41 mAb NC-1 from Dr. Shibo Jiang (Cat# 11482), 126-7 from Dr. Susan Zolla-Pazner (Cat# 9967), 2F5 from Dr. Hermann Katinger (Cat# 130220), 4E10 from Dr. Herman Katinger (Cat# 10091), Z13e1 from Dr. Michael Zwick (Cat# 11557) and 10E8 from Dr. Mark Connors (Cat# 12294). This work was supported by a grant from the NIH, NIAID (P01 AI074286) grant. MWC has an equity interest in NeoVaxSyn Inc., and serves as the CEO/President. NeoVaxSyn Inc. did not contribute to this work or the interpretation of the data.

2.6 Figures

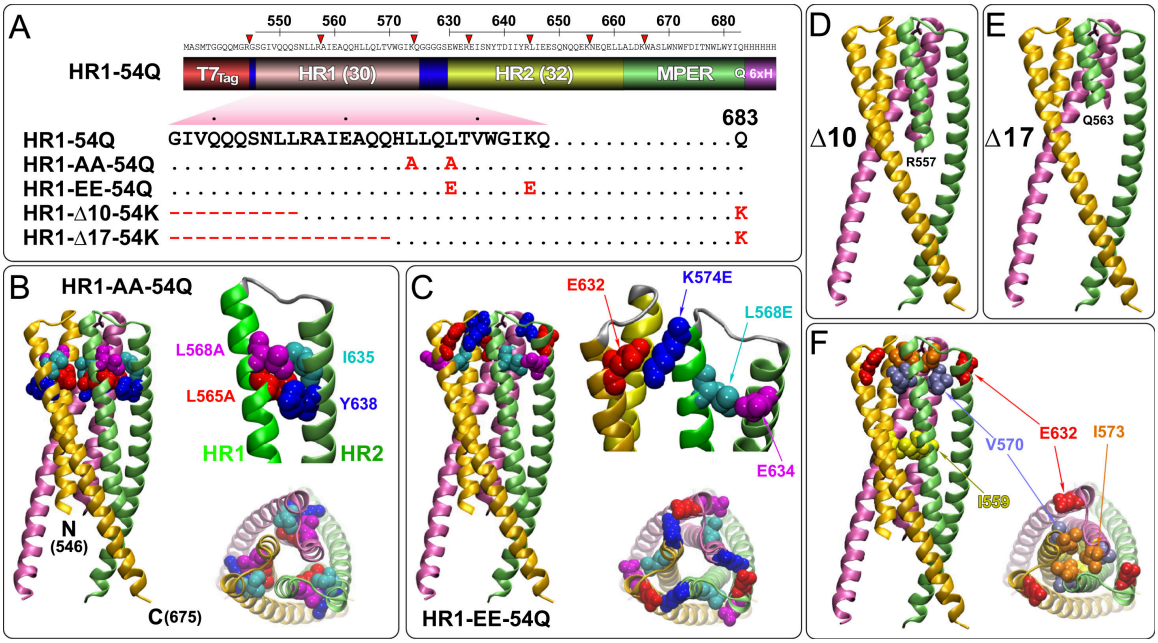

Fig 1. Design of putative fusion intermediates of gp41-HR1-54Q. (A) A domain structure of gp41-HR1-54Q consisting of the $T7_{Tag}$, heptad repeat region 1 (HR1), GGGGS linker, heptad repeat region 2 (HR2), membrane-proximal external region (MPER) and the 6x His tag is shown. The HR1 domain sequence, along with the terminal 683Q residue, is indicated for gp41-HR1-54Q. Point mutations and deletions introduced into the HR1 domain to generate variants HR1-AA-54Q, HR1-EE-54Q, HR1-Δ10-54K and HR1-Δ17-54K are indicated. The terminal 683Q residue was reverted back to 683K in HR1-Δ10-54K and HR1-Δ17-54K. (B) The mutations introduced in HR1-AA-54Q (L565A and L568A) are plotted on the gp41-HR1-54Q crystal structure (pdb: 3K9A) (9, 19) to highlight the proximity of these residues to the neighboring I635 and Y638 residues located on the HR2 domain. Structures of the unmutated amino

acids are shown. (C) The mutations introduced in HR1-EE-54Q (L568E and K574E) are plotted on the gp41-HR1-54Q crystal structure to display the proximity of these residues to E632 and E634 residues. The truncations introduced at the N-terminal end of the HR1 domain are plotted onto the gp41-HR1-54Q structure simply to show the point of deletion for (D) HR1-Δ10-54K and (E) HR1-Δ17-54K.

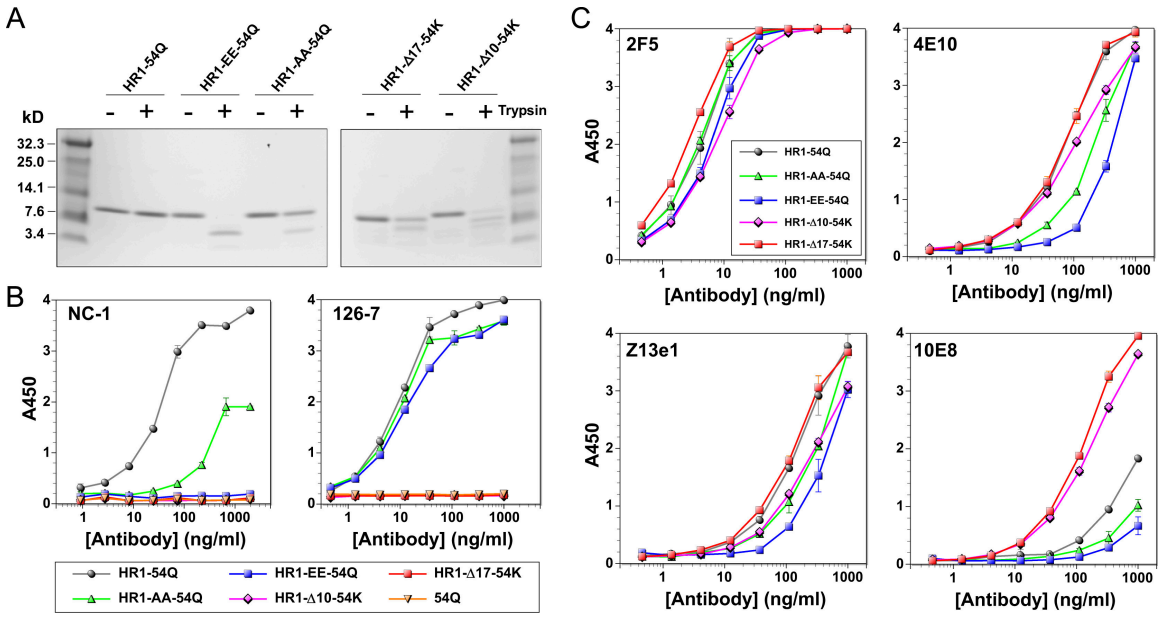

Fig 2. Biochemical and antigenic properties of putative fusion intermediates. (A) Evaluation of trypsin sensitivity of PFIs in comparison to gp41-HR1-54Q. (B) ELISA with mAbs NC-1 and 126-7 to monitor effects of the mutation on six-helix bundle formation. gp41-HR1-54Q was used as a positive control, while another protein (gp41-54Q) that lacks the HR1 domain was used as a negative control. (C) The antigenic integrity of the variants was tested by performing ELISA with bnAbs 2F5, 4E10, Z13e1 and 10E8.

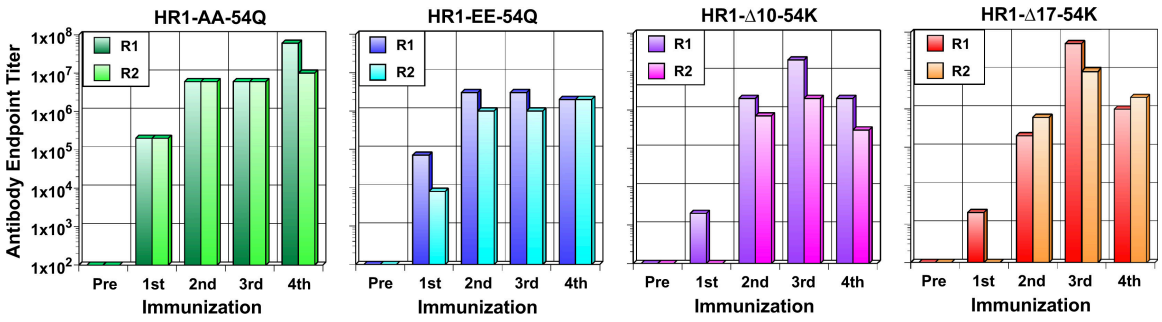

Fig 3. Antibody end-point titers induced by putative fusion intermediates. Serum samples collected two weeks after each immunization were evaluated by ELISA to determine the end-point antibody titers against autologous antigens. Pre-immune serum was used as a negative control.

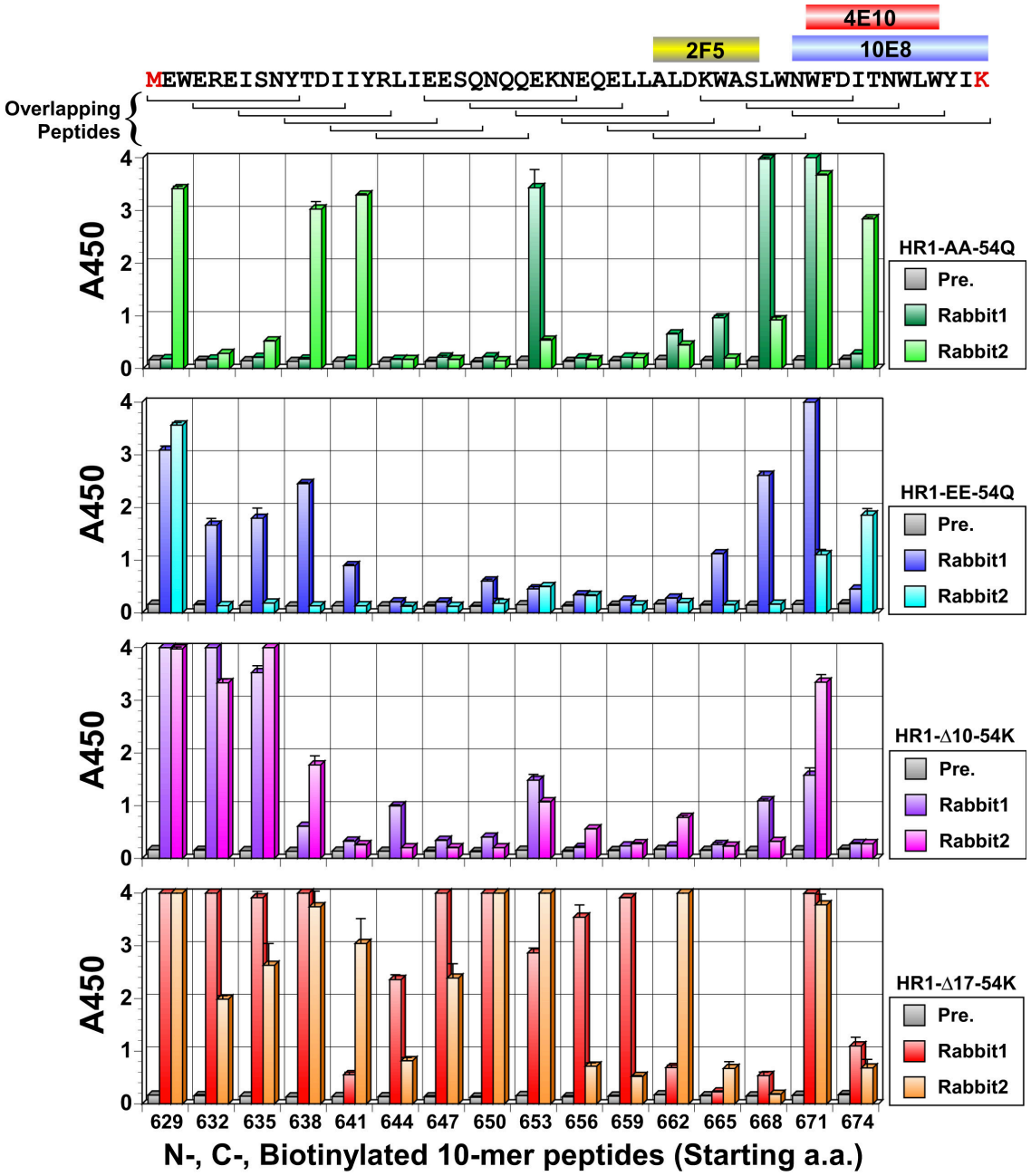

Fig 4. PepScan analyses using linear, overlapping peptides. Serum samples collected after the fourth immunization were evaluated for reactivity against biotinylated 10-mer peptides (a mixture of peptides biotinylated at the N-terminal and C-terminal ends) spanning both HR2 and MPER domains. The amino acid sequence of each peptide is indicated by horizontal brackets. The first and last residue in the peptide

panel is indicated in red as they can differ from the immunogens. The core binding epitopes for 2F5, 4E10 and 10E8 bnAbs are also indicated. Pre-immune serum was used as a negative control.

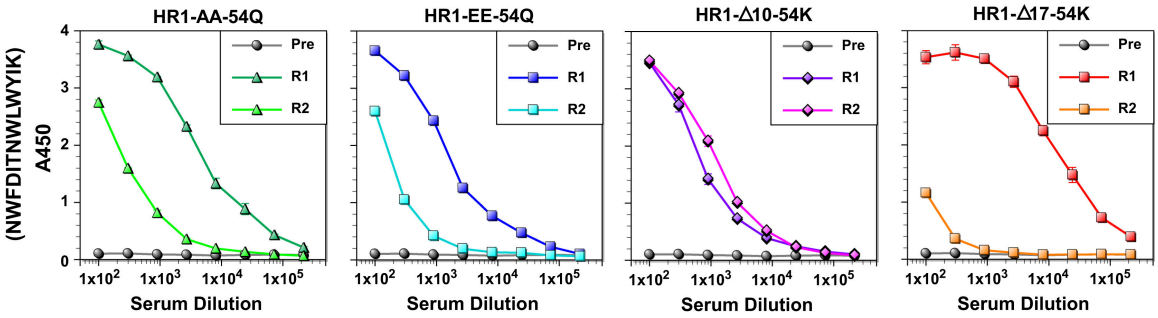

Fig 5. Antibody titers against a wild type peptide containing C-terminal 13 amino acids. Serum samples collected after the fourth immunization was evaluated for binding biotinylated 13-mer peptide (^{671}NWFDITNWLWYIK^{683}) that contains 4E10/10E8 epitopes. Pre-immune serum was used as negative control.

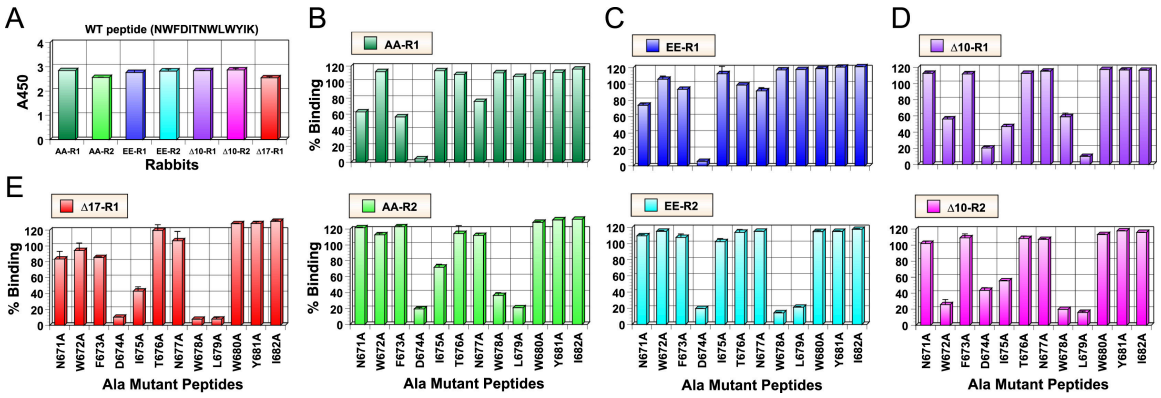

Fig 6. Detailed epitope mapping analysis of antibodies against the C-terminal 13 amino acid residues using alanine-scanning mutant peptides. (A) Serum samples after the fourth immunization were examined for binding biotinylated 13-mer peptide (^{671}NWFDITNWLWYIK^{683}). The analyses were done using normalized serum samples to yield comparable binding signal (AA-R1 at 1:2000 dilution; AA-R2 and EE-R2 at 1:100 dilution; EE-R1 at 1:600 dilution; Δ10-R1 at 1:300 dilution; Δ10-R2 at 1:400 dilution; and Δ17-R1 at 1:5000 dilution). (B-E) The same dilutions were tested for binding to mutant peptides. Results are shown as the percentage of binding to the wild type peptide observed in panel (A).

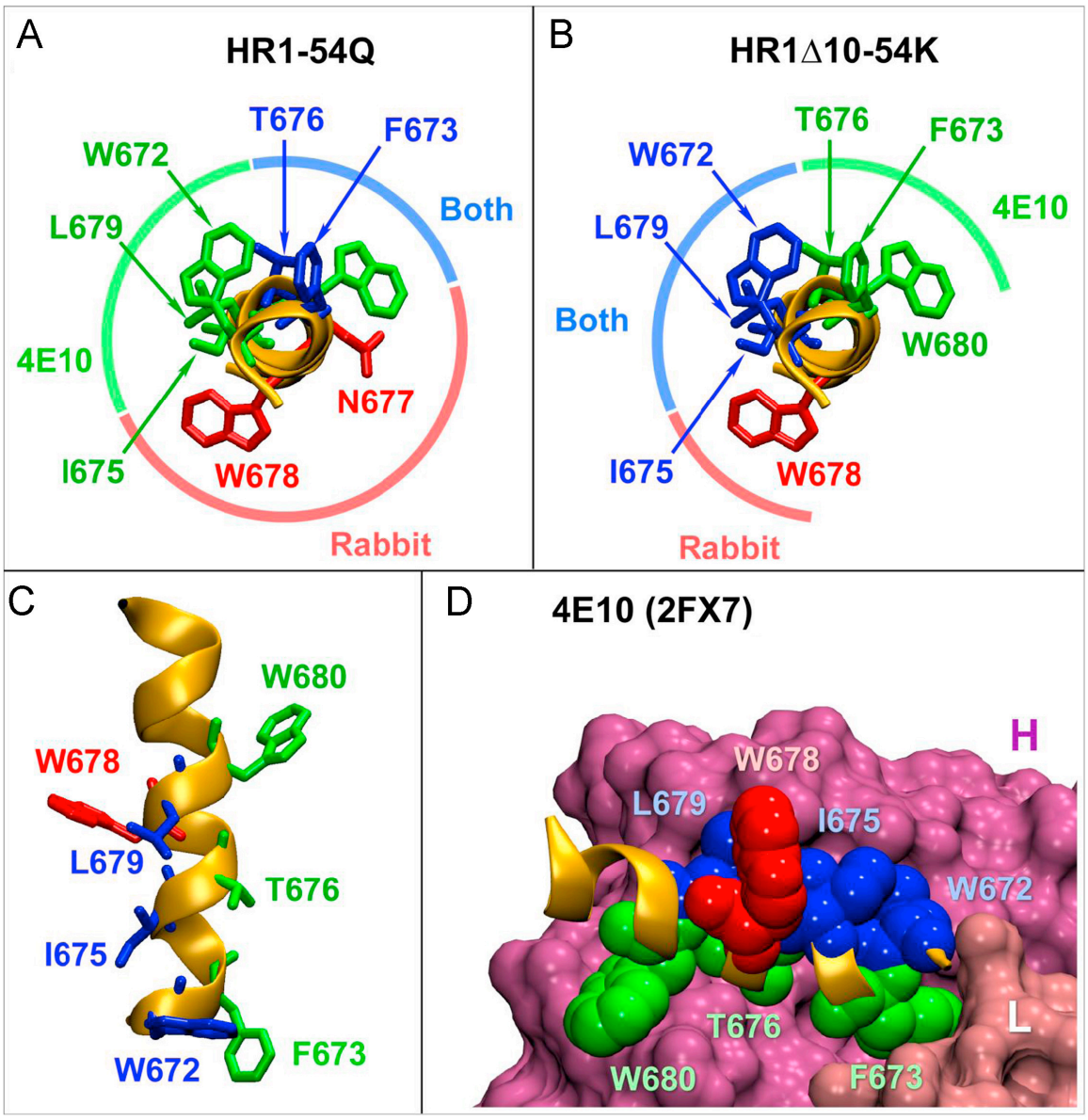

Fig 7. Comparison of critical binding residues for antibodies induced by HR1-Δ10-54K and gp41-HR1-54Q relative to 4E10. (A) Critical binding residues for 4E10 and antibodies induced by gp41-HR1-54Q are plotted onto the peptide co-crystalized with 4E10 (pdb: 2FX7) (17, 34-36). Residues critical for 4E10 or rabbit antibody only are shown in green and red, respectively. Residues important for both are shown in blue. (B) Critical binding residues for antibodies induced by HR1-Δ10-54K were also plotted onto the same peptide revealing significant difference from the pattern observed

with gp41-HR1-54Q. (C) A lateral view of the peptide displaying critical binding residues for 4E10 and HR1-Δ10-54K-induced antibodies. (D) Position of all the HR1-Δ10-54K critical residues in context of the 4E10 bound peptide. The heavy and light chains for the antibody are indicated as H and L.

2.7 References

1. **Li Y**, **Migueles SA**, **Welcher B**, **Svehla K**, **Phogat A**, **Louder MK**, **Wu X**, **Shaw GM**, **Connors M**, **Wyatt RT**, **Mascola JR**. 2007. Broad HIV-1 neutralization mediated by CD4-binding site antibodies. Nat Med **13**:1032–1034.

2. **Simek MD**, **Rida W**, **Priddy FH**, **Pung P**, **Carrow E**, **Laufer DS**, **Lehrman JK**, **Boaz M**, **Tarragona-Fiol T**, **Miiro G**, **Birungi J**, **Pozniak A**, **McPhee DA**, **Manigart O**, **Karita E**, **Inwoley A**, **Jaoko W**, **Dehovitz J**, **Bekker L-G**, **Pitisuttithum P**, **Paris R**, **Walker LM**, **Poignard P**, **Wrin T**, **Fast PE**, **Burton DR**, **Koff WC**. 2009. Human immunodeficiency virus type 1 elite neutralizers: individuals with broad and potent neutralizing activity identified by using a high-throughput neutralization assay together with an analytical selection algorithm. journal of virology **83**:7337–7348.

3. **Mouquet H**. 2014. Antibody B cell responses in HIV-1 infection. Trends Immunol **35**:549–561.

4. **Montero M**, **van Houten NE**, **Wang X**, **Scott JK**. 2008. The membrane-proximal external region of the human immunodeficiency virus type 1 envelope: dominant site of antibody neutralization and target for vaccine design. Microbiol Mol Biol Rev **72**:54–84– table of contents.

5. **Huang J**, **Ofek G**, **Laub L**, **Louder MK**, **Doria-Rose NA**, **Longo NS**, **Imamichi H**, **Bailer RT**, **Chakrabarti B**, **Sharma SK**, **Alam SM**, **Wang T**, **Yang Y**, **Zhang B**, **Migueles SA**, **Wyatt R**, **Haynes BF**, **Kwong PD**, **Mascola JR**, **Connors M**. 2012. Broad and potent neutralization of HIV-1 by a gp41-specific human antibody. Nature **491**:406–412.

6. **Purtscher M**, **Trkola A**, **Gruber G**, **Buchacher A**, **Predl R**, **Steindl F**, **Tauer C**, **Berger R**, **Barrett N**, **Jungbauer A**. 1994. A broadly neutralizing human monoclonal antibody against gp41 of human immunodeficiency virus type 1. AIDS Res Hum Retroviruses **10**:1651–1658.

7. **Stiegler G**, **Kunert R**, **Purtscher M**, **Wolbank S**, **Voglauer R**, **Steindl F**, **Katinger H**. 2001. A potent cross-clade neutralizing human monoclonal antibody against a novel epitope on gp41 of human immunodeficiency virus type 1. AIDS Res Hum Retroviruses **17**:1757–1765.

8. 2001. Broadly neutralizing antibodies targeted to the membrane-proximal external region of human immunodeficiency virus type 1 glycoprotein gp41. **75**:10892–10905.

9. **Habte HH**, **Banerjee S**, **Shi H**, **Qin Y**, **Cho MW**. 2015. Immunogenic properties of a trimeric gp41-based immunogen containing an exposed membrane-proximal external region. Virology **486**:187–197.

10. 2013. A gp41 MPER-specific Llama VHH Requires a Hydrophobic CDR3 for Neutralization but not for Antigen Recognition **9**:e1003202.

11. **Krebs SJ**, **McBurney SP**, **Kovarik DN**, **Waddell CD**, **Jaworski JP**, **Sutton WF**, **Gomes MM**, **Trovato M**, **Waagmeester G**, **Barnett SJ**, **DeBerardinis P**, **Haigwood NL**. 2014. Multimeric scaffolds displaying the HIV-1 envelope MPER induce MPER-specific antibodies and cross-neutralizing antibodies when co-immunized with gp160 DNA. PLoS ONE **9**:e113463.

12. **Lai RPJ**, **Hock M**, **Radzimanowski J**, **Tonks P**, **Hulsik DL**, **Effantin G**, **Seilly DJ**, **Dreja H**, **Kliche A**, **Wagner R**, **Barnett SW**, **Tumba N**, **Morris L**, **LaBranche CC**, **Montefiori DC**, **Seaman MS**, **Heeney JL**, **Weissenhorn W**. 2014. A fusion intermediate gp41 immunogen elicits neutralizing antibodies to HIV-1. J Biol Chem **289**:29912–29926.

13. **Ye L**, **Wen Z**, **Dong K**, **Wang X**, **Bu Z**, **Zhang H**, **Compans RW**, **Yang C**. 2011. Induction of HIV neutralizing antibodies against the MPER of the HIV envelope protein by HA/gp41 chimeric protein-based DNA and VLP vaccines. PLoS ONE **6**:e14813.

14. **Yi G**, **Lapelosa M**, **Bradley R**, **Mariano TM**, **Dietz DE**, **Hughes S**, **Wrin T**, **Petropoulos C**, **Gallicchio E**, **Levy RM**, **Arnold E**, **Arnold GF**. 2013. Chimeric rhinoviruses displaying MPER epitopes elicit anti-HIV neutralizing responses. PLoS ONE **8**:e72205.

15. 2008. Structural details of HIV-1 recognition by the broadly neutralizing monoclonal antibody 2F5: epitope conformation, antigen-recognition loop mobility, and anion-binding site. **384**:377–392.

16. 2004. Structure and mechanistic analysis of the anti-human immunodeficiency virus type 1 antibody 2F5 in complex with its gp41 epitope. **78**:10724–10737.

17. 2007. Structural basis of enhanced binding of extended and helically constrained peptide epitopes of the broadly neutralizing HIV-1 antibody 4E10. **365**:1533–1544.

18. **Pancera M**, **Zhou T**, **Druz A**, **Georgiev IS**, **Soto C**. 2014. Structure and immune recognition of trimeric pre-fusion HIV-1 Env. Nature.

19. 2010. Structural characterization of HIV gp41 with the membrane-proximal external region. **285**:24290–24298.

20. 2002. Antigenic properties of the human immunodeficiency virus transmembrane glycoprotein during cell-cell fusion. **76**:12123–12134.

21. 2004. Binding of the 2F5 monoclonal antibody to native and fusion-intermediate forms of human immunodeficiency virus type 1 gp41: implications for fusion-inducing conformational changes. **78**:2627–2631.

22. 2007. Exposure of the membrane-proximal external region of HIV-1 gp41 in the course of HIV-1 envelope glycoprotein-mediated fusion. **46**:1398–1401.

23. 2008. A fusion-intermediate state of HIV-1 gp41 targeted by broadly neutralizing antibodies. **105**:3739–3744.

24. 2011. Antibody mechanics on a membrane-bound HIV segment essential for GP41-targeted viral neutralization. **18**:1235–1243.

25. 2011. Direct antibody access to the HIV-1 membrane-proximal external region positively correlates with neutralization sensitivity. **85**:8217–8226.

26. **Julien JP**, **Lee JH**, **Cupo A**, **Murin CD**. 2013. Asymmetric recognition of the HIV-1 trimer by broadly neutralizing antibody PG9. *In*.

27. **Markosyan RM**, **Leung MY**, **Cohen FS**. 2009. The six-helix bundle of human immunodeficiency virus Env controls pore formation and enlargement and is initiated at residues proximal to the hairpin turn. journal of virology **83**:10048–10057.

28. 1998. A conformation-specific monoclonal antibody reacting with fusion-active gp41 from the human immunodeficiency virus type 1 envelope glycoprotein. **72**:10213–10217.

29. 2009. Oligomer-specific conformations of the human immunodeficiency virus (HIV-1) gp41 envelope glycoprotein ectodomain recognized by human monoclonal antibodies. **25**:319–328.

30. 1989. Generation of human monoclonal antibodies to human immunodeficiency virus. **86**:1624–1628.

31. 1990. Identification of sites within gp41 that serve as targets for antibody-dependent cellular cytotoxicity by using human monoclonal antibodies. **145**:3276–3282.

32. 1991. Epitope mapping of two immunodominant domains of gp41, the transmembrane protein of human immunodeficiency virus type 1, using ten human monoclonal antibodies. **65**:4832–4838.

33. 1991. Two immunodominant domains of gp41 bind antibodies which enhance human immunodeficiency virus type 1 infection in vitro. **65**:4169–4176.

34. 2010. Distinct conformational states of HIV-1 gp41 are recognized by neutralizing and non-neutralizing antibodies **17**:1486–1491.

35. 2008. Human immunodeficiency virus type 1 gp41 antibodies that mask membrane proximal region epitopes: antibody binding kinetics, induction, and potential for regulation in acute infection. **82**:115–125.

36. 1997. Neutralization of HIV-1 primary isolates by polyclonal and monoclonal human antibodies. **9**:1281–1290.

37. 2006. Structure-function analysis of the epitope for 4E10, a broadly neutralizing human immunodeficiency virus type 1 antibody. **80**:1680–1687.

38. 2014. Mechanism of HIV-1 neutralization by antibodies targeting a membrane-proximal region of gp41. **88**:1249–1258.

39. 2005. Broadly neutralizing anti-HIV antibody 4E10 recognizes a helical conformation of a highly conserved fusion-associated motif in gp41. **22**:163–173.

40. **Qin Y**, **Banasik M**, **Kim S**, **Penn-Nicholson A**, **Habte HH**, **LaBranche C**, **Montefiori DC**, **Wang C**, **Cho MW**. 2014. Eliciting neutralizing antibodies with gp120 outer domain constructs based on M-group consensus sequence. Virology **462-463**:363–376.

41. 2002. Emergence of resistant human immunodeficiency virus type 1 in patients receiving fusion inhibitor (T-20) monotherapy. **46**:1896–1905.

42. 2005. Human immunodeficiency virus type 1 env clones from acute and early subtype B infections for standardized assessments of vaccine-elicited neutralizing antibodies. **79**:10108–10125.

CHAPTER 3

EVALUATION OF A NOVEL MULTI-IMMUNOGEN VACCINE STRATEGY FOR TARGETING 4E10/10E8 NEUTRALIZING EPITOPES ON HIV-1 GP41 MEMBRANE PROXIMAL EXTERNAL REGION.

Saikat Banerjee[†], Heliang Shi[†], Marisa Banasik[†], Hojin Moon[§], William Lees[§], Yali Qin, Andrew Harley , Adrian Shepherd and Michael W. Cho

[†]These authors contributed equally.

[§]These authors contributed equally.

Abstract

The membrane proximal external region (MPER) of HIV-1 gp41 is targeted by broadly neutralizing antibodies (bnAbs) 4E10 and 10E8. In this proof-of-concept study, we evaluated a novel multi-immunogen vaccine strategy referred to as Incremental, Phased Antigenic Stimulation for Rapid Antibody Maturation (IPAS-RAM) to induce 4E10/10E8-like bnAbs. Rabbits were immunized sequentially, but in a phased manner, with three immunogens that are progressively more native (gp41-28x3, gp41-54CT, and rVV-gp160$_{DH12}$). Although nAbs were not induced, epitope-mapping analyses indicated that IPAS-RAM vaccination was better able to target antibodies towards the 4E10/10E8 epitopes than homologous prime-boost immunization using gp41-28x3 alone. MPER-specific rabbit monoclonal antibodies were generated, including 9F6. Although it lacked neutralizing activity, the target epitope profile of 9F6 closely resembled those of 4E10 and 10E8 (^{671}NWFDITNWLWYIK^{683}). B-cell repertoire analyses suggested the

importance of co-immunizations for maturation of 9F6, which warrants further evaluation of our IPAS-RAM vaccine strategy using an improved priming immunogen.

3.1 Introduction

To date, dozens of human monoclonal antibodies (mAbs) have been isolated from virus-infected patients that can neutralize a large number of HIV-1 variants from multiple clades (Huang et al., 2014; 2012; Pejchal et al., 2011; Scheid et al., 2011; Walker et al., 2011; 2009; Wu et al., 2010; Zwick et al., 2001). These broadly neutralizing antibodies (bnAbs) target a few select conserved sites of vulnerability on viral envelope glycoproteins gp120 and gp41 (for reviews, see (Georgiev et al., 2013; Haynes et al., 2014; Kwong et al., 2013; Mascola and Haynes, 2013; Mascola and Montefiori, 2010; McCoy and Weiss, 2013; van Gils and Sanders, 2013)). One of these targets is the membrane proximal external region (MPER), a highly conserved domain of ~22 amino acid residues situated at the C-terminal end of the gp41 ectodomain. The MPER is thought to play a critical role during the fusion between viral and cellular membranes (Muñoz-Barroso et al., 1999; Salzwedel et al., 1999). It is targeted by bnAbs 2F5, Z13e1, 4E10 and 10E8 (Huang et al., 2012; Purtscher et al., 1994; Stiegler et al., 2001; Zwick et al., 2001). 4E10 and 10E8 are particularly notable as they have been shown to neutralize ~98% of the HIV-1 isolates tested (Huang et al., 2012). 4E10 and 10E8 epitopes lie within the C-terminal 13 residues of the MPER (^{671}NWFDITNWLWYIK^{683}) and their crystal structures have been determined (Cardoso et al., 2007; Huang et al., 2012).

Despite having short, linear, simple alpha-helical epitopes, efforts to develop a vaccine that can induce 4E10/10E8-like bnAbs have been unsuccessful (see (Banerjee et al., 2016; Habte et al., 2015) and references therein). Since all of the immunogens evaluated could bind 4E10/10E8, the failure to induce similar bnAbs is not because antigens could not assume the correct epitope structures. Rather, it is likely due to the difficulty in inducing high levels of MPER-directed antibodies that can bind the neutralizing epitopes in the context of a whole trimeric gp120/gp41 complex. Unfortunately, this problem cannot be remedied simply by using a trimeric envelope complex because the MPER is immunorecessive compared to other epitopes that elicit type-specific or non-neutralizing antibodies. An additional challenge in inducing MPER neutralizing antibodies is that the structure of the gp41 subunit is highly dynamic as it undergoes significant structural changes to mediate fusion between viral and cellular membranes (Melikyan, 2008). The conformation, orientation and accessibility of the MPER neutralizing epitopes likely vary significantly at different stages of the fusion process, about which little is known at the present time. Further, the orientation of the MPER relative to the membrane surface or to the rest of the native gp120/gp41 trimeric complex is unknown. These factors make it difficult to design small, sub-domain immunogens.

Taken together, the major challenge in developing an MPER-based vaccine is designing immunogens and/or developing vaccine strategies that both force the immune system to focus antibody responses towards the MPER and also guide antibody evolution so that mature antibodies bind neutralizing epitopes on trimeric envelope spikes on the virus particles. Considering that antibody maturation will have to occur

during a relatively short timeframe, we postulated that inducing bnAbs against the MPER would be impossible to accomplish with a single immunogen using typical vaccine approaches. To address this problem, we devised a novel vaccine strategy, referred to as Incremental, Phased Antigenic Stimulation for Rapid Antibody Maturation, or IPAS-RAM. The basic concept is to prime the immune system using a small MPER-derived peptide to stimulate a broad spectrum of antibodies against the MPER, then selectively amplify those that bind the native structure by boosting with progressively more native immunogens. Although a number of studies recently reported sequential immunization with different immunogens, they used only a single immunogen during each immunization (Briney et al., 2016; Escolano et al., 2016; Haynes et al., 2012; Sok et al., 2016; Tian et al., 2016). What makes our IPAS-RAM strategy unique is that the immune system is exposed to different, but related, immunogens simultaneously in a phased manner, such that B cells stimulated by a smaller immunogen can concurrently engage common epitopes on a larger, more native immunogen. By repeating this process using incrementally more native antigens, we hypothesized that the immune system can better target MPER neutralizing epitopes and that antibodies could undergo the maturation process more efficiently.

In this proof-of-concept study, we evaluated the IPAS-RAM vaccine strategy using three immunogens in rabbits: An MPER-based polypeptide, a membrane-bound gp41 mini-protein, and a full-length gp160. We hypothesized that a peptide-based priming antigen would be highly effective in eliciting antibodies against the MPER, and that boosting with progressively larger and more “native” antigens would enable select antibodies to mature into bnAbs capable of binding gp41 as it appears in the native

trimer. Although we did not succeed in eliciting neutralizing antibodies, results of our study demonstrate proof-of-principle for the IPAS-RAM vaccine strategy and identify ways to improve it.

3.2 Results

3.2.1 Immunogens.

To focus antibody responses towards the MPER, we generated an immunogen designated gp41-28x3, which consists of three tandem copies of the C-terminal 28 a.a. of gp41 ectodomain (Fig. 1A). The immunogen was produced initially as a fusion protein (HR1-HR2-28x3) by adjoining 28x3 to the HR1-HR2 regions of gp41, which forms a six-helix bundle (6HB) (Shi et al., 2010). This was done because we had observed that HR1-HR2 6HB allows high level protein expression in *E. coli* (Fig. 1B; unpublished data). The HR1-HR2 portion was subsequently removed by thrombin digestion (Fig. 1C). Three 28-mer peptides were linked together to increase its immunogenicity without requiring conjugation to a heterologous carrier protein. All of the bnAbs tested (2F5, Z13e1, 4E10 and 10E8) bound gp41-28x3 (Fig. 1D). However, 10E8 binding was about ~50-fold weaker than 2F5 and ~10-fold weaker than Z13e1 and 4E10, suggesting that the conformation of the 10E8 epitope may not be optimal.

The conformation, orientation, and/or accessibility of the neutralizing epitopes on the MPER are likely affected by the membrane surface as well as other proximal gp41 domains. To present the MPER in a more native-like setting, we generated a second immunogen designated gp41-54CT. It comprises the C-terminal 54 a.a. of the gp41 ectodomain that includes the HR2 domain and the MPER, along with the

transmembrane domain and the cytoplasmic tail (CT). We hypothesized that this immunogen would selectively amplify a subset of antibodies induced by gp41-28x3 that could bind the MPER in the context of the membrane surface and HR2. HEK-293T cells transfected with the plasmid encoding gp41-54CT could be detected using 2F5, Z13e1 and 4E10 by flow cytometry analyses (Fig. 1E), indicating cell surface expression of the protein and correct conformation of neutralizing epitopes. Both gp41-54CT and gp41-28x3 are based on M group consensus sequence (Gao et al., 2005). For the final boost, we generated a recombinant vaccinia virus expressing the full-length gp160 of the HIV-1_{DH12} strain (rVV-$gp160_{DH12}$). A schematic diagram showing the relative sizes of the three immunogens are illustrated in Fig. 1F.

3.2.2 Immunization and evaluation of antibody responses.

To evaluate IPAS-RAM vaccine strategy, rabbits (R1, R2 and R3) were first immunized with gp41-28x3 only (Fig. 2A). Four weeks later, a combination of gp41-28x3 and gp41-54CT was administered, with the latter delivered by DNA electroporation. Instead of immunizing with just gp41-54CT, gp41-28x3 was also included because we postulated that immunizing with both immunogens would preferentially stimulate antibody responses towards epitopes present on both immunogens (*i.e*. the C-terminal 28 a.a.). Similarly, on week 11, a combination of gp41-54CT and rVV-$gp160_{DH12}$ was administered. After the first immunization, the antibody titers reached $>10^3$ and the titers increased by about 100 fold after the second immunization (Fig. S1A). However, because significant increases in antibody responses were not observed after the third immunization, a fourth immunization was

given on week 29 using gp41-28x3 and rVV-gp160$_{DH12}$. The long resting period was given so that antiviral immune responses against the vaccinia virus vector induced after the third immunization would subside. Antibody responses increased only slightly after the fourth immunization.

To identify immunogenic linear epitopes, ELISAs were conducted with overlapping 10-mer peptide sets biotinylated either at the N- or C-terminal ends as we previously reported (Banerjee et al., 2016; Habte et al., 2015). After the first immunization, very little response was detected (Fig. 2B). However, after the second immunization, strong antibody responses were detected against peptides in the cluster II immunodominant region just upstream of the 2F5 epitope (peptides 653 and 656; also peptide 650 for rabbit R2). After the third immunization with gp41-54CT and rVV-gp160$_{DH12}$, the cluster II region still remained immunodominant. However, antibody responses appeared against other linear epitopes. R2 showed good binding to peptide ^{662}ALDKWASLWN^{671} containing the 2F5 epitope. This rabbit also showed low level reactivity against other C-terminal peptides (665, 668, 671 and 674). Antiserum from R3 bound to peptides ^{668}SLWNWFDITN^{677} and ^{671}NWFDITNWLW^{680} that contain parts of the 4E10 and 10E8 epitopes. R1 and R2 also recognized additional peptides within the HR2 domain (peptides 629 and 638). Except for peptides 653 and 656, antibody responses against peptides exhibited animal-to-animal variation. The fourth immunization with gp41-28x3 and rVV-gp160$_{DH12}$ further enhanced anti-MPER antibodies in R2 (peptides 668 and 671) and R3 (peptides 671 and 674). Interestingly, antibody responses against many of the upstream peptides (629, 638, 650, 653 and 656) diminished significantly for R2.

To examine whether antibodies induced by the IPAS-RAM vaccine strategy are different from those induced by a typical homologous prime-boost immunization, another group of rabbits (R4, R5 and R6) was immunized with gp41-28x3 (Fig. 2C). Strong antibody responses were induced after the first immunization (~10^4 to ~10^5; Fig. S1B). Since these rabbits received the same vaccination as the animals in the IPAS-RAM group, the higher antibody titers suggest animal-to-animal variations. Antibody titers continued to increase with each successive immunization (Fig. S1B). Not surprisingly, 10-mer peptides 653 and 656 were most immunogenic (Fig. 2D), which might be the reason why they were also immunogenic in the IPAS-RAM group. After the fourth immunization, high-level antibodies were detected against peptide 674 all three animals. R4 and R6 mounted strong antibody responses against the 671 peptide that contains 4E10/10E8 epitopes. Overall, the gp41-28x3 homologous prime-boost immunization induced more robust antibody responses against gp41 MPER than the IPAS-RAM vaccine. Interestingly, antibodies from rabbit R6 reacted strongly to peptides 647 even though gp41-28x3 does not extend this far out into the N-terminus. One likely possibility is that antibody responses against ^{656}NEQE^{659} could be cross-reacting against ^{651}NQQE^{654}. Another possibility is that our gp41-28x3 preparation could have minor contamination of the HR1-HR2 fragment or undigested HR1-HR2-28x3.

3.2.3 Antibody responses against the C-terminal 13 a.a. residues of gp41 MPER.

Despite the overwhelming immunodominance of the cluster II region, antibody responses were induced towards the C-terminal end of the gp41 ectodomain, where

4E10/10E8 epitopes are located, in five of the six rabbits. However, no serum neutralizing activity was detected (data not shown). Regardless, we were interested in determining whether antibodies against the C-terminal 13 a.a. were qualitatively different between the two vaccine groups. We examined which amino acid residues were targeted by ELISA using two panels of alanine mutant peptides. First, we used peptides that were biotinylated at the C-terminal K683 (Fig. 3A and 3B) as previously described (Banerjee et al., 2016; Habte et al., 2015). As noted previously, the D674A mutation affects the helical conformation of the peptide (Banerjee et al., 2016; Habte et al., 2015); thus, results should be interpreted with some caution. For R2, the four residues that were most affected besides D674 were F673, N677, W678 and L679. Residues important in R3 were N671, W672, F673 and N677; the first three of which are absolutely critical for 10E8. Of all the immunogens we evaluated to date (Banerjee et al., 2016; Habte et al., 2015), this is the first time we targeted all three of these residues. For R4, the pattern was similar to R3, except that W6762 was not targeted. The epitope targeting profile for R5 was quite unique in that antibodies targeted W680 and Y681. The overall pattern for R6 resembled that of R2, except for not targeting T676 and N677.

Because peptides that were biotinylated at the C-terminal K683 might not allow binding of some antibodies (*e.g.* those that approach the peptide from the C-terminal end), the analyses were repeated using peptides that were biotinylated at the N-terminus with a two glycine spacer (Fig 3C and 3D). As expected, somewhat different patterns were observed, especially for R3, which showed marked difference (>40%) for residues N671, N677, W678, W680, Y681 and K683. For rabbits R4, R5 and R6,

however, many of the targeted residues identified using one peptide set are simply a subset of residues identified using the other peptide set. For example, for R6, the important residues were F673, I675, W678 and L679 using peptides biotinylated at the C-terminus while additional N677 and W680 were identified using peptides biotinylated at the N-terminus.

3.2.4 Generation and characterization of MPER-specific monoclonal antibodies.

Although target epitope profiling with whole serum provides some sense of which residues are critical for antibody binding, it does not provide clear information due to the polyclonal nature of antibodies. To better evaluate MPER-directed antibodies in R3 that might be targeting the ^{671}NWF^{673} residues, and to gain insights into how we might improve immunogens and/or vaccine strategies, MPER-specific mAbs were generated and characterized. Hybridomas were screened initially with gp41-28x3, and then with 15-mer peptides containing epitopes for 2F5 (^{657}EQELLALDKWASLWN^{671}) and 4E10/10E8 (^{669}LWNWFDITNWLWYIK^{683}). Based on the level of reactivity and epitope targeting profiles, three hybridomas (6C10, 9F6 and 21B5) were selected for detailed characterization. Based on the binding to various 15-mer peptides, the epitope for 6C10 was mapped approximately to ^{657}EQELLAL^{663}, just upstream of the 2F5 epitope (Fig. 4A). Since 6C10 epitope is situated within the two 10-mer peptides that were most reactive (^{653}QEKNEQELLA and ^{656}NEQELLALDK; Fig. 2), 6C10 could represent the most predominant antibody induced.

Although 9F6 and 21B5 strongly bound ^{669}LWNWFDITNWLWYIK^{683}, both failed to bind upstream (^{665}KWASLWNWFDITNWL^{679}) and downstream

(^{673}FDITNWLWYIKIFIM^{686}) peptides (Fig. 4A). This result suggested that the critical residues are also present within the four terminal residues at either ends of the 669-peptide (*i.e.* LWNW and WYIK). Alternatively, but not exclusively, the presence of four residues on 665- and 673-peptides at the N- or C-terminal ends, respectively (*i.e*. KWAS and IFIM), could be interfering with antibody binding. Additional mapping analyses with shorter peptides indicated that 9F6 epitope resided within ^{671}NWFDITNWLW^{680}, whereas 21B5 epitope extended further downstream. 21B5 bound the N-terminally biotinylated $_{GG}$NWFDITNWLWYIK^{683} peptide, but not the ^{671}NWFDITNWLWYIK^{683} peptide biotinylated at the C-terminal lysine, suggesting that the antibody binds near K683 and that biotin sterically interferes with access. Interestingly, we isolated twelve 21B5-like hybridomas that bound both the N- and C-terminally biotinylated 13-mer peptide.

3.2.5 Detailed epitope mapping of 9F6 and 21B5 mAbs.

To better define the 9F6 and 21B5 epitopes, ELISAs were performed using the two panels of scanning alanine mutant peptides as described in Fig. 3. For comparison, 4E10 and 10E8 were also analyzed. As shown in Fig. 4B, mutating N671, W672, F673, T676 and N677 residues to alanine markedly diminished 9F6 binding when tested with C-terminally biotinylated peptides. To a lesser degree, mutating W680 also reduced binding. This target profile was remarkably similar to that of 4E10, except for I675 and L679, which were recognized by 4E10 but not by 9F6. Conversely, 9F6 targeted N677 whereas 4E10 did not. A similar targeting profile was observed when N-terminally biotinylated peptides were used (Fig. 4C). However, one major difference was the

W680A mutation, which showed a severe binding defect for both 9F6 and 4E10 using the N-terminally biotinylated peptide. It should be noted that while mutating W680 only moderately reduced peptide binding (Brunel et al., 2006), it is important for neutralization by 4E10 (Brunel et al., 2006; Zwick et al., 2005). Interestingly, the K683A mutation reduced 9F6 binding to 60% of the wild type even though the 671 10-mer peptide (^{671}NWFDITNWLW^{680}) was sufficient to bind the mAb efficiently. The results are graphically summarized in Fig. 4D, which show significant overlap between 9F6 and 4E10/10E8 binding.

Overall, the antibody targeting profile of 10E8 was fairly similar to that of 4E10, although there were some notable differences (*e.g.* L679). Mutated residues that most severely diminished 10E8 binding were N671, W672, F673, I675, T676 and W680. To a lesser extent, mutating N677 and W678 also reduced recognition. The reduced binding by the W678A mutation was somewhat unexpected since it resides on the opposite side of the neutralization face, although Huang *et al.* (Huang et al., 2012) reported that mutating this residue reduced peptide inhibition of 10E8 neutralization by ~20%. Also unexpected was the reduction of only 40% for the K683A mutation, when R683 was critical (Huang et al., 2012). Similarly, the L679A mutation resulted in only a slight reduction when it had previously shown to exhibit a moderate effect (~50% reduction in neutralization inhibition) (Huang et al., 2012).

The five most critical residues for 21B5 binding were W672, N677, W678, W680 and Y681. Thus, in contrast to 9F6, 21B5 targeted W678 and Y681, both which are located directly on the opposite side of the neutralization face (Fig. 4D). Conversely,

21B5 failed to target N671 and only weakly bound F673, both of which are targeted by 9F6, as well as 4E10 and 10E8. Overall, 9F6 better resembled 4E10/10E8 than 21B5.

3.2.6 Sequence analysis of MPER-binding mAbs

To determine the origin of MPER-specific mAbs, antibody genes were PCR-amplified from hybridomas, cloned and sequenced. Analyses indicated that 6C10 heavy chain was derived from IGHV1S45*01 germline, while 9F6 and 21B5 originated from IGHV1S40*01(Fig. S2A-C). 6C10 and 21B5 kappa chains were derived from the IGKV1S32*01 and IGKV1S44*01 germlines, respectively. Unfortunately, we have not been able to recover a productive light chain from the 9F6 hybridoma using known primer sets for either kappa or lambda chains, despite exhaustive efforts. We suspect our primers may not bind the 9F6 light chain, due to either somatic mutations in the primer-binding site or suboptimal design based on incomplete information about the rabbit germline repertoire. Among the three mAbs, the heavy chain of 9F6 was the most conserved to germline, sharing 96.53% nucleotide identity, followed by 6C10 and 21B5 (91.32% and 87.85%, respectively). Similarly, the kappa chain of 6C10 was less divergent than that of 21B5 (90.32% and 89.25%, respectively). Surprisingly, high levels of mutations were observed in FR1 for 6C10 and 21B5. The importance of these mutations in antibody function has not been determined.

Of the twelve 21B5-like mAbs that could bind C-terminally biotinylated 671-13mer peptide, five representative hybridomas (21B6, 6F9, 31A4, 27A1 and 17-4H2) were selected for further characterization. Sequence analyses showed that all five mAbs belonged to the same clonal family as 21B5, as determined by predicted germline,

junction, and CDR3 identity (Fig. S2D). It should be noted, however, that IMGT predicted 21B6 heavy chain might be derived from the IGHV1S45*01, which is highly related to IGHV1S40*01 and differ only at the very 5' end of the V-gene. Although convergent affinity maturation of antibodies derived from different germlines is a possibility, we suspect a more likely scenario is a gene conversion event between the two V genes following V-D-J recombination. The heavy chains of 21B5-like mAbs were less divergent from the germline than 21B5 (91.93%-97.22% identity). Differences from germline were most highly concentrated in the CDR2. The kappa chains were more divergent than the heavy chains, with germline identity ranging from 87.81% to 89.96%. Of note, when comparing similarities versus differences of 21B5 to the other clonal family members (Fig. S2D, cyan highlights versus yellow), mutations in the kappa chain were much more conserved across the family than the heavy chain, where 21B5 represents a more distinctive sequence.

3.2.7 Bioinformatics analyses of B cell repertoire and antibody maturation.

To gain insights into possible maturation pathways of the MPER-specific mAbs, the antibody repertoire was analyzed by NGS of PBMC collected four days after the first three immunizations (A1, A2, A3), as well as terminal PBMC (TP) and spleen (TS). A total of 13.3M raw reads were obtained, from which 290,848 unique productive heavy chain and 89,760 kappa chain sequences were derived (Table S1). The lambda chain was not sequenced, as it is underutilized in rabbits (Appella et al., 1974). Clonally related heavy chain sequences were inferred, and the full-length sequences related to MPER-specific mAbs were determined (Table S2). A total of 19,133 clonally related sequence clusters were found. Of these, 10,924 were represented in more than one

sample (Fig. S3). Heavy chains used primarily two V-genes (IGHV1S40 and IGHV1S45) and J-gene IGHJ4 (Fig. S4). D-gene usage was more diverse. The most commonly used V-gene for the kappa chain was IGKV1S32 (Fig. S5). Note that the apparent predominant usage of kappa J-gene IGKJ1-2 is an artifact from amplification with a 3' primer that closely resembles the IGKJ1-2*01 sequence.

The interrelationships between all sequenced antibody heavy chains ("heavy chain antibodyome") are graphically shown in Fig. 5A, and the clonal families of mAbs 6C10, 9F6 and 21B5 are highlighted. CDR3 spectratypes of the heavy and kappa chains are plotted in Fig. 5B. The non-Gaussian distribution of the CDR3 length suggests active antibody developments. While kCDR3 length remained fairly constant, peaking at 12 a.a., significant variation in HCDR3 length was observed after each immunization, indicating fluctuations in antibody repertoires generated in response to different antigenic stimulations. After the second immunization, when a strong anamnestic response against gp41-28x3 was observed (Fig. 2B), the peak HCDR3 length changed from 13 to 14 a.a. (Fig. 5B). It is worth noting that, 6C10 has a HCDR3 length of 14 a.a. (ARDLDDVIGWNFGW), suggesting that the shift in the peak HCDR3 length could be due to a strong antibody response against the cluster II immunodominant region (peptides 653 and 656; Fig. 3). Consistent with this notion, IGHV1S45 and IGKV1S32, from which 6C10 heavy and light chains were derived, became the most dominant V-genes used after the second immunization (Fig. S4, S5). After the third immunization, there was a notable shift from 14 back to 13 a.a, which could be due to antibody responses directed against new antigens (*e.g*. vaccinia virus

antigens). The dual peaks in the terminal PBMC sample (TP) may be the result of the fourth immunization with gp41-28x3 and rVV-gp160$_{DH12}$.

The development of the B cell repertoire after the third immunization was dominated by novel clonal families (Fig. 5C). This dominance could also be seen in Fig. 5A, where a large number of distinct clusters are predominantly associated with the A3 sample. The vast majority of the novel clonal families are expected to be antibodies against vaccinia virus antigens. Consequently, the proportions of clonal family sizes of the MPER-specific mAbs all contracted. Evaluation of the size of clonal family sizes showed clear immunodominance of 6C10 over both 21B5 and 9F6 epitopes (Fig. 5C). Importantly, the development of the 6C10 lineage had already begun after the first immunization, whereas significant levels of 9F6 and 21B5 clonal families were observed only in terminal samples (TP and TS; Figs. 5C and 5D). Alternatively, 9F6 and 21B5 could have begun to expand after the fourth immunization with gp41-28x3 and rVV-gp160$_{DH12}$, a sample we were not able to analyze.

Since 9F6 best targeted the 4E10/10E8 binding site, its maturation process was examined in greater detail (Fig. 6). A total of 12 substitutions were observed compared to the inferred naïve germline sequence, of which only six were within the CDRs (Fig. 6A). Two of these mutations are potentially the result of a gene conversion event with IGHV1S45 (G14E and A15G; Fig. 6B). Phylogenetic lineage analyses indicated that there was another mutation at an early stage of antibody evolution (T102A), which reverted back to the germline sequence (A102T). All substitutions except for V106L and T108I are observed in reads from the A1 sample point onwards. V106L and T108I

are observed only in the terminal sample. The sequence identity/divergence plot of 9F6 heavy chain illustrates good coverage of the development history of the mAb (Fig. 6C).

3.3 Discussion

The antigenic repertoire of HIV-1 envelope proteins in virus-infected patients is vast, largely due to chronic virus replication continuously generating antigens with different amino acid sequences and variable glycosylation patterns. In addition, a large number of protein fragments is likely generated by proteolytic degradation. This antigenic complexity is further increased by the drastic conformational changes that occur when the protein mediates fusion between viral and cellular membranes. The transient nature of the gp41 structure makes targeting the MPER neutralizing epitopes even more challenging. Despite this unfavorable antigenic environment, the isolation of bnAbs against the MPER from HIV-1-infected patients, albeit rare, demonstrates the possibility of inducing such antibodies through vaccination. An important question is whether similar bnAbs can be induced using a single immunogen or whether it requires multiple immunogens used in concert. If the latter, then, how many and which immunogens? Would a single cocktail of all of the immunogens together work, or would administering them in a particular, sequential order be better?

During the past three decades, the vast majority of HIV-1 vaccine efforts focused on designing immunogens that can induce bnAbs. The failures to elicit bnAbs using immunogens that are deemed antigenically native dictate that more efforts should be made in exploring novel vaccine strategies, regimens, and/or formulations. Typically, vaccines consist of the same immunogen(s) administered multiple times (*i.e.*

"homologous prime-boost" immunization; Fig. 7A). However, as discussed earlier, it might be difficult, if not impossible, to induce MPER bnAbs using this strategy with a single immunogen. While a trimeric envelope complex may present the MPER in the native conformation, specifically targeting 4E10/10E8 epitopes may be like looking for the proverbial needle in a haystack due to the presence of multiple highly immunogenic epitopes on gp120. Even if naïve B cells that encode precursor B cell receptors against the neutralizing epitopes can be stimulated, they may fail to expand efficiently due to the competitive environment of germinal centers. Successive immunizations with the same immunogen would preferentially amplify antibodies targeting the immunodominant epitopes, resulting in the induction of non- or type-specific-neutralizing antibodies.

In this report, we described our initial effort to explore a novel IPAS-RAM vaccine strategy designed to enhance antibody responses against non-immunogenic target epitopes and expedite the antibody maturation process. The strategy is different from typical vaccinations in that the immune system is exposed to different immunogens sequentially, starting with a small immunogen to focus immune responses towards the desired region (*e.g.* the MPER) and induce a large antibody repertoire against all possible epitope conformations (Fig. 7B). Subsequent immunizations are carried out with antigenically distinct, but related, immunogens that are progressively more "native" to guide the antibody maturation process. The final immunization is done with a whole envelope protein complex (*i.e.* trimeric gp120/gp41 on membrane) to increase the likelihood that mature antibodies can bind MPER epitopes in the context of native envelope spikes. During the course of the immunization, B cell clones that encode antibodies against epitopes that are absent or inaccessible on boosting immunogens

are not expected to expand and are filtered out. The major hallmark of the IPAS-RAM vaccine strategy that is uniquely different from a simple “heterologous prime-boost” immunization is that immunogens are administered in a phased manner. Because the antigens are phased, anamnestic immune responses against previously administered immunogens could rapidly boost antibody responses against related epitopes on new immunogens (“concurrent boosting”). We hypothesized this strategy of “antigenic relays” and “antigenic filters” would expedite clonal expansion of target-epitope-specific B cells and accelerate antibody maturation, while minimizing antibody responses against non-conserved epitopes. We further hypothesized that the presence of different, but related immunogens simultaneously would allow “immunological crosstalks” (*i.e.* B cells stimulated by one immunogen engaging common/similar epitopes on different immunogens) that could facilitate antibody maturation.

To demonstrate proof-of-concept, we examined the IPAS-RAM vaccine strategy using three immunogens, gp41-28x3, gp41-54CT, and rVV-gp160$_{DH12}$. Although we failed to induce neutralizing antibodies, we identified one mAb (9F6) with an epitope targeting profile that closely resembled those of 4E10 and 10E8. As typical for any exploratory study, we acknowledge that our study raised more questions than it answered. They include (1) what led to the induction of 9F6 and why does it fail to neutralize HIV-1; (2) did we use optimal immunogens; (3) is the complex IPAS-RAM vaccine regimen any better than traditional vaccine strategies with simpler regimens; and (4) can the IPAS-RAM vaccine strategy be improved?

Based on NGS analyses (Fig. 5C), we did not see significant levels of either 9F6 or 21B5-related antibodies until the terminal samples (TP and TS). While we cannot be

certain as to what exactly led to the induction of 9F6, bioinformatics analyses provided some insights. In the A1 sample, % identities of the heavy chain V-gene of 9F6 clonal family members to germline ranged from 76.7% to 97.22%, which suggests that repertoire diversification had started. Since we did not have pre-immune samples to examine, it is possible that they existed prior to immunization with gp41-28x3. However, further expansion of all three mAb clonal families after the second immunization (A2) suggests that they were induced by gp41-28x3. The relative expansion of the 9F6 clonal family, as well as that of 21B5 and 6C10, shrank in A3, most likely due to dominant immune responses against vaccinia virus antigens. TP and TS samples were taken at the conclusion of the study. Thus, the marked expansion of the 9F6 and 21B5 clonal families could be due to the co-immunization with rVV-$gp160_{DH12}$ and gp41-28x3 (fourth immunization) and/or the final immunization with gp41-28x3 alone (administered to increase efficiency of generating MPER-specific hybridomas). However, three lines of evidence suggest that co-immunization might be responsible: (1) We did not see many 9F6 clonal family members in the A2 sample when a strong anamnestic response against gp41-28x3 was induced. (2) In phylogenetic analyses, the closest sequences to the evolutionary development of 9F6 (shown as intermediate sequences in Figs. 6B and 6C) all came from the TP sample. Further, the substitutions V106L and T108I were observed only in the terminal samples. (3) Epitope targeting profiles of antibodies induced by homologous prime-boost immunization using gp41-28x3 only, did not resemble that of 9F6 (compiled in Fig. 4E). Notably, none of the three animals mounted significant antibody responses towards W672 and R683. Thus, while gp41-28x3 may have initiated repertoire diversification,

subsequent exposure to both gp41-28x3 and rVV-gp160$_{DH12}$ simultaneously likely have guided the evolution of an 9F6 intermediate to its mature form. Several studies have reported the potential benefits of co-immunizations by combining different vaccine modalities (*e.g*. protein and DNA vaccination; (Jaworski et al., 2012; Krebs et al., 2014; J. Li et al., 2013; Patel et al., 2013)). However, potential benefits could also arise from synergy between different immunogens, regardless of the vaccine modality. At the present time, the contribution of gp41-54CT/ rVV-gp160$_{DH12}$ co-immunization is unclear.

The major difference in epitope targeting between 21B5 and 4E10/10E8 is that 21B5 recognizes W678 and Y681, which lie on the opposite side of the neutralizing face. Thus, the lack of neutralizing activity is not unexpected. In contrast, all of the residues critical for 9F6 binding lie within the footprint of 4E10/10E8 (with a possible exception of N677). Therefore, it is not readily apparent why 9F6 fails to exhibit neutralizing activity. It is possible there are steric clashes between 9F6 and other parts of the trimeric envelope complex and/or the membrane surface that prevent 9F6 binding. Comparing a co-crystal structure of 9F6 bound to an MPER peptide to those of 4E10 and 10E8 could provide critical insights.

Linear epitope mapping analyses identified two peptides highly reactive to immune sera (^{653}QEKNEQELLA^{662} and ^{656}NEQELLALDK^{665}; Fig. 2). This region corresponds to the cluster II immunodominant domain of gp41 (Xu et al., 1991), suggesting that our vaccine regimen tested in rabbits recapitulated antibody responses mounted against gp41 in humans. A large fraction of the antibodies that bind these peptides likely represents clonal family members of 6C10 mAb, which targets ^{657}EQELLAL^{663} (Fig. 4A). Immunodominance of the 6C10 epitope was also evident

from the antibodyome data. After the first immunization, the repertoire of antibodies clonally related to 6C10 was about 25-fold greater than those of 9F6 and 21B5 combined (Fig. 5C). That increased to about 62-fold after the second immunization. Despite partial epitope overlap with 2F5, 6C10 lacked neutralizing activity. Interestingly, the llama mAb 2H10 that recognizes ^{657}EQELLELDK^{665} has neutralizing activity (Lutje Hulsik et al., 2013), suggesting the importance of targeting further downstream of the 6C10 epitope for inhibiting gp41 function. Other factors such as antibody footprint size, angle of approach, affinity, *etc*. could also play important roles. Detailed mapping analyses and structural studies of 6C10 could provide additional insights.

All of the MPER bnAbs (2F5, Z13e1, 4E10 and 10E8) bound gp41-28x3 (Fig. 1D). However, 2F5 binding was significantly better than 10E8 (~50-fold) and 4E10/Z13e1 (~3- to 5-fold). This might explain why the nearby 6C10 epitope was so immunogenic. Thus, retrospectively, gp41-28x3 may not have been an ideal priming immunogen to use for eliciting 4E10/10E8-like antibodies. An immunogen that has better exposure and/or affinity to 10E8, as well as one devoid of the cluster II region, might have elicited stronger antibody responses towards 4E10/10E8 epitopes. In this regard, a 4E10 scaffold-based immunogen, T88, which induced non-neutralizing antibodies with a target epitope profile similar to 4E10 (Correia et al., 2010), could have been a better priming immunogen for our IPAS-RAM vaccine strategy.

Our NGS analyses of antibody genes largely focused on heavy chain sequences related to three mAbs against a few select linear epitopes. These sequences likely represent a tiny fraction of the vast range of clonal families and superfamilies whose development were stimulated by immunizations. We were unable to employ single-cell

PCR amplification of antibody genes from FACS-sorted antigen-specific B cells due to the lack of suitable markers on rabbit plasmablasts and non-specific binding of MPER immunogens to cells. Isolation of additional mAbs could facilitate and expand the scope of analyses. Alternatively, the NGS analyses could be conducted in conjunction with proteomics analyses of antibodies purified using immunogens (Cheung et al., 2012), although we would not be able to pair authentic heavy and light chains. While further work is needed to dissect the specific developmental stimuli and mechanisms, these analyses have clearly demonstrated the utilities of repertoire sequencing towards better understanding of antibody responses in vaccine settings. Importantly, we have demonstrated that rabbits can be a useful animal model to evaluate antigen-specific B cell population dynamics using NGS data despite the fact that they use gene conversion, in addition to somatic hypermutation, to generate antibody repertoire.

3.4 Conclusions

In this study, we evaluated a novel IPAS-RAM vaccine strategy to induce 4E10/10E8-like bnAbs. Despite our unsuccessful efforts, which we attribute largely to the use of a suboptimal priming immunogen, results of the study demonstrate potential benefits of sequential immunization with multiple immunogens in a phased manner. Thus, further evaluation of the IPAS-RAM vaccine strategy is warranted, perhaps using a better priming immunogen. The target epitope profile of the 9F6 mAb we isolated closely resembled those of 4E10 and 10E8. To our knowledge, 9F6 may be the best vaccine-induced mAb that mimics 10E8. Comparative structural analyses of 9F6 with

10E8 could provide insights into why 9F6 lacks neutralizing activity and how we might be able to improve our vaccine strategy.

3.5 Materials and methods

3.5. Immunogen generation.

The priming immunogen, gp41-28x3, was produced from a pHR1-HR2-28x3 construct. This plasmid was constructed from pGEX-2T gp41(30) (Penn-Nicholson et al., 2008) through multiple subcloning steps and site-directed mutagenesis, details of which will be described elsewhere. The pHR1-HR2-28x3 construct is based on a pET-21a vector. HR1-HR2-28x3 was expressed in *E. coli*, solubilized with 8M urea, and purified using Ni-NTA as we previously described (Banerjee et al., 2016; Habte et al., 2015; Qin et al., 2014). The fusion protein was cleaved with thrombin (GE Healthcare; cat#27-0846-01) and the gp41-28x3 fragment was purified using Ni-NTA, dialyzed into PBS (pH 8.0) and stored at -80°C.

To clone gp41-54CT construct, a region containing the gene (^{630}EWEREISN…..QGLERALL^{856}; numbering based on HXB2) was PCR amplified from pcDNA-MCON6gp160 (kindly provided by Dr. Beatrice Hahn (Gao et al., 2005)) using a sense primer 5'-cgcggatcc GAG TGG GAG CGC GAG ATC-3’ and an antisense primer 5’-cggaagc TTA atg gtg atg atg gtg atg CAG CAG GGC GCG CTC CAG-3’. Six-histidine tag was incorporated into the primer. The resulting PCR product was digested with BamHI and HindIII (underlined) and inserted into the corresponding sites on a vector based on pcDNA*MCON6-gp120-OD described previously (Qin et al., 2014).

The recombinant vaccinia virus expressing gp160 of the HIV-1_{DH12} isolate (rVV-$gp160_{DH12}$) was generated from a single virus variant that was plaque-purified from the Dryvax vaccine obtained from the CDC. First, the plaque-purified virus was attenuated by disrupting the B8R gene by homologous recombination using a pB8R-IFN-g plasmid that encodes a human IFN-g gene in between two segments of the B8R gene, resulting in rVVΔB8R-IFN-g. Next, a second plasmid pNVVDHenv (Cho et al., 2001; 1998), which encodes gp160, was used to insert the gene into a thymidine kinase gene of rVVΔB8R-IFN-g. The resulting virus was propagated and purified as described previously (Cho et al., 2001; 1998). The details of pB8R-IFN-g construction, and biological properties of rVVΔB8R-IFN-g and rVV-$gp160_{DH12,}$ will be described elsewhere.

3.5.2 Rabbit immunization,

New Zealand white female rabbits (2.5 to 3 kg) were purchased from Charles River (USA) and housed under specific pathogen free environments. All animals were tested in compliance with the animal protocol approved by IACUC of Iowa State University. For the IPAS-RAM vaccine group, three rabbits were immunized and samples were taken as shown in Fig. 2A. Rabbits were primed with 200 μg of gp41-28x3 subcutaneously with zinc-chitosan as an adjuvant as previously described (Habte et al., 2015; Qin et al., 2014). For the second immunization, rabbits were injected subcutaneously with 50 μg of gp41-28x3. Animals were also injected intradermally with 200 μg of gp41-54CT DNA in PBS, followed by electroporation using the AgilePulse In Vivo System (BTX, Harvard Apparatus). For the third immunization, rabbits were

injected with 200 µg of gp41-54CT DNA as described above and with rVV-gp160$_{DH12}$ (1x10^8 PFUs) through intradermal injection. For the fourth immunization, both 50 µg of gp41-28x3 and rVV-gp160$_{DH12}$ were administered. For the homologous prime-boost vaccine group, three rabbits were immunized subcutaneously with 200 µg of gp41-28x3 as shown in Fig. 2C. Serum samples were collected prior to immunization (pre-immune) as well as two weeks post-immunizations. PBMC samples were also collected 4 days after each immunization. Samples were stored at -80 °C until used.

3.5.3 Hybridoma generation

Rabbit R3 was injected intravenously using 200 µg of soluble gp41-28x3 in PBS without any adjuvant on week 35. Four days later, the animal was euthanized and the spleen was harvested. Splenocytes were fused to 240E-1 cells (kindly provided by Dr. Katherine L. Knight (Spieker-Polet et al., 1995)), as previously described (Qin et al., 2015; Spieker-Polet et al., 1995). Hybridoma cell culture media were screened by ELISA for specific binding to gp41-28x3 and individual peptides of interest as described below.

3.5.4 Enzyme-linked immunosorbent assays (ELISA)

All ELISA were performed using a protocol described previously (Banerjee et al., 2016; Habte et al., 2015), except for the use of an alternate blocking buffer containing 2.5% milk and 5% calf sera in PBS (pH 7.5). Various coating antigens were used, including gp41-28x3 (30 ng/ml), a mixture of 10-mer peptides biotinylated either at the N- or C-terminus (Habte et al., 2015), and HIV-1 M group consensus Env peptides (15-

mer; 100 ng/well) obtained from the NIH AIDS Reagent Program (Cat# 9487). For the fine epitope mapping analyses, two sets of scanning alanine mutant peptides were used: (1) 13-mer 671 peptide (^{671}NWFDITNWLWYIK^{683}) biotinylated at the C-terminal lysine, as previously described (Banerjee et al., 2016; Habte et al., 2015); and (2) 13-mer 671 peptide biotinylated at the N-terminal amine of two glycine linker (GG^{671}NWFDITNWLWYIK^{683}). For all ELISA testing hybridoma binding, goat anti-rabbit, horseradish peroxidase (HRP)-conjugated antibody (Thermo Scientific; Cat# 31430) was used as secondary antibody.

3.5.5 Neutralization assays

Neutralization assays were performed in TZM-bl cells as previously described (M. Li et al., 2005; Qin et al., 2014; Wei et al., 2002). Viruses tested included SF162 (tier 1A, clade B), MW965.26 (tier 1A, clade C), and MN.3 (tier 1A, clade B). Murine leukemia virus Env-pseudotyped virus was used as a negative control.

3.5.6 Hybridoma gene sequence analysis

Total RNA was extracted from hybridoma cells using the RNeasy Mini kit (Qiagen). RNA samples were treated with DNAse (Invitrogen) according to a manufacturer's protocol. Samples were subjected to cDNA synthesis using random hexamers and SuperScript III Reverse Transcriptase (Invitrogen). Briefly, 1μL of random hexamers (Roche) and 1μL of 10mM dNTPs were added to 11μL of DNAse treated RNA. The mixture was heated to 65^{o}C for 5min, then cooled briefly on ice. Subsequently, 4μL of 5x First-Strand Buffer, 1uL of 0.1M DTT, 1uL of RNaseOUT^{TM}

(Invitrogen), and 1uL of SuperScript III were added. Reaction was incubated at 25^{o}C for 5min, 45^{o}C for 45mins, and 70^{o}C for 15mins.

Heavy and kappa chain sequences were amplified with Platnium Pfx (Invitrogen) according to manufacturer's recommendations. Primers were based on a previous publication (Lightwood et al., 2006) and were designed for leader sequences (5' primer) or the C-terminus of the junction. Primers used for heavy chain amplification were 5'-GATATCAAGCTTACGCTCACCATGGAGACTGGGC-3' and 5'-CGCGCGCTCGAGACGGTGACSAGGGTSCCYKGGCCCC-3'. Primers used for kappa chain amplification were 5'-GATATCAAGCTTCGAATCGACATGGACACGAGGGCCCCC-3' and 5'-TCTAGACGTACGTTTGACCACCACCTCGGTCCCTC-3'. Cycling conditions were as follows: Initial denaturation at 94^{o}C for 5mins; followed by 40 cycles of 94^{o}C for 30sec, 68^{o}C for 1.5mins; final extension at 68^{o}C for 7mins; hold at 4^{o}C. PCR products were examined via agarose gel to confirm specific amplification. Products were subcloned for DNA sequencing.

For the 9F6 hybridoma, alternate light chain primers were tested as well. For the kappa chain, 5' and 3' primers from a second publication were tested (Rader et al., 2000). The two 5' primers used bound the beginning of the FR1 region: 5'-GGGCCCAGGCGGCCGAGCTCGTGMTGACCCAGACT-3' and GGGCCCAGGCGGCCGAGCTCGATMTGACCAGACT-3'. An alternate 3' primer, 5'-AGATGGTGCAGCCACAGTTCGTAGGATCTCCAGCTCGGTCCC-3', bound a similar region in the junction as did our original primer. An additional 3' primer which bound the 3' untranslated region immediately after the stop codon, 5'-

TCACTGGCGGTGCCCTGGCAGGCGTCT-3', was designed in-house based on NCBI nucleotide sequences. The lambda chain primer set described in Rader *et al.* (Rader et al., 2000) was also tested.

3.5.7 PBMC isolation

For PBMC isolation, EDTA-treated blood was initially spun at 2000xg for 10 min, with brakes turned off, to remove plasma and RBC. The collected cell layer was diluted 1:3 with PBS. Diluted blood was layered over Ficoll-Paque^{TM} PLUS (GE Healthcare) in a 4:3 (blood:Ficoll) ratio. The layered sample was centrifuged at 400xg for 45mins, with brakes turned off. PBMC layer was removed, and washed twice with 10 mL of PBS, followed by centrifugation at 600xg for 10min, with brakes. PBMC were stored in freezing media (90% fetal bovine serum and 10% DMSO) at -140^{o}C until RNA isolation.

3.5.8 Antibody gene RT-PCR for NGS analysis

Total RNA extraction and cDNA synthesis from PBMCs (~6x10^{6} per timepoint) was done as described for hybridomas. cDNA synthesis was scaled up by 6-fold to accommodate this amount of starting material. Resulting cDNA was split into two aliquots and subjected to separate antibody heavy and light chain gene amplification using Platinum Pfx polymerase (Invitrogen). Reaction mixture (150uL volume, split into 3 tubes) was: 60µL of cDNA template, 15µL of 10x buffer, 4.5µL of 10mM dNTPs, 3µL of $MgSO_4$, 6µL of each primer (10µM concentration), and 1.2 µL of polymerase. Primers and cycling conditions were the same as described for hybridoma sequencing. PCR products were examined via agarose gel to confirm specific amplification. Extra

primers and nucleotides were removed using QIAquick PCR Purification Kit (Qiagen). Following purification, heavy and light chain concentrations were calculated, and equal amounts of each were pooled for library preparation.

3.5.9 Next-generation sequencing

The TruSeq DNA LT sample prep kit (Illumina) was used to make NGS libraries, consisting of both heavy and kappa chain variable region fragments, for a 2x250bp read. The low sample (LS), gel free method was performed according to manufacturer's protocol, with minor modifications. Following end repair of the cDNA, indexing adapters were added using DNA adapter tubes (Illumina). The library quality was determined by Bioanalyzer High Sensitivity DNA chip (Agilent), and quantification was performed using a Qubit (Life Technologies). The sample libraries were run on an Illumina MiSeq according to manufacturer's recommendations.

3.5.10 NGS Read Processing and Quality Control

Reads from the two sequencing runs were demultiplexed by Illumina software, following which they were pre-processed using the Repertoire Sequencing Toolkit (pRESTO) (Vander Heiden et al., 2014). Processing consisted of the following steps: (1) Reads having a mean Phred quality score <20 were removed. (2) Paired reads were identified from Illumina headers, and aligned. Unpaired reads, and reads with a significance threshold calculated by pRESTO of <0.1, or mismatch rate <0.2, were removed. (3) Forward and reverse primers were matched and removed from the reads. Reads not matching primers, or which matched with error rate >0.2, were removed. (4)

Duplicate reads were merged. (5) Reads were split into heavy and light chain sets, based on 5' primer identification. (6) Where the same sample was processed in multiple sequencing runs, the resulting sets were merged to provide a single heavy and light chain set per sample. (7) Resulting read sets were clustered to a minimum identity of 97% using uparse (Edgar, 2010). The NGS data have been deposited to the NCBI Short Reads Archives (SRA) as study SRP094044.

3.5.11 Junction Identification and Germline Assignment

Sequences were parsed with IgBLAST v1.4.0 (Ye et al., 2013) using the IMGT germline library for the rabbit (Oryctolagus cuniculus) (M. P. Lefranc and G. Lefranc, 2001), downloaded 14 March 2015. Subsequent alignments were converted to IMGT output format using IgBlastPlus v1.0(Lees and Shepherd, 2015).

3.5.12 Clonal Inference and Analysis

The cluster diagram (Fig. 5A) was rendered in Gephi v0.9.1 (Bastian et al., 2009). For the Streamgraph view (Fig. 5C), heavy chain reads from all samples were combined, together with the sequences for the isolated mAbs. Clonally related sequences were inferred by clustering junction sequences of the same length and with nucleotide identity ≥ 90% using a single-linkage algorithm. The identity threshold was determined by consideration of the nearest-neighbor distribution (Fig. S6). Where the same CDR3 sequence was observed in multiple samples, it was assigned to the earliest sample (or leftmost in the diagram for terminal samples). Code for the Streamgraph view was as previously described (Laserson et al., 2014). Bands represent clusters of 2 or more

CDR3 sequences and the height at each sample point is proportional to the number of CDR3 sequences in the cluster. The overall band height at the sample point reflects the number of unique CDR3 sequences observed in the sample. Because the sequencing depth was uneven, equal samples of 100,000 reads of functional v-regions were drawn from each sample, and the number of unique CDR3 sequences was determined. The average was determined over 100 random draws. Overall band heights are proportional to that average number of unique CDR3 sequences.

3.5.13 Phylogenetic Analysis

Full-length sequences from blood plasma samples whose CDR3s clustered with mAbs of interest were reviewed, and those which did not match the mAb V- and J-germline assignments were rejected. To obtain a computationally tractable data set, random downsampling was used to limit the maximum number of sequences from each sample point to 200. Sequences were codon-aligned using TranslatorX (Abascal et al., 2010) following which phylogenetic trees were inferred by IQ-TREE v1.2.2 (Minh et al., 2013). The selection model GTR+G4 was selected by the program from consideration of the log-likelihoods of the initial parsimony tree for all available models. Trees were rendered using the ETE Toolkit (Huerta-Cepas et al., 2010). Intermediate sequences in the development of mAb 9F6, and its germline junction sequence, were inferred by PHYLIP (Felsenstein and Churchill, 1996) and the logo diagram was rendered with

Berkley WebLogo (Crooks et al., 2004). Percentage identity to the mAb and divergence from the germline was determined using in-house software.

3.5.14 In-House Software

Software developed in-house for the analysis (clustering, charting of germline attribution, spectratypes, germline identity) is available as part of TRIgS (http://cimm.ismb.lon.ac.uk/trigs/) (Lees and Shepherd, 2015).

3.6 Acknowledgments

We are grateful to Dr. Beatrice Hahn for providing pcDNA-MCON6gp160. The following reagents were obtained through the NIH AIDS Reagent Program, Division of AIDS, NIAID, NIH: HIV-1 Consensus group M Env peptides (Cat# 9487), Z13e1 from Dr. Michael Zwick (Cat# 11557), 2F5 from Dr. Hermann Katinger (Cat# 130220) and 4E10 from Dr. Herman Katinger (Cat# 10091). This work was supported by a Grant from the HHS/NIH/NIAID (P01 AI074286) and funding from Iowa State University. MWC has an equity interest in NeoVaxSyn Inc., and serves as the CEO/President. NeoVaxSyn Inc. did not contribute to this work or the interpretation of the data.

3.7 **Figures**

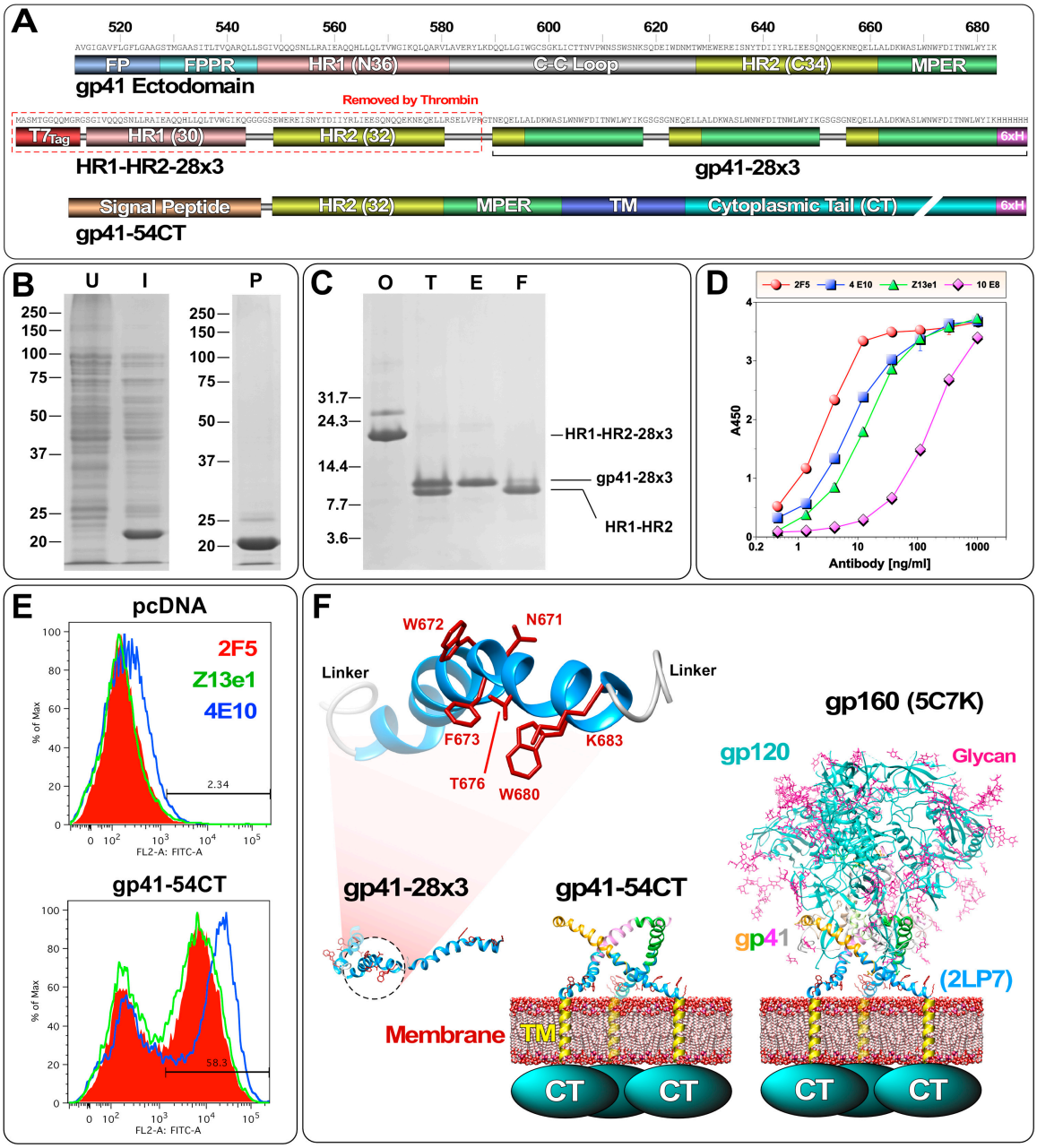

Fig. 1. Immunogens used in the study. (A) Schematic diagrams of gp41-28x3 and gp41-54CT. The entire gp41 ectodomain is shown on the top as a reference. To generate gp41-28x3, the $T7_{Tag}$ and HR1-HR2 portion of HR1-HR2-28x3 is removed with thrombin. (B) Expression and purification of HR1-HR2-28x3. Coomassie stained SDS-

PAGE gels (U: uninduced; I: induced; P: purified). (C) Cleavage of HR1-HR2-28x3 with thrombin (O: original sample; T: thrombin cleaved; E: eluted from Ni-NTA column; F: flow through). (D) ELISA of gp41-28x3 with 2F5, 4E10, Z13e1 and 10E8. (E) Flow cytometry analyses of cell surface expression of gp41-54CT. 2F5, Z13e1 and 4E10 were used to probe the antigen. (F) Structural models of three immunogens are shown to illustrate their relative size. NMR structure of 28-mer peptide (2LP7; (Reardon et al., 2014)) and BG505 SOSIP gp140 structure (5C7K; (Kong et al., 2015)) were used. Structures were rendered using Chimera (Pettersen et al., 2004).

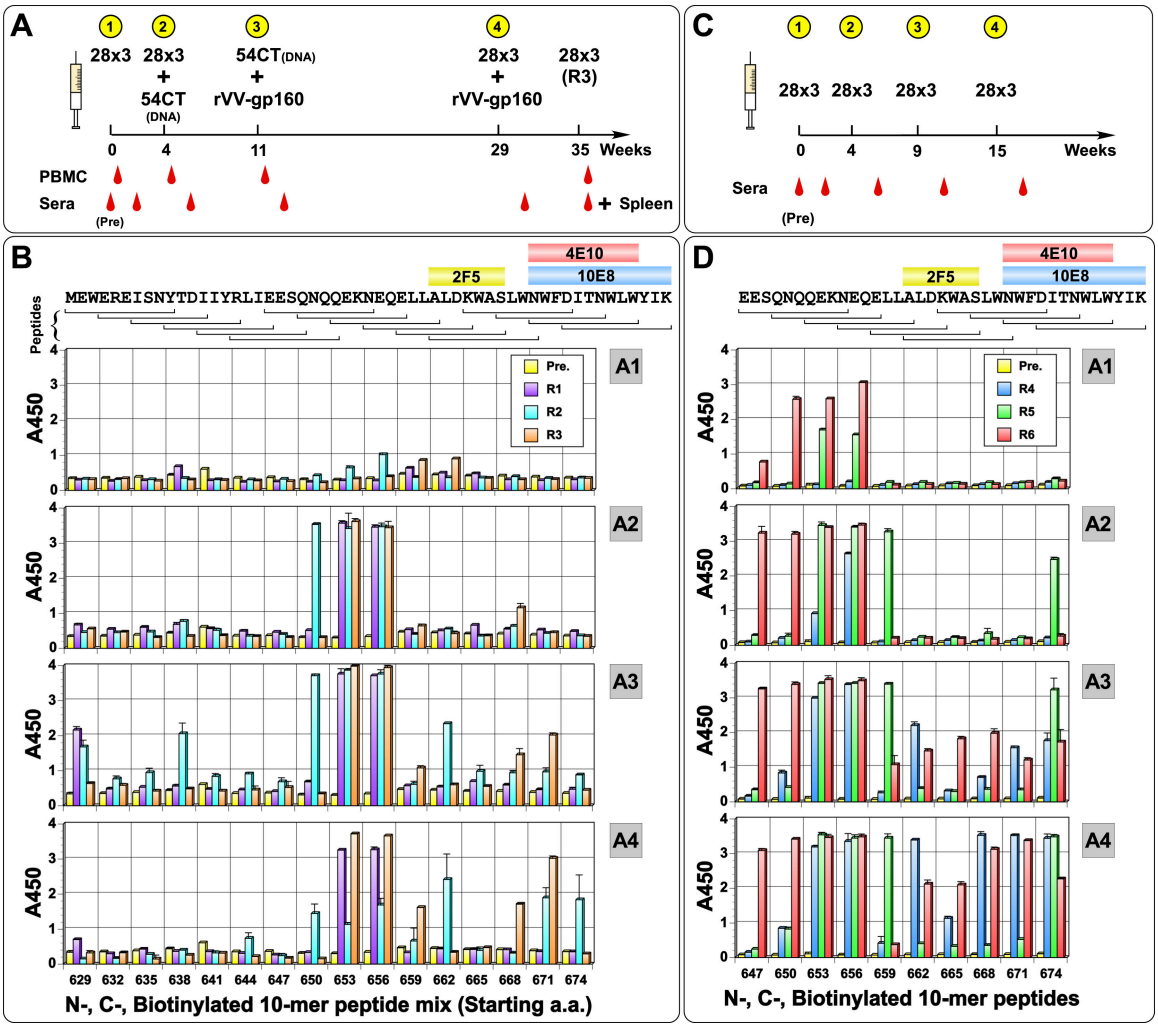

Fig. 2. Immunization schedule and linear epitope mapping analyses. (A) Timeline for IPAS-RAM immunization and sampling. Rabbits were immunized on weeks 0, 4, 11 and 29. Pre-immune, as well as post-immune sera (two weeks post each immunization) were taken. PBMC were collected four days after immunization (except after 4th) for antibody repertoire analyses. To generate hybridomas, Rabbit R3 was immunized intravenously on week 35 with gp41-28x3 and spleen was collected four days later. (B) Immunogenic, linear epitopes were identified by ELISA using overlapping "10-mer" peptides. Serum samples collected two weeks after first (A1), second (A2), third (A3)

and fourth (A4) immunizations were analyzed. A mixture of N- and C-terminally biotinylated peptides spanning the C-terminal 54 a.a. of gp41 ectodomain was used (B_{GG}XXXXXXXXXX and XXXXXXXXXX$_{GGK}$B, respectively; B=biotin, X=gp41 sequence). Pre-immune serum was used as a negative control. Horizontal brackets on top indicate the sequence for each peptide and core-binding epitopes for bnAbs 2F5, 4E10 and 10E8 are shown. The numbers on the X-axis indicate the starting a.a. position of “10-mer” peptides. Timeline (C) and epitope mapping analyses (D) for the homologous prime-boost immunization group with gp41-28x3.

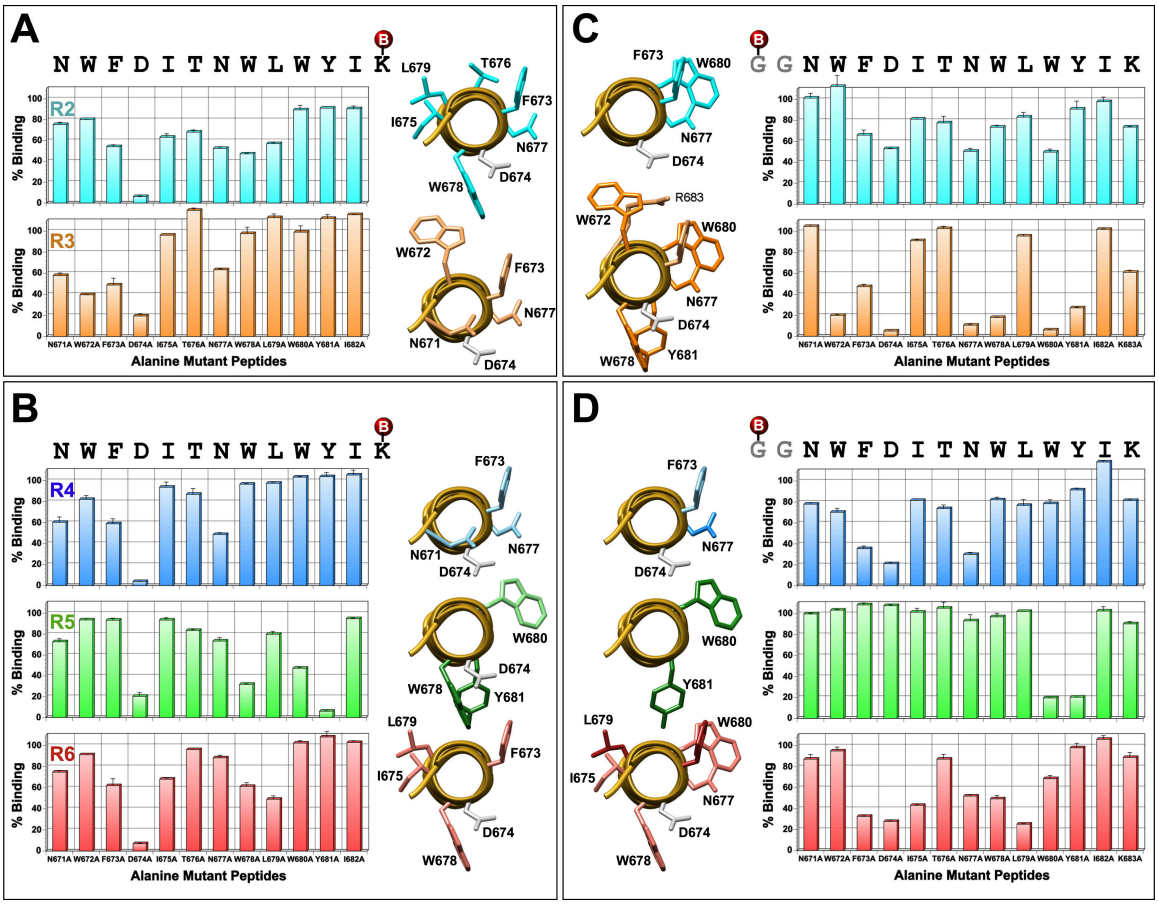

Fig. 3. Detailed epitope mapping analyses of immune sera. ELISAs were conducted using a set of C- (panels A and B) or N-terminally (panels C and D) biotinylated 13-mer alanine mutant peptides to assess impact of the mutation on antibody binding in comparison to the unmutated peptide. Serum samples collected after the fourth immunization were used. Peptides were biotinylated (shown as red spheres) at the primary amine of C-terminal lysine or N-terminal glycine. Graphic views of residues that affected binding when mutated are shown. An axial view of N671 to R683 segment of the 28-mer peptide co-crystalized with 10E8 is shown (PDB: 4G6F). Residues that showed <70% binding are shown. Residues that are more critical (<35%) are shown in darker tone. D674 is shown in grey.

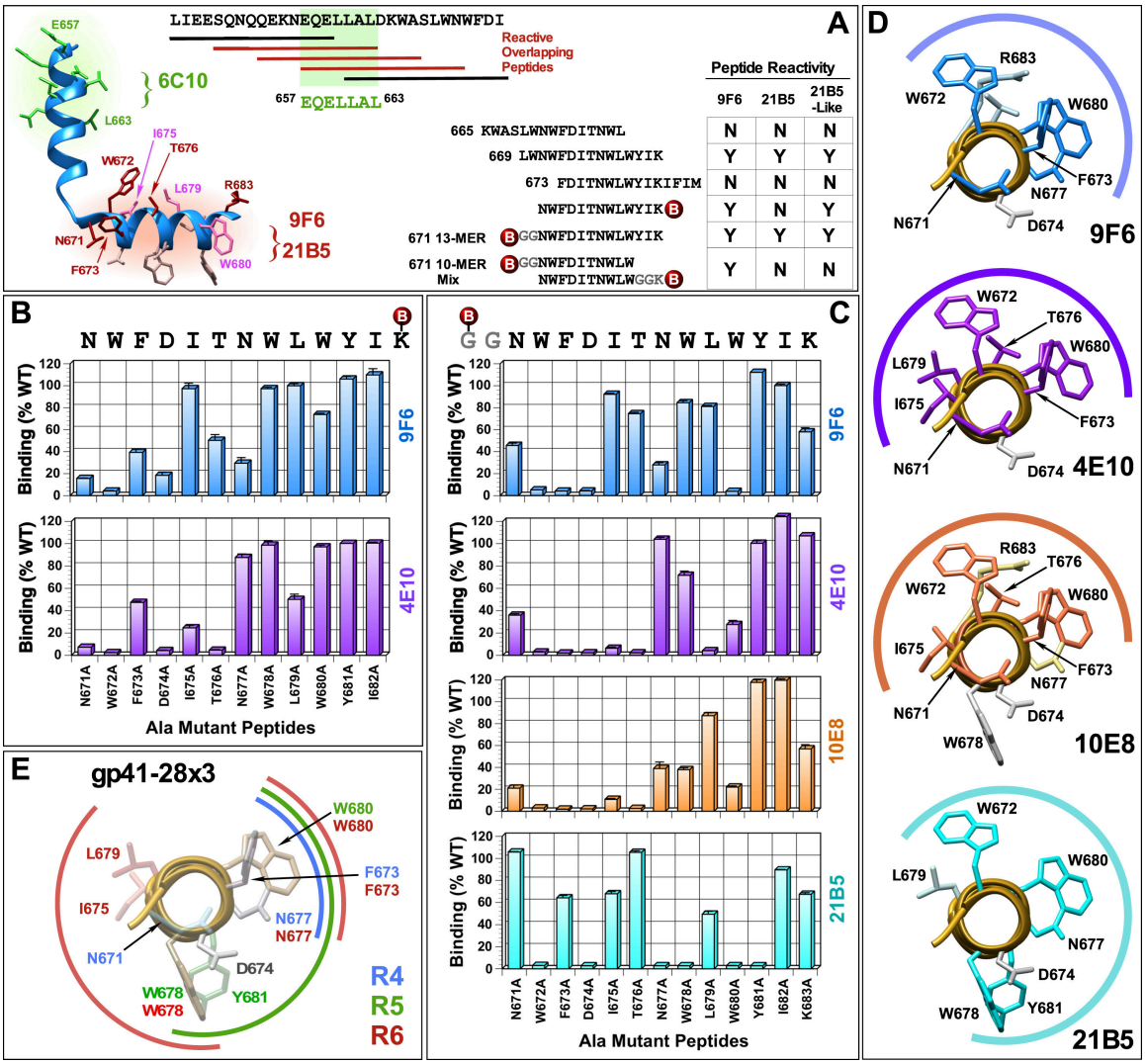

Fig. 4. Epitope mapping analyses of mAbs. (A) Linear epitope mapping of 6C10, 9F6, 21B5 and twelve 21B5-like mAbs using overlapping peptides. A structure of a 28-mer peptide co-crystallized with 10E8 (PDB: 4G6F) is shown to illustrate general locations of the epitopes. (B and C) Fine-epitope mapping analyses using C- or N-terminally biotinylated alanine mutant peptides. (D) Graphic presentations of residues important for 9F6 and 21B5 binding. 4E10 and 10E8 are included for comparison. (E) A summary of all residues targeted by antibodies induced by immunization with gp41-28x3 alone. Data from all three rabbits (R4, R5 and R6) are compiled.

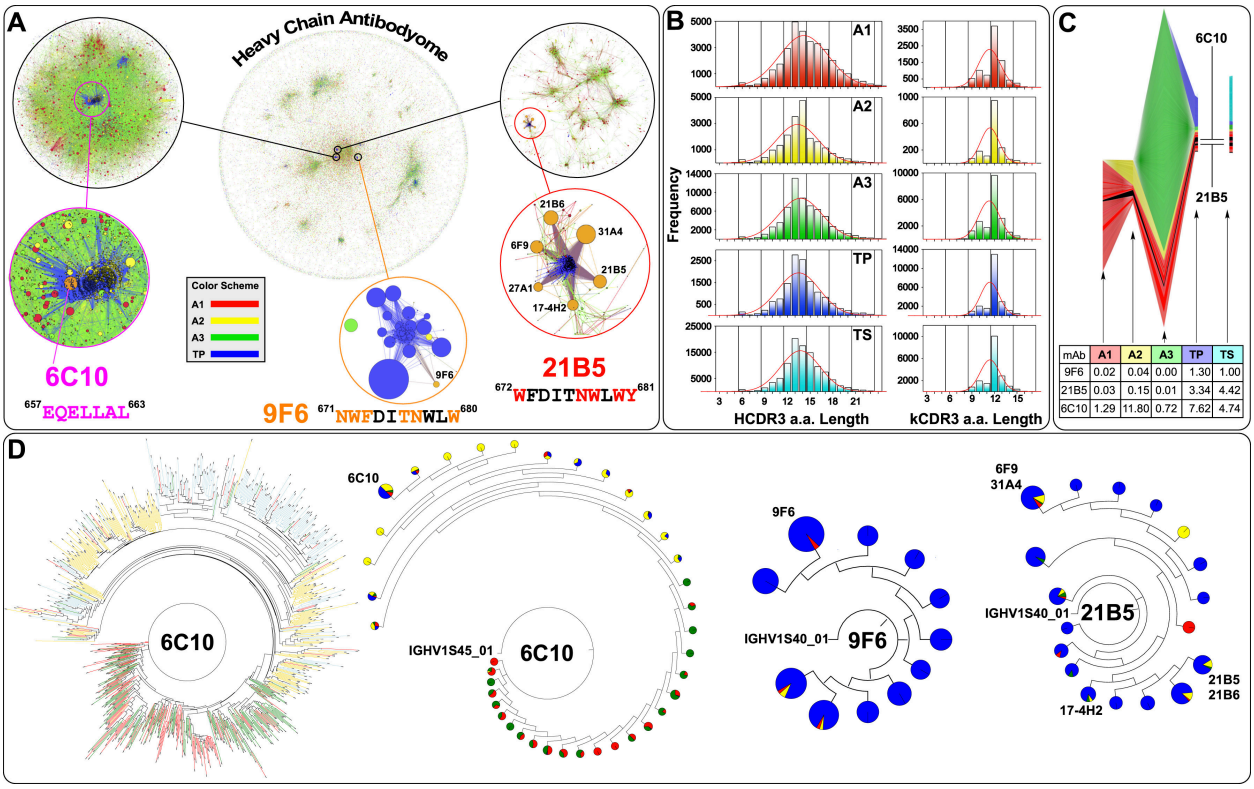

Fig. 5. **Clonal and phylogenetic analysis of the heavy chain repertoire.** (A) Heavy chain antibodyome: Each unique CDR3 sequence is represented as a point, colored by its sampling time (A1-A3: PBMC after the first, second and third immunizations; TP: terminal PBMC). Due to large amount of data, the analyses was limited to antibodies recovered from PBMC only. The same color-coding is used throughout the figure. Sequences forming clonal families are joined by lines. The approximate locations of 6C10, 9F6 and 21B5 mAb families are shown in expanded detail (Note: they are zoomed at different levels). The presence of multiple large clonal families indicates diverse responses to immunizations. Families remote from the isolated mAbs are dominated by reads from the A3 sample. (B) Heavy and kappa chain CDR3 spectratypes are shown for each sample. The distribution shows distinct and different

skews in samples A2 and A3, and a composite of the two in sample TP, suggesting that clonal populations stimulated at the earlier timepoints have been integrated into longer term memory. (C) Antibody clone dynamics colored by sampling time and earliest identification of clonotypes. The overall band height at each sample point is proportional to the number of novel CDR3 nucleotide sequences in that sample, normalized to account for variation in sequencing depth. The height of each individual band (representing a single clonotype) is proportional to the number of novel CDR3 nucleotide sequences identified in that clonotype. Black bands indicate the 6C10 and 21B5 clonal families. The underlying table shows the percentage of novel CDR3s that are clonally related to the indicated mAb. In both cases, CDR3s found in multiple timepoints are counted only in the earliest (or leftmost) sample and clonal relationship is inferred by CDR3 sequence identity alone. (D) Dendrograms of mAb clonal families (inferred by CDR3 sequence identity and V-/J-gene origin), in which pie charts indicate by size the number of descendants at each node, and by color their timepoint origin (only PBMC samples are included). The dendrogram for 6C10 is shown for comparison in both the conventional and accumulated (pie chart) form. The limited representation of earlier timepoints in the 21B5 and 9F6 charts suggests that development of these clonal families occurred almost exclusively after the A3 sampling time. By contrast, the dendrogram for the Ab 6C10 clonal family shows development at each timepoint, suggesting that refinements have been introduced by successive immunizations. Dendrograms for the three mAbs are shown at different zoom levels. Thus, their size should not be compared between different mAbs.

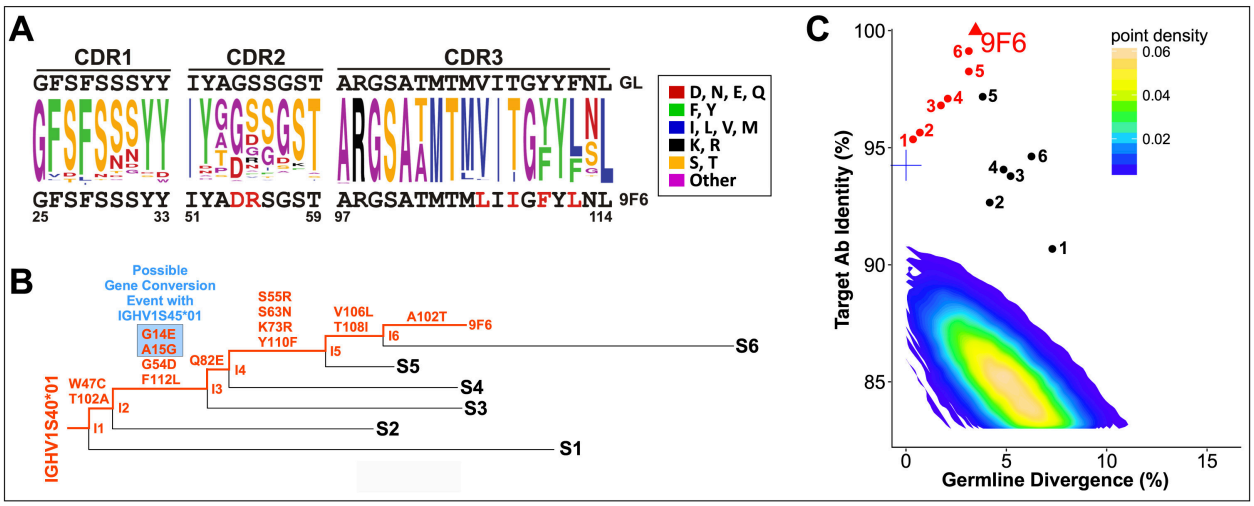

Fig. 6. Detailed analyses of 9F6 maturation. (A) Sequence variation of the HCDRs of the 9F6 clonal family. The germline sequence is shown at the top. The 9F6 sequence is shown at the bottom, with mutated residues indicated in red. (B) Lineage tree derived from representative sequences from the sample, selected to illustrate the inferred development of the antibody. Amino acid changes as a result of a possible gene conversion event are highlighted in blue. (C) Sequence identity/divergence plot of heavy chain sequences sharing the 9F6 germline. Clones shown in panel B are indicated, with inferred intermediates marked in red. The intermediates indicate good coverage of the development history. The contour plot encompasses all full-length productive sequences of the IGHV1S40 germline extracted from PBMC samples. The inferred 9F6 germline is marked by a blue cross.

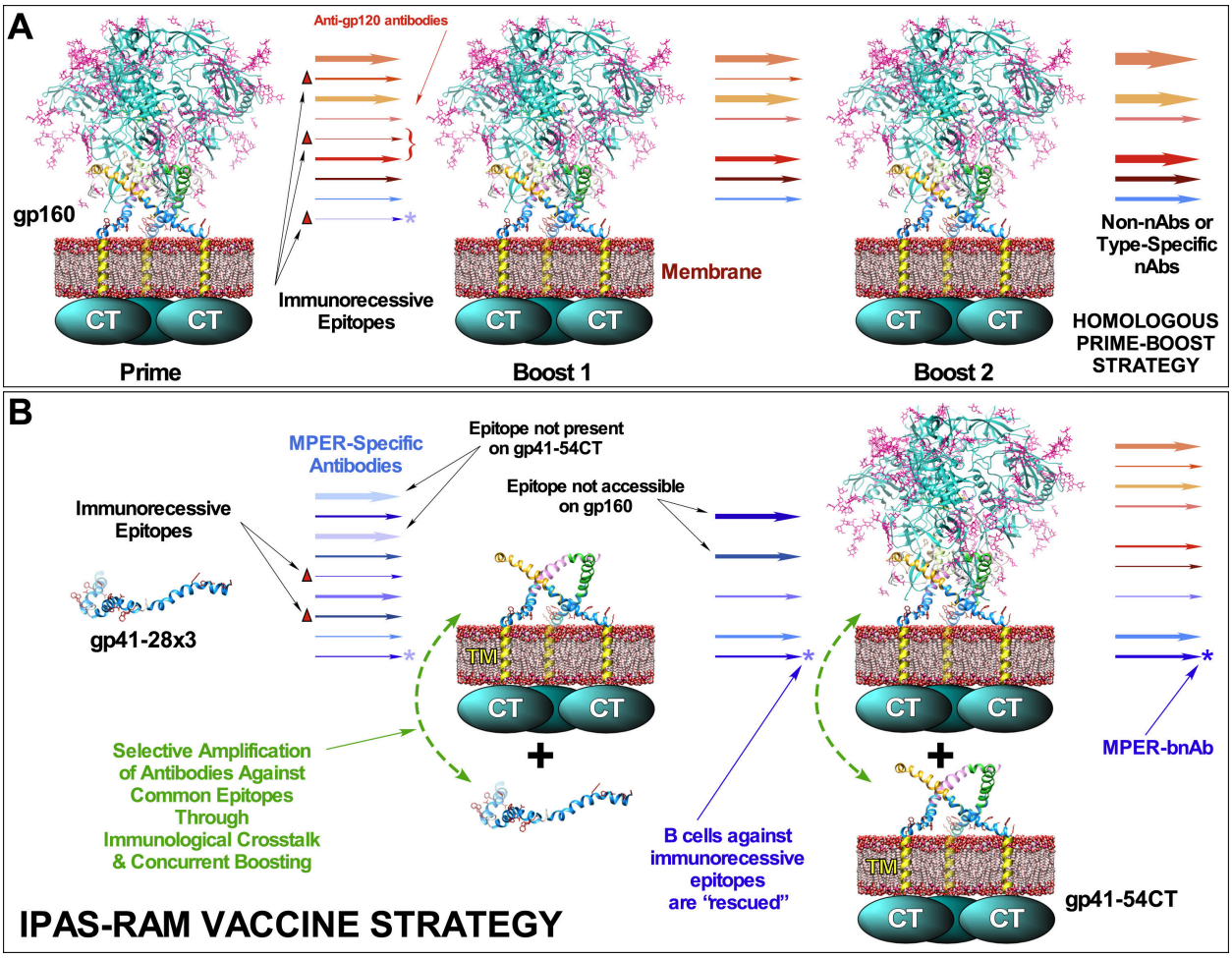

Fig. 7. Hypothetical comparison of two vaccine strategies. (A) A homologous prime-boost vaccine strategy using a whole, trimeric gp120/gp41 complex. Because MPER neutralizing epitopes (indicated by asterisks) are immunorecessive, B cells that target them are not stimulated sufficiently and are eventually eliminated. (B) In addition to sequential immunization with different immunogens (heterologous prime-boost), the unique and a novel feature of the IPAS-RAM vaccine strategy is co-immunization in a phased manner, which we expect will allow selective amplification of antibodies against common epitopes through immunological crosstalk and concurrent boosting. It is

hypothesized that this could potentially rescue B cells against immunorecessive neutralizing epitopes.

3.7 References

Abascal, F., Zardoya, R., Telford, M.J., 2010. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. Nucleic Acids Res. 38, W7–13. doi:10.1093/nar/gkq291

Appella, E., Chersi, A., Rejnek, J., Reisfeld, R., Mage, R., 1974. Rabbit immunoglobulin lambda chains: isolation and amino acid sequence of cysteine-containing peptides. Immunochemistry 11, 395–402.

Banerjee, S., Shi, H., Habte, H.H., Qin, Y., Cho, M.W., 2016. Modulating immunogenic properties of HIV-1 gp41 membrane-proximal external region by destabilizing six-helix bundle structure. Virology 490, 17–26. doi:10.1016/j.virol.2016.01.002

Bastian, M., Heymann, S., Jacomy, M., 2009. Gephi: an open source software for exploring and manipulating networks. ICWSM.

Briney, B., Sok, D., Jardine, J.G., Kulp, D.W., Skog, P., Menis, S., Jacak, R., Kalyuzhniy, O., de Val, N., Sesterhenn, F., Le, K.M., Ramos, A., Jones, M., Saye-Francisco, K.L., Blane, T.R., Spencer, S., Georgeson, E., Hu, X., Ozorowski, G., Adachi, Y., Kubitz, M., Sarkar, A., Wilson, I.A., Ward, A.B., Nemazee, D., Burton, D.R., Schief, W.R., 2016. Tailored Immunogens Direct Affinity Maturation toward HIV Neutralizing Antibodies. Cell 166, 1459–1470.e11. doi:10.1016/j.cell.2016.08.005

Brunel, F.M., Zwick, M.B., Cardoso, R.M.F., Nelson, J.D., Wilson, I.A., Burton, D.R., Dawson, P.E., 2006. Structure-function analysis of the epitope for 4E10, a broadly neutralizing human immunodeficiency virus type 1 antibody. J Virol 80, 1680–1687. doi:10.1128/JVI.80.4.1680-1687.2006

Cardoso, R.M.F., Brunel, F.M., Ferguson, S., Zwick, M., Burton, D.R., Dawson, P.E., Wilson, I.A., 2007. Structural basis of enhanced binding of extended and helically constrained peptide epitopes of the broadly neutralizing HIV-1 antibody 4E10. J Mol Biol 365, 1533–1544. doi:10.1016/j.jmb.2006.10.088

Cheung, W.C., Beausoleil, S.A., Zhang, X., Sato, S., Schieferl, S.M., Wieler, J.S., Beaudet, J.G., Ramenani, R.K., Popova, L., Comb, M.J., Rush, J., Polakiewicz, R.D., 2012. A proteomics approach for the identification and cloning of monoclonal antibodies from serum. Nat. Biotechnol. 30, 447–452. doi:10.1038/nbt.2167

Cho, M.W., Kim, Y.B., Lee, M.K., Gupta, K.C., Ross, W., Plishka, R., Buckler-White, A., Igarashi, T., Theodore, T., Byrum, R., Kemp, C., Montefiori, D.C., Martin, M.A., 2001.

Polyvalent envelope glycoprotein vaccine elicits a broader neutralizing antibody response but is unable to provide sterilizing protection against heterologous Simian/human immunodeficiency virus infection in pigtailed macaques. J Virol 75, 2224–2234. doi:10.1128/JVI.75.5.2224-2234.2001

Cho, M.W., Lee, M.K., Carney, M.C., Berson, J.F., Doms, R.W., Martin, M.A., 1998. Identification of determinants on a dualtropic human immunodeficiency virus type 1 envelope glycoprotein that confer usage of CXCR4. J Virol 72, 2509–2515.

Correia, B.E., Ban, Y.-E.A., Holmes, M.A., Xu, H., Ellingson, K., Kraft, Z., Carrico, C., Boni, E., Sather, D.N., Zenobia, C., Burke, K.Y., Bradley-Hewitt, T., Bruhn-Johannsen, J.F., Kalyuzhniy, O., Baker, D., Strong, R.K., Stamatatos, L., Schief, W.R., 2010. Computational Design of Epitope-Scaffolds Allows Induction of Antibodies Specific for a Poorly Immunogenic HIV Vaccine Epitope. Structure 18, 1116–1126. doi:10.1016/j.str.2010.06.010

Crooks, G.E., Hon, G., Chandonia, J.-M., Brenner, S.E., 2004. WebLogo: a sequence logo generator. Genome Res. 14, 1188–1190. doi:10.1101/gr.849004

Edgar, R.C., 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26, 2460–2461. doi:10.1093/bioinformatics/btq461

Escolano, A., Steichen, J.M., Dosenovic, P., Kulp, D.W., Golijanin, J., Sok, D., Freund, N.T., Gitlin, A.D., Oliveira, T., Araki, T., Lowe, S., Chen, S.T., Heinemann, J., Yao, K.-H., Georgeson, E., Saye-Francisco, K.L., Gazumyan, A., Adachi, Y., Kubitz, M., Burton, D.R., Schief, W.R., Nussenzweig, M.C., 2016. Sequential Immunization Elicits Broadly Neutralizing Anti-HIV-1 Antibodies in Ig Knockin Mice. Cell 166, 1445–1458.e12. doi:10.1016/j.cell.2016.07.030

Felsenstein, J., Churchill, G.A., 1996. A Hidden Markov Model approach to variation among sites in rate of evolution. Mol. Biol. Evol. 13, 93–104.

Gao, F., Weaver, E.A., Lu, Z., Li, Y., Liao, H.-X., Ma, B., Alam, S.M., Scearce, R.M., Sutherland, L.L., Yu, J.-S., Decker, J.M., Shaw, G.M., Montefiori, D.C., Korber, B.T., Hahn, B.H., Haynes, B.F., 2005. Antigenicity and immunogenicity of a synthetic human immunodeficiency virus type 1 group m consensus envelope glycoprotein. J Virol 79, 1154–1163. doi:10.1128/JVI.79.2.1154-1163.2005

Georgiev, I.S., Gordon Joyce, M., Zhou, T., Kwong, P.D., 2013. Elicitation of HIV-1-neutralizing antibodies against the CD4-binding site. Curr Opin HIV AIDS 8, 382–392. doi:10.1097/COH.0b013e328363a90e

Habte, H.H., Banerjee, S., Shi, H., Qin, Y., Cho, M.W., 2015. Immunogenic properties of a trimeric gp41-based immunogen containing an exposed membrane-proximal external region. Virology 486, 187–197. doi:10.1016/j.virol.2015.09.010

Haynes, B.F., Kelsoe, G., Harrison, S.C., Kepler, T.B., 2012. B-cell-lineage immunogen design in vaccine development with HIV-1 as a case study. Nat. Biotechnol. 30, 423–433. doi:10.1038/nbt.2197

Haynes, B.F., Moody, M.A., Alam, M., Bonsignori, M., Verkoczy, L., Ferrari, G., Gao, F., Tomaras, G., Liao, H.-X., Kelsoe, G., 2014. Progress in HIV-1 vaccine development. Journal of Allergy and Clinical Immunology 134, 3–10. doi:10.1016/j.jaci.2014.04.025

Huang, J., Kang, B.H., Pancera, M., Lee, J.H., Tong, T., Feng, Y., Imamichi, H., Georgiev, I.S., Chuang, G.-Y., Druz, A., Doria-Rose, N.A., Laub, L., Sliepen, K., van Gils, M.J., la Peña, de, A.T., Derking, R., Klasse, P.J., Migueles, S.A., Bailer, R.T., Alam, M., Pugach, P., Haynes, B.F., Wyatt, R.T., Sanders, R.W., Binley, J.M., Ward, A.B., Mascola, J.R., Kwong, P.D., Connors, M., 2014. Broad and potent HIV-1 neutralization by a human antibody that binds the gp41–gp120 interface. Nature 515, 138–142. doi:10.1038/nature13601

Huang, J., Ofek, G., Laub, L., Louder, M.K., Doria-Rose, N.A., Longo, N.S., Imamichi, H., Bailer, R.T., Chakrabarti, B., Sharma, S.K., Alam, S.M., Wang, T., Yang, Y., Zhang, B., Migueles, S.A., Wyatt, R., Haynes, B.F., Kwong, P.D., Mascola, J.R., Connors, M., 2012. Broad and potent neutralization of HIV-1 by a gp41-specific human antibody. Nature 491, 406–412. doi:10.1038/nature11544

Huerta-Cepas, J., Dopazo, J., Gabaldón, T., 2010. ETE: a python Environment for Tree Exploration. BMC Bioinformatics 11, 24. doi:10.1186/1471-2105-11-24

Jaworski, J.P., Krebs, S.J., Trovato, M., Kovarik, D.N., Brower, Z., Sutton, W.F., Waagmeester, G., Sartorius, R., D'Apice, L., Caivano, A., Doria-Rose, N.A., Malherbe, D., Montefiori, D.C., Barnett, S., De Berardinis, P., Haigwood, N.L., 2012. Co-immunization with multimeric scaffolds and DNA rapidly induces potent autologous HIV-1 neutralizing antibodies and CD8+ T cells. PLoS ONE 7, e31464. doi:10.1371/journal.pone.0031464

Kong, L., Torrents de la Peña, A., Deller, M.C., Garces, F., Sliepen, K., Hua, Y., Stanfield, R.L., Sanders, R.W., Wilson, I.A., 2015. Complete epitopes for vaccine design derived from a crystal structure of the broadly neutralizing antibodies PGT128 and 8ANC195 in complex with an HIV-1 Env trimer. Acta Crystallogr. D Biol. Crystallogr. 71, 2099–2108. doi:10.1107/S1399004715013917

Krebs, S.J., McBurney, S.P., Kovarik, D.N., Waddell, C.D., Jaworski, J.P., Sutton, W.F., Gomes, M.M., Trovato, M., Waagmeester, G., Barnett, S.J., DeBerardinis, P., Haigwood, N.L., 2014. Multimeric scaffolds displaying the HIV-1 envelope MPER induce MPER-specific antibodies and cross-neutralizing antibodies when co-immunized with gp160 DNA. PLoS ONE 9, e113463. doi:10.1371/journal.pone.0113463

Kwong, P.D., Mascola, J.R., Nabel, G.J., 2013. Broadly neutralizing antibodies and the search for an HIV-1 vaccine: the end of the beginning. Nat. Rev. Immunol. 13, 693–701. doi:doi:10.1038/nri3516

Laserson, U., Vigneault, F., Gadala-Maria, D., Yaari, G., Uduman, M., Vander Heiden, J.A., Kelton, W., Taek Jung, S., Liu, Y., Laserson, J., Chari, R., Lee, J.-H., Bachelet, I., Hickey, B., Lieberman-Aiden, E., Hanczaruk, B., Simen, B.B., Egholm, M., Koller, D., Georgiou, G., Kleinstein, S.H., Church, G.M., 2014. High-resolution antibody dynamics of vaccine-induced immune responses. Proceedings of the National Academy of Sciences 111, 4928–4933. doi:10.1073/pnas.1323862111

Lees, W.D., Shepherd, A.J., 2015. Utilities for High-Throughput Analysis of B-Cell Clonal Lineages. J Immunol Res 2015, 323506. doi:10.1155/2015/323506

Lefranc, M.P., Lefranc, G., 2001. The immunoglobulin factsbook.

Li, J., Valentin, A., Kulkarni, V., Rosati, M., Beach, R.K., Alicea, C., Hannaman, D., Reed, S.G., Felber, B.K., Pavlakis, G.N., 2013. HIV/SIV DNA vaccine combined with protein in a co-immunization protocol elicits highest humoral responses to envelope in mice and macaques. Vaccine 31, 3747–3755. doi:10.1016/j.vaccine.2013.04.037

Li, M., Gao, F., Mascola, J.R., Stamatatos, L., Polonis, V.R., Koutsoukos, M., Voss, G., Goepfert, P., Gilbert, P., Greene, K.M., Bilska, M., Kothe, D.L., Salazar-Gonzalez, J.F., Wei, X., Decker, J.M., Hahn, B.H., Montefiori, D.C., 2005. Human immunodeficiency virus type 1 env clones from acute and early subtype B infections for standardized assessments of vaccine-elicited neutralizing antibodies. J Virol 79, 10108–10125. doi:10.1128/JVI.79.16.10108-10125.2005

Lightwood, D.J., Carrington, B., Henry, A.J., McKnight, A.J., Crook, K., Cromie, K., Lawson, A.D.G., 2006. Antibody generation through B cell panning on antigen followed by in situ culture and direct RT-PCR on cells harvested en masse from antigen-positive wells. J. Immunol. Methods 316, 133–143. doi:10.1016/j.jim.2006.08.010

Lutje Hulsik, D., Liu, Y.-Y., Strokappe, N.M., Battella, S., Khattabi, El, M., McCoy, L.E., Sabin, C., Hinz, A., Hock, M., Macheboeuf, P., Bonvin, A.M.J.J., Langedijk, J.P.M., Davis, D., Forsman Quigley, A., Aasa-Chapman, M.M.I., Seaman, M.S., Ramos, A., Poignard, P., Favier, A., Simorre, J.-P., Weiss, R.A., Verrips, C.T., Weissenhorn, W., Rutten, L., 2013. A gp41 MPER-specific llama VHH requires a hydrophobic CDR3 for neutralization but not for antigen recognition. PLoS Pathog 9, e1003202. doi:10.1371/journal.ppat.1003202

Mascola, J.R., Haynes, B.F., 2013. HIV-1 neutralizing antibodies: understanding nature's pathways. Immunol Rev 254, 225–244. doi:10.1111/imr.12075

Mascola, J.R., Montefiori, D.C., 2010. The role of antibodies in HIV vaccines. Annu. Rev. Immunol. 28, 413–444. doi:10.1146/annurev-immunol-030409-101256

McCoy, L.E., Weiss, R.A., 2013. Neutralizing antibodies to HIV-1 induced by immunization. J. Exp. Med. 210, 209–223. doi:10.1084/jem.20121827

Melikyan, G.B., 2008. Common principles and intermediates of viral protein-mediated fusion: the HIV-1 paradigm. Retrovirology 5, 111. doi:10.1186/1742-4690-5-111

Minh, B.Q., Nguyen, M.A.T., Haeseler, von, A., 2013. Ultrafast approximation for phylogenetic bootstrap. Mol. Biol. Evol. 30, 1188–1195. doi:10.1093/molbev/mst024

Muñoz-Barroso, I., Salzwedel, K., Hunter, E., Blumenthal, R., 1999. Role of the membrane-proximal domain in the initial stages of human immunodeficiency virus type 1 envelope glycoprotein-mediated membrane fusion. J Virol 73, 6089–6092.

Patel, V., Jalah, R., Kulkarni, V., Valentin, A., Rosati, M., Alicea, C., Gegerfelt, von, A., Huang, W., Guan, Y., Keele, B.F., Bess, J.W., Piatak, M., Lifson, J.D., Williams, W.T., Shen, X., Tomaras, G.D., Amara, R.R., Robinson, H.L., Johnson, W., Broderick, K.E., Sardesai, N.Y., Venzon, D.J., Hirsch, V.M., Felber, B.K., Pavlakis, G.N., 2013. DNA and virus particle vaccination protects against acquisition and confers control of viremia upon heterologous simian immunodeficiency virus challenge. Proceedings of the National Academy of Sciences 110, 2975–2980. doi:10.1073/pnas.1215393110

Pejchal, R., Doores, K.J., Walker, L.M., Khayat, R., Huang, P.-S., Wang, S.-K., Stanfield, R.L., Julien, J.-P., Ramos, A., Crispin, M., Depetris, R., Katpally, U., Marozsan, A., Cupo, A., Maloveste, S., Liu, Y., McBride, R., Ito, Y., Sanders, R.W., Ogohara, C., Paulson, J.C., Feizi, T., Scanlan, C.N., Wong, C.-H., Moore, J.P., Olson, W.C., Ward, A.B., Poignard, P., Schief, W.R., Burton, D.R., Wilson, I.A., 2011. A potent and broad neutralizing antibody recognizes and penetrates the HIV glycan shield. Science 334, 1097–1103. doi:10.1126/science.1213256

Penn-Nicholson, A., Han, D.P., Kim, S.J., Park, H., Ansari, R., Montefiori, D.C., Cho, M.W., 2008. Assessment of antibody responses against gp41 in HIV-1-infected patients using soluble gp41 fusion proteins and peptides derived from M group consensus envelope. Virology 372, 442–456. doi:10.1016/j.virol.2007.11.009

Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., Ferrin, T.E., 2004. UCSF Chimera--a visualization system for exploratory research and analysis. J Comput Chem 25, 1605–1612. doi:10.1002/jcc.20084

Purtscher, M., Trkola, A., Gruber, G., Buchacher, A., Predl, R., Steindl, F., Tauer, C., Berger, R., Barrett, N., Jungbauer, A., 1994. A broadly neutralizing human monoclonal antibody against gp41 of human immunodeficiency virus type 1. AIDS Res Hum Retroviruses 10, 1651–1658. doi:10.1089/aid.1994.10.1651

Qin, Y., Banasik, M., Kim, S., Penn-Nicholson, A., Habte, H.H., LaBranche, C., Montefiori, D.C., Wang, C., Cho, M.W., 2014. Eliciting neutralizing antibodies with gp120 outer domain constructs based on M-group consensus sequence. Virology 462-463, 363–376. doi:10.1016/j.virol.2014.06.006

Qin, Y., Banerjee, S., Agrawal, A., Shi, H., Banasik, M., Lin, F., Rohl, K., LaBranche, C., Montefiori, D.C., Cho, M.W., 2015. Characterization of a Large Panel of Rabbit Monoclonal Antibodies against HIV-1 gp120 and Isolation of Novel Neutralizing Antibodies against the V3 Loop. PLoS ONE 10, e0128823. doi:10.1371/journal.pone.0128823

Rader, C., Ritter, G., Nathan, S., Elia, M., Gout, I., Jungbluth, A.A., Cohen, L.S., Welt, S., Old, L.J., Barbas, C.F., 2000. The rabbit antibody repertoire as a novel source for the generation of therapeutic human antibodies. J Biol Chem 275, 13668–13676.

Reardon, P.N., Sage, H., Dennison, S.M., Martin, J.W., Donald, B.R., Alam, S.M., Haynes, B.F., Spicer, L.D., 2014. Structure of an HIV-1-neutralizing antibody target, the lipid-bound gp41 envelope membrane proximal region trimer. Proceedings of the National Academy of Sciences 111, 1391–1396. doi:10.1073/pnas.1309842111

Salzwedel, K., West, J.T., Hunter, E., 1999. A conserved tryptophan-rich motif in the membrane-proximal region of the human immunodeficiency virus type 1 gp41 ectodomain is important for Env-mediated fusion and virus infectivity. J Virol 73, 2469–2480.

Scheid, J.F., Mouquet, H., Ueberheide, B., Diskin, R., Klein, F., Oliveira, T.Y.K., Pietzsch, J., Fenyo, D., Abadir, A., Velinzon, K., Hurley, A., Myung, S., Boulad, F., Poignard, P., Burton, D.R., Pereyra, F., Ho, D.D., Walker, B.D., Seaman, M.S., Bjorkman, P.J., Chait, B.T., Nussenzweig, M.C., 2011. Sequence and structural convergence of broad and potent HIV antibodies that mimic CD4 binding. Science 333, 1633–1637. doi:10.1126/science.1207227

Shi, W., Bohon, J., Han, D.P., Habte, H., Qin, Y., Cho, M.W., Chance, M.R., 2010. Structural characterization of HIV gp41 with the membrane-proximal external region. Journal of Biological Chemistry 285, 24290–24298. doi:10.1074/jbc.M110.111351

Sok, D., Briney, B., Jardine, J.G., Kulp, D.W., Menis, S., Pauthner, M., Wood, A., Lee, E.-C., Le, K.M., Jones, M., Ramos, A., Kalyuzhniy, O., Adachi, Y., Kubitz, M., MacPherson, S., Bradley, A., Friedrich, G.A., Schief, W.R., Burton, D.R., 2016. Priming HIV-1 broadly neutralizing antibody precursors in human Ig loci transgenic mice. Science 353, 1557–1560. doi:10.1126/science.aah3945

Spieker-Polet, H., Sethupathi, P., Yam, P.C., Knight, K.L., 1995. Rabbit monoclonal antibodies: generating a fusion partner to produce rabbit-rabbit hybridomas. Proc. Natl. Acad. Sci. U.S.A. 92, 9348–9352.

Stiegler, G., Kunert, R., Purtscher, M., Wolbank, S., Voglauer, R., Steindl, F., Katinger, H., 2001. A potent cross-clade neutralizing human monoclonal antibody against a novel epitope on gp41 of human immunodeficiency virus type 1. AIDS Res Hum Retroviruses 17, 1757–1765. doi:10.1089/08892220152741450

Tian, M., Cheng, C., Chen, X., Duan, H., Cheng, H.-L., Dao, M., Sheng, Z., Kimble, M., Wang, L., Lin, S., Schmidt, S.D., Du, Z., Joyce, M.G., Chen, Y., DeKosky, B.J., Chen, Y., Normandin, E., Cantor, E., Chen, R.E., Doria-Rose, N.A., Zhang, Y., Shi, W., Kong, W.-P., Choe, M., Henry, A.R., Laboune, F., Georgiev, I.S., Huang, P.-Y., Jain, S., McGuire, A.T., Georgeson, E., Menis, S., Douek, D.C., Schief, W.R., Stamatatos, L., Kwong, P.D., Shapiro, L., Haynes, B.F., Mascola, J.R., Alt, F.W., 2016. Induction of HIV Neutralizing Antibody Lineages in Mice with Diverse Precursor Repertoires. Cell 166, 1471–1484.e18. doi:10.1016/j.cell.2016.07.029

van Gils, M.J., Sanders, R.W., 2013. Broadly neutralizing antibodies against HIV-1: templates for a vaccine. Virology 435, 46–56. doi:10.1016/j.virol.2012.10.004

Vander Heiden, J.A., Yaari, G., Uduman, M., Stern, J.N.H., O'Connor, K.C., Hafler, D.A., Vigneault, F., Kleinstein, S.H., 2014. pRESTO: a toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. Bioinformatics 30, 1930–1932. doi:10.1093/bioinformatics/btu138

Walker, L.M., Huber, M., Doores, K.J., Falkowska, E., Pejchal, R., Julien, J.-P., Wang, S.-K., Ramos, A., Chan-Hui, P.-Y., Moyle, M., Mitcham, J.L., Hammond, P.W., Olsen, O.A., Phung, P., Fling, S., Wong, C.-H., Phogat, S., Wrin, T., Simek, M.D., Protocol G Principal Investigators, Koff, W.C., Wilson, I.A., Burton, D.R., Poignard, P., 2011. Broad neutralization coverage of HIV by multiple highly potent antibodies. Nature 477, 466–470. doi:10.1038/nature10373

Walker, L.M., Phogat, S.K., Chan-Hui, P.-Y., Wagner, D., Phung, P., Goss, J.L., Wrin, T., Simek, M.D., Fling, S., Mitcham, J.L., Lehrman, J.K., Priddy, F.H., Olsen, O.A., Frey, S.M., Hammond, P.W., Protocol G Principal Investigators, Kaminsky, S., Zamb, T., Moyle, M., Koff, W.C., Poignard, P., Burton, D.R., 2009. Broad and potent neutralizing antibodies from an African donor reveal a new HIV-1 vaccine target. Science 326, 285–289. doi:10.1126/science.1178746

Wei, X., Decker, J.M., Liu, H., Zhang, Z., Arani, R.B., Kilby, J.M., Saag, M.S., Wu, X., Shaw, G.M., Kappes, J.C., 2002. Emergence of resistant human immunodeficiency virus type 1 in patients receiving fusion inhibitor (T-20) monotherapy. Antimicrob. Agents Chemother. 46, 1896–1905. doi:10.1128/AAC.46.6.1896-1905.2002

Wu, X., Yang, Z.-Y., Li, Y., Hogerkorp, C.-M., Schief, W.R., Seaman, M.S., Zhou, T., Schmidt, S.D., Wu, L., Xu, L., Longo, N.S., McKee, K., O'Dell, S., Louder, M.K., Wycuff, D.L., Feng, Y., Nason, M., Doria-Rose, N., Connors, M., Kwong, P.D., Roederer, M., Wyatt, R.T., Nabel, G.J., Mascola, J.R., 2010. Rational design of

envelope identifies broadly neutralizing human monoclonal antibodies to HIV-1. Science 329, 856–861. doi:10.1126/science.1187659

Xu, J.Y., Gorny, M.K., Palker, T., Karwowska, S., Zolla-Pazner, S., 1991. Epitope mapping of two immunodominant domains of gp41, the transmembrane protein of human immunodeficiency virus type 1, using ten human monoclonal antibodies. J Virol 65, 4832–4838.

Ye, J., Ma, N., Madden, T.L., Ostell, J.M., 2013. IgBLAST: an immunoglobulin variable domain sequence analysis tool. Nucleic Acids Res. 41, W34–40. doi:10.1093/nar/gkt382

Zwick, M.B., Jensen, R., Church, S., Wang, M., Stiegler, G., Kunert, R., Katinger, H., Burton, D.R., 2005. Anti-human immunodeficiency virus type 1 (HIV-1) antibodies 2F5 and 4E10 require surprisingly few crucial residues in the membrane-proximal external region of glycoprotein gp41 to neutralize HIV-1. J Virol 79, 1252–1261. doi:10.1128/JVI.79.2.1252-1261.2005

Zwick, M.B., Labrijn, A.F., Wang, M., Spenlehauer, C., Saphire, E.O., Binley, J.M., Moore, J.P., Stiegler, G., Katinger, H., Burton, D.R., Parren, P.W., 2001. Broadly neutralizing antibodies targeted to the membrane-proximal external region of human immunodeficiency virus type 1 glycoprotein gp41. J Virol 75, 10892–10905. doi:10.1128/JVI.75.22.10892-10905.2001

CHAPTER 4

EVALUATION OF A NOVEL, RAPID HETEROLOGOUS PRIME-BOOST STRATEGY FOR TARGETING CD4 BINDING SITE NEUTRALIZING EPITOPES ON HIV-1 GP120

Heliang Shi, Saikat Banerjee, Aditi Agrawal, Michael Cho

Abstract

The CD4 binding site (CD4bs) of HIV-1 gp120 is an attractive broadly neutralizing antibody (bnAbs) target. In this proof-of-concept study, we evaluated a novel, rapid heterologous prime-boost vaccine strategy to induce CD4bs-directed Abs, which we hypothesized would be neutralizing. Rabbits were immunized alternatively with either a native trimer-like immunogen (SOSIP gp140) or a small CD4bs-based immunogen (eOD-GT6) with short intervals between doses. The induced antibodies could neutralize sensitive tier 1 viruses and compete with VRC01 binding. Although the sera antibodies exhibited similar neutralizing activities to those of the homologous prime-boost immunization using SOSIP gp140 alone, the sera from the alternating immunization better recognized epitopes on eOD-GT6. Epitope mapping analyses indicated that nAbs primarily targeted the V3 loop. eOD-GT6 efficiently binds VRC01-class bnAbs and induced antibodies competing with VRC01 binding on eOD-GT6, but the antibodies failed to recognize SOSIP gp140 or exhibit neutralizing activities. Further exploration of this novel immunization strategy could facilitate the development of HIV-1 vaccines that induce CD4bs-directed bnAbs.

4.1 Introduction

Broadly neutralizing antibody (bnAb) induction is necessary for an effective HIV-1 vaccine. However, due to extensive variability and complex immune-evasion strategies, the majority of attempts to elicit bnAbs with HIV-1 immunogens have failed.(1-4) Recent advances in bnAb isolation from HIV-1 infected patients have renewed interest in vaccine strategy development.(5-8) The CD4 binding site (CD4bs) is an attractive target because it contains the epitopes of several bnAbs.

To date, various strategies have been employed for vaccination studies with CD4bs-based immunogens (reviewed in (9)). However, efforts to develop a vaccine that can induce CD4bs targeting bnAbs have been unsuccessful. Although all of the immunogens evaluated could bind VRC01-class bnAbs, they failed to induce similar bnAbs, which was possibly due to difficulties in inducing high levels of CD4bs-directed antibodies that can bind HIV-1 virions (10, 11). Native-like envelope trimers have been developed (e.g. SOSIP gp140), (12-16) but the induced neutralizing activities were mainly limited to tier 1 viruses and homologous tier 2 viruses.(16) The major challenge in developing a CD4bs-based vaccine is designing and/or developing novel vaccine strategies that could force the immune system to focus antibody responses towards the appropriate CD4bs and recognize this epitope on native HIV-1 virions. (10)

Although the studies of sequential immunization with different immunogens have been reported recently, they used relatively long dose intervals and different order of immunogens administration. (17-20) We propose that a heterologous prime-boost with short dosage intervals and the more optimal selection of immunogens would improve this strategy. In this study, we devised a novel rapid heterologous prime-boost vaccine

strategy. The basic concept is to prime the immune system using a native-like envelope trimer immunogen to stimulate antibodies against that could bind epitopes on trimeric envelope spikes on the virus particles. Then antibodies that bind the CD4bs would be selectively amplified by boosting with a smaller CD4bs-based immunogen that lacks distracting non-neutralizing regions within short dose intervals. This prime-boost cycle would be repeated to stimulate high levels of antibodies that can bind CD4bs on native virus particles and minimize responses against the non-neutralizing immunodominant epitopes.

In this proof-of-concept study, we evaluated our vaccine strategy using SOSIP gp140 (15)and eOD-GT6 in rabbits. eOD-GT6, as described elsewhere (21), is an engineered outer domain lacking the V3 loop and large portions of V4 and V5 regions. This protein could efficiently bind multiple VRC01-class bnAbs but not V3 specific antibodies, and thus seems to be a promising immunogen to induce CD4bs specific antibodies. However, the immunogecity analysis of eOD-GT6 has not been reported yet. Here we evaluated the immunogecity of eOD-GT6 and also SOSIP gp140 immunogens alone as comparison to our combined vaccine strategy.

4.2 Results

4.2.1 Immunization and evaluation of antibody responses

To evaluate the immunogenic properties of eOD-GT6, three rabbits were immunized as previously described. To evaluate the rapid *heterologous* prime-boost strategy, as shown in Fig.1, three rabbits were first immunized with SOSIP gp140, then alternatively boosted with either eOD-GT6 and SOSIP gp140. A control group of three

rabbits were immunized with SOSIP 140 only. Serum samples were collected at the indicated time points.

The antibody responses were evaluated against both eOD-GT6 and SOSIP gp140 by ELISA. As shown in Fig.2, the eOD-GT6 specific antibody responses induced in eOD-GT6 group was greater than that of the heterologous prime-boost group, whereas the heterologous prime-boost group induced greater eOD-GT6 specific antibodies than that of the SOSIP gp140 group. It suggests that more antibodies were induced to target the eOD-GT6 by using prime-boost strategy. The SOSIP gp140 specific antibody responses induced in SOSIP gp140 and prime-boost groups were at a similar level, while that of eOD-GT6 group were much weaker. Although many antibodies were induced in the eOD-GT6 group, they could not recognize the corresponding epitopes on SOSIP gp140 trimer spikes. In summary, these results indicate that while eOD-GT6 specific antibody responses were induced in all three groups, SOSIP gp140 specific antibody responses were primarily induced in only the SOSIP gp140 and prime-boost groups. This difference in antigen recognition may be due to steric hindrance present on the larger SOSIP gp140 antigen that prevent the binding of eOD-GT6 elicited antibodies to their epitopes. Neutralizing activity was induced in both the prime-boost and SOSIP gp140 groups, but they were mainly limited to sensitive tier 1 viruses. No neutralization was observed in the eOD-GT6 group (data not shown).

4.2.2 Competition analyses with VRC01

Since there was a significant level of antibodies that bound eOD-GT6 in all immunization groups, we evaluated the induction of antibodies at or near the CD4bs.

Antibody competition assays with VRC01 were conducted using eOD-GT6 and SOSIP gp140. When tested on eOD-GT6, antibodies from both eOD-GT6 and prime-boost groups could compete away approximately 90% of VRC01 at a 1:10 dilution of antisera collected after the third immunization and second set of immunizations (after fourth) respectively. Antibodies from two rabbits in SOSIP gp140 group could compete VRC01 at 1:10 dilution of antisera after the third immunization, but with a much lower percentage. When tested on SOSIP gp140, antibodies from both the SOSIP gp140 and prime-boost groups could compete away a significantly higher percent of VRC01 than antibodies from eOD-GT6 group. The most probable reason again could be the inability of antibodies induced by eOD-GT6 to access epitopes on SOSIP gp140. Nevertheless, these results indicated antibodies were induced targeting near the CD4bs.

4.2.3 Identification of immunogenic linear epitopes within the outer domain

To better characterize antibody responses and identify immunogenic linear epitopes in the outer domain region, we conducted ELISA with overlapping peptides for all three groups using immune sera collected after the final immunization. As shown in Fig.4, no linear epitope was immunogenic on eOD-GT6. This is not surprising since eOD-GT6 lacks the V3 loop and most of the V4 and V5 regions, thus the induced antibodies may mostly target the conformational epitopes. For antibodies from SOSIP gp140 and prime-boost groups, the most immunogenic linear epitope was the V3 loop, especially the N-terminal half of the loop (peptides 9047-CTRPNNNTRKSIRIG, 9048-NNNTRKSIRIGPGQA and 9049-RKSIRIGPGQAFYAT). It is important to note that antibody responses against peptides exhibited animal-to-animal variation. Rabbit 1 from

the prime-boost group recognized peptides within C4 region (9078-MWQGVGQAMYAPPIE and 9079- VGQAMYAPPIEGKIT), while no anti-C4 antibodies were induced in SOSIP gp140 immunized animals.

4.3 Discussion

In this study, we described the immunogenic properties of two immunogens (eOD-GT6 and SOSIP gp140), and evaluated a rapid heterologous prime-boost vaccine strategy using these two immunogens. Although a significant amount of antibodies were induced near the CD4bs, eOD-GT6 failed to elicit nAbs. SOSIP gp140 successfully elicited nAbs, but the neutralizing activity was limited largely to Tier 1 viruses, which is consistent with a previous study. The heterologous prime-boost strategy induced nAbs, and the neutralizing activity was similar to that of SOSIP gp140 alone. Although we failed to induce true bnAbs, results described are still meaningful as no report has described the immunogenicity of eOD-GT6 or have evaluated this particular rapid heterologous prime-boost vaccine strategy.

That eOD-GT6 failed to induce nAbs may be primarily due to steric hindrance surrounding the epitopes of the antibodies on native HIV-1 virions. Since eOD-GT6 is a small immunogen and lacks a large portion of the variable regions, the epitopes on eOD-GT6 may not represent a true native structure as seen on HIV-1 virion. Antibodies induced by eOD-GT6 could not access the epitopes on virions, thus they did not exhibit neutralizing activity. This was supported by the results that antibodies induced from eOD-GT6 could not recognize native-like SOSIP gp140.

The most probable reason why SOSIP gp140 failed to elicit bnAbs is the presence of distracting factors. The epitopes on SOSIP gp140 are native-like, therefore antibodies induced against the CD4bs should have great potential to exhibit broad and potent neutralization. However, SOSIP gp140 contains immunodominant variable regions (i.e. V3 loop), which direct the immune responses away from the broadly neutralizing epitope on the CD4bs. Although antibodies were induced that competed with VRC01 binding, they may target epitopes around CD4bs and prevent VRC01 binding via steric hindrance. Therefore, although SOSIP gp140 contains native-like CD4bs epitopes, it failed to induce potent bnAbs against CD4bs because of immunodominant epitopes and imperfect targeting.

The prime-boost vaccine regimen induced antibody responses that could recognize both eOD-GT6 and SOSIP gp140, and compete with VRC01 binding and neutralize Tier 1 sensitive viruses. One possible reason why this strategy could not induce bnAbs is that the eOD-GT6 may not be an optimal boosting immunogen. Although eOD-GT6 could efficiently bind various VRC01-class antibodies, it failed to induce antibodies that could recognize SOSIP gp140 and exhibit neutralizing activity. Substituting eOD-GT6 with another immunogen that lacks distracting regions but preserves native conformational epitopes, such as the more recent eOD-GT8, may enhance the immunogenicity of CD4bs and elicit VRC01-like nAbs. Another possible reason why we were unable to induce fairly potent nAbs with our prime-boost vaccine strategy is the order of the immunogens administrated. Priming with SOSIP gp140 may have driven initial antibody responses against the immunodominant V3 region, which is dominant to the immune system and could distract responses from targeting the CD4bs.

Perhaps priming with an immunogen lacking immunodominant regions (e.g. eOD-GT6) may enhance the immunogenicity toward CD4bs.

4.4 Conclusions

In this study, we described the immunogenicity of eOD-GT6 and evaluated a rapid heterologous prime-boost vaccine strategy. Additional studies will be needed to better evaluate this rapid vaccination strategy (e.g. substituting eOD-GT6 and changing the order of immunogen administration). Developing novel heterologous prime-boost vaccine strategies that incorporate multiple immunogens to focus anti-body responses to critical neutralizing epitopes should facilitate the development of HIV-1 vaccine.

4.5 Materials and methods

4.5.1 Immunogen generation

eOD-GT6 fused to Lumazine Synthase was synthesized commercially (Life Technologies) with restriction sites AgeI and KpnI introduced at 5’ and 3’ ends, respectively. The synthesized gene was amplified by PCR with EcoR1 and AgeI sites at 5’ and 3’ ends, respectively. The PCR fragment was cloned into pcDNA3.1 expression vector through corresponding restriction enzyme sites. eOD-GT6 expression and purification was described previously. The SOSIP gp140 trimer plasmid bearing a His-tag at the C-terminus and Furin plasmid were provided by Dr. John P. Moore. The SOSIP gp140 protein was expressed in 293 F cells with SOSIP gp140 plasmid and Furin plasmid at 4:1 mass ratio and purified from cell culture medium by a single-step affinity chromatography using Ni-NTA resin.

4.5.2 Rabbit immunization

Nine New Zealand white female rabbits (2.5 to 3 kg) were purchased from Charles River (USA), housed under specific pathogen free environments. All animals were tested in compliance with the animal protocol approved by IACUC of Iowa State University. The rabbits were randomly divided into three groups of three rabbits each.

For groups 1 and 2 (Fig.1A), rabbits were subcutaneously injected with 200 μg of eOD-GT6 or SOSIP gp140 three times (weeks 0, 4, and 9) respectively. Zn-chitosan was used as an adjutant. Sera were collected two weeks post each immunization. For group 3 (Fig.1A), rabbits were primed subcutaneously with 50 μg of SOSIP gp140, then boosted with 10 μg of eOD-GT6 after three and half days. Then repeated this prime-boost process one more time with 10 μg immunogens. After three weeks, the rabbits were alternatively immunized twice with 10 μg of SOSIP gp140 and 10 μg of eOD-GT6. Sera were collected according to the immunization schedule (Fig.1).

4.5.3 Enzyme-linked immunosorbent assays (ELISA)

All ELISAs were performed as described elsewhere, but with an alternate blocking buffer of PBS (pH 7.5) with 2.5% skim milk and 5% calf sera.

4.5.4 Neutralization assays

Neutralization assays were performed in TZM-bl cells as described. A panel of pseudoviruses was tested, including SF162 (tier 1A, clade B), MW965.26 (tier 1A, clade C), and MN.3 (tier 1A, clade B).

4.5.5 Acknowledgments

The following reagents were obtained through the NIH AIDS Reagent Program, Division of AIDS, NIAID, NIH: HIV-1 Consensus Group M Env peptides (Cat# 9487); HIV-1 gp120 MAb (VRC01) from Dr. John Mascola (Cat# 12033); We are grateful to Dr. E. Yvonne Jones and Dr. John P. Moore for providing pHLsec and BG505 SOSIP gp140 constructs, respectively. This work was supported by the NIH grant P01AI074286 and Iowa State University.

4.6 Figures

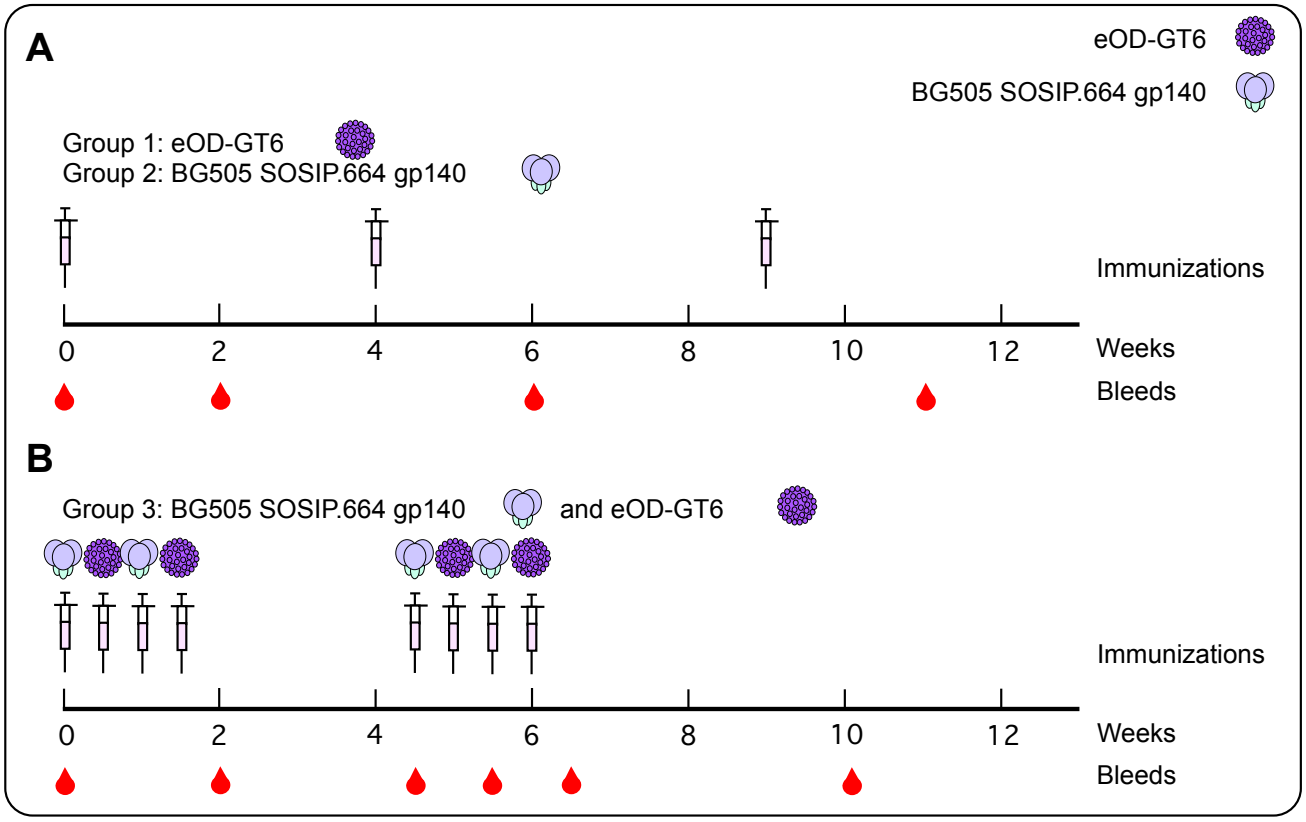

Fig .1. Immunization schedule. (A) Timeline for eOD-GT6 (group 1) and SOSIP gp140 (group 2) immunization and sampling. Rabbits were immunized on weeks 0, 4, and 9. Pre-immune, as well as post-immune sera (two weeks post each immunization) were taken. (B). Immunization schedule for rapid heterologous prime-boost (group 3) immunization. Rabbits were alternatively immunized with SOSIP gp140 and eOD-GT6 on weeks 0, 0.5, 1, 1.5, 4.5, 5, 5.5, and 6. Sera were collected on weeks 0, 2, 4.5, 5.5, 6.5, and 10.

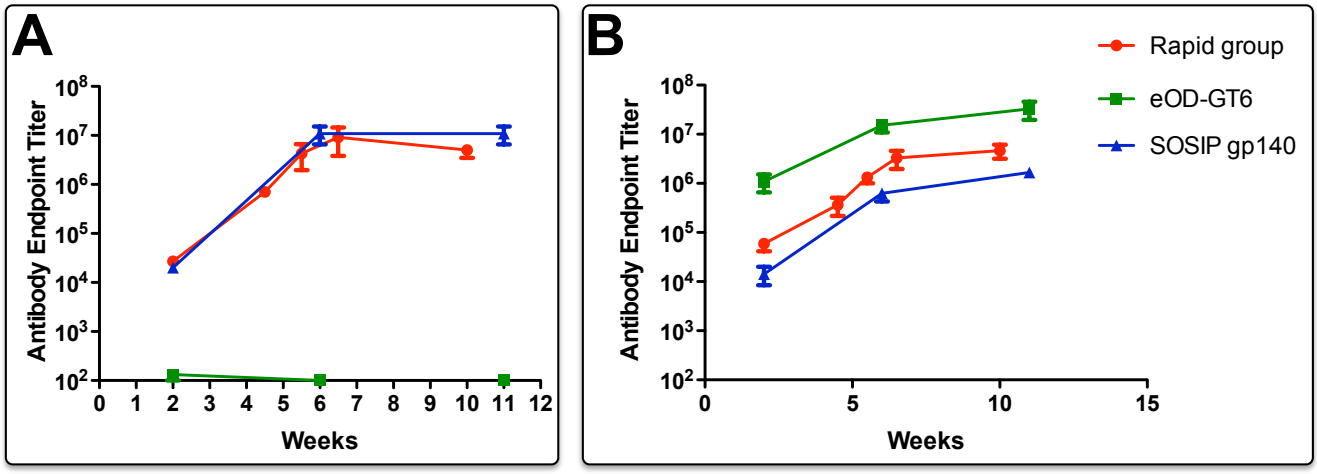

Fig. 2. Endpoint titers of cross-reactive antibodies. The collected sera were monitored for SOSIP gp140 (A) and eOD-GT6 (B) cross-reactive antibodies by ELISA. The results are presented as average endpoint titers of three rabbits within each group with standard deviation.

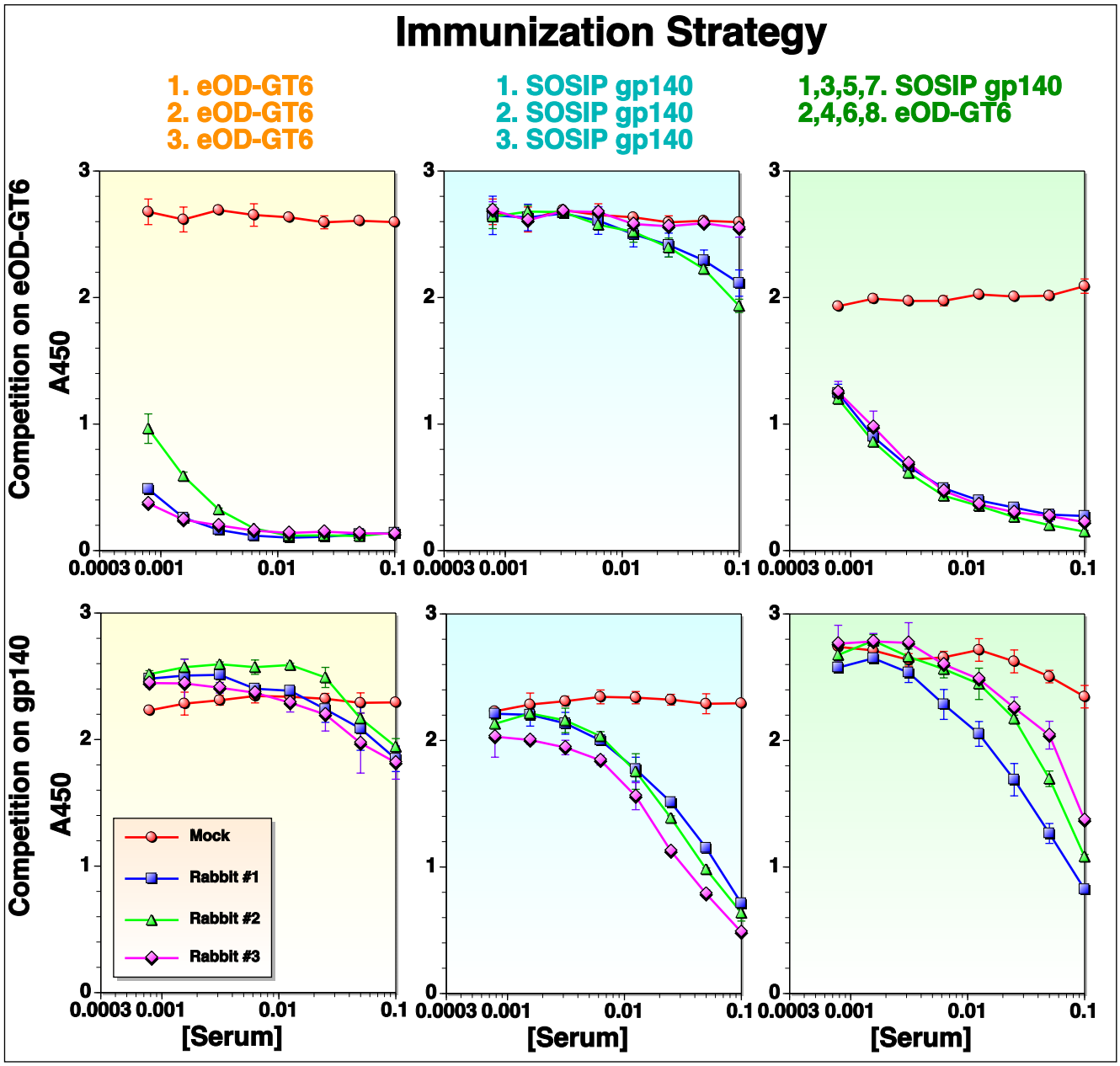

Fig. 3. Competition assay against VRC01. Binding of VRC01 to eOD-GT6 or SOSIP gp140 was competed with immune sera after the final immunization from the immunized rabbits or mock-immunized rabbit (PBS).

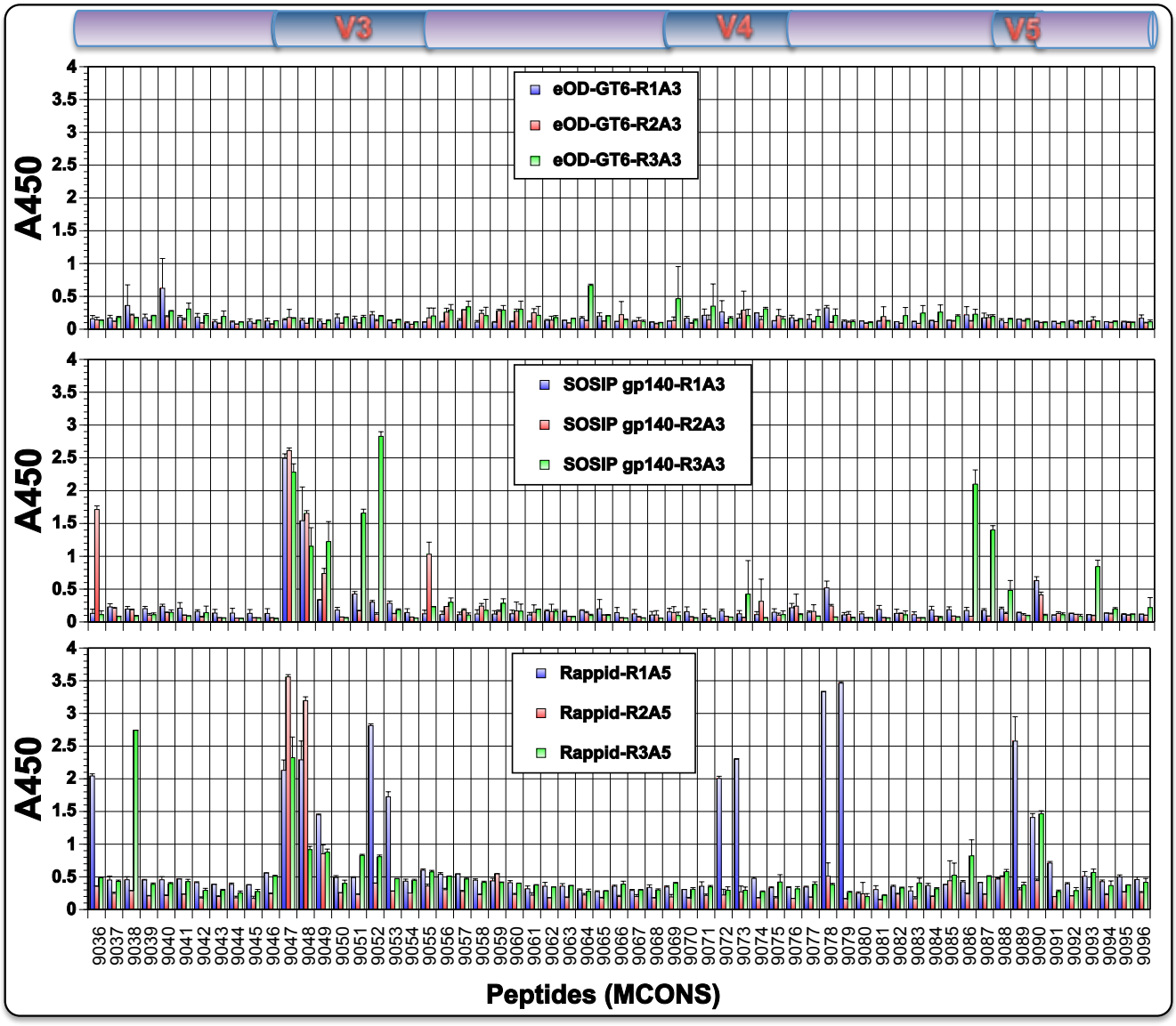

Fig. 4. Identification of immunogenic linear epitopes using overlapping peptides. ELISA was performed with the immune sera after the final immunization. On the top of the figure shows a schematic diagram of gp120 outer domain. Peptide numbers stand for catalog numbers from the NIH AIDS Region Program.

4.7 References

1. **Kwong PD**, **Mascola JR**, **Nabel GJ**. 2011. Rational design of vaccines to elicit broadly neutralizing antibodies to HIV-1. Cold Spring Harb Perspect Med **1**:a007278.

2. **Burton DR**, **Desrosiers RC**, **Doms RW**, **Koff WC**, **Kwong PD**, **Moore JP**, **Nabel GJ**, **Sodroski J**, **Wilson IA**, **Wyatt RT**. 2004. HIV vaccine design and the neutralizing antibody problem. Nat Immunol **5**:233–236.

3. **Cohen YZ**, **Dolin R**. 2013. Novel HIV vaccine strategies: overview and perspective. Ther Adv Vaccines **1**:99–112.

4. **Schiffner T**, **Sattentau QJ**, **Dorrell L**. 2013. Development of prophylactic vaccines against HIV-1. Retrovirology **10**:72.

5. **Diskin R**, **Scheid JF**, **Marcovecchio PM**, **West AP**, **Klein F**, **Gao H**, **Gnanapragasam PNP**, **Abadir A**, **Seaman MS**, **Nussenzweig MC**, **Björkman PJ**. 2011. Increasing the potency and breadth of an HIV antibody by using structure-based rational design. Science **334**:1289–1293.

6. **Scheid JF**, **Mouquet H**, **Ueberheide B**, **Diskin R**, **Klein F**, **Oliveira TYK**, **Pietzsch J**, **Fenyo D**, **Abadir A**, **Velinzon K**, **Hurley A**, **Myung S**, **Boulad F**, **Poignard P**, **Burton DR**, **Pereyra F**, **Ho DD**, **Walker BD**, **Seaman MS**, **Björkman PJ**, **Chait BT**, **Nussenzweig MC**. 2011. Sequence and structural convergence of broad and potent HIV antibodies that mimic CD4 binding. Science **333**:1633–1637.

7. **Walker LM**, **Phogat SK**, **Chan-Hui P-Y**, **Wagner D**, **Phung P**, **Goss JL**, **Wrin T**, **Simek MD**, **Fling S**, **Mitcham JL**, **Lehrman JK**, **Priddy FH**, **Olsen OA**, **Frey SM**, **Hammond PW**, **Protocol G Principal Investigators**, **Kaminsky S**, **Zamb T**, **Moyle M**, **Koff WC**, **Poignard P**, **Burton DR**. 2009. Broad and potent neutralizing antibodies from an African donor reveal a new HIV-1 vaccine target. Science **326**:285–289.

8. **Purtscher M**, **Trkola A**, **Gruber G**, **Buchacher A**, **Predl R**, **Steindl F**, **Tauer C**, **Berger R**, **Barrett N**, **Jungbauer A**. 1994. A broadly neutralizing human monoclonal antibody against gp41 of human immunodeficiency virus type 1. AIDS Res Hum Retroviruses **10**:1651–1658.

9. **Georgiev IS**, **Gordon Joyce M**, **Zhou T**, **Kwong PD**. 2013. Elicitation of HIV-1-neutralizing antibodies against the CD4-binding site. Curr Opin HIV AIDS **8**:382–392.

10. **Qin Y**, **Shi H**, **Banerjee S**, **Agrawal A**, **Banasik M**, **Cho MW**. 2014. Detailed characterization of antibody responses against HIV-1 group M consensus gp120 in rabbits. Retrovirology **11**:125.

11. **Qin Y**, **Banasik M**, **Kim S**, **Penn-Nicholson A**, **Habte HH**, **LaBranche C**, **Montefiori DC**, **Wang C**, **Cho MW**. 2014. Eliciting neutralizing antibodies with gp120 outer domain constructs based on M-group consensus sequence. Virology **462-463**:363–376.

12. **Sanders RW**, **Vesanen M**, **Schuelke N**, **Master A**, **Schiffner L**, **Kalyanaraman R**, **Paluch M**, **Berkhout B**, **Maddon PJ**, **Olson WC**, **Lu M**, **Moore JP**. 2002. Stabilization of the soluble, cleaved, trimeric form of the envelope glycoprotein complex of human immunodeficiency virus type 1. journal of virology **76**:8875–8889.

13. **Klasse PJ**, **Depetris RS**, **Pejchal R**, **Julien J-P**, **Khayat R**, **Lee JH**, **Marozsan AJ**, **Cupo A**, **Cocco N**, **Korzun J**, **Yasmeen A**, **Ward AB**, **Wilson IA**, **Sanders RW**, **Moore JP**. 2013. Influences on trimerization and aggregation of soluble, cleaved HIV-1 SOSIP envelope glycoprotein. journal of virology **87**:9873–9885.

14. **Hoffenberg S**, **Powell R**, **Carpov A**, **Wagner D**, **Wilson A**, **Kosakovsky Pond S**, **Lindsay R**, **Arendt H**, **Destefano J**, **Phogat S**, **Poignard P**, **Fling SP**, **Simek M**, **LaBranche C**, **Montefiori D**, **Wrin T**, **Phung P**, **Burton D**, **Koff W**, **King CR**, **Parks CL**, **Caulfield MJ**. 2013. Identification of an HIV-1 clade A envelope that exhibits broad antigenicity and neutralization sensitivity and elicits antibodies targeting three distinct epitopes. journal of virology **87**:5372–5383.

15. **Sanders RW**, **Derking R**, **Cupo A**, **Julien J-P**, **Yasmeen A**, **de Val N**, **Kim HJ**, **Blattner C**, **la Peña de AT**, **Korzun J**, **Golabek M**, **de Los Reyes K**, **Ketas TJ**, **van Gils MJ**, **King CR**, **Wilson IA**, **Ward AB**, **Klasse PJ**, **Moore JP**. 2013. A next-generation cleaved, soluble HIV-1 Env trimer, BG505 SOSIP.664 gp140, expresses multiple epitopes for broadly neutralizing but not non-neutralizing antibodies. PLoS Pathog **9**:e1003618.

16. **Sanders RW**, **van Gils MJ**, **Derking R**, **Sok D**, **Ketas TJ**, **Burger JA**, **Ozorowski G**, **Cupo A**, **Simonich C**, **Goo L**, **Arendt H**, **Kim HJ**, **Lee JH**, **Pugach P**, **Williams M**, **Debnath G**, **Moldt B**, **van Breemen MJ**, **Isik G**, **Medina-Ramírez M**, **Back JW**, **Koff WC**, **Julien J-P**, **Rakasz EG**, **Seaman MS**, **Guttman M**, **Lee KK**, **Klasse PJ**, **LaBranche C**, **Schief WR**, **Wilson IA**, **Overbaugh J**, **Burton DR**, **Ward AB**, **Montefiori DC**, **Dean H**, **Moore JP**. 2015. HIV-1 VACCINES. HIV-1 neutralizing antibodies induced by native-like envelope trimers. Science **349**:aac4223.

17. **Zhang M**, **Zhang L**, **Zhang C**, **Hong K**, **Shao Y**, **Huang Z**, **Wang S**, **Lu S**. 2012. DNA prime-protein boost using subtype consensus Env was effective in eliciting neutralizing antibody responses against subtype BC HIV-1 viruses circulating in China. Human Vaccines & Immunotherapeutics **8**:1630–1637.

18. **Vaine M**, **Wang S**, **Hackett A**, **Arthos J**, **Lu S**. 2010. Antibody responses elicited through homologous or heterologous prime-boost DNA and protein vaccinations differ in functional activity and avidity. Vaccines **28**:2999–3007.

19. **Lu S**. 2009. Heterologous prime-boost vaccination. Curr Opin Immunol **21**:346–351.

20. **Escolano A**, **Steichen JM**, **Dosenovic P**, **Kulp DW**, **Golijanin J**, **Sok D**, **Freund NT**, **Gitlin AD**, **Oliveira T**, **Araki T**, **Lowe S**, **Chen ST**, **Heinemann J**, **Yao K-H**, **Georgeson E**, **Saye-Francisco KL**, **Gazumyan A**, **Adachi Y**, **Kubitz M**, **Burton DR**, **Schief WR**, **Nussenzweig MC**. 2016. Sequential Immunization Elicits Broadly Neutralizing Anti-HIV-1 Antibodies in Ig Knockin Mice. Cell **166**:1445–1458.e12.

21. **Jardine J**, **Julien J-P**, **Menis S**, **Ota T**, **Kalyuzhniy O**, **McGuire A**, **Sok D**, **Huang P-S**, **MacPherson S**, **Jones M**, **Nieusma T**, **Mathison J**, **Baker D**, **Ward AB**, **Burton DR**, **Stamatatos L**, **Nemazee D**, **Wilson IA**, **Schief WR**. 2013. Rational HIV immunogen design to target specific germline B cell receptors. Science **340**:711–716.

CHAPTER 5

EVALUATION OF A MULTI-IMMUNOGEN VACCINE STRATEGY FOR TARGETING CD4BS NEUTRALIZING EPITOPES ON HIV-1 GP120

Heliang Shi, Saikat Banerjee, Aditi Agrawal, Michael Cho

Abstract

The CD4 binding site (CD4bs) of gp120 is a prime target for the elicitation of broadly neutralizing antibodies (bnAbs). In this study, we evaluated a vaccine strategy applying a sequential and phased mannered immunization approach with related but antigenically distinct immunogens. Rabbits were immunized with a small CD4bs-based immunogen (eOD-GT6) and boosted with progressively more native immunogens (gp120 and SOSIP gp140). Although bnAbs were not induced, antibody responses analysises indicated our vaccination elicited antibodies that could efficiently recognize all three administered immunogens. Antibodies could compete VRC01 binding on all three immunogens. We may induce antibodies that could recognize the conformational consensus of the administered immunogens. Further exploration of our vaccine strategy using improved immunogens would enhance the possibility to induce bnAbs against CD4bs.

5.1 Introduction

Development of a safe and effective HIV vaccine remains a global public health priority. Recent discoveries of potent and broad neutralizing antibodies (bnAbs) from

HIV-1 infected patients have renewed interest in bnAbs induction(1-4). However, the elicitation of bnAbs has not been successfully achieved by vaccination in standard animal models or humans(5-8).

The CD4 binding site (CD4bs) of gp120 is a particularly attractive target for the elicitation of bnAbs because it is highly conserved and contains epitopes of bnAbs. To date, much effort has been made in order to elicit CD4bs-directed bnAbs (reviewed in (9)). However, it has not been successful to elicit such bnAbs by vaccination.

The major challenge for the development of a vaccine that is able to induce bnAbs towards CD4bs is developing novel immuogens and/or vaccine strategies that could enhance the CD4bs specific immune responses and recognize neutralizing epitopes on native envelope spikes.(10, 11) We previously described a rapid heterologous prime boost vaccine strategy using SOSIP gp140 and eOD-GT6, we induced strong antibody responses that could recognize both immunogens and compete VRC01 binding, but neutralize mainly tier 1 viruses, which we believed was large to suboptimal immunogens or immunogens administration order. We have reported another vaccine strategy, IPAS-RAM, using a sequential and phased mannered immunization with three immunogens that are progressively more native. (12)Although nAbs were not induced, antibodies whose epitopes closely resembled those of 4E10 and 10E8 were induced. We believed this strategy could be improved with optimal priming immunogen. Inspired by these studies, we developed a vaccine strategy to induce bnAbs towards CD4bs.

In this study, we evaluated a heterologous prime boost vaccine strategy using three related but antigenically distinct immunogens. The basic concept is to prime the

immune system using a small CD4bs-based immunogen (eOD-GT6)(13) to produce a large antibody repertoire against CD4bs, then boost with progressively more native immunogens (gp120 and SOSIP gp140) (11, 14) to specifically amplify antibodies that bind the native structure. Although the studies of heterologous prime boost vaccine strategy have been reported recently, we applied different set of immunogens along with sequential and phase mannered immunization approach. Here we evaluated our vaccine strategy in rabbits.

5.2 Results

5.2.1 Immunization schedule

The production of all immunogens used in this study has been described before. (10, 11, 14)To evaluate our vaccine strategy, six rabbits were used in the study. Three rabbits were first immunized with eOD-GT6 only. Four weeks later, a combination of eOD-GT6 and Mcon6 gp120 was administered. The eOD-GT6 was also included, rather than immunizing Mcon6 gp120 alone, because we hypothesized that immunizing with both immunogens would preferentially induce antibody responses targeting epitopes present on both immunogens (I.e the CD4bs). Similarly, on week 11, a combination of Mcon6 gp120 and SOSIP gp140 was administered. Three more rabbits were immunized with eOD-GT6 alone twice with four weeks interval. We hypothesized that priming with eOD-GT6 would initialize antibody responses towards the CD4bs and boost one more time would make those antibodies dominant the immune system which may help reduce the antibody responses against the immunodominant variable loops in the following immunogens. The eOD-GT6 and Mcon6 gp120 combination and Mcon6 gp120 and

SOSIP gp140 combination were administered on weeks 11 and 15 respectively. On week 23, SOSIP gp140 was administered, because we hypothesized that boosting with SOSIP gp140 alone would selectively amplify the CD4bs-directed antibodies that could recognize the epitopes on native HIV-1 virions.

5.2.2 Evaluation of antibody responses

The immunogenicity studies of eOD-GT6, Mcon6 gp120, and SOSIP gp140 have been described before. To examine whether antibodies induced by our vaccine strategies are different from those induced by homologous prime-boost immunization, we performed ELISA to evaluate the antibody responses of all groups on different immunogens. The eOD-GT6 induced high antibody responses targeting eOD-GT6 after first immunization, and the antibody responses were almost saturated after second immunization. However, the antibodies bound Mcon6 gp120 very poorly and could not recognize SOSIP gp140. These were possibly due to the steric hindrance of antibodies on binding on more native immunogens. The epitopes on eOD-GT6 may be more similar to those on Mcon6 gp120 than on the native SOSIP gp140, thus they could weakly recognize Mcon6 gp120. The antibodies from both Mcon6 gp120 and SOSIP gp140 groups could recognize all three immunogens. For Mcon6 gp120 group, the antibodies level that recognized SOSIP gp140 was lower than those for Mcon6 gp120 and eOD-GT6 except one rabbit, whose antibodies failed to bind to eOD-GT6 and was possibly due to the animal-to-animal variation. These results indicate there are some antibodies that could recognize the epitopes on both eOD-GT6 and Mcon6 gp120 failed

to recognize the epitopes on SOSIP gp140, which could be mainly due to the epitopes conformation differences. The levels of antibodies induced from SOSIP gp140 recognizing SOSIP gp140 and Mcon6 gp120 are similar.

For rabbits in group 1, the antibodies induced after first immunization only recognized eOD-GT6. After second immunization, the antibodies could recognize all of three immunogens, and the titer of antibodies against eOD-GT6 was highest and those against SOSIP gp140 were lowest. Antibody titers against all three immunogens continued to increase after third immunization, among which the antibody titer against SOSIP gp140 was slightly lower. For rabbits in group 2, strong antibody responses against eOD-GT6 were induced after the first immunization, but no antibody responses were detected against neither Mcon6 gp120 nor SOSIP gp140. Antibody titers against eOD-GT6 continued to increase after second immunization, and weak antibody responses were induced against Mcon6 gp120. After the immunization with a combination of eOD-GT6 and Mcon6 gp120, strong antibody responses were induced against for both eOD-GT6 and Mcon6 gp120, which was similar to that of group 1 after second immunization. But the antibody titer against SOSIP gp140 was lower than that of group 1. One possible reason is the existence of strong antibody responses against eOD-GT6 may suppress the immune responses against the immunodominant regions on SOSIP gp140. Then the antibody titers against eOD-GT6 began to decrease, while those against Mcon6 gp120 and SOSIP gp140 continued to increase. Then significant changes in antibody responses were not observed after the final immunization. The final antibodies binding profiles from group 2 were similar to those from groups 1. In summary, these results suggest that strong antibody responses recognizing all three

immunogens were induced using our vaccine strategies. The neutralizing activities were induced in both groups; however, they were limited to tier 1 viruses (data not shown).

5.2.3 Competition analyses with bnAbs VRC01

Since strong antibody responses induced from our immunized rabbits could recognize all three immunogens, we would like to evaluate the induction of antibodies at or near the CD4bs. Antibody competition assays with VRC01 were conducted using eOD6-GT6, Mcon6 gp120, and SOSIP gp140. When tested on eOD-GT6, antibodies from group 1, group 2, and eOD-GT6 group could compete away approximately over 70% of VRC01 at a 1:10 dilution of antisera, while antibodies from Mcon6 gp120 could not compete away VRC01 and antibodies from SOSIP gp140 group competed VRC01 weakly.

One possible reason is the antibodies from eOD-GT6 group have higher affinity to eOD-GT6 than those from gp120 and SOSIP gp140 groups since rabbits were immunized with the same eOD-GT6. When tested on gp120, antibodies from eOD-GT6 group failed to compete VRC01 binding, while antibodies from all other groups could efficiently compete away VRC01, and the competing antibody level in SOSIP gp140 was lowest. The reason why antibodies from eOD-GT6 group failed to compete VRC01 on gp120 is antibodies could not recognize epitopes on gp120. The lower competing antibody levels from SOSIP gp140 is possibly due to the lower binding affinity to gp120 than other groups since rabbits were immunized with gp120. Similar competing results were observed when tested on SOSIP gp140, and competing antibody level in SOSIP gp140, instead of gp120, was highest.

5.2.4 Identification of immunogenic linear epitopes within the outer domain

To identify immunogenic linear epitopes, ELISAs were conducted with overlapping MCONS peptide sets as we previously described. Although there are some differences in amino acid sequences between MCONS and our immunogens, the variant residues are mostly in the variable loops. Although there were animal-to-animal variations within the group, most of the immunoreactive peptides were within the V3, C4, and C5 regions.

The most immunogenic region was the N-terminal of V3 loop (peptides 9047-CTRPNNNTRKSIRIG, 9048-NNNTRKSIRIGPGQA and 9049- RKSIRIGPGQAFYAT). Two peptides in the C4 region (9079-VGQAMYAPPIEGKIT and 9080-MYAPPIEGKITCKSN) were highly reactive to antibodies induced in group 2 rabbits, and modest level of antibodies against peptide 9080 was observed from group 1 rabbits. These anti-C4 antibodies could be resulted from Mcon6 gp120 immunization, because no anti-C4 antibodies were induced in both eOD-GT6 and SOSIP gp140 immunized rabbits. The C5 region (peptides 9090, 9093 and 9094 sequences) was reactive to antibodies induced in the rabbits from group 1, and the peptides 9093 and 9094 was reactive to antibodies induced from group 2 immunized rabbits. This is not surprising since both Mcon6 gp120 and SOSIP gp140 could induce antibodies recognizing the C5 region.

5.3 Discussions

In this study, we described a heterologous multi-immunogen prime-boost vaccine strategy to enhance antibody responses against non-immunogenic neutralizing

epitopes. Multiple related, but antigenically distinct, immuogens were used. We primed the immune system with a small immuogen (eOD-GT6) to focus immune responses to the desired epitopes (e.g. CD4bs) and produce a large antibody repertoire that could recognize all possible epitope conformations. The boost immunizations were performed with more “native” immunogens (gp120) to direct the antibody maturation process. The final immunizations were carried out with native-like SOSIP gp140 trimer to enhance the possibility that mature antibodies can recognize epitopes on native envelope spikes. The immunnogens were immunized in a phased manner, because we hypothesized that anamnestic immune responses against previously injected immunogens would boost related epitopes on new immunogens rapidly and force the immune responses to identify a conformational consensus among all immunogens administered. We induced antibodies that efficiently compete VRC01 binding and exhibit neutralizing activities to Tier 1 viruses.

Both of our vaccine regimens induced strong antibody responses that could efficiently recognize all three immunogens. After immunization with eOD-GT6 alone, antibodies barely recognized gp120 and SOSIP gp140. This is possibly due to the differences in eptiope conformations among immunogens. After subsequent immunizations with more native immunogens, antibodies recognizing all immunogens were induced. The ideal epitopes of these antibodies would be the conformational consensus of all immunogens, CD4bs. The competitive ELISA with VRC01 was performed to assess the antibodies targeting at or near CD4bs. The competing antibodies from our vaccine regimens could efficiently compete VRC01 on all three immunogens, while antibodies from gp120 and SOSIO gp140 alone groups could not

compete away VRC01 on eOD-GT6. This suggested that our vaccine regimens could induce antibodies that recognize the common conformational epitopes on all immunogens. However, the neutralizing activities were mainly limited to tier 1 viruses. It indicated that the common conformational epitopes identified by our vaccine regimens may not be CD4bs, and the efficiently competition with VRC01 may be due to the steric hindrance of antibody binding to other epitopes rather than CD4bs. Linear epitope mapping analysis indicated V3 loop highly reactive to immune sera.

The most possible reason why we did not elicit bnAbs is the selected immunogens are not optimal. Although eOD-GT6 is a small immunogens that could bind VRC01-class bnAbs efficiently, it failed to induce antibodies recognizing more native immunogens (gp120 and SOSIP gp140) and exhibiting neutralizing activities. Another small immunogen with higher binding affinity to VRC01-class bnAbs and more native conformation may improve our vaccine regimens. Since V3 loop was very immunogenic, substitution of gp120 with a immunogen lacking V3 loop (e.g. gp120-ΔV3) may reduce the immune responses to the immunodominant V3 region and focus immune responses to the common epitopes (e.g. CD4bs).

5.4 Conclusions

In this study, we evaluated a heterologous prime-boost vaccine strategy to induce VRC01-like bnAbs. We induced antibodies that could recognize all administered immunogens , competed with VRC01 binding on all immungens, and exhibited neutralizing activities to tier 1 viruses. We did not induce bnAbs, which was largely due to the use of suboptimal immunogens. Additional work will be needed to improve our

vaccine strategy (e.g. use of optimal priming and boosting immunogens). Developing novel heterologous prime-boost vaccine strategies that force the immune system to identify the common critical neutralizing epitopes would facilitate the development of HIV-1 vaccine.

5.5 Materials and methods

5.5.1 Rabbit immunization

The production of all three immunogens was described previously(10, 11, 14). Six New Zealand white female rabbits (2.5 to 3 kg) were purchased from Charles River (USA) and randomly divided into two groups of three rabbits each. For group 1 (Fig.1A), rabbits were subcutaneously injected with 200 μg of eOD-GT6 on week 0, then boosted with mixture of 20 μg of eOD-GT6 and 200 μg of Mcon6 gp120 on week 5 and mixture of 20 μg of Mcon6 gp120 and 200 μg of SOSIP gp140 on week 11. For group 2 (Fig.1B), rabbits were primed subcutaneously with 200 μg eOD-GT6 on weeks 0 and 4, then boosted subsequently with 20 μg OD-GT6 and 200 μg of Mcon6 gp120 on week 9, 20 μg of Mcon6 gp120 and 200 μg of SOSIP gp140 on week 15, and 200 μg of SOSIP gp140 on week 23. All immune sera were collected two weeks post each immunization.

5.5.2 Enzyme-linked immunosorbent assays (ELISA)

All ELISAs were performed as described elsewhere.

5.5.3 Neutralization assays

Neutralization assays were performed in TZM-bl cells as described previously in other chapters.

5.6 Acknowledgments

The following reagents were obtained through the NIH AIDS Reagent Program, Division of AIDS, NIAID, NIH: HIV-1 Consensus Group M Env peptides (Cat# 9487); HIV-1 gp120 MAb (VRC01) from Dr. John Mascola (Cat# 12033); We are grateful to Dr. E. Yvonne Jones and Dr. John P. Moore for providing pHLsec and BG505 SOSIP gp140 constructs, respectively. This work was supported by the NIH grant P01AI074286 and Iowa State University.

5.7 Figures

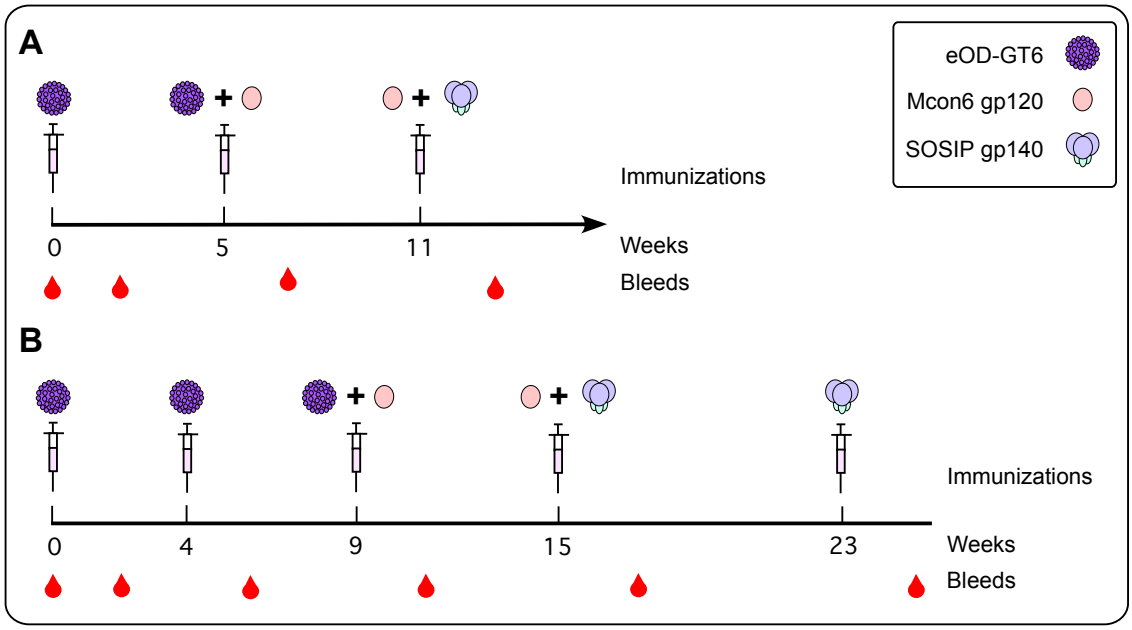

Fig .1. Immunization schedule. Timeline for immunization and sampling. Rabbits in group 1 were immunized with on weeks 0, 5, and 11. . Rabbits in group 2 were immunized with on weeks 0, 4, 9, 15, and 23. The immunogens immunized at each time were illustrated as above. Pre-immune, as well as post-immune sera (two weeks post each immunization) were taken.

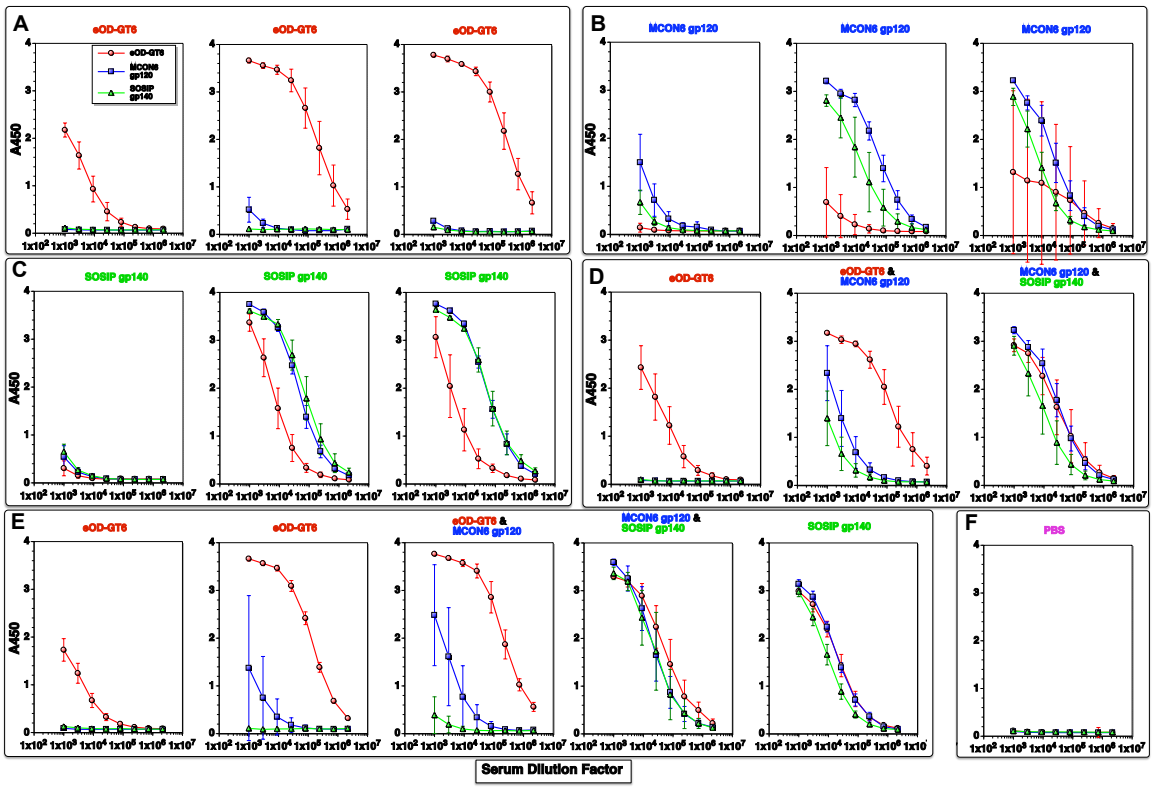

Fig. 2. Cross-reactivity analysis of antibodies. The collected sera were evaluated for eOD-GT6, Mcon6 gp120, and SOSIP gp140 cross-reactive antibodies by ELISA. A to E represents different immunization studies. The results are presented as average value of three rabbits within each group with standard deviation.

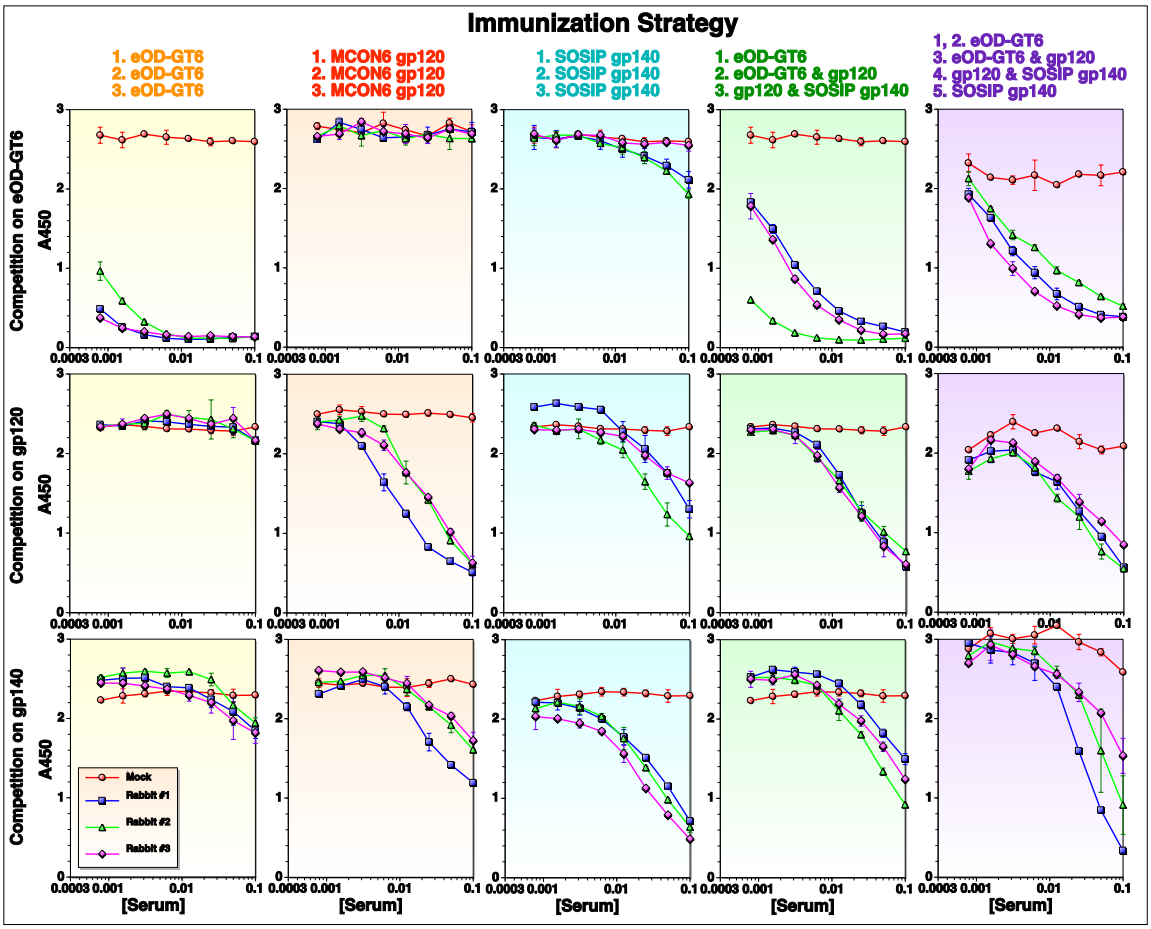

Fig. 3. Competition assay against VRC01. Binding of VRC01 to eOD-GT6, Mcon6 gp120, or SOSIP gp140 was competed. The immune sera after 5th immunization for group 2 and 3rd immunization for other groups from the immunized rabbits or mock-immunized rabbit (PBS) were used.

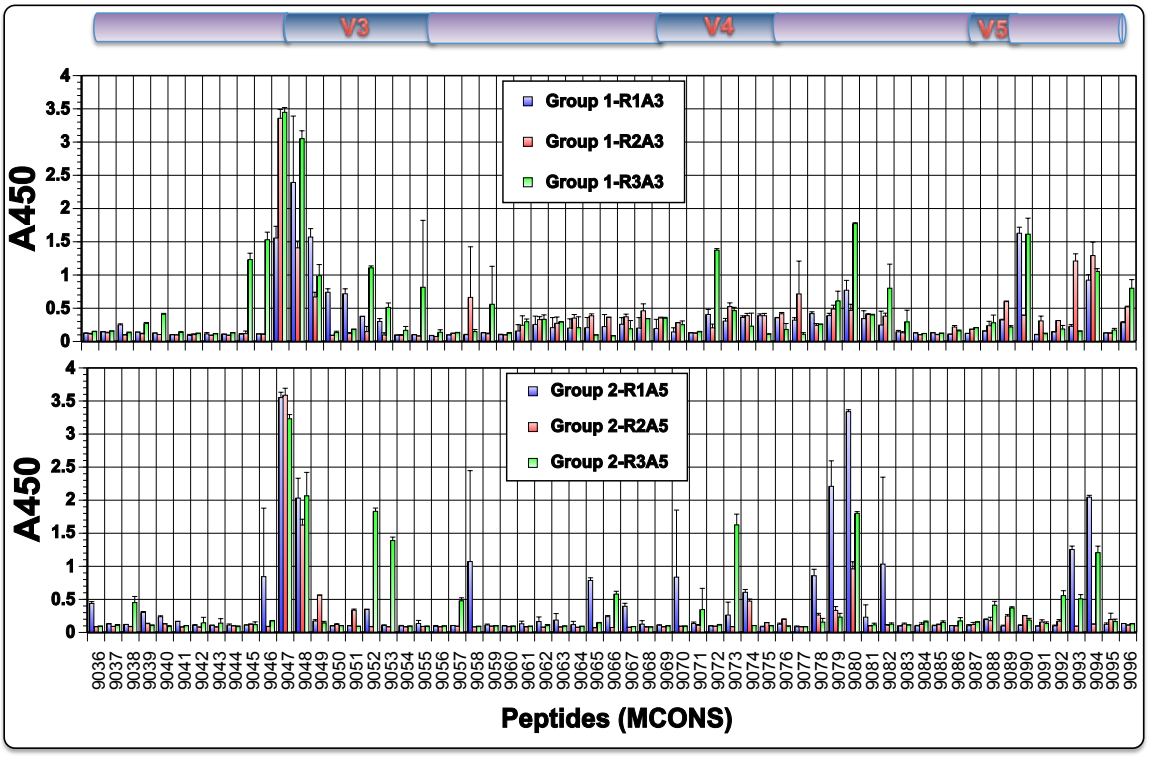

Fig. 4. Identification of immunogenic linear epitopes. ELISA was performed using overlapping peptides. The immune sera used were after the final immunization. Peptide numbers stand for catalog numbers from the NIH AIDS Region Program.

5.8 References

1. **Diskin R**, **Scheid JF**, **Marcovecchio PM**, **West AP**, **Klein F**, **Gao H**, **Gnanapragasam PNP**, **Abadir A**, **Seaman MS**, **Nussenzweig MC**, **Björkman PJ**. 2011. Increasing the potency and breadth of an HIV antibody by using structure-based rational design. Science **334**:1289–1293.

2. **Scheid JF**, **Mouquet H**, **Ueberheide B**, **Diskin R**, **Klein F**, **Oliveira TYK**, **Pietzsch J**, **Fenyo D**, **Abadir A**, **Velinzon K**, **Hurley A**, **Myung S**, **Boulad F**, **Poignard P**, **Burton DR**, **Pereyra F**, **Ho DD**, **Walker BD**, **Seaman MS**, **Björkman PJ**, **Chait BT**, **Nussenzweig MC**. 2011. Sequence and structural convergence of broad and potent HIV antibodies that mimic CD4 binding. Science **333**:1633–1637.

3. **Walker LM**, **Phogat SK**, **Chan-Hui P-Y**, **Wagner D**, **Phung P**, **Goss JL**, **Wrin T**, **Simek MD**, **Fling S**, **Mitcham JL**, **Lehrman JK**, **Priddy FH**, **Olsen OA**, **Frey SM**, **Hammond PW**, **Protocol G Principal Investigators**, **Kaminsky S**, **Zamb T**, **Moyle M**, **Koff WC**, **Poignard P**, **Burton DR**. 2009. Broad and potent neutralizing antibodies from an African donor reveal a new HIV-1 vaccine target. Science **326**:285–289.

4. **Purtscher M**, **Trkola A**, **Gruber G**, **Buchacher A**, **Predl R**, **Steindl F**, **Tauer C**, **Berger R**, **Barrett N**, **Jungbauer A**. 1994. A broadly neutralizing human monoclonal antibody against gp41 of human immunodeficiency virus type 1. AIDS Res Hum Retroviruses **10**:1651–1658.

5. **Kwong PD**, **Mascola JR**, **Nabel GJ**. 2011. Rational design of vaccines to elicit broadly neutralizing antibodies to HIV-1. Cold Spring Harb Perspect Med **1**:a007278.

6. **Burton DR**, **Desrosiers RC**, **Doms RW**, **Koff WC**, **Kwong PD**, **Moore JP**, **Nabel GJ**, **Sodroski J**, **Wilson IA**, **Wyatt RT**. 2004. HIV vaccine design and the neutralizing antibody problem. Nat Immunol **5**:233–236.

7. **Cohen YZ**, **Dolin R**. 2013. Novel HIV vaccine strategies: overview and perspective. Ther Adv Vaccines **1**:99–112.

8. **Schiffner T**, **Sattentau QJ**, **Dorrell L**. 2013. Development of prophylactic vaccines against HIV-1. Retrovirology **10**:72.

9. **Georgiev IS**, **Gordon Joyce M**, **Zhou T**, **Kwong PD**. 2013. Elicitation of HIV-1-neutralizing antibodies against the CD4-binding site. Curr Opin HIV AIDS **8**:382–392.

10. **Qin Y**, **Shi H**, **Banerjee S**, **Agrawal A**, **Banasik M**, **Cho MW**. 2014. Detailed characterization of antibody responses against HIV-1 group M consensus gp120 in rabbits. Retrovirology **11**:125.

11. **Qin Y**, **Banasik M**, **Kim S**, **Penn-Nicholson A**, **Habte HH**, **LaBranche C**, **Montefiori DC**, **Wang C**, **Cho MW**. 2014. Eliciting neutralizing antibodies with gp120 outer domain constructs based on M-group consensus sequence. Virology **462-463**:363–376.

12. **Banerjee S**, **Shi H**, **Banasik M**, **Moon H**, **Lees W**, **Qin Y**, **Harley A**, **Shepherd A**, **Cho MW**. 2017. Evaluation of a novel multi-immunogen vaccine strategy for targeting 4E10/10E8 neutralizing epitopes on HIV-1 gp41 membrane proximal external region. Virology **505**:113–126.

13. **Jardine J**, **Julien J-P**, **Menis S**, **Ota T**, **Kalyuzhniy O**, **McGuire A**, **Sok D**, **Huang P-S**, **MacPherson S**, **Jones M**, **Nieusma T**, **Mathison J**, **Baker D**, **Ward AB**, **Burton DR**, **Stamatatos L**, **Nemazee D**, **Wilson IA**, **Schief WR**. 2013. Rational HIV immunogen design to target specific germline B cell receptors. Science **340**:711–716.

14. **Sanders RW**, **Derking R**, **Cupo A**, **Julien J-P**, **Yasmeen A**, **de Val N**, **Kim HJ**, **Blattner C**, **la Peña de AT**, **Korzun J**, **Golabek M**, **de Los Reyes K**, **Ketas TJ**, **van Gils MJ**, **King CR**, **Wilson IA**, **Ward AB**, **Klasse PJ**, **Moore JP**. 2013. A next-generation cleaved, soluble HIV-1 Env trimer, BG505 SOSIP.664 gp140, expresses multiple epitopes for broadly neutralizing but not non-neutralizing antibodies. PLoS Pathog **9**:e1003618.

CHAPTER 6

A NOVEL SAMPLE INFERENCE METHOD FOR ILLUMINA AMPLICON DATA

Heliang Shi, Xiyu Peng, Karin Dorman

Abstract

Illumina amplicon sequencing is an important and widely used tool for the identification and quantification of species or variants in mixed samples, but the presence of errors makes it a challenging problem. Many denoising algorithms have been developed, but most completely discard the highly informative quality scores or reduce them to summary statistics to reduce data size. In this study, we develop *ampliclust*, a fully probabilistic error modeling approach using uncompressed sequences and quality scores to denoise Illumina amplicon data. Using artificial, real, and simulated data sets, the analyses show that our approach has comparable accuracy on data with well-separated clusters and better accuracy on data with overlapping clusters than DADA2, a popular state-of-the-art denoising tool.

6.1 Introduction

Current Next Generation Sequencing (NGS) technologies provide an enormous volume of data at low cost and allow high resolution analysis of genetic diversity, but the associated error rate is higher than traditional Sanger sequencing [1, 2]. These errors can interfere with the identification and quantification of species [3] or variants [4, 5] in mixed sample studies. The high coverage of NGS enables the detection of many low-frequency species or variants, but the increasing depth also increases the number of error-containing sequences, making it difficult to distinguish genuine minority variants from errors. Thus, NGS data needs be denoised to better inference the genuine genetic diversity [6].

This chapter focuses on error correction methods for amplicon sequencing, where several different approaches have been proposed to minimize the effect of errors on the detection of minority variants. The typical analysis of amplicon sequencing data is to construct OTUs (operational taxonomic units) based on Hamming distance. One unique sequence represents each OTU, and abundance is estimated as the number of reads associating with each OTU. Unfortunately, the OTU method has high false positive and false negative rates, especially when evaluating fine-scale diversity [7, 8]. A variety of approaches have been proposed to improve accuracy, but they have shortcomings, such as slow running speed and dependence on ad hoc parameters, sometimes estimated from required training data [9, 10].

The Hamming distance ignores known complexity in the error process, so some have proposed to improve accuracy by modeling the error process and evaluating the individual reads via a probability model. Recently, the Divisive Amplicon Denoising Algorithm (DADA) [6] and DADA2 [11] combined an error model with heuristic, hierarchical clustering techniques to achieve fast clustering of reads into species/variant clusters without use of OTUs. The error-containing reads assigned to a single cluster are presumed to derive from the same haplotype sequence. The divisive clustering algorithm and inference of the error parameters are performed alternatively until convergence. DADA2 extends and improves the DADA algorithm for Illumina amplicon data by incorporating quality scores in the model. DADA2 has shown to infer sample sequences from Illumina amplicon data with high resolution [11].

However, DADA2 assumes all reads of the same sequence originate from the same amplicon sequence, which can be violated in practice. To compress the data, DADA2 initially groups all amplicon reads by their unique nucleotide sequence and associates each unique sequence with an abundance and average quality score profile. By using only average quality score information, DADA2 underutilizes the information they provide. To illustrate the loss, suppose there are two amplicon sequences ACCTA and AGCAA, and we have observed five copies of (errored) read $s =$ AGCTA as shown in Table 1. Based on the raw quality score information, the most possible assign-

Table 1: Example data demonstrates how DADA2 compression can result in misclassification of reads to clusters.

Read	**Quality Scores**					**Assignment?**
$\boldsymbol{r}_1$	33	12	40	40	40	ACCTA
$\boldsymbol{r}_2$	40	39	38	8	39	AGCAA
$\boldsymbol{r}_3$	35	7	40	40	40	ACCTA
$\boldsymbol{r}_4$	38	12	40	40	40	ACCTA
$\boldsymbol{r}_5$	38	40	35	15	40	AGCAA
$\boldsymbol{s}$	38	12	40	40	40	ACCTA

ments for $\{\boldsymbol{r}_1, \boldsymbol{r}_3, \boldsymbol{r}_4\}$ and $\{\boldsymbol{r}_2, \boldsymbol{r}_5\}$ are ACCTA and AGCAA, respectively. The sites with low quality score are likely misreads of the true amplicon nucleotide. However, the compressed data would force all five reads to be assigned to the ACCTA amplicon sequence. Clearly, the raw, uncompressed data would provide more accurate estimates of the abundance of true variants, especially when there are very similar variant sequences in the original sample.

More technically, the DADA2 model heuristically maximizes the conditional joint distribution for each cluster of the sequence abundances given the total abundance and at least one observation of each sequence. To facilitate the calculations, they assume, using the law of rare events, this conditional distribution (scaled multinomial) is equivalent, up to a constant, to the distribution (scaled Poisson) dropping the condition on the sum. However, the constant of proportionality actually depends on the model parameters, and it is not obvious that the constant is insignificant. Thus, there are several structural flaws in the DADA2 statistical model, in addition to the discarded quality score information.

In this study, we describe *ampliclust*, a fully consistent, model-based approach inspired by DADA2 for the inference of haplotypes and their abundances from Illumina amplicon data using uncompressed quality information. Out method is shown to outperform DADA2 in accuracy on several data sets.

6.2 Methods

6.2.1 Model

Our goal is to identify the number and abundance of unique DNA molecules, *haplotypes*, in a sample. The observed data consist of a set of amplicon reads $\mathcal{R} = \{\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_n\}$ with the same length l and corresponding quality scores $\mathcal{Q} = \{\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_n\}$. Let $\mathcal{H} = \{\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_K\}$ be the collection of *unobserved* haplotypes, with h_{kj} the nucleotide at position j of the amplicon. The object of our analysis is the number of haplotypes K, the identity of haplotypes $\mathcal{H}$ and $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_K)$, their relative abundance in the original sample. Let $\boldsymbol{C} = (C_1, C_2, \cdots, C_n)$ be the unknown class assignment of all reads. The complete data likelihood is

$$L(\boldsymbol{\theta} \mid \mathcal{R}, \mathcal{Q}, \boldsymbol{C}) = \prod_{i=1}^{n} \prod_{k=1}^{K} \left[\pi_k P(\boldsymbol{R}_i = \boldsymbol{r}_i, \boldsymbol{Q}_i = \boldsymbol{q}_i \mid C_i = k)\right]^{\mathbb{1}\{C_i = k\}},$$

where $\boldsymbol{\theta}$ are the model parameters, including $\mathcal{H}$ and $\boldsymbol{\pi}$. We assume pairs (R_{ij}, Q_{ij}) of a read base call in $\boldsymbol{R}_i$ and quality score in $\boldsymbol{Q}_i$ are generated independently as

$$P(\boldsymbol{R}_i = \boldsymbol{r}_i, \boldsymbol{Q}_i = \boldsymbol{q}_i \mid C_i = k) = \prod_{j=1}^{l} P(R_{ij} = r_{ij}, Q_{ij} = q_{ij} \mid C_i = k).$$

We parameterize the joint probability of a single base and quality score as the product of

$$P(R_{ij} = b \mid C_i = k, H_{kj} = h) = \delta_j^{\mathbb{1}\{h=b\}} \left[(1 - \delta_j)\gamma_{hb}\right]^{\mathbb{1}\{h \neq b\}}$$

and

$$P(Q_{ij} = q \mid R_{ij} = b, C_i = k) = \begin{cases} \lambda_{0jq} & h_{kj} = b \\ \lambda_{1jq} & h_{kj} \neq b, \end{cases}$$

where $b, h \in \{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{T}\}$ are bases and $q \in \mathbb{Q}$ is a discrete quality score in the set $\mathbb{Q}$, which depends on the NGS chemistry. Here, δ_j is the probability the observed

nucleotide at read position j is generated from the haplotype without error and γ_{hb} is the probability of misreading haplotype nucleotide h as read nucleotide $b \neq h$ when there is an error. Both δ_j and $\lambda_{\cdot jq}$ currently assume a read position-specific effect.

The complete data likelihood becomes

$$L(\boldsymbol{\theta} \mid \mathcal{R}, \mathcal{Q}, \boldsymbol{C}) = \prod_{i=1}^{n} \prod_{k=1}^{K} \left\{ \pi_k \prod_{j=1}^{l} \left[\delta_j \lambda_{0jq_{ij}}\right]^{\mathbb{1}\{h_{kj}=r_{ij}\}} \left[(1-\delta_j)\gamma_{h_{kj}r_{ij}} \lambda_{1jq_{ij}}\right]^{\mathbb{1}\{h_{kj}\neq r_{ij}\}} \right\}^{\mathbb{1}\{C_i=k\}},$$

for parameters $\boldsymbol{\theta} = \{\boldsymbol{\delta}, \Gamma, \Lambda, \boldsymbol{\pi}, \mathcal{H}\}$, where

$$\begin{aligned} \boldsymbol{\delta} &= (\delta_1, \delta_2, \ldots, \delta_l) & \Gamma &= \{\gamma_{hb} : b \neq h \in \{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{T}\}\} \\ \boldsymbol{\pi} &= (\pi_1, \pi_2, \ldots, \pi_K) & \Lambda &= \{\lambda_{0jq}, \lambda_{1jq} : 1 \leq j \leq l, q \in \mathbb{Q}\} \\ \mathcal{H} &= (\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_K). \end{aligned}$$

We sum over the hidden $\mathcal{C}$ in the complete data likelihood to obtain the observed data likelihood as

$$L(\boldsymbol{\theta} \mid \mathcal{R}, \mathcal{Q}) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k \prod_{j=1}^{l} \left[\delta_j \lambda_{0jq_{ij}}\right]^{\mathbb{1}\left\{h_{kj}=r_{ij}\right\}} \left[(1-\delta_j)\gamma_{h_{kj}r_{ij}} \lambda_{1jq_{ij}}\right]^{\mathbb{1}\left\{h_{kj}\neq r_{ij}\right\}}.$$

6.2.2 Inference

We will apply the alternating expectation-conditional maximization (AECM) algorithm [12] to update $\boldsymbol{\theta}$. The idea of the AECM algorithm is to partition $\boldsymbol{\theta}$ into $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ when it is easy to maximize the likelihood for $\boldsymbol{\theta}_1$ given $\boldsymbol{\theta}_2$ and vice versa. In this case, let $\boldsymbol{\theta}_1 = \mathcal{H}$ and $\boldsymbol{\theta}_2 = \{\boldsymbol{\delta}, \Gamma, \Lambda, \boldsymbol{\pi}\}$. The AECM consists of two cycles, each containing an E-step and a conditional maximization (CM)-step. These two cycles will be alternated until convergence. Details of the steps are given below.

In the first E-step, the conditional expected complete data log likelihood at iteration $t+1$ is

$$
\begin{aligned}
Q_1(\mathcal{H}; \mathcal{H}^{(t)}, \boldsymbol{\theta}_2^{(t)}) &= \mathbb{E}\left[l(\mathcal{H}, \boldsymbol{\theta}_2^{(t)} \mid \mathcal{R}, \mathcal{Q}, \mathcal{C}) \mid \mathcal{R}, \mathcal{Q}, \mathcal{H}^{(t)}, \boldsymbol{\theta}_2^{(t)} \right] \\
&= \sum_{k=1}^{K} \sum_{i=1}^{n} P(C_i = k \mid \mathcal{R}, \mathcal{Q}, \mathcal{H}^{(t)}, \boldsymbol{\theta}_2^{(t)}) \Bigg\{ \ln \pi_k^{(t)} \\
&\quad + \sum_{j=1}^{l} \left(\mathbb{1}\left\{ h_{kj}^{(t)} = r_{ij} \right\} \ln \left[\delta_j^{(t)} \lambda_{0jq_{ij}}^{(t)} \right] + \mathbb{1}\left\{ h_{kj}^{(t)} \neq r_{ij} \right\} \ln \left[(1-\delta_j^{(t)}) \gamma_{h_{kj} r_{ij}}^{(t)} \lambda_{1jq_{ij}}^{(t)} \right] \right) \Bigg\},
\end{aligned}
$$

which requires we compute

$$
\begin{aligned}
e_{ik}^{(t)} &:= P(C_i = k \mid \mathcal{R}, \mathcal{Q}; \mathcal{H}^{(t)}, \boldsymbol{\theta}_2^{(t)}) \\
&\propto P(\boldsymbol{R} = \boldsymbol{r}_i, \boldsymbol{Q}_i = \boldsymbol{q}_i \mid C_i = k; \mathcal{H}^{(t)}, \boldsymbol{\theta}_2^{(t)}) P(C_i = k; \mathcal{H}^{(t)}, \boldsymbol{\theta}_2^{(t)}) \\
&= \pi_k^{(t)} \prod_{j=1}^{l} \left[\delta_j^{(t)} \lambda_{0jq_{ij}}^{(t)} \right]^{\mathbb{1}\left\{ h_{kj}^{(t)} = r_{ij} \right\}} \left[(1-\delta_j^{(t)}) \gamma_{h_{kj}^{(t)} r_{ij}}^{(t)} \lambda_{1jq_{ij}}^{(t)} \right]^{\mathbb{1}\left\{ h_{kj}^{(t)} \neq r_{ij} \right\}}.
\end{aligned}
$$

In the first CM-step, we maximize $Q_1(\mathcal{H}; \mathcal{H}^{(t)}, \boldsymbol{\theta}_2^{(t)})$ for $\mathcal{H}$ given $\boldsymbol{\theta}_2^{(t)} = \left\{ \boldsymbol{\delta}^{(t)}, \ .\Gamma^{(t)}, \Lambda^{(t)}, \boldsymbol{\pi}^{(t)} \right\}$. The maximization of this likelihood yields

$$
h_{kj}^{(t+1)} = \underset{h \in \{\mathsf{A},\mathsf{C},\mathsf{G},\mathsf{T}\}}{\arg\max} \sum_{i=1}^{n} e_{ik}^{(t)} \left(\mathbb{1}\left\{ h = r_{ij} \right\} \ln \left[\delta_j^{(t)} \lambda_{0jq_{ij}}^{(t)} \right] + \mathbb{1}\left\{ h \neq r_{ij} \right\} \ln \left[(1-\delta_j^{(t)}) \gamma_{h r_{ij}}^{(t)} \lambda_{1jq_{ij}}^{(t)} \right] \right)
$$

for all $k \in \{1, 2, \ldots, K\}$ and $j \in \{1, 2, \ldots, l\}$.

In the second E-step, the calculations are virtually identical. We need the same conditional expectation, now viewed as a function of $\boldsymbol{\theta}_2$ and conditioning on updated $\mathcal{H}^{(t+1)}$. Specifically,

$$
Q_2(\boldsymbol{\theta}_2; \mathcal{H}^{(t+1)}, \boldsymbol{\theta}_2^{(t)}) = \mathbb{E}\left[l(\mathcal{H}^{(t+1)}, \boldsymbol{\theta}_2 \mid \mathcal{R}, \mathcal{Q}, \mathcal{C}) \mid \mathcal{R}, \mathcal{Q}, \mathcal{H}^{(t+1)}, \boldsymbol{\theta}_2^{(t)} \right],
$$

which requires

$$
\begin{aligned}
e_{ik}^{(t+0.5)} &:= P(C_i = k \mid \mathcal{R}, \mathcal{Q}; \mathcal{H}^{(t+1)}, \boldsymbol{\theta}_2^{(t)}) \\
&\propto P(\boldsymbol{R}_i = \boldsymbol{r}_i, \boldsymbol{Q}_i = \boldsymbol{q}_i \mid C_i = k; \mathcal{H}^{(t+1)}, \boldsymbol{\theta}_2^{(t)}) P(C_i = k; \mathcal{H}^{(t+1)}, \boldsymbol{\theta}_2^{(t)}) \\
&= \pi_k^{(t)} \prod_{j=1}^{l} \left[\delta_j^{(t)} \lambda_{0jq_{ij}}^{(t)} \right]^{\mathbb{1}\left\{h_{kj}^{(t+1)} = r_{ij}\right\}} \left[(1-\delta_j^{(t)}) \gamma^{(t)}_{h_{kj}^{(t+1)} r_{ij}} \lambda_{1jq_{ij}}^{(t)} \right]^{\mathbb{1}\left\{h_{kj}^{(t+1)} \neq r_{ij}\right\}}.
\end{aligned}
$$

In the second CM-step, we need to maximize $Q(\boldsymbol{\theta}_2;\ \mathcal{H}^{(t+1)}, \boldsymbol{\theta}_2^{(t)})$ for $\boldsymbol{\theta}_2 = \{\boldsymbol{\delta}\ , \Gamma, \Lambda, \boldsymbol{\pi}\}$ given $\mathcal{H}^{(t+1)}$ The update equations are

$$
\begin{aligned}
\delta_j^{(t+1)} &= \frac{\sum_{i=1}^{n} \sum_{k=1}^{K} e_{ik}^{(t+0.5)} \mathbb{1}\left\{h_{kj}^{(t+1)} = r_{ij}\right\}}{n} \\
\gamma_{hb}^{(t+1)} &= \frac{\sum_{i=1}^{n} \sum_{k=1}^{K} e_{ik}^{(t+0.5)} \sum_{j=1}^{l} \mathbb{1}\left\{h_{kj}^{(t+1)} = h, r_{ij} = b\right\}}{\sum_{i=1}^{n} \sum_{k=1}^{K} e_{ik}^{(t+0.5)} \sum_{j=1}^{l} \mathbb{1}\{h_{kj}^{(t+1)} = h\}} \\
\lambda_{0jq}^{(t+1)} &= \frac{\sum_{i=1}^{n} \sum_{k=1}^{K} e_{ik}^{(t+0.5)} \mathbb{1}\left\{h_{kj}^{(t+1)} = r_{ij}, q_{ij} = q\right\}}{\sum_{i=1}^{n} \sum_{k=1}^{K} e_{ik}^{(t+0.5)} \mathbb{1}\left\{h_{kj}^{(t+1)} = r_{ij}\right\}} \\
\lambda_{1jq}^{(t+1)} &= \frac{\sum_{i=1}^{n} \sum_{k=1}^{K} e_{ik}^{(t+0.5)} \mathbb{1}\left\{h_{kj}^{(t+1)} \neq r_{ij}, q_{ij} = q\right\}}{\sum_{i=1}^{n} \sum_{k=1}^{K} e_{ik}^{(t+0.5)} \mathbb{1}\left\{h_{kj}^{(t+1)} \neq r_{ij}\right\}} \\
\pi_k^{(t+1)} &= \frac{\sum_{i=1}^{n} e_{ik}^{(t+0.5)}}{n}
\end{aligned}
$$

for all $b \neq h \in \mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{T}$ and $q \in \mathbb{Q}$.

6.2.3 Data

We evaluated our model using two real datasets, a mock 16S mixture and an HIV-1 sample, as well as simulated data.

6.2.3.1 Mock data

We compare our method and DADA2 on the Extreme mock dataset used to validate DADA2 [11] (Table 2). This dataset contains 2,040,085 reads of an unbalanced mixture of 27 bacterial species/strains, 24 of which are distinguishable in the amplicon

Table 2: **Extreme data.** Dilution levels, number of species or strains at each dilution, and expected proportion of reads at each dilution level.

Dilution	No.	Proportion
10^0	2	0.4265302
10^{-1}	3	0.0426530
10^{-2}	4	0.0042653
10^{-3}	4	0.0004265
10^{-4}	4	0.0000427
10^{-5}	10	0.0000043

region. The authors have provided 61 reference sequences, including various strains for these species. Thirty-three, including three distinct strains each for *B. ovatus* and *B. vulgatus*, are distinct in the amplicon region. To eliminate contaminant reads, the log likelihood of generating each read from each of the 33 reference sequences was computed assuming all substitutions are equally likely and the quality score is the true error probability. Reads with maximum log likelihoods across all reference strains smaller than -10 were removed. Finally, 1,641,443 reads remained (80.4$\%$), and 10 random subsets of 3,000 reads were extracted for detailed analysis.

6.2.3.2 HIV data

We also apply both methods to a real data set of amplicon reads from a single HIV-1-infected patient. The data are 2×300 paired-end reads of the V1–V3 envelope region, generated on an Illumina MiSeq [13]. Specifically, we use the reverse read data because they contain a 9-nucleotide barcode, which may be useful to assess the quality of clustering. These data were processed as described in Appendix $\S$, resulting in a dataset of size 10,344 reads.

6.2.3.3 Simulated data

We also applied the methods to simulated data. We simulate data under a model that matches neither DADA2 nor our own model. Specifically, given haplotypes $\mathcal{H}$ and a clustering $\mathcal{C}$, we simulate pairs (r_{ij}, q_{ij}) of read nucleotides and quality scores

independently as

$$P(R_{ij} = r_{ij}, Q_{ij} = q_{ij}) = \lambda_{jq_{ij}} \mathbb{1}\{C_i = k\} \times \begin{cases} \left(1 - 10^{-q_{ij}/10}\right) & r_{ij} = h_{kj} \\ 10^{-q_{ij}/10} \gamma_{h_{kj} r_{ij}} & r_{ij} \neq h_{kj}. \end{cases}$$

The parameters are estimated from existing data given haplotypes $\mathcal{H}$ and clustering $\mathcal{C}$ as

$$\begin{aligned} \hat{\lambda}_{jq} &= \frac{\sum_{i=1}^{n} \mathbb{1}\{Q_{ij} = q\}}{n} \\ \hat{\gamma}_{hb} &= \frac{\sum_{i=1}^{n} \sum_{j=1}^{l} \sum_{k=1}^{K} \mathbb{1}\{C_i = k, H_{kj} = h, r_{ij} = b\}}{\sum_{i=1}^{n} \sum_{j=1}^{l} \sum_{k=1}^{K} \mathbb{1}\{C_i = k, H_{kj} = h\}}, \end{aligned}$$

where n is the number of sequences in the dataset. We simulate under two settings of λ_{jq} and γ_{hb}, one obtained from the first $n = 3{,}000$ subset of the mock data, and another obtained from the $n = 10{,}344$ HIV-1 reads. For the mock data, we used the hard clustering from the $K = 12$ seeded solution (see §); for the HIV data, we used the $K = 8$ solution.

To control the difficulty of clustering, we also simulate data where we alter the haplotypes to control the amount of separation between the clusters. In this case, the haplotypes are simulated independently from an "ancestral haplotype" on a star-shaped phylogenetic tree, where all branches have the same length given as the expected number of nucleotide changes. When the expected number of changes η is small, the expected number of changes between haplotypes should be 2η. To match the simulated nucleotide content to the existing real datasets, we used the amplicon region of accession KF99671 (see Appendix §) as the ancestral haplotype for the HIV-1 simulation and the consensus sequence of the 33 reference sequences for the mock simulation. In addition to simulations that used the inferred haplotypes, we simulated data using two branch lengths (0.005 and 0.001) for the mock data and four branch lengths (0.0025, 0.005, 0.01, and 0.015) for the HIV-1 data. Finally, all simulated datasets were the same size as the original real datasets prepared in § and §.

6.3 Running the methods

6.3.1 Initialization and convergence

The choice of initial values is of great importance in EM-based algorithms [14]. It can influence the convergence rate and the algorithm's ability to locate the global optimum. Given K clusters, we obtain initial guesses for the indicators $\boldsymbol{C} = (C_1, C_2, \cdots, C_n)$ using k-modes [15], with the Hamming distance as the distance metric between reads. The kth mode is used to initialize the kth haplotype $\boldsymbol{H}_k^{(0)}$, and the size of the corresponding cluster initializes $\pi_k^{(0)}$. We estimate the remaining parameters in $\boldsymbol{\theta}_2^{(0)}$ using the update equations from the second CM-step of AECM, where $e_{ik}^{(0.5)} = 1$ if the ith read is assigned to cluster k and 0 otherwise. To initialize k-modes, we implement several different methods. In Rnd-Ini(I) initialization, we select K distinct reads as the initial modes, run k-modes for no more than 100 iterations, and select the best of I such random initializations as judged by the k-modes criterion to initialize AECM. In Rnd-EM(I, J) initialization [16, 17], we select K distinct reads as the initial modes, run k-modes for no more than 100 iterations followed by AECM for exactly J iterations, and selected the best of I such random initializations as judged by the log likelihood achieved after J iterations to initialize AECM.

We repeat the chosen initialization, possibly multiple times, followed by AECM iteration until the rate of change in the log likelihood value between iterations is small. Specifically, let $\epsilon_{t+1} = \frac{l\left(\boldsymbol{\theta}^{(t+1)}|\mathcal{R},\mathcal{Q}\right) - l\left(\boldsymbol{\theta}^{(t)}|\mathcal{R},\mathcal{Q}\right)}{l\left(\boldsymbol{\theta}^{(t)}|\mathcal{R},\mathcal{Q}\right)}$ be the relative change in the log likelihood at iteration $t+1$, where $\boldsymbol{\theta}^{(t)}$ is the parameter estimate after the tth iteration. We stop the AECM iterations when $\epsilon < 1 \times 10^{-6}$. The estimates yielding the maximum log likelihood across initializations are considered the MLEs.

To maximize the *ampliclust* model on the mock datasets, we took advantage of the known reference sequences to avoid the many random initializations required to find the global maximum. Instead, we provide *ampliclust* with K reference haplotypes and then partition the reads by assigning each read to the most likely haplotype sequence. A read likelihood is computed for each possible haplotype by interpreting quality scores

as the probability of an error in the read nucleotide and assuming all substitutions are equally likely, *i.e.* $10^{-q/10}$ is the probability of error for observed quality score q, and $\gamma_{hb} = \frac{1}{3}$ for all $b \neq h$. We call such an initialization of the AECM the "seeded initialization." We may also seed with both hapotypes *and* a partition, in which case the cluster assignment step is not needed. Given the initial haplotypes and partition, the parameters are initialized as described in §. For each mock dataset, we first analyze it with $K = 33$, providing all 33 reference sequences and without updating the haplotypes during AECM. The resulting clustering is referred to later as the $K33^*$ solution. To select K, we first drop all references with empty clusters in the $K33^*$ solution. Then, iteratively, we remove the reference sequence with the smallest estimated abundance and seed initialize with the remaining reference sequences, now allowing haplotype updating in the first M step. If more than one reference sequence has the smallest abundance, each is removed in turn, and the solution achieving the highest log likelihood identifies the next reference to be discarded. The final chosen K is given by the solution that achieves the minimum Bayesian information criterion (BIC).

For the analysis of the HIV-1 data, 100 Rnd-Ini(1) initializations were performed followed by an additional 0–700 Rnd-Ini(10) initializations. For some K, we also tried 100 Rnd-EM(10) initializations. More initializations were run when the solution included null clusters (clusters with no reads assigned to it in the hard clustering) or the maximum achieved log likelihood was lower than that of a solution with smaller K. The solution with the maximum log likelihood at a given K is retained as the final solution for that K.

6.3.2 Details

Implementation. We have implemented the proposed AECM algorithm in C under the C11 standard. We call our software and method *ampliclust* in this chapter.

Filtering and trimming. Unlike the recommended DADA2 pipeline, we did not filter or trim the reads for low quality data. The intent is to handle such noise via the model.

We did, as described in Methods §, strongly filter the data for contaminants or non-amplicon reads, since neither model accounts for such reads in the data.

Running DADA2. Except for not filtering or trimming the reads, DADA2 is run per published instructions in the DADA2 software manual [11].

Clustering solution. DADA2 produces a hard cluster directly. To obtain a hard clustering from *ampliclust*, we compute the posterior probability that a read belongs to each cluster and assign the read to the cluster with the highest posterior probability.

6.4 Results

To measure the performance of our model relative to DADA2, we apply both methods to mock, real, and simulated data sets.

6.4.1 Mock data

We analyzed 10 random subsets of the mock data, each of size $n =$ 3,000 reads. With only 3,000 reads, we expect to detect 9 of the 24 species/strains mixed in the mock sample, but we might also detect distinct strains within the more common species or one of the more highly diluted species.

While DADA2 estimates the value of K automatically, our method requires the user to provide K. We use BIC as criteria for choosing K. The best choice of K for each dataset is marked red in Fig. 1. All the predicted haplotypes from both DADA2 and *ampliclust* at the chosen K perfectly match a subset of the reference sequences. Table 3 shows the average and range of K found by both methods across the 10 datasets. On the right, we report the number of the original 24 species/strains mixed in the mock sample found by each method. There appears to be no significant difference in the number of haplotypes found by the two methods or the number of species/strains recovered.

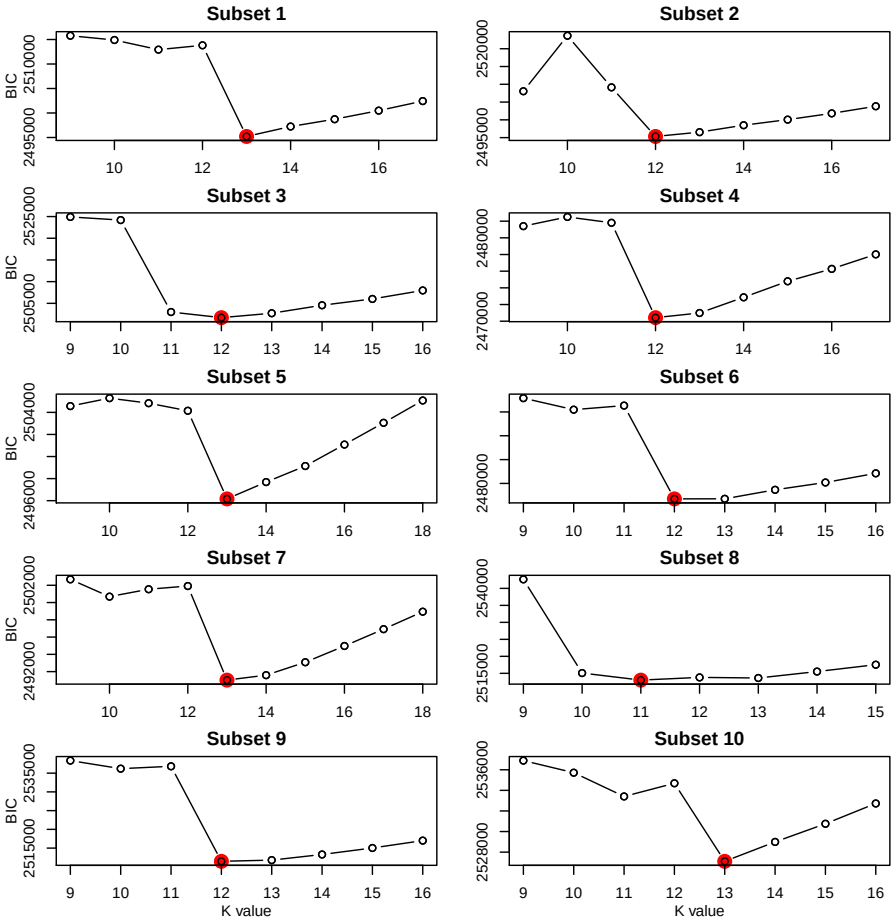

Figure 1: Plots of BIC against the number of clusters for 10 subsets of the mock data. The K yielding the smallest BIC value is marked red.

Table 3: Comparison of the inferred haplotypes and identified species/strains for *ampliclust* and DADA2 on the mock data. The mean over 10 subsets of the data is shown, with the standard error in parentheses.

	Haplotypes			Species/Strains		
Method	**Min.**	**Mean**	**Max.**	**Min.**	**Mean**	**Max.**
DADA2	11	12.4 (0.221)	13	8	9.4 (0.221)	10
ampliclust	11	12.3 (0.213)	13	8	9.3 (0.213)	10

Table 4: The average Adjusted RAND indices comparing the clustering results of DADA2, *ampliclust*, and $K33^*$ across 10 subsets of the mock data. Standard errors in parentheses.

	Ampliclust	**K33***
DADA2	0.993 (0.001)	0.988 (0.001)
Ampliclust	–	0.990 (0.001)

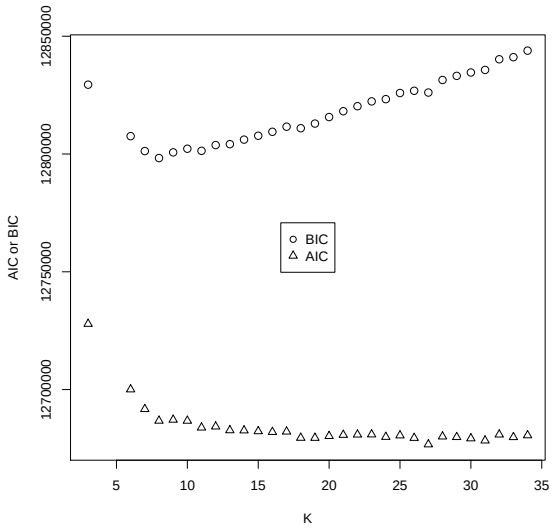

Figure 2: AIC and BIC computed for HIV data analyzed with *ampliclust* using $K = 2, 4, 5, \ldots, 34$.

We further compare the solutions estimated by our methodology and DADA2 by examining the number of reads assigned to each reference sequence (Table 5). In general, DADA2 and *ampliclust* provided very similar clustering results, which is consistent with the adjusted RAND indices [18] (Table 4). Two strains of *P. distasonis*, differing by a single nucleotide, were mixed at highly disparate dilutions in the sample, and both DADA2 and *ampliclust* sometimes misidentify some strain JCM 13401 reads as originating from strain JCM 13400. Both methods also detect distinct strains of the *B. ovatus* and *B. vulgatus* species. The $K33^*$ solutions detected several highly diluted (dilutions 10^{-3}, 10^{-4}, and 10^{-5}) reference species, including the JCM 13400 strain. Other than JCM 13400, these may be true reads that were generally not detected by DADA2 nor *ampliclust*, although DADA2 predicted the existence of *R. inulinivorans*, in subset 5. Both methods do reliably identify 13 common species/strains, including the misidentified *P. distasonis* JCM 13400, in near equivalent proportions.

We conclude there is no reproducible difference in the performance of DADA2 and *ampliclust* on the mock datasets. Both methods perform very well at identifying the major species/strains in the sample and have very low rates of misclassification.

Table 5: The dilution (Dil.), expected number of reads (Exp.), and average number of reads assigned to each strain in the $K33^*$, DADA2, and *ampliclust* solutions. The mock mixture was formed by applying dilution 10^k for the integer k given in the column to these species (and others not shown). The true proportions of each strain from species *B. ovatus* and *B. vulgatus* are not known. The expected number of reads is based on the dilution factor and assuming perfect experimental methods with no bias or contamination. $K33^*$ is the $K = 33$ solution with fixed haplotypes obtained as described in the Methods. DADA2 is the DADA2 solution. *Ampliclust* is our solution seeded with reference sequences chosen as described in the Methods.

Species	**Dil.**	**Exp.**	$K33^*$	**DADA2**	***Ampliclust***
Bacteroides ovatus 1	0	1279.6	282.3	281.6	283.9
Bacteroides ovatus 2			1.1	0	0
Bacteroides ovatus 3			1117.8	1118.8	1115.9
Bacteroides ovatus total			1401.2	1400.4	1399.8
Bacteroides vulgatus 1	0	1279.6	839.6	842.3	844.1
Bacteroides vulgatus 2			170.3	163.7	172.9
Bacteroides vulgatus 3			144.6	151	140.1
Bacteroides vulgatus total			1154.5	1157	1157.1
Bacteroides cellulosilyticus DSM 14838	-1	128	136.5	136.2	139.5
Clostridium xylanovorans	-1	128	128.7	131.4	133.6
Parabacteroides distasonis JCM 13401	-1	128	58.7	52.9	58.6
Parabacteroides distasonis JCM 13400	-5	0	65.7	73.6	67.9
Bacteroides uniformis	-2	12.8	13.8	13.2	9.9
Parabacteroides merdae	-2	12.8	15.9	15.5	15.5
Clostridium cocleatum	-2	12.8	10.3	9.6	10.1
Coprococcus comes ATCC 27758	-2	12.8	10.7	10	8
Bacteroides fragilis	-3	1.3	0.9	0	0
Clostridium celatum JCM 1394	-3	1.3	0.9	0	0
Roseburia inulinivorans DSM 16841	-3	1.3	0.8	0.2	0
Bacteroides thetaiotaomicron DSM 2079	-3	1.3	1.2	0	0
Bacteroides massiliensis JCM 12982	-4	0.1	0.1	0	0
Ruminococcus gnavus ATCC 29149	-5	0	0.1	0	0

```
D3          GCCTGCCCCC TGTGCCGT 166
A7          .......... ........ 159
D5          A.T.A.TA.. ..C..... 132
A5          A.T.A.TA.. ........ 149
D1          ATTCAT.GTT .....TA. 4132
A4          ATTCAT.GTT .....TA. 2185
D6          .TTCA..GTT .....TA. 549
A3          .TTCAT.GTT .....TA. 2567
D2          .TTCA..AT. GA..TT.C 1218
A6          .TTCA..AT. GA..TT.C 1217
D4          ATTCA..AT. GA..TT.C 1136
A2          ATTCA..AT. GA..TT.C 1133
A0          .TTCA..AT. G..TTT.C 1416
D0          ATTCA..AT. G..TTT.C 3011
A1          ATTCA..AT. G..TTT.C 1518
```

Figure 3: Alignment showing only the segregating sites of *ampliclust* inferred haplotypes (A0–A7) and DADA2 inferred centers (D0–D6). After each sequence is the number of reads assigned to each haplotype.

6.4.2 HIV data analysis

The performance of *ampliclust* and DADA2 were further evaluated by analyzing the real HIV-1 data. DADA2 estimated $K = 7$, while our method found $K = 8$ is the best choice based on the BIC criteria (Fig. 2).

Fig. 3 is an alignment of the DADA2 and *ampliclust* inferred haplotypes, showing only the segregating sites. The methods agree on five haplotypes and disagree by a single nucleotide difference on two others. In addition, *ampliclust* identifies an eighth haplotype A0 that differs at one position from D0, which is itself equivalent to A1. This difference is not a trivial disagreement, as *ampliclust* assigns 1,416 reads to this distinct haplotype using the maximum posterior probability. The 3,011 reads DADA2 assigns to D0 are split largely between A0 and A1, but a total of 160 reads are disbursed elsewhere, with at least one read going to *each* other *ampliclust* cluster, and the two *ampliclust* haplotypes, A0 and A1, attract 83 additional reads from DADA2 clusters D3, D5, and D7. One of the other haplotype mismatches (A3 vs. D6) leads to a partitioning of the reads assigned to D6, largely to A3 and A4, but two reads each go to A2 and A5. All together the adjusted RAND index comparing the two solutions is 0.556, showing substantial disagreement between the two methods.

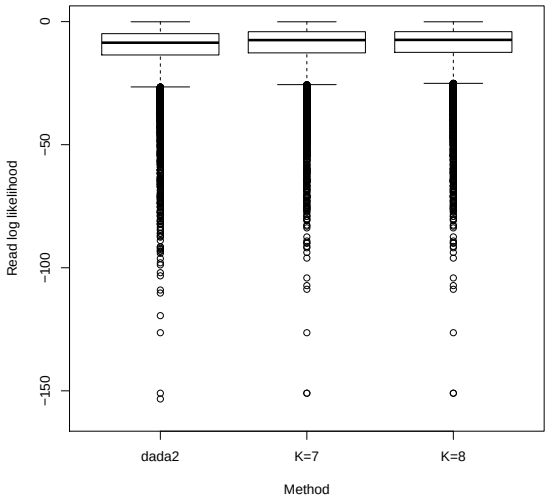

Figure 4: Box plots of the 10,344 read log likelihoods computed under the DADA2 or *ampliclust*, either $K = 7$ or $K = 8$, clustering solution. The y-axis is truncated below at -160, and the log likelihood of one very unlikely read (log likelihood around -300) is thus not shown.

We do not know the true clusters in the HIV dataset, but we can compute the log likelihood of each read assuming the cluster solutions provided by DADA2 or *ampliclust* and assuming all substitutions are equally likely and the quality scores can be treated literally. Quality scores are quite accurate representations of the true error probability (Dorman, personal communication), and although substitutions are not uniform, this modeling choice does not preferentially match either of the competing methods. Figure 4 shows that the read log likelihoods are slightly improved under the *ampliclust* clustering solution, and the difference is not merely attributable to the increase in K from 7 to 8. The median log likelihood increases from -8.559 for the DADA2 clustering solution to -7.506 ($K = 7$) or -7.415 ($K = 8$) under the *ampliclust* solution.

We can also assess the quality of the *ampliclust* solution by considering the barcodes included in the read data. The adapter and barcodes were trimmed before the reads were clustered (see Appendix §), so the barcodes represent separate sequence information that was not used by either DADA2 or *ampliclust* to cluster the reads. Reads with the same barcode are supposed to be replicate reads of the same molecule. Unfortunately, the barcodes themselves are observed with error, and there

is evidence of recombination during PCR that disrupted the connection between the barcode and the amplicon sequence. Nevertheless, a preliminary analysis showed that reads with the same barcode tended to be more similar than reads with different barcodes (data not shown). Thus, there remains valid clustering information in the barcodes, which we use to test whether the *ampliclust* solution is reasonable. In particular, we examine the DADA2 D0 cluster, which is split largely into *ampliclust* clusters A0 and A1. Of the 3,011 reads in cluster D0, 1,366 went into cluster A0. Of the barcodes associated with these 1,366 reads, only 27 were also associated with reads left in cluster A1. Given the distribution of the 639 barcodes in D0 to start, this event is incredibly unlikely. In 1,000 random selections of 1,366 reads from D1, the *minimum* number of such split barcodes was 125. Similarly, when 1,982 reads from cluster D1 were moved into A3, only 96 barcodes were split between A3 and A4. The minimum number of split barcodes in 1,000 random selections of 1,982 reads from D1 is 183. Thus, we conclude that the *ampliclust* solution partitions DADA2 clusters in a way that respects the clustering of barcodes.

6.4.3 Simulated data analysis

In addition, the two methods were compared on simulated datasets where the true clustering solution was known. We simulated data similar to both the mock data (subset 1) and the HIV-1 data (see Methods §). To compare the methods, we computed the adjusted RAND index [18] between the estimated clustering solution and the true clustering. The ARI means and standard errors over three replications for the mock data and 15 replications for the HIV-1 data are reported in Table 6. We report two *ampliclust* results, one obtained from just 10 Rnd-Ini(10) initializations, and another obtained by initializing with the true haplotypes and partition.

The mean ARI when initializing with the truth is higher than the mean ARI of the random initialization, indicating that AECM is generally not finding the global maximum of the log likelihood in just 10 initializations. Nevertheless, *ampliclust* was better able to recover the true clustering even without finding the global maximum under nearly all

Table 6: Adjusted RAND index (ARI) for simulation results. Separation is the expected number of nucleotide substitutions along the branch separating the haplotypes from the ancestral haplotype. The expected number of differences between each haplotype is roughly twice this number. Rows with – given for separation simulated from the inferred haplotypes for the respective dataset. DADA2 is the DADA2 solution. *Ampliclust* is the *ampliclust* solution using 10 random Rnd-Ini(10) initializations. *Ampliclust** is the *ampliclust* solution using the true haplotypes *and* cluster assignments for initialization. The standard error of ARI is shown in parentheses.

		Adjusted RAND Index		
Based on Data	**Separation**	**DADA2**	***Ampliclust***	***Ampliclust****
Mock data (K = 12)	0.005	0.805 (0.184)	0.842 (0.183)	0.853 (0.187)
	0.01	0.959 (0.034)	0.959 (0.028)	0.975 (0.032)
	-	0.996 (0.002)	0.918 (0.070)	0.997 (0.002)
HIV data (K = 8)	0.0025	0.295 (0.257)	0.627 (0.302)	0.645 (0.297)
	0.005	0.517 (0.267)	0.724 (0.223)	0.760 (0.226)
	-	0.723 (0.076)	0.987 (0.034)	0.997 (0.001)
	0.01	0.970 (0.048)	0.984 (0.027)	0.998 (0.005)
	0.015	0.978 (0.069)	0.990 (0.010)	1.000 (0.001)

simulation settings, but especially for overlapping clusters where the haplotypes were not well-separated. Notably, for data simulated directly from 13 reference sequences of the mock dataset, DADA2 performed better than the randomly initialized *ampliclust*. This simulation produced the easiest data to cluster among all simulations, as indicated by overall ARI levels, suggesting that the clusters in the mock data are very well-separated. Indeed, the minimum proportion of mutated sites between any pair of haplotypes is 0.004, the average is 0.21 and the maximum is 0.60, which is large because of insertions/deletions in 16S. Some real datasets, such as that observed in the HIV-1 sample, are far more difficult to cluster.

The haplotypes estimated from the $K = 8$ HIV-1 solution differ at a fraction of sites between 0.004 and 0.04. This amount of separation appears to confuse DADA2, as the average ARI is only 0.723, yielding significantly worse recovery of the true clusters than either *ampliclust* solution. Both methods suffer as the average amount of separation drops below this level, but *ampliclust* continues to perform significantly better. As the separation increases, both methods improve and when the *ampliclust* solution is not properly optimized, DADA2 can outperform *ampliclust*.

6.5 Discussion

In this study, we introduce *ampliclust*, a fully probabilistic model for denoising Illumina amplicon data with the goal of identifying true sequences and their abundances in mixed samples, such as microbiome samples sequenced in 16S. This method is inspired by DADA2 [11], but it assumes each read is generated independently and uses raw quality score information rather than averages on compressed data. It also corrects some logical inconsistencies in the DADA2 model formulation. We utilize the AECM algorithm to estimate model parameters and implement it in the C programming language. Our method *ampliclust* is shown to better recover the true sequences and abundances when the sequences are less well-separated.

In the analysis of mock data where various species/strains of bacteria were mixed in known proportions, the *ampliclust* solutions are highly similar to those of DADA2. The adjusted RAND index (ARI) between the *ampliclust* and DADA2 partitions are around 0.99. Simulations reveal that the mock data represent well-separated clusters, a data situation where both methods perform quite well. In fact, DADA2 can outperform *ampliclust* in this situation because the latter tends to return local maxima. Indeed, we conclude that DADA2 is a very efficient algorithm for finding well-separated clusters, much better AECM with random initialization.

Different clustering results are observed for *ampliclust* and DADA2 on the real HIV dataset, and the *ampliclust* solution has some characteristics that suggest it may be a better solution. Although *ampliclust* and DADA2 identify five identical haplotypes and two that differ by a single nucleotide, their estimated abundances are different and the ARI comparing the solutions is small. The read likelihoods evaluated under a simple error model that requires no parameter estimation are slightly better for the *ampliclust* clustering compared to that of DADA2, suggesting that *ampliclust* may better partition the data. In addition, the *ampliclust* solution splits two of the DADA2 clusters in a way that respects the barcodes attached to the amplicon sequences better than random splitting. Thus, there is limited evidence that the *ampliclust* clustering solution for the

HIV-1 data is better than the DADA2 solution, but there is no conclusive support for *ampliclust* in the absence of a true clustering.

In the simulation study, *ampliclust* outperforms DADA2 even after just 10 RndIni(10) initializations in almost all simulation scenarios, but especially for data with clusters that are not well-separated. The simulation study reveals that the clusters in the mock dataset are well-separated, and considerably more separated than the clusters in the HIV-1 dataset. Apparently, there is little ability to distinguish the accuracy of clustering performance for such well-separated datasets. The major distinguishing feature of the two methods is speed, where DADA2 excels.

The difference in the accuracy of *ampliclust* and DADA2 could be attributed to DADA2's underutilization of the quality scores. Compressing the data to unique sequences and average quality scores accelerates DADA2, but distorts the error signal and obscures read-level information that can be used for better cluster placement. Furthermore, DADA2 uses a greedy algorithm for estimation, so it is not guaranteed to find a globally optimal solution. *Ampliclust* may also get trapped in local optima, but repeated random initializations can help it find a better optimum than DADA2 at the cost of a longer run time. During the initial stage of the DADA2 algorithm, all reads are combined in a single cluster, and the error model estimated at this stage necessarily overestimates error probabilities. DADA2 can nevertheless detect separate clusters if there is a replicated read that is "unusual" in this large cluster, but when errors are common (as is routine for long reads of diverse samples), there may be few true haplotypes read without error. Indeed, the HIV dataset consisted of 8,631 unique sequences out of 10,344 reads, so relatively few sequences were replicated. Further, true clusters that start absorbed in a larger cluster may not appear unusual if the error rates are severely overestimated, especially when there is little separation between haplotypes. For both these reasons, DADA2 may have difficulty forking off clusters. Indeed in the HIV-1 real data, DADA2 found one less cluster than *ampliclust*, and in the simulated data based on HIV, the mean number of clusters for DADA2 was 7.5 for separation 0.015, 7.1 for separation 0.01, 4.1 for separation 0.005, and 2.8 for sepa-

ration 0.0025. *Ampliclust* always found all 8 haplotypes even after just 10 Rnd-Ini(10) initializations. Of course, *ampliclust* spent much more time in the effort. Although we did not do timed runs, anectodally DADA2 took about a minute and *ampliclust* with 10 Rnd-Ini(10) initializations took about 20 minutes. In addition to possibly missing the global optimum, DADA2 computes a p-value using a likelihood that is assumed to be proportional to the actual conditional likelihood, which conditioned on the current cluster sizes. Unfortunately, the neglected proportionality constant is not actually constant in the parameters, and it is not clear how much this fact can alter the calculations. These facts also make it more difficult for a user to select an appropriate p-value cutoff for DADA2, since the computed number is not a true p-value. In summary, because DADA2 does not optimize the actual likelihood imposed by its modeling assumptions, it achieves speed at the cost of accuracy, and the accuracy difference is most visible for overlapping clusters.

We were surprised by the level of contamination in all real datasets. The mock data contained many high quality reads of apparently unrelated 16S sequences. Another positive of DADA2 is that it effectively ignores such contamination by refusing to initiate a cluster unless the founding sequence has been observed at least twice. This is an algorithmic solution to the problem; the DADA2 model does not actually accommodate such contamination. The current *ampliclust* would handle such contamination by finding solutions with singleton clusters (clusters with one member) to explain the contaminant outliers. To avoid this complication, we removed most contaminants from the mock dataset before analysis. Extensions to both models are needed to properly account for such contamination if it is to be universally expected in 16S datasets. For example, *ampliclust* could be extended to handle "scatter" reads that are produced under some other generative model, rather than as error from the haplotypes. Contamination with replicated reads from the same contaminating species, such as can be observed in Fig. 1 of the DADA2 publication [11], would be hard to detect as contamination except in mock datasets. Technically, such contamination is accommodated

by both DADA2 and *ampliclust* as additional clusters. It would be a subsequent challenge, then, to identify which clusters were contaminants.

We had hoped that the mock datasets would provide a unique opportunity to compare the performance of the two methods. Unfortunately, the problem of contamination and the presence of unrecognized variation within species/strains means that even in mock datasets, the true number and abundance of clusters is not known. Thus, good simulation techniques are still needed to compare different clustering solutions. Simulation is also useful for calibrating the difficulty of clustering in real data. For example, through simulation we reveal that the mock dataset was substantially easier to cluster than the HIV dataset, even though we could not know the true clusters in the latter dataset. We have implemented a simulation method in our *ampliclust* program, but there is still much room for improvement. Methods to simulate with known levels of overlap/difficulty and more varied haplotype structures would be desirable. Also, simulation methods must accurately mimic the true error properties of the Illumina sequencing machines.

One major challenge to apply *ampliclust* is the choice of initialization for AECM. We bypassed this problem in the mock data by initializing with the true haplotypes as a shortcut to what we hope is the global maximum. The k-modes initialization schemes we applied to the HIV and simulation data have not been thoroughly tested. In particular, this initialization method ignores the information in the quality scores, and thus may ignore critical information for obtaining good starting points for AECM. A simple solution would be to use a quality-based distance metric instead of the Hamming distance, but there are also many clever initialization algorithms for k-modes, for example [19]. Thus, *ampliclust* could further benefit from better initialization methods.

AECM is an EM algorithm, which are notoriously slow to converge. There are many methods to accelerate EM that could be applied to our AECM algorithm [20, 21, 22], and the EM can be parallelized [23, 24]. However, the very large datasets that characterize NGS data may not be amenable to any of these techniques, especially when K is very large. In this case, it is also possible to incrementally handle the data

such that the whole dataset need not be processed at once [25]. All of these ideas can be used to speed up *ampliclust* to make it computationally competitive with DADA2, while retaining its enhanced accuracy.

In summary, *ampliclust* is better able to detect the true haplotypes and abundances when haplotype error clouds are not well-separated. The most important future development for *ampliclust* is to speed it up, particularly with respect to the multiple initializations that are currently needed to find the global maximum. The singe most important advantage of other methods is their reduced computational complexity compared to *ampliclust*. A benefit of our approach is that the fully probabilistic model can be extended in several directions. To handle contamination, the model may need to be extended with a scatter component, but we may also want to improve clustering by utilizing auxiliary information, most importantly the barcode data that are routinely sequenced along with the amplicons. Since barcodes are uniquely attached to the original cDNA molecules, they contain valuable information for clustering.

6.6 Appendix

6.6.1 Processing HIV data

To process the HIV data, we propose a model to identify and trim the adapter/primer and barcodes from the reads. Each valid read begins with 0–2 random nucleotides followed by the 52bp adapter/primer

GCCTTGCCACACGCTCAGNNNNNNNNNGTTGTAAYTTCTAGRTCCCCTCCTG,

including the 9bp barcode and two ambiguous nucleotides at positions $\{35, 42\}$. The last 25bp are homologous to the C3 region of the HIV-1 env gene and serve as the primer for reverse transcription. The first 18bp is an adapter for subsequent PCR amplification and sequencing. We will refer to this 52bp pattern as the (ambiguous) "primer," ambiguous because of the unspecified barcode as well as the Y and R. We will refer to the 43bp pattern excluding the barcode as the (ambiguous) primer sans

barcode. If we can neglect indels in the reads, then we expect a primer to start at read position 1, 2, or 3. Let $Z_i \in \{0, 1, 2\}$ be the unknown number of random nucleotides at the start of read $\boldsymbol{r}_i$. Further, let $Y_i \in \{0, 1, 2, 3\}$ be the unknown state of the *unambiguous* primer sans barcode with the Y and R resolved. The combined hidden state indicator $X_i \in \{0, 1, \ldots, 12\}$, defined via the one-to-one mapping where $Z_i = \lfloor X_i/4 \rfloor$ and $Y_i = X_i \bmod 4$.

Dropping read index i, let the probability of read nucleotide r at position j given $X = x$ be $p_{jr}(x)$. This probability will vary according to the type of template nucleotide. For all non-barcode primer positions $j \in \{z+1, z+2, \ldots, z+52\} \backslash \{z+19, z+20, \ldots, z+27\}$, we have

$$p_{jr}(x) := s_{\mathsf{P}_m r} = \begin{cases} (1-\delta_j)\gamma_{\mathsf{P}_m r} & r \neq \mathsf{P}_m \\ \delta_j & r = \mathsf{P}_m, \end{cases} \tag{1}$$

where P_m is the unambiguous nucleotide at primer position $m = z - j$. For the barcode positions $j \in \{z + 19, z + 20, \ldots, z + 27\}$, we assume

$$p_{jr}(x) = q_{br}, \sum_{r \in \{\mathsf{A,C,G,T}\}} q_{br} = 1.$$

For positions outside the primer in the read, we assume $p_{jr}(x) = q_{sr}$, $\sum_{r \in \{\mathsf{A,C,G,T}\}} q_{sr} = 1$, another distribution over nucleotides that reflects nucleotide content in the virus.

Finally, some reads may not contain the primer anywhere. In this case, we generated the entire read as iid Multinomial$(1, \boldsymbol{q}_s)$ and define $X = 12$, $Z = 3$ and ignore the now meaningless Y. As a results, there are $3 \times 4 + 1 = 13$ possible hidden states X_i or (Z_i, Y_i) for every read $\boldsymbol{r}_i$.

6.6.1.1 Complete data likelihood

Assuming $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ are iid, z_i and y_i are a priori independent, $\boldsymbol{\pi}$ give the probabilities of each possible $z_i \in \{0, 1, 2, 3\}$ value and $\boldsymbol{\eta}$ give the probabilities of each

possible $y_i \in \{0, 1, 2, 3\}$ when $z_i > 0$, then the complete data likelihood is

$$L(\boldsymbol{\theta} \mid \mathcal{R}, \boldsymbol{x}) = \prod_{i=1}^{n} \prod_{x=0}^{12} \left[\pi_{\lfloor x/4 \rfloor} \eta_{x \bmod 4}^{\mathbb{1}\{x<12\}} \prod_{j=1}^{l_i} p_{jr_{ij}}(x) \right]^{\mathbb{1}\{x_i=x\}},$$

where l_i is the length of read i. For each read i and x_i, partition the read positions $\{1, 2, \ldots, l_i\}$ into the set of positions $\mathcal{J}_{ix_iP}$ corresponding to the primer sans barcode, the set of positions $\mathcal{J}_{ix_iB}$ corresponding to the barcode, and the set of all other positions $\mathcal{J}_{ix_iS}$. Let m_{xj} be the position in the primer given the read position j and primer position $\lfloor x \rfloor$. Then,

$$L(\boldsymbol{\theta} \mid \mathcal{R}, \boldsymbol{x}) = \prod_{i=1}^{n} \prod_{z=0}^{3} \left\{ \pi_z \prod_{j \in \mathcal{J}_{ix_iB}} q_{br_{ij}} \prod_{j \in \mathcal{J}_{ix_iS}} q_{sr_{ij}} \right.$$
$$\left. \prod_{y \in \mathbb{N}(\mathcal{J}_{ix_iP})} \left(\eta_{x_i \bmod 4} \prod_{j \in \mathcal{J}_{ix_iP}} \delta_j^{\mathbb{1}\{r_{ij}=p_{yj}\}} \left[(1-\delta_j)\gamma_{p_{yj}r_{ij}} \right]^{\mathbb{1}\{r_{ij} \neq p_{yj}\}} \right)^{\mathbb{1}\{x_i \bmod 4 = y\}} \right\}^{\mathbb{1}\{\lfloor x_i/4 \rfloor = z\}},$$

where $\mathbb{N}(\mathcal{J}_{ix_iP})$ is the index set of all possible primers without ambiguity except at the barcode $\{\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots\}$ consistent with the fully ambiguous primer and p_{yj} is the jth nucleotide in yth primer. We define $\mathbb{N}(\mathcal{J}_{ix_iP}) = \varnothing$ when the read contains no primer site ($x_i = 12$).

6.6.1.2 E-step

In the E-step, we merely need to compute

$$e_{ix} = P(x_i = x \mid \boldsymbol{r}_i) \propto \prod_{j=1}^{l} p_{jr_{ij}}(x).$$

6.6.1.3 M-step

In the M-step, the update formulae at iteration $t+1$ are

$$\begin{aligned}
\pi_z^{(t+1)} &= \frac{\sum_{i=1}^n \sum_{x=0}^{12} \mathbb{1}\left\{\lfloor x/4 \rfloor = z\right\} e_{ix}^{(t)}}{\sum_{i=1}^n \sum_{x=0}^{12} e_{ix}^{(t)}} \\
\eta_y^{(t+1)} &= \frac{\sum_{i=1}^n \sum_{x=0}^{11} \mathbb{1}\left\{x \bmod 4 = y\right\} e_{ix}^{(t)}}{\sum_{i=1}^n \sum_{x=0}^{11} e_{ix}^{(t)}} \\
\delta_j^{(t+1)} &= \frac{\sum_{i=1}^n \sum_{x=0}^{11} \mathbb{1}\left\{r_{ij} = p_{x \bmod 4,j}\right\} e_{ix}^{(t)}}{\sum_{i=1}^n \sum_{x=0}^{11} e_{ix}^{(t)}}, j \in \mathcal{J}_{ix_iP} \\
\gamma_{\mathsf{N}_1\mathsf{N}_2}^{(t+1)} &= \frac{\sum_{i=1}^n \sum_{x=0}^{11} \mathbb{1}\left\{p_{x \bmod 4,j} = \mathsf{N}_1, r_{ij} = \mathsf{N}_2\right\} e_{ix}^{(t)}}{\sum_{i=1}^n \sum_{x=0}^{11} \mathbb{1}\left\{p_{x \bmod 4,j} = \mathsf{N}_1, r_{ij} \neq \mathsf{N}_1\right\} e_{ix}^{(t)}}, \mathsf{N}_1 \neq \mathsf{N}_2, j \in \mathcal{J}_{ix_iP} \\
q_{br}^{(t+1)} &= \frac{\sum_{i=1}^n \sum_{x=0}^{11} \sum_{j \in \mathcal{J}_{ix_iB}} \mathbb{1}\left\{r_{ij} = r\right\} e_{ix}^{(t)}}{\sum_{i=1}^n \sum_{x=0}^{11} |\mathcal{J}_{ix_iB}| e_{ix}^{(t)}} \\
q_{sr}^{(t+1)} &= \frac{\sum_{i=1}^n \sum_{x=0}^{12} \sum_{j \in \mathcal{J}_{ix_iS}} \mathbb{1}\left\{r_{ij} = r\right\} e_{ix}^{(t)}}{\sum_{i=1}^n \sum_{x=0}^{12} |\mathcal{J}_{ix_iS}| e_{ix}^{(t)}},
\end{aligned}$$

where $\mathcal{J}_{ix_iS}$ consists of all read positions when $x_i = 12$.

6.6.1.4 Scoring the primer sequence

The EM algorithm is initialized by a random partition of the reads into one of thirteen classes indexed by X_i. Multiple initializations converged to the same maximum. After convergence of the EM algorithm and estimation of the MLEs $\hat{\boldsymbol{\theta}}$, we scored the primer sequence by identifying the most likely position of the primer

$$\hat{z}_i = \underset{z \in \{0,1,2\}}{\arg\max} \sum_{x \in \{0,1,\ldots,11\}:\lfloor x/4 \rfloor = z} l(\hat{\boldsymbol{\theta}} \mid \boldsymbol{r}_i, x)$$

and computing the score as the summed log likelihood over all possible primers

$$S_{ip} = \sum_{y \in \{0,1,2,3\}} \sum_{x = 4\hat{z}_i + y} l(\hat{\boldsymbol{\theta}} \mid \boldsymbol{r}_i x).$$

6.6.1.5 Scoring the 3' end of the read

Because there was extensive mispriming in these data, we also scored the 3' end of the read. These data were isolated from a patient in South Africa between 2005 and 2009. To identify the expected sequence at the 3' end of the read, we identified an HIV-1 subtype C reference sequence from Zambia in the 2010 Env reference alignment provided by the HIV Sequence Database hosted at `https://www.hiv.lanl.gov/content/`. We used bwa [26] with default parameters except a penalty of 100 for 5' and 3' end clipping to align the reads to the subtype C reference sequence. We selected the first 10,000 of the aligned reads (out of 30,266 total aligned reads from 58,071 total reads), and used muscle [27] with default parameters to align them to the reverse complement of the primer sequence and the next 300bp in the reference sequence. We computed a consensus sequence from the alignment by removing all positions with an insertion in more than 40% of the reads. We blasted the resulting consensus sequence against the nucleotide collection at NCBI using megablast [28], and found several 100% identical South African sequences. We selected the sequence with accession KF996710 as our new reference. The 89bp located 160bp downstream of the primer site in KF996710 (in the reverse complement) are $s =$ TCTTATACTTTTTCTTGTATTATTGTTGGGTCTTGTACAATTAATCC CTACAGATTCATTGAGATGTACTATTATTGTTTTGACATTGT. These 89bp should be contained in all reads of the correctly primed and amplified amplicon. Assuming an error rate of 0.02 at all read positions and the MLEs $\hat{\gamma}_{\mathsf{N}_1\mathsf{N}_2}$ estimated from the EM, we computed a score for the 3' end of the amplicon as

$$S_{i3} = \sum_{j=\hat{z}_i+52+160}^{l_i} \left[\mathbb{1}\left\{r_{ij} = s_j\right\} \log(0.98) + \mathbb{1}\left\{r_{ij} \neq s_j\right\} \left(\log(0.02) + \log \gamma_{s_j r_{ij}}\right)\right].$$

6.6.1.6 Filtering and trimming the reads

We filtered the reads by discarding reads shorter than 296 bp, requiring $\hat{z}_i < 3$, a primer score above -2 and a 3' end score above -45. Lastly, we trimmed $\hat{z}_i + 27$

nucleotides from all the 5' end of all remaining reads. Of the 61,711 reads, 83% were discarded by this process, leaving 10,344 reads for analysis.

6.7 References

[1] S. Goodwin, J. D. McPherson, and W. R. McCombie. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.*, 17:333–351, 2016.

[2] H. Y. K. Lam, M. J. Clark, R. Chen, R. Chen, G. Natsoulis, M. O'Huallachain, F. E. Dewey, L. Habegger, E. A. Ashley, and M. B. and Gerstein. Performance comparison of whole-genome sequencing platforms. *Nat. Biotechnol.*, 30:78–82, 2012.

[3] Lee C. K., Herbold C. W., Polson S. W., K. E. Wommack, Williamson S. J., and I. R. and McDonald. Groundtruthing next-gen sequencing for microbial ecology-biases and errors in community structure estimates from PCR amplicon pyrosequencing. *PLoS One*, 7:e44224, 2012.

[4] X. Chen, J. B. Listman, F. J. Slack, J. Gelernter, and H. Zhao. Biases and errors on allele frequency estimation and disease association tests of next-generation sequencing of pooled samples. *Genet. Epidemiol.*, 36:549–560, 2012.

[5] NGS-eval: NGS error analysis and novel sequence VAriant detection tooL.

[6] Rosen MJ, Callahan BJ, Fisher DS, and Holmes SP. Denoising pcr-amplified metagenome data. *BMC Bioinformatics*, 13:283–299, 2012.

[7] Kunan V, Engelbrektson A, Ochman H, and Hugenholtz P. Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environ Microbiol*, 12:118123, 2010.

[8] Zhou J, Wu L, Deng Y, Zhi X, Jiang Y, Tu Q, Xie J, Nostrand JDV, He Z, and Yang Y. Reproducibility and quantitation of amplicon sequencing-based detection. *ISME J*, 5:13031313, 2011.

[9] Quince C, Lanzen A, Curtis TP, Davenport RJ, Hall N, Head IA, Read LF, and Sloan WT. Accurate determination of microbial diversity from 454 pyrosequencing data. *Nat methods*, 6:639641, 2009.

[10] Quince C, Lanzen A, Davenport RJ, and Turnbaugh PJ. Removing noise from pyrosequenced amplicons. *BMC Bioinf*, 12:38–44, 2011.

[11] Callahan BJ, Mcmurdie PJ, Rosen MJ, Han AW, Johnson AJ, and Holmes SP. Dada2: High resolution sample inference from amplicon data. *Nature Method*, 13:581–587, 2015.

[12] Xiao-Li Meng and David van Dyk. The em algorithm–an old folk-song sung to a fast new tune. *Journal of the Royal Statistical Society. Series B (Methodological)*, 59(3):511–567, 1997.

[13] Jinal N Bhiman, Colin Anthony, Nicole A Doria-Rose, Owen Karimanzira, Chaim A Schramm, Thandeka Khoza, Dale Kitchin, Gordon Botha, Jason Gorman, Nigel J Garrett, Salim S Abdool Karim, Lawrence Shapiro, Carolyn Williamson, Peter D Kwong, John R Mascola, Lynn Morris, and Penny L Moore. Viral variants that initiate and drive maturation of V1V2-directed HIV-1 broadly neutralizing antibodies. *Nat Med*, 21(11):1332–1336, 2015.

[14] Dimitris Karlis and Evdokia Xekalaki. Choosing initial values for the em algorithm for finite mixtures. *Computational Statistics and Data Analysis*, 41:577–590, 2003.

[15] Zhexue Huang. A fast clustering algorithm to cluster very large categorical data sets in data mining. In *In Research Issues on Data Mining and Knowledge Discovery*, pages 1–8, 1997.

[16] Ranjan Maitra. Initializing partition-optimization algorithms. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 6(1):144–157, 2009.

[17] Volodymyr Melnykov Ranjan Maitra. Simulating data to study performance of finite mixture modeling and clustering algorithms. *Journal of Computational and Graphical Statistics*, 19(2):354–376, 2010.

[18] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2:193–218, 1985.

[19] Fuyuan Cao, Jiye Liang, and Liang Bai. A new initialization method for categorical data clustering. *Expert Systems with Applications*, 36(7):10223 – 10228, 2009.

[20] A. Berlinet and C. Roland. Parabolic acceleration of the em algorithm. *Statistics and Computing*, 19(1):35–47, 2009.

[21] Ravi Varadhan and Christophe Roland. Simple and globally convergent methods for accelerating the convergence of any em algorithm. *Scandinavian Journal of Statistics*, 35(2):335 – 353, 2008.

[22] Hua Zhou, David Alexander, and Kenneth Lange. A quasi-newton acceleration for high-dimensional optimization algorithms. *Stat Comput*, 21(2):261–273, Jan 2011.

[23] Chen W.-C. Model-based clustering of regression time series data via apecman aecm algorithm sung to an even faster beat. *Statistical Analysis and Data Mining*, 4:567, 2011.

[24] Wei chen Chen, George Ostrouchov, David Pugmire, Prabhat, and Michael Wehner. A parallel em algorithm for model-based clustering applied to the exploration of large spatio-temporal data. *Technometrics*, 55(4):513–523, 2013.

[25] Ranjan Maitra. Initializing partition-optimization algorithms. *IEEE/ACM Trans Comput Biol Bioinform*, 6(1):144–157, 2009.

[26] Heng Li. Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. In *arXiv:1303.3997v2*, pages [q–bio.GN], 2013.

[27] Robert C. Edgar. Muscle: multiple sequence alignment with high accuracy and high throughput. In *Nucleic Acids Res.*, pages 32(5):1792–1797, 2004.

[28] Zheng Zhang, Scott Schwartz, Lukas Wagner, and Webb Miller. A greedy algorithm for aligning dna sequences. In *J Comput Biol*, pages 7(1–2):203–14., 2000.

[29] A. Ratan, W. Miller, J. Guillory, J. Stinson, S. Seshagiri, and S. C. Schuster. Comparison of sequencing platforms for single nucleotide variant calls in a human sample. *PLoS One*, 8:e55089, 2013.

[30] Jabara CB, Jones CD, Roach J, Anderson JA, and Swanstrom R. Accurate sampling and deep sequencing of the hiv-1 protease gene using a primer id. *Proc Natl Acad Sci U S A*, 108:201662017, 2011.

[31] Sohn KA and Xing EP. Spectrum: joint bayesian inference of population structure and recombination eventss. *BMC Bioinformatics*, 23:479–489, 2007.

[32] K. A. Sohn, Z. Ghahramani, , and E. P. Xing. Robust estimation of local genetic ancestry in admixed populations using a nonparametric bayesian approach. *Genetics*, 191:12951308, 2012.

[33] P J Green. On use of the EM algorithm for penalized likelihood estimation. *Journal of the Royal Statistical Society, Series B*, 52(3):443–452, 1990.

[34] Shelley B. Bull, Juan Pablo Lewinger, and Sophia S. F. Lee. Confidence intervals for multinomial logistic regression in sparse data. *Statistics in Medicine*, 26(4):903–918, 2007.

[35] http://www.hiv.lanl.gov/.

CHAPTER 7

CONCLUSION

Elicitation of bnAbs is a major goal of HIV-1 vaccine development. The HIV-1 Env are the only target for bnAbs. In this dissertation, we evaluated different vaccine approaches to induce bnAbs. The approaches include use of putative fusion intermediate state of gp41 and novel heterologous prime-boost strategies. Several interesting conclusions have been drawn from our studies.

The MPER of gp41 is an attractive target for HIV-1 vaccine development. We previously described that gp41-HR1-54Q, a post-fusion form of gp41, induced strong antibody responses against non-neutralizing face of the helix. To refocus immune responses to the neutralizing face of the helix and better target 4E10/10E8 epitopes, we generated putative fusion intermediates by destabilizing six-helix bundle structure through the introduction of mutations or deletions in the HR1. One immunogen induced antibodies that targeted the residues on the neutralizing face of the helix, which are crucial for 4E10/10E8 recognition. The study suggested that the destabilization of six-helix bundle could influence the immunogenicity of the MPER.

The major challenge in inducing bnAbs is developing immunogens and/or immunization strategies that direct the immune responses towards neutralizing epitopes and guide the antibody evolution and maturation so that antibodies can recognize neutralizing epitopes on the native spikes. It is believed that multiple immunizations using a single antigen would be difficult to induce bnAbs. We devised two novel strategies to induce bnAbs. One strategy was designed to

sequentially immunize animals with progressively more native immunogens. Different sets of immunogens have been used to induce anti-MPER bnAbs (Chapter 3) and anti-CD4bs bnAbs (Chapter 5). In Chapter 3, this strategy failed to induce bnAbs, but was better able to direct antibodies towards the anti-MPER bnAbs epitopes than homologous prime-boost immunization using a single immunogen. In Chapter 5, this strategy was tested in two groups using gp120 based immunogens induced strong antibody responses that compete VRC01 binding and neutralizing activity against Tier 1 viruses, but no bnAbs were elicited. Another vaccine strategy was designed to alternatively immunize a native like immunogen and a small CD4bs-based immunogen using a rapid schedule. In Chapter 4, this strategy was tested using two different immunogens, it induced neutralizing antibodies that were limited to Tier 1 viruses. Although we failed to induce bnAbs, which was possibly due to the suboptimal immunogens, our vaccine strategies showed their potential strength. Additional studies will be needed to better evaluate this vaccine strategy using optimal immunogens.

Current Next Generation Sequencing (NGS) technologies could provide high throughout data at low cost, but the associated errors can influence the downstream analyses. Thus, it is critically important for error correction of sequencing data to provide accurate, high-quality sequencing data for further downstream analysis. Most error correction methods so far have been designed for substitution errors, possibly due to the dominance of Illumina platform in the market. Illumina quality scores are related to error rates and informative for error correction. However, most Illumina denoiser algorithms underutilize quality score,

they usually discard quality score or turn to compression, which may affect the downstream analyses and lead to undesirable consequences. We proposed an ampliclust, an error modeling approach using uncompressed data for denoising Illumina amplicon data. The study showed that our method has higher accuracy than DADA2 for data sets that are not well separated. Our denoising method could be further improved using the barcode data, because a barcode is uniquely attached to a cDNA molecule, which should be informative for clustering.

APPENDIX

HIV-1 GP41-HR1-HR2 SIX-HELIX BUNDLE AS A NOVEL FUSION PROTEIN PARTNER FOR EFFICIENT RECOMBINANT PROTEIN EXPRESSION

Heliang Shi, Andrew Harley, Michael Cho

Abstract

The demand for recombinant proteins is growing globally due to their great application values in medicine, research laboratories, and biopharmaceutical industries. Efficient strategies that produce higher yields of recombinant proteins at lower costs are needed to achieve commercially viable recombinant proteins production. Here we describe a novel method to produce large quantities of recombinant proteins by using an N-terminally fused protein partner derived from HIV-1 gp41, and contains the heptad repeat 1(HR1) and heptad repeat 2 (HR2) regions. We introduced a thrombin cleavage site between gp41-HR1-HR2 and the recombinant proteins to efficiently remove the fusion partner through His-tag affinity purification. To purify non-His-tagged recombinant proteins from thrombin-cleaved fusion proteins, we further incorporated a 6xHis-tag between HR1 and HR2 regions. By linking the fusion partner to the N-terminus of target proteins, we demonstrated significant enhancement of both His-tag and non His-tag recombinant protein expression yield in *Escherichia coli*. This novel fusion expression system could be used for any application that requires large quantities of recombinant proteins.

A.1 Introduction

Recombinant proteins have been widely used in modern applications, including treating a number of diseases including cancer, diabetes, and heart failure.(1-4) The demand for recombinant proteins has increased significantly, especially in biopharmaceutical industries. As of 2015, nearly 400 recombinant protein-based products were approved as biopharmaceuticals, and over 1000 protein candidates were under development. (5) Thus, the future applications for recombinant proteins are very promising.

Given the increasing demand for recombinant proteins, developing efficient strategies for their production will be extremely important. Higher expression yield and lower costs are needed to achieve commercially viable large-scale production. Currently, bacterial expression is the most common production system for recombinant proteins. Among them, *Escherichia coli* (*E. coli*) has always been preferred as it grows fast, is easy to handle, cost-effective, and usually provides high yield. (6) A variety of studies in the *E. coli* expression system show that the target protein's expression level could be improved by linking to a highly expressed fusion partner(7-11). In some cases, fusion partners could direct the target proteins to the inclusion bodies and simply downstream purification.(7, 9, 12).

In this study, we described a novel method to enhance the expression yield of recombinant proteins by using a N-terminally fused protein partner, gp41-HR1-HR2, which consists of the heptad regions 1 (HR1) and 2 (HR2) of HIV-1 glycoprotein 41 (gp41) and could form a stable six-helix bundle (6HB). Two versions of the fusion partner were generated. First, we incorporated a thrombin cleavage site between gp41-

HR1-HR2 and target proteins to efficiently cleave the fusion protein and liberate the target proteins. Second, we adjusted the location of a His-tag from the C-terminus of target proteins to the middle of the HR1 and HR2 regions to produce non His-tagged proteins. We successfully generated large quantities of several HIV-1 derived recombinant proteins using our fusion partners. The fused proteins were expressed at a significantly higher level in *E. coli* as compared to unfused proteins. Target recombinant proteins, including both His-tagged and non His-tagged, were efficiently liberated from the fusion partners.

A.2 Results and discussion

A.2.1 Expression and purification of HR1-HR2-54Q fusion protein

To detect the effect of gp41-HR-HR2 on protein expression, we compared the expression levels of HR1-HR2-54Q and 54Q in parallel. As shown in Fig.1A, the yield of 54Q alone was poor after induction with IPTG (2-4mg/l). In contrast, HR1-HR2-54Q was expressed at much higher levels (>120mg/l). The amount of HR1-HR2-54Q is estimated to be ~40% of the total bacterial protein content according to the SDS-PAGE. The yield of fusion protein was ~40-fold higher than that the original 54Q target protein. The enhanced expression observed in HR1-HR2-54Q was possibly due to the N-terminal fused HR1-HR2, which could form a stable six-helix bundle (6HB) structure. From these experiments, we concluded that the gp41-HR1-HR2 fragment could be a useful fusion partner to direct target proteins to inclusion bodies and increase expression.

A.2.2 Cleavage and purification of 54Q peptide

In our initial construct, trypsin digestion was necessary to remove the HR1-HR2 portion from the purified fusion protein. A potential problem with trypsin cleavage was the presence of cleavage sites within 54Q. As shown in Fig.1B, three major bands were observed after trypsin cleavage. From top to bottom, respectively, they represent the HR1-HR2 portion, 54Q peptide, and a fragment digested from 54Q. The small fragment may contain the 6xHis tag, so it could not be differentiated from 54Q during the Ni-NTA affinity chromatography purification process and contaminated the purified 54Q. Thus, to better cleave and purify the target protein, a more specific cleavage site should be introduced between fusion partner and target protein.

A.2.3 Expression and purification of HR1-HR2-TH based fusion proteins

To specifically cleave and purify the target protein, we replaced the trypsin site with a thrombin site at the C-terminal of HR1-HR2, called HR1-HR2-TH, and the expression of a fusion protein was evaluated. As shown in Fig.2A, the yield of HR1-HR2-TH-54Q (>200mg/ml) is even higher than that of HR1-HR2-54Q. The amount of HR1-HR2-TH-54Q was estimated to be ~60%-70% of the total protein content. The enhanced expression compared to HR1-HR2-54Q was possibly due to mutated residues resulting from the introduction of thrombin cleavage site.

To evaluate the generality of the exhibited high expression of the HR1-HR2-TH fusion partner, we generated a second fusion protein examined the expression compared to the unfused protein. The protein selected was gp41-28x3, which consists of three tandem repeats of the C-terminal 28 amino acids of the extracellular region. As

shown in Fig.2C, the expression of gp41-28x3 was confirmed by Western blotting, but the yield was very poor as it could not be detected through Coomassie stained SDS-PAGE. However, the expression of gp41-28x3 was significantly enhanced after N-terminal fusion to HR1-HR2-TH (Fig.2D). Both HR1-HR2-TH based fusion proteins described exhibited similarly high expression, suggesting the generality of HR1-HR2-TH as an efficient fusion partner for enhancing target protein expression.

A.2.4 Cleavage and purification of target proteins from HR1-HR2-TH based fusion proteins

To liberate target proteins from the HR1-HR2-TH fusion peptide, purified fusion proteins were subjected to thrombin cleavage and Ni-NTA affinity chromatography. As shown in Fig.2B, HR1-HR2-TH-54Q could be efficiently digested by thrombin and purified through His-tag affinity. The eluted 54Q protein was >99% pure based on SDS-PAGE estimation. Above the 54Q band, two adjacent protein bands were observed in the SDS-PAGE lane of thrombin-cleaved HR1-HR2-TH-54Q. There may be another thrombin digestion site between the T7-tag and HR1-HR2-TH, but it is not as sensitive to cleavage as at the C-terminal site following HR1-HR2-TH. These two protein fragments would not affect the purification of the target protein because they do not contain His-tag. Similarly, the HR1-HR2-TH portion was successfully removed from HR1-HR2-TH-28x3 by thrombin digestion (Fig.2E). Thus, the introduction of a thrombin cleavage site to the C-terminal of HR1-HR2 efficiently increased the target protein expression and permitted release of the target proteins from the HR1-HR2-TH fusion peptide.

A.2.5 Expression of HR1-6H-HR2-TH based fusion proteins

In some cases, the presence of a His-tag could affect protein structure and/or activity, thus non-His-tagged target proteins would be preferred. To generate non-His-tagged target proteins, an HR1-6H-HR2-TH fusion partner was constructed by adjusting the location of 6xHis-tag from the C-terminal of target proteins to between the HR1 and HR2 regions. Post-thrombin digestion, the fusion proteins would be subjected to His-tag affinity chromatography, and the target proteins would be recovered from the flow through due to the absence of the His-tag.

HR1-6H-HR2-TH-54 fusion protein was constructed in order to generate large amount of non-His-tag gp41-54 protein, which consists of the C-terminal 54 amino acids of the extracellular portion of HIV-1 gp41. As shown in Fig.4, HR1-6H-HR2-TH-54 was strongly expressed, which is consistent with that of HR1-HR2-TH based fusion proteins, indicating the location of His-tag does not significantly affect expression. These data suggest that HR1-6H-HR2-TH could serve as an efficient fusion partner to effectively generate large amounts of non-His-tagged target proteins.

A.3 Conclusions

In this study, we described a strategy for generating large quantities of purified target proteins by fusing novel fusion peptide partners to the N-terminal end of target proteins. We developed two novel fusion partners, gp41-HR1-HR2-TH and gp41-HR1-6H-HR2-TH, and have successfully applied these fusion partners for the production of His-tag and non His-tag target proteins, respectively, derived from HIV-1 gp41. The

expression of fusion proteins was significantly higher than that of the unfused target proteins, and the fusion partner portion was efficiently removed via thrombin cleavage followed by His-tag affinity chromatography. These fusion partners should be applicable for the production of other recombinant proteins important for biopharmaceuticals, including cytokines, enzymes, vaccines, etc.. Overall, the novel fusion partners described in this study will be very attractive for any applications that require significant quantities of pure proteins.

A.4 Materials and methods

A.4.1 Construct generation

The gp41-54Q and gp41-HR1-54Q proteins were described previously. (13)The HR2 fragment was amplified by PCR from gp41-HR1-54Q.The amplified HR2 fragment was digested by BamH1 and BglII and then inserted into the pET-gp41-HR1-54Q BamH1 site to yield pET-HR1-HR2-54Q. To generate gp41-HR1-HR2-TH-54Q, site-directed mutagenesis was performed to introduce a thrombin site between the HR2 and 54Q regions. To construct gp41-HR1-6H-HR2-TH-54, the HR1-6H fragment was amplified by PCR from gp41-HR1-54Q. The amplified HR1-6H fragment was digested by BamH1 and BglII and then inserted into the pET-HR1-HR2-TH-54TM (unpublished data) BamH1 site to yield pET-HR1-6H-HR2-TH-54TM. The 54 fragment was amplified by PCR from gp41-54Q with a Q683K (HxB2 numbering) introduced. The amplified 54 fragment was inserted into pET-HR1-6H-HR2-TH-54TM through Kpn1 and EcoR1 sites to yield gp41-HR1-6H-HR2-TH-54.

The gp41-28x3 protein was generated based on pGEX-gp41-C30x3, which contains a C-terminal 6xHis tag (unpublished data). C30x3 was the template to generate pGEX-gp41-C30-28x2 to introduce the GSGSG linker between repeats. The QuikChange®XL Site directed mutagenesis kit was used per the manufacturer's instructions with primers 5'-GGCTGTGGTACATCAAGGGATCGGGATCGGGAAACGAGCAGGAGCTGCTGG-3' and 5'-GCCAGCAGCTCCTGCTCGTTTCCCGATCCCGATCCCTTGATGTACCACAGCC-3'. The C30-28x2 fragment was then transferred to pET-21a (Novagen; cat#69740-3) via BamH1/EcoR1 digestion. To delete two asparagine residues and introduce a Kpn1 site, site directed mutagenesis was performed using primers 5'-GGACAGCAAATGGGTCGCGGTACCAACGAGCAGGAGCTGCTGGC-3' and 5'-GCCAGCAGCTCCTGCTCGTTGGTACCGCGACCCATTTGCTGTCC-3'. The resulting gp41-28x3 fragment was transferred into another pET construct containing HR1-HR2-TH.

A.4.2 Protein expression, purification, and characterization

All constructs were expressed in *E. coli* and purified through Ni-NTA affinity chromatography as we previously described(13-15). The fusion proteins were cleaved with thrombin (GE Healthcare; cat#27-0846-01), and the resulting fragments were purified via Ni-NTA. The His-tagged proteins gp41-28x3 and gp41-54Q were collected in the elution fractions, and the non His-tagged gp41-54 protein was collected in flow through fraction. The final purified proteins were dialyzed into PBS (pH 8.0) and stored

at −80°C. The purified proteins were subjected to SDS-PAGE or Western blotting analysis. Western blotting was performed to detect the expression of gp41-28x3 using gp41-54Q immunized rabbit serum (unpublished data) as primary antibodies and goat anti-rabbit secondary antibodies.

A.5 Figures

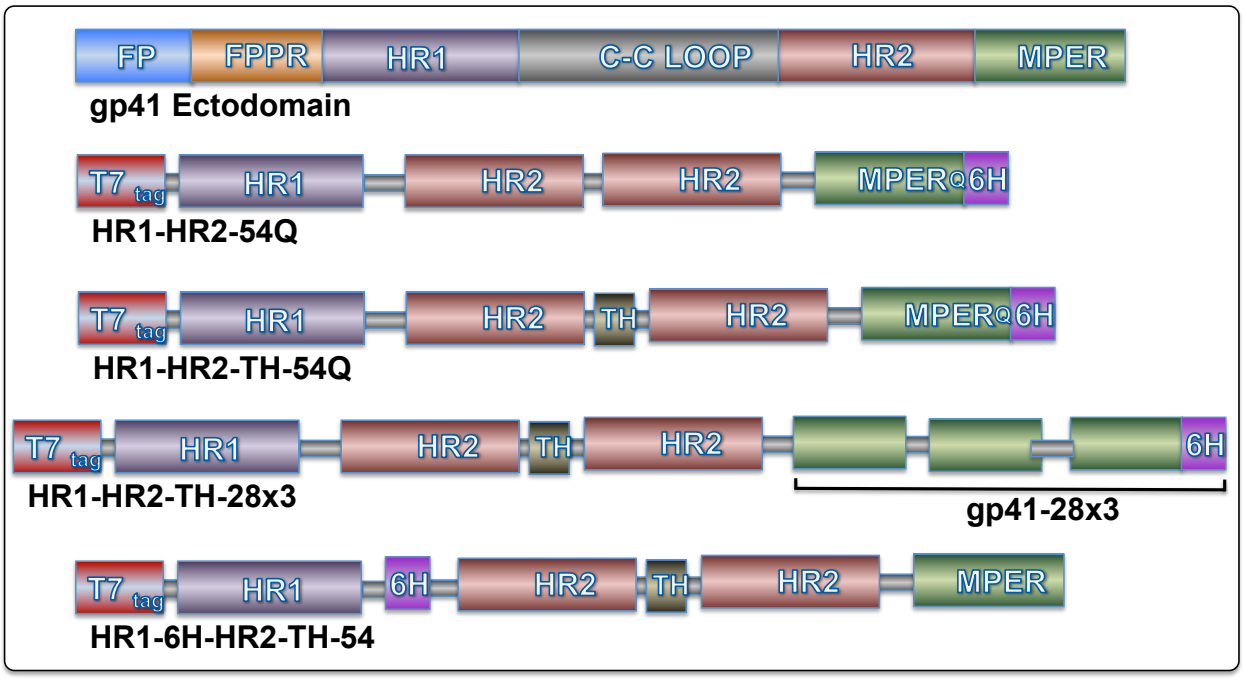

Fig.1. Schematic representation of all constructs. The entire gp41 ectodomain is shown on the top as a reference. TH represents the thrombin cleavage site and 6H represents 6xHis tag.

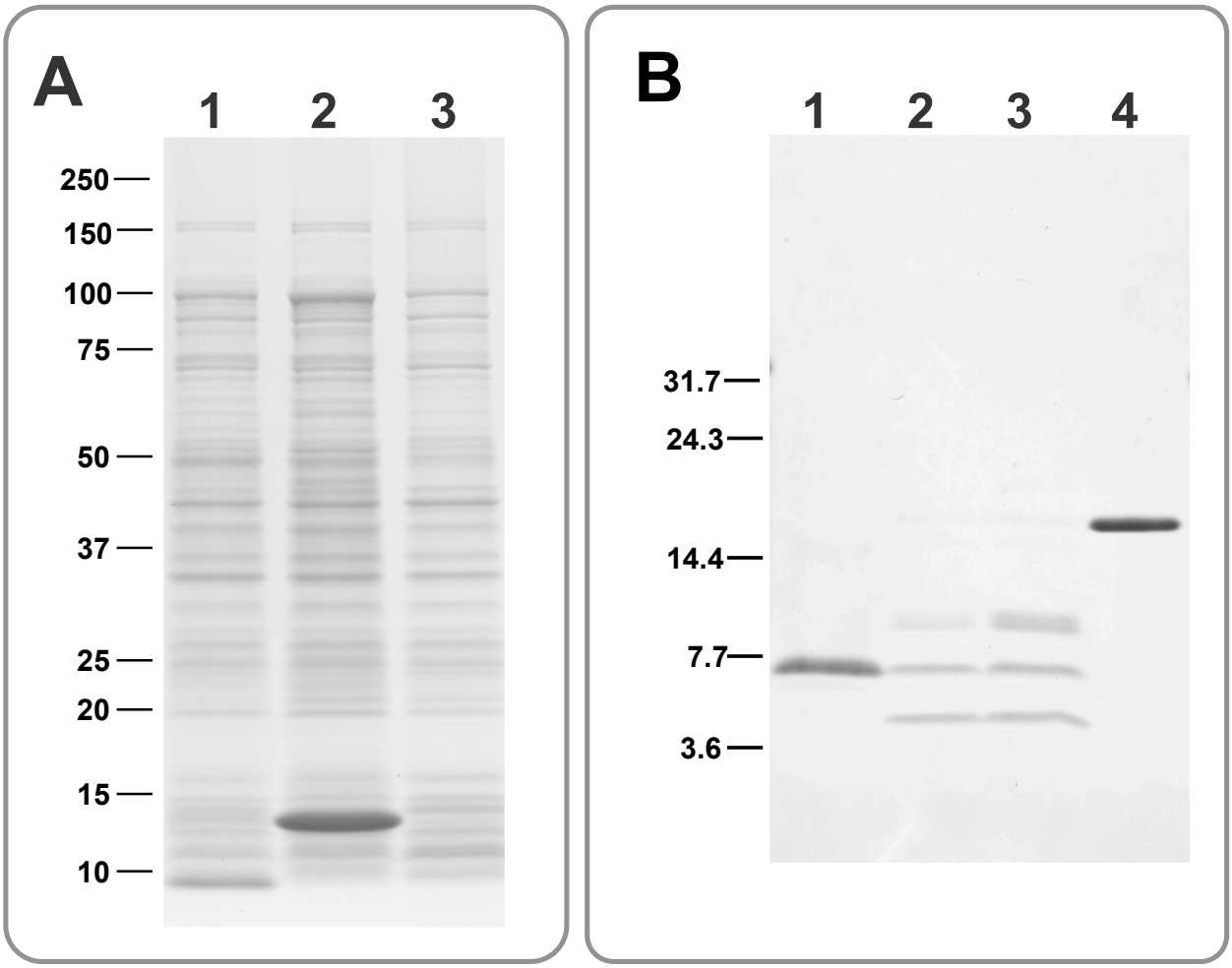

Fig.2. Comparison of HR1-HR2-54Q and 54Q expression yield. Expression and purification of HR1-HR2 fusion protein. (A) Expression of HR1-HR2-54Q and 54Q. Coomassie stained SDS-PAGE gels (Lane 1: mock sample without induction; Lane 2: induced HR1-HR2-54Q; Lane 3: induced 54Q.) (B) Cleavage of HR1-HR2-54Q with trypsin (Lane 1: original sample; Lane 2: trypsin cleaved; Lane 3: eluted from Ni-NTA column; Lane 4: 54Q).

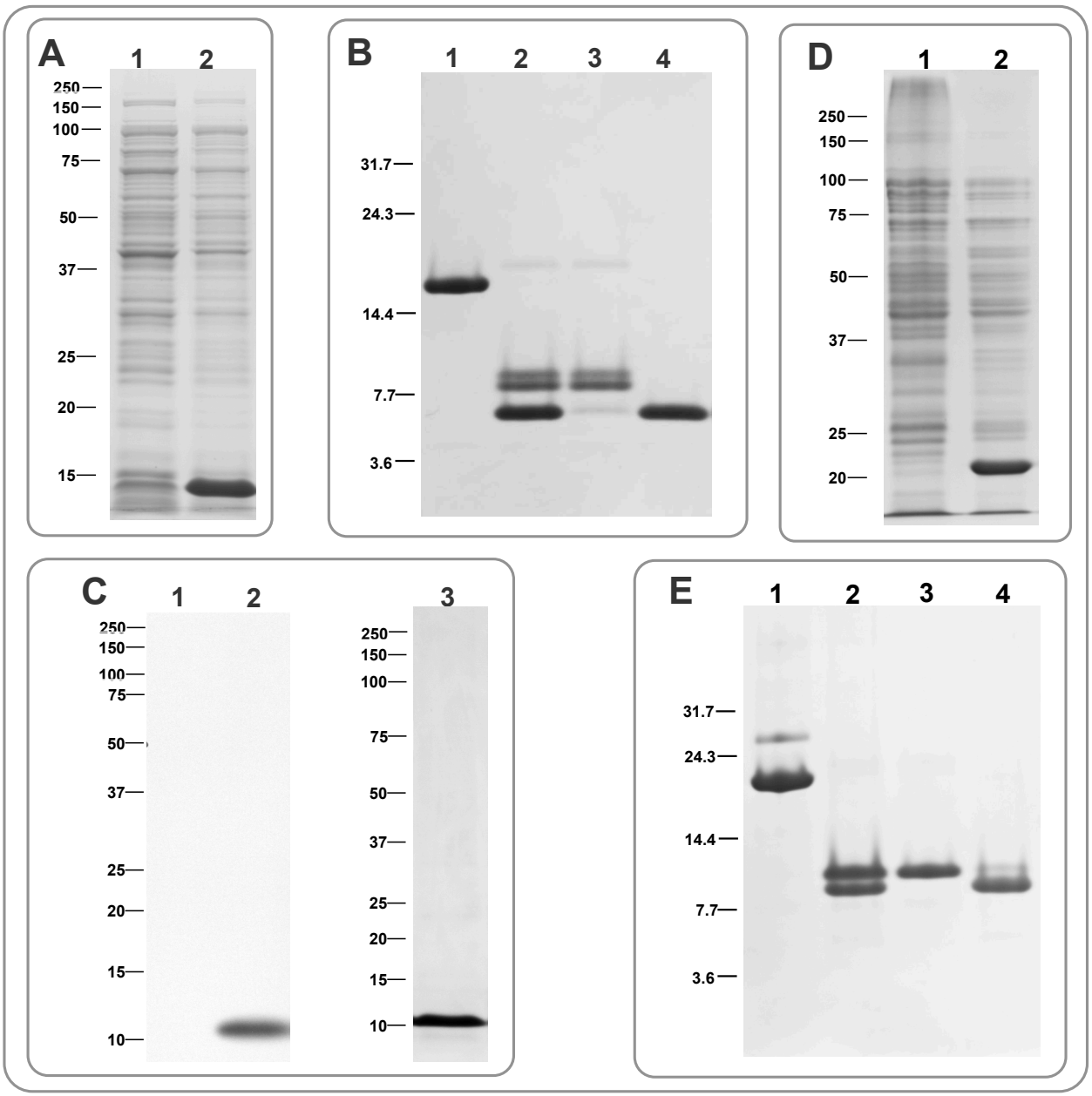

Fig.3. Expression and purification of HR1-HR2-TH fusion proteins. (A) Expression of HR1-HR2-TH-54Q. Coomassie stained SDS-PAGE gels (Lane 1: uninduced; Lane 2: induced.) (B) Cleavage of HR1-HR2-TH-54Q with thrombin (Lane 1: original sample; Lane 2: thrombin cleaved; Lane 3: flow through; Lane 4: eluted from Ni-NTA column.). (C) Expression and purification of gp41-28x3. Western blotting (Lane 1: uninduced; Lane 2: induced) and coomassie stained SDS-PAGE gel (lane 3: purified). (D) (B)

Expression of HR1-HR2-TH-28×3. Coomassie stained SDS-PAGE gel (Lane 1: uninduced; Lane 2: induced). (E) Cleavage of HR1-HR2-28×3 with thrombin (Lane 1: original sample; Lane 2: thrombin cleaved; Lane 3: eluted from Ni-NTA column; Lane 4: flow through).

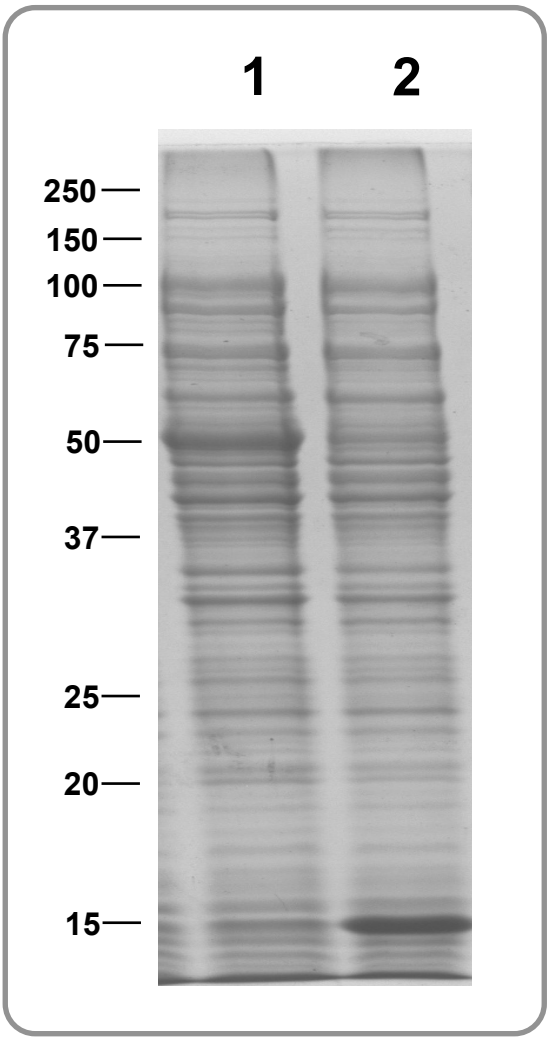

Fig.4. Expression HR1-6H-HR2-TH fusion protein. Expression of HR1-6H-HR2-TH-54. Coomassie stained SDS-PAGE gels (Lane 1: uninduced; Lane 2: induced.).

A.6 References

1. **Tyrberg B**, **Levine F**. 2001. Current and future treatment strategies for type 2 diabetes: the beta-cell as a therapeutic target. Curr Opin Investig Drugs **2**:1568–1574.

2. **Nathisuwan S**, **Talbert RL**. 2002. A review of vasopeptidase inhibitors: a new modality in the treatment of hypertension and chronic heart failure. Pharmacotherapy **22**:27–42.

3. **Disis ML**, **Knutson KL**, **McNeel DG**, **Davis D**, **Schiffman K**. 2001. Clinical translation of peptide-based vaccine trials: the HER-2/neu model. Crit Rev Immunol **21**:263–273.

4. **Osborne MJ**, **Su Z**, **Sridaran V**, **Ni F**. 2003. Efficient expression of isotopically labeled peptides for high resolution NMR studies: application to the Cdc42/Rac binding domains of virulent kinases in Candida albicans. J Biomol NMR **26**:317–326.

5. **Sanchez-Garcia L**, **Martín L**, **Mangues R**, **Ferrer-Miralles N**, **Vázquez E**, **Villaverde A**. 2016. Recombinant pharmaceuticals from microbial cells: a 2015 update. Microb Cell Fact **15**:33.

6. **Baeshen NA**, **Baeshen MN**, **Sheikh A**, **Bora RS**, **Ahmed MMM**, **Ramadan HAI**, **Saini KS**, **Redwan EM**. 2014. Cell factories for insulin production. Microb Cell Fact **13**:141.

7. **Kuliopulos A**, **Walsh CT**. 1994. Production, purification, and cleavage of tandem repeats of recombinant peptides. Journal of the American Chemical ….

8. **Baker RT**. 1996. Protein expression using ubiquitin fusion and cleavage. Curr Opin Biotechnol **7**:541–546.

9. **Majerle A**, **Kidric J**, **Jerala R**. 2000. Production of stable isotope enriched antimicrobial peptides in Escherichia coli: an application to the production of a 15N-enriched fragment of lactoferrin. J Biomol NMR **18**:145–151.

10. **Sharon M**, **Görlach M**, **Levy R**, **Hayek Y**, **Anglister J**. 2002. Expression, purification, and isotope labeling of a gp120 V3 peptide and production of a Fab from a HIV-1 neutralizing antibody for NMR studies. Protein Expr Purif **24**:374–383.

11. **Fairlie WD**, **Uboldi AD**, **De Souza DP**, **Hemmings GJ**, **Nicola NA**, **Baca M**. 2002. A fusion protein system for the recombinant production of short disulfide-containing peptides. Protein Expr Purif **26**:171–178.

12. **Jones DD**, **Stott KM**, **Howard MJ**, **Perham RN**. 2000. Restricted motion of the lipoyl-lysine swinging arm in the pyruvate dehydrogenase complex of Escherichia coli. Biochemistry **39**:8448–8459.

13. **Habte HH**, **Banerjee S**, **Shi H**, **Qin Y**, **Cho MW**. 2015. Immunogenic properties of a trimeric gp41-based immunogen containing an exposed membrane-proximal external region. Virology **486**:187–197.

14. **Banerjee S**, **Shi H**, **Habte HH**, **Qin Y**, **Cho MW**. 2016. Modulating immunogenic properties of HIV-1 gp41 membrane-proximal external region by destabilizing six-helix bundle structure. Virology **490**:17–26.

15. **Qin Y**, **Banasik M**, **Kim S**, **Penn-Nicholson A**, **Habte HH**, **LaBranche C**, **Montefiori DC**, **Wang C**, **Cho MW**. 2014. Eliciting neutralizing antibodies with gp120 outer domain constructs based on M-group consensus sequence. Virology **462-463**:363–376.