UNIVERSITY OF CALIFORNIA,
IRVINE

An expanded genetic code for the characterization and directed evolution of tyrosine-sulfated proteins

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILSOPHY

in Biomedical Engineering

by

Xiang Li

Dissertation Committee:
Assistant Professor Chang C. Liu, Chair
Associate Professor Wendy Liu
Associate Professor Jennifer A. Prescher

2018

Dedication

To

My parents Audrey Bai and Yong Li

and

My brother Joshua Li

Table of Content

List of Figures

List of Tables

CURRICULUM VITAE

Xiang Li

Education: **PhD** Biomedical Engineering, University of California, Irvine (2018)
Thesis: An expanded genetic code for the characterization and directed evolution of tyrosine-sulfated proteins

MS Biomedical Engineering, University of California, Irvine (2016)
BS Biochemistry and Molecular Biology, University of California, Santa Cruz (2012)

Professional Positions:

PhD Candidate
Professor Chang C. Liu's Laboratory, UC Irvine April 2012 – March 2018

- Developed an upgraded expanded genetic code system for sulfotyrosine and optimized expression conditions to improve the yield of sulfated antibodies by more than 20-fold, resulting in a publication in *Methods in Molecular Biology*.
- Helped to discover a bacterial sulfated peptide responsible for activating rice plant immunity in a collaboration with Professor Pamela Ronald's laboratory at UC Davis, resulting in publications in *Science Advances* and *Integrative Biology*.
- Characterized the binding contribution of individual sulfated tyrosines in a sulfated anti-HIV antibody using an expanded genetic code (publication is in preparation).
- Evolved sulfated anti-HIV antibodies against HIV's coat protein gp120 in multiple rounds of phage display.

Teaching Assistant and Guest Lecturer
UC Irvine January 2015 – September 2017

- Taught lectures and led discussions on molecular engineering (BME50A/B) or biochemistry (BioSci98) to over 100 undergraduate students per class.
- Invited as a guest lecturer to present cutting-edge technologies and technical concepts to undergraduate and high school students.

Intern
Siemens, Beijing, China July 2012 – September 2012

- Supported Siemens management institute team to organize executive assessment events.
- Completed the translation of a 30-page document from Mandarin to English in 2 months.

Undergraduate Researcher
Professor Nader Pourmand's Laboratory, UC Santa Cruz January 2009 – June 2012

- Developed a metal ion sensor using a glass pipette with a chitosan-coated nanopore, resulting in a publication in *Langmuir*.
- Developed a biosensor for the detection of cancer biomarkers by conjugating antibodies on the nanopore of glass pipettes.

Research Intern
MagArray Incorporation, Sunnyvale, CA January 2011 – June 2011

- Developed immunoassays utilizing paramagnetic particles for the detection of multiple cancer biomarkers in patient serum by conjugating antibodies on GMR chips.

Awards and Honors:

- Svetlana Bershadsky Community Award (2017)
- Best Poster Award – SynBERC Fall Symposium (2014)
- NSF GRFP honorable mention (2014)

Certifications:

Professional Development (UCI GPS-BIOMED)
Business of Industry (SciPhD)
Public Speaking (Activate to Captivate)
Mentoring (UCI Mentor Excellence Program)

Mentoring:

Undergraduate students:
2016: Jimmy Duong (Professor Chang C. Liu's Laboratory)
2013-2015: Justin Hitomi (Professor Chang C. Liu's Laboratory)
2011-2012: Alex Salaza (Professor Nader Pourmand's Laboratory)
2011-2012: Jose Morales (Professor Nader Pourmand's Laboratory)

Laboratory Technician:
2017: Muaeen Obedi (Professor Chang C. Liu's Laboratory)

Professional Organizations:

UCI GPS-BIOMED (NIH-funded professional development program) Member
Cheeky Scientist Association Member
UCI GABES (Graduate Association for Biomedical Engineering Students) Co-President

Publications:

Paolo Actis, Boaz. Vilozny, Adam Ronald Seger, **Xiang Li**, Marguerite Rinaudo, Nader Pourmand. Voltage-Controlled Metal Binding on Polyelectrolyte-Functionalized Nanopores. *Langmuir* **27,** 6528-6533 (2011).

Xiang Li, Chang C. Liu. Biological Applications of Expanded Genetic Codes. *ChemBioChem* **15,** 2335-2341 (2014).

Roy N. Pruitt, Benjamin Schwessinger, Anna Joe, Nicholas Thomas, Furong Liu, Markus Albert, Michelle R. Robinson, Leanne Jade G. Chan, Dee Dee Luu, Huamin Chen, Ofir Bahar, Aralan

Daudi, David De Vleesschauwer, Daniel Caddell, Weiguo Zhang, Xiuxiang Zhao, **Xiang Li**, Joshua L. Heazlewood, Deling Ruan, Dipali Majumder, Mawsheng Chern, Hubert Kalbacher, Samriti Midha, Prabhu B. Patil, Ramesh V. Sonti, Christopher J. Petzold, Chang C. Liu, Jennifer S. Brodbelt, Georg Felix, Pamela C. Ronald. The rice immune receptor XA21 recognizes a tyrosine-sulfated protein from a Gram-negative bacterium. *Science Advances* **1,** e1500245 (2015).

Benjamin Schwessinger, **Xiang Li**, Thomas L. Ellinghaus, Leanne Jade G. Chan, Tony Wei, Anna Joe, Nicholas Thomas, Roy Pruitt, Paul D. Adams, Maw Sheng Chern, Christopher J. Petzold, Chang C. Liu, Pamela C. Ronald. A second-generation expression system for tyrosine sulfated proteins and its application in crop protection. *Integrative Biology* **8**, 542-545 (2016).

Xiang Li, Chang C. Liu. Site-specific incorporation of sulfotyrosine using an expanded genetic code. *Methods in Molecular Biology* **1728**, 191-200 (2018).

Xiang Li, Justin Hitomi, Chang C. Liu. Characterization of a sulfated anti-HIV antibody using an expanded genetic code. **(In preparation)**

Presentations:

Oral Presentations:

Xiang Li, "Unnatural antibodies against HIV", UCI AGS Symposium, UC Irvine, 2016.

Xiang Li, "Outsmart HIV using unnatural antibodies", UC Grad Slam, UC Irvine, 2016.

Xiang Li, "Engineered antibodies against HIV", AGS & GPS Biomed Pitch competition, UC Irvine, 2016.

Xiang Li, Chang C. Liu. Directed evolution of anti-HIV antibodies using an expanded genetic code. 15th Annual UC Systemwide Bioengineering Symposium, UC Irvine, 2014.

Poster Presentations:

Xiang Li, Chang C. Liu. Characterization of Sulfated Anti-HIV Antibodies Using an Expanded Genetic Code. SynBERC Fall Symposium, Northwestern University, 2016.

Xiang Li, Justin Hitomi, Chang C. Liu. Antibody Evolution Using an Expanded Genetic Code. SynBERC Spring Symposium, UC Berkeley, 2015.

Xiang Li, Justin Hitomi, Chang C. Liu. Directed Evolution of HIV Antibodies Using an Expanded Genetic Code. SynBERC Spring Symposium, UC Berkeley, 2014.

Xiang Li, Justin Hitomi, Chang C. Liu. Directed Evolution of HIV Antibodies Using an Expanded Genetic Code. SynBERC Fall Symposium, MIT, 2013.

Acknowledgements

I can't thank my academic advisor Dr. Chang C. Liu enough for all of his guidance and encouragements throughout my graduate studies. Chang has invested so much of his time in my development as a scientist, writer, and communicator; he caters his mentorship style to fit my needs, constantly challenges me to improve, and has been extremely patient and encouraging when I struggled. I also really appreciate the fact that Chang genuinely cares about every one of his students, not just about their research progress but also their well-being. I couldn't have asked for a better mentor for my time in graduate school.

I would like to thank my committee members, Dr. Wendy Liu, Dr. Jennifer Prescher, Dr. Elliot Hui, and Dr. Han Li, for their constructive feedback and advise on my thesis. I would also like to thank my collaborators: Dr. Pamela Ronald and Dr. Benjamin Schwessinger for the discovery of RaxX peptide, Dr. Matthew Gardner, Justin Hitomi, and Muaeen Obedi for the anti-HIV antibody projects, and Dr. Benjamin Katz for his assistance on mass spectrometry.

I am extremely grateful to have worked with talented and caring labmates. Whether it's talking about science, watching sports games in lab, or biking to San Diego on the weekends, we have always enjoyed each other's company and personal quirks. Even though we spent long nights and weekends in lab, we have always found ways to make fun a milestone.

The class of 2012 BME friends I have bonded with over the past few years have been my strongest support group throughout all of my highs and lows. Our first year together struggling with course work and partying outside of lab was the best year of my life so far. I really appreciate their lifelong friendships and look forward to going on many more awesome adventures with them.

Finally, I have the most amazing family. My parents sacrificed everything to move to the United States for me to receive the best education. Thank you for your sacrifices, love and support throughout all of my life. I would also like to thank my brother for making me smile on days when I didn't feel like I deserved it and for being one of my best friends. I hope that I have shown you that you can achieve your dreams through perseverance and hard-work.

Abstract

An expanded genetic code for the characterization and directed evolution of tyrosine-sulfated proteins

By

Xiang Li

Doctor of Philosophy in Biomedical Engineering

University of California, Irvine, 2018

Assistant Professor Chang C. Liu, Chair

Tyrosine sulfation, a post-translational modification that enhances extracellular protein-protein binding, plays a crucial role in various bacterial and viral infections. However, studying tyrosine-sulfated proteins can be difficult, because tyrosine sulfation of proteins often leads to mixtures of different sulfation patterns. In addition, efforts to engineer tyrosine-sulfated proteins is limited, because sulfation of the correct tyrosine is coupled to its local amino acid sequence contexts, termed sulfation motifs, which cannot be mutated without changing the sulfation efficiency.

These complications can be resolved using an expanded genetic code system. A cell with an expanded genetic code for sulfotyrosine (sY) decouples sulfation motifs from the expression of homogenously sulfated proteins; this is accomplished through the site-specific incorporation of sY as a 21st amino acid at amber stop codons. Here, we use an upgrade expanded genetic code system to express high amounts of tyrosine-sulfated proteins for several applications. First, we expressed homogenously sulfated, bacterial RaxX peptide in order to validate its role in triggering immune response in rice plants. Second, we expressed all 32 possible sulfoforms of E51, one of the best

sulfated anti-HIV antibodies at neutralizing HIV infections, and characterized their ability to bind to HIV's coat protein, gp120. We found several E51 sulfoforms that bind to gp120 with similar affinities as naturally sulfated E51 and determined key E51 residues for binding gp120. Lastly, we conducted multiple directed evolution experiments to improve the gp120-binding of E51 by repurposing E51's sulfation motifs with residues that bind to gp120.

Chapter 1.Introduction

1.1. Introduction

The universal genetic code encodes for 20 amino acids that serve as the building blocks of proteins. To access functional groups beyond those of the 20 amino acids, cells employ enzymes to post-translationally modify protein side chains with new chemical groups. These post-translational modifications (PTMs) are involved in regulation of gene expression[1–3], cellular signaling pathways[4–6], and extracellular interactions[5,7]. PTMs occur when enzymes locate and modify the side chain of a target residue by recognizing its local sequence context. For example, tyrosine sulfation – a PTM that alters extracellular protein-protein interactions and is best known for its crucial role in HIV infection – occurs when tyrosylprotein sulfotransferases (TPSTs) recognize a tyrosine surrounded by tyrosine sulfation motifs consisting of small or acidic residues[7].

Although PTMs expand the functional repertoire of proteins without rewriting the genetic code, there are several challenges in studying and engineering proteins with PTMs. First, some PTM processes such as glycosylation and tyrosine sulfation yield proteins with heterogenous PTMs, which complicates the characterization of functional roles of individual PTMs and occasionally reduces protein activity[8,9]. In the case of tyrosine sulfation, TPSTs recognize a rather broad set of sequence residues compared to other PTMs. Thus, when a target region contains multiple tyrosines, the resulting protein is a mixture composed of isoforms with different number of sulfated tyrosines[9]. Second, many PTM processes such as tyrosine sulfation are absent in standard bacterial cells, which are often the most economical protein expression strains and are ideal for the optimization of protein function through techniques such as phage display. Third, the

sequence constraint of PTM sequence motifs limits the evolution of proteins with PTMs, and it is possible that PTM sequence motifs themselves could contribute additional protein function if allowed to evolve more freely. Finally, it is extremely difficult to engineer existing enzymes and biosynthesis pathways to modify proteins with PTM analogues or novel functional groups *in vivo*.

Expansion of the genetic code provides solutions to studying and engineering proteins with PTMs. In a cell with an expanded genetic code, codons that do not encode for one of the 20 amino acids ($codon_{BL}$) are repurposed to encode non-canonical amino acids (ncAA) with user-defined functional groups[10]. To incorporate a ncAA at $codon_{BL}$, aminoacyl-tRNA synthetase (aaRS) and tRNA from a distant species (archaeal aaRS/tRNA pairs for bacterial hosts; bacterial aaRS/tRNA pairs for eukaryotic hosts) are engineered to recognize the ncAA and decode $codon_{BL}$, respectively, and expressed in a desired host cell. The new translation machinery is chosen and engineered to prevent the imported aaRS from charging ncAA onto any of the host's endogenous tRNAs and to prevent endogenous aaRSs from charging any of the natural amino acids onto the imported tRNA, ensuring that the incorporation of ncAA only occurs at $codon_{BL}$. Through this orthogonal decoding of a $codon_{BL}$ with a ncAA of interest, new functional groups can be site-specifically introduced into proteins, resulting in homogenous products without dependency on special enzymes and sequence contexts surrounding the $codon_{BL}$. Furthermore, owing to the ease of engineering the aaRS to accommodate a plethora of ncAAs with unique side chains, the genetic codes of both bacterial and eukaryotic cells have been expanded to incorporate over 100 different ncAAs including various PTMs[11].

1.2. Overview of dissertation

My dissertation describes the use of an upgraded expanded genetic code to study and evolve tyrosine-sulfated proteins. Chapter 2 provides an overview of an expanded genetic code

and discusses its biological applications including studying PTMs and protein evolution. Chapter 3 describes an optimized method to incorporate sulfotyrosine (sY) into proteins. Chapter 4 details our upgraded expanded genetic code system for the expression of a sulfated bacterial peptide, RaxX, in high purity and validation of its ability to trigger rice plant immunity. Chapter 5 describes the incorporation of up to five sYs into a high-neutralizing sulfated anti-HIV antibody, E51, for the expression and systematic characterization of all 32 possible sulfoforms of E51. We used these homogenously sulfated E51s to determine key residues for binding to HIV's coat protein gp120, as naturally sulfated E51 is composed of a mixture of different sulfoforms. Chapter 6 details our hypothesis that E51's sulfation motifs are suboptimal for binding gp120 and our attempts to discover sulfated antibodies by replacing sulfation motifs with residues that bind to gp120. After multiple selections, we conclude that E51's sulfation motifs sufficiently contribute to gp120-binding and highlight different phage display selection strategies that are synergistic with an expanded genetic code.

1.2. References

1. Elsässer, S. J., Ernst, R. J., Walker, O. S. & Chin, J. W. Genetic code expansion in stable cell lines enables encoded chromatin modification. *Nat. Methods* **advance on,** (2016).
2. Strahl, B. D. & Allis, C. D. The language of covalent histone modifications. *Nature* **403,** 41–45 (2000).
3. Heintzman, N. D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459,** 108–112 (2009).
4. Chen, Z. J. & Sun, L. J. Nonproteolytic Functions of Ubiquitin in Cell Signaling. *Mol. Cell* **33,** 275–286 (2009).
5. Boscher, C., Dennis, J. W. & Nabi, I. R. Glycosylation, galectins and cellular signaling.

Curr. Opin. Cell Biol. **23,** 383–392 (2011).

6. Olsen, J. V. *et al.* Global, In Vivo, and Site-Specific Phosphorylation Dynamics in Signaling Networks. *Cell* **127,** 635–648 (2006).

7. Stone, M. J., Chuang, S., Hou, X., Shoham, M. & Zhu, J. Z. Tyrosine sulfation: an increasingly recognised post-translational modification of secreted proteins. *N. Biotechnol.* **25,** 299–317 (2009).

8. Jefferis, R. Glycosylation as a strategy to improve antibody-based therapeutics. *Nat. Rev. Drug Discov.* **8,** 226–234 (2009).

9. Seibert, C., Cadene, M., Sanfiz, A., Chait, B. T. & Sakmar, T. P. Tyrosine sulfation of CCR5 N-terminal peptide by tyrosylprotein sulfotransferases 1 and 2 follows a discrete pattern and temporal sequence. *Proc. Natl. Acad. Sci. U. S. A.* **99,** 11031–6 (2002).

10. Li, X. & Liu, C. C. Biological Applications of Expanded Genetic Codes. *Chembiochem* 1–8 (2014). doi:10.1002/cbic.201402159

11. Liu, C. C. & Schultz, P. G. Adding new chemistries to the genetic code. *Annu. Rev. Biochem.* **79,** 413–44 (2010).

Chapter 2. Biological Applications of Expanded Genetic Codes

Xiang Li[1] and Chang C. Liu[1,2,3]

1. Department of Biomedical Engineering, University of California, Irvine
2. Department of Chemistry, University of California, Irvine
3. Department of Molecular Biology and Chemistry, University of California, Irvine

ChemBioChem **15,** 2335-2341 (2014).
DOI: 10.1002/cbic.201402159

Abstract

Substantial efforts in the past decade have resulted in the systematic expansion of genetic codes, allowing for the direct ribosomal incorporation of ~100 unnatural amino acids into bacteria, yeast, mammalian cells, and animals. Here, we illustrate the versatility of expanded genetic codes in biology and bioengineering, focusing on the application of expanded genetic codes to problems in protein, cell, synthetic, and experimental evolutionary biology. As the expanded genetic code field continues to develop, its place as a foundational technology in the whole of biological sciences will solidify.

2.1. Introduction

With few exceptions, nature uses a 20 amino acid genetic code for protein synthesis. While the question "why 20?" remains largely unanswered and perhaps unanswerable,[1–3] the related question "is it possible to go beyond 20?" has been successfully addressed with the synthetic expansion of genetic codes for the direct translational incorporation of custom unnatural amino acids (uAAs).[4] To date, ~ 100 uAAs representing physicochemical properties not found in the natural 20 amino acid repertoire, have been added to the genetic codes of bacteria,[5] yeast,[6] mammalian cells,[7,8] and animals.[9–11]

The synthetic addition of uAAs to the genetic code follows a systematic method – reviewed in detail elsewhere[4] – that has become streamlined and highly accessible in the last few years. First, one hijacks a "blank" codon ($codon_{BL}$) – usually the amber stop codon, UAG – to specify the desired uAA; second, one adopts an engineered orthogonal aminoacyl-tRNA synthetase (aaRS)/tRNA pair that decodes $codon_{BL}$ and does not crossreact with host aaRS/tRNA pairs; and third, one evolves the orthogonal aaRS/tRNA pair, using a well-established positive/negative selection system, toward specificity for the uAA of interest. After this process, the resulting aaRS/tRNA pair achieves the $codon_{BL}$-templated incorporation of the desired uAA in a manner identical to the genetic specification of natural amino acids, thus seamlessly expanding the genetic code of the host organism.

This ability to incorporate custom uAAs into proteins *in vivo* represents an unprecedented level of chemical and synthetic control over biology's workhorse macromolecule. Expanded genetic codes therefore facilitate a versatile array of applications in the biological sciences. In this review, we discuss four application areas for expanded genetic codes in biology. First among these

is protein biology, where the primary challenge is often the site-specific introduction of new chemistry, ranging from biophysical probes to biologically important posttranslational modifications, into proteins. Though chemical modification, solid-phase peptide synthesis, native chemical ligation, and expressed protein ligation have made significant strides toward the production of specialized proteins, expanded genetic codes yield site-specifically modified proteins simply through recombinant expression. They therefore offer a degree of efficiency and ease unavailable in other methods. Second among application areas is cell biology. Expanded genetic codes access new protein chemistries *in vivo*. Therefore, expanded genetic codes can modify and probe cellular processes under native conditions, with minimal disruption to the living cell. Third, we argue that expanded genetic codes act both as motivation and as an enabling tool for synthetic biology. This is because uAA incorporation – especially the simultaneous addition of multiple uAAs in response to multiple codon$_{BL}$s – has challenged the synthetic biology community to dramatically redesign organisms through systematic ribosome engineering,[12–15] genome-wide codon replacement,[16–18] and large-scale modification of codon usage.[19] At the same time, expanded genetic codes have been repurposed for gene regulation by uAA inputs in a way that can achieve forward engineering of custom gene regulatory functions.[20] Finally, we discuss the promising area of evolution with uAAs where new chemistries enter into the process of mutation, amplification, and functional selection. [21–23] This range of applications outlines the far-reaching implications and utility that expanded genetic codes have for the whole of biology.

2.2. Protein Biology

The modern study of proteins relies on two fundamental technologies: mutagenesis and recombinant expression. The former allows researchers to change, through simple manipulations of the corresponding genetic material, amino acid sequences of natural proteins; and the latter

allows for their efficient production and purification. However powerful, the degree to which these technologies facilitate the study and modification of protein structure and function is limited by the chemistries available in the natural 20 amino acids. With expanded genetic codes, this limitation is removed.

2.2.1 Therapeutic Proteins through Bioconjugation

One clear area that realizes the resulting potential is the production of therapeutic proteins through bioorthogonal conjugation. Here, the incorporation of uAAs with reactive functional groups becomes the critical first step in the custom site-specific derivatization of proteins for therapeutic purposes. For example, Cho *et al.* reported the recombinant expression of human growth hormone (hGH) containing a site-specifically incorporated para-acetylphenylalanine (pAcF), which served as a chemical handle for conjugation to polyethylene glycol (PEG).[24] The resulting homogeneously mono-PEGylated hGH demonstrated favorable pharmacodynamics and is being developed clinically. Several other ongoing efforts in this area may share similar success for therapy. These include the conjugation of small molecules onto antibodies to recruit T cells to cancer markers,[25] the attachment of charged polymers onto antibodies to promote siRNA delivery into target cells,[26] and the site-specific conjugation of oligonucleotides, facilitating the generation of antibody dimers and multimers.[27] In all these cases, conjugation is facilitated by a site-specifically incorporated bioorthogonal uAA, and therefore yields homogeneous therapeutic proteins rather than mixtures that are less potent and more sensitive to production conditions.

2.2.2 Addressing Posttranslationally Modified Proteins

Another area that highlights the potential of mutagenesis and recombinant expression with expanded genetic codes is the study of posttranslationally modified proteins. Posttranslational modifications (PTMs) impart an additional layer of control over protein function beyond expression, often playing central roles in binding, catalysis, signaling, and epigenetics. However,

the traditional study of posttranslationally modified proteins faces major challenges because the fundamental methods of mutagenesis and recombinant expression upon which protein biology relies do not easily address PTMs. These challenges are resolved by expanding genetic codes for the translational incorporation of uAAs corresponding to PTMs.

The first complete demonstration of this idea came through a number of studies on tyrosine sulfation. Sulfation is a ubiquitous PTM found in higher eukaryotes whereby a sulfate group is added to tyrosine, functioning to strengthen protein-protein interactions involved in processes ranging from clotting to chemokine signaling.[28] The study of sulfation, however, is limited by organismal constraints (sulfation does not occur in the common lab microbes *Escherichia coli* or *Saccharomyces cerevisiae*), homogeneity constraints (the enzymes responsible for sulfation are expressed only in certain parts of the cell), and sequence constraints (only specific amino acid motifs get sulfated). By expanding the genetic code of *E. coli* for the direct incorporation of sulfotyrosine, these constraints have been removed.[29] For example, the leech protein sulfohirudin is a therapeutically important anticoagulant, but is clinically used only in a lower-activity unsulfated recombinant form (desulfohirudin), because extracting sulfohirudin from leeches is highly inefficient and yields various isoforms. *E. coli* with their genetic codes expanded for sulfotyrosine incorporation allowed the production of sulfohirudin recombinantly, in yields viable for therapeutic development.[29] This recombinant sulfohirudin, which displays a 10-fold enhancement in affinity for thrombin compared to desulfohirudin, was also used to solve the crystal structure of the sulfohirudin-thrombin complex, which shows the structural contribution of tyrosine-sulfate to sulfohirudin's activity.[30] In other instances, Liu, Schultz *et al.* were able to conduct studies on sulfated proteins that required mutagenesis beyond the sequence motifs normally required for sulfation.[21,31] Taken together, these experiments illuminate new paths for

the study and application of sulfated proteins and more generally, outline a set of principles for addressing PTMs using expanded genetic codes.

These principles extend to PTMs beyond tyrosine sulfation, including the important class of lysine PTMs central to eukaryotic biology and histone-based epigenetics. For example, Neumann *et al.* expanded the genetic code of *E. coli* for the direct incorporation of acetyllysine, providing a recombinant expression system for site-specifically acetylated proteins.[32] This system has been used to reveal the role of Lys120 acetylation in modulating p53's DNA-binding specificity,[33] to kinetically and structurally characterize the regulatory roles of cyclophilin A acetylation in immunosuppression and HIV isomerization,[34] to dissect the effects of His3/Lys56 acetylation in nucleosome organization,[35] to understand the function of His4/Lys16 acetylation in gene silencing by interaction with Sir proteins,[36] and to analyze the change in global gene expression in response to acetylation of six lysines in histone H3 *in vivo*[37]. Several efforts have also aimed to expand genetic codes for the site-specific incorporation of methyllysines, by incorporating a precursor methyllysine that can be deprotected with light,[38,39] pH,[40] or transition metal chemistries[41]. In model studies, these expanded genetic codes have been used to test the effects of monomethylation or dimethylation of His3/Lys9 in the interaction with a known partner, heterochromatin protein 1.[40,42] Finally, expanded genetic code have been developed to address ubiquitination, another important lysine modification, through the site-specific incorporation of 1,2-aminothiols that can be ligated to ubiquitin.[43] A variation of this approach that employs site-specifically protected lysines, introduced as uAAs, allowed for the synthesis of an atypical diubiquitin, elucidation of its structure, and the discovery that TRABID has preferential deubiquitinase activity on the atypical Lys29-linked ubiquitin.[44] This growing set of genetically

encoded lysine modifications may contribute to the larger goal of understanding and possibly controlling epigenetic states.

A valuable feature of genetically encoded lysine modifications is that they predominantly rely on engineered variants of aaRS/tRNA pairs specific for pyrollysine (PylRS/$tRNA^{Pyl}$). Since the parent PylRS/$tRNA^{Pyl}$ pairs have been shown to be compatible in a wide range of hosts, uAAs incorporated through engineered variants automatically function in *E. coli* and mammalian cells.[45] Therefore, expanded genetic codes for acetyllysine and variously caged methyllysines can be directly ported into mammalian cells, providing a platform for the study of lysine modification on proteins *in vivo*. Indeed the use of expanded genetic codes for recombinant expression and mutagenesis with a growing list of PTMs, including the phosphoserine and phosphothreonine modifications critical to kinase signaling,[46–48] has become a mainstay of protein biology.

2.3. Cell Biology

The development of genetically encoded GFPs revolutionized biology by providing a way to image specific proteins non-invasively within living cells and tissues. Since then, a major thrust in cell biology has been the elaboration of new genetically encodable chemistries for observing, perturbing, and controlling proteins in vivo. Toward this end, expanded genetic codes that site-specifically incorporate uAAs acting as small sensitive fluorophores, context-dependent IR, PET, or NMR probes, and photoactive switches for optogenetics have been developed.[4] This has led to a large suite of applications for uAAs in cell biology.

2.3.1. Light-activated Crosslinking and Control

Among the most widely applied uAAs in cell biology are ones containing benzophenone, azides, and diazirines.[49–54] These uAAs are incorporated site-specifically into proteins in order to trap, through UV-triggered covalent crosslinking, specific protein-protein interactions and protein

conformational changes. Numerous examples of this use are available and include studies on chaperone interactions with substrates,[55] protein interactions that regulate the cell cycle,[56] nascent peptide recognition inside the ribosome tunnel,[57] spontaneous conformational changes in RNA polymerase,[58] and protein interactions at various intracellular and cell membrane locations in E. coli, yeast, and mammalian cells.[53,59,60] In many of these examples, crosslinking was triggered in vivo, thus favoring the isolation of native intra- or intermolecular interactions in their biologically most relevant conditions.

Photoactive uAAs can also be used to modulate protein function directly inside cells in real time (Figure 1). For example, Lemke et al. added 4,5-dimethoxy-2-nitrobenzylserine, a photocaged serine, to the genetic code of Saccharomyces cerevisiae and incorporated this uAA into various serine phosphorylation sites in the transcription factor Pho4.[61] Upon decaging through laser irradiation, serines were revealed and the kinetics of their subsequent phosphorylation and resulting localization were studied. More recently, Gautier et al. encoded a photocaged lysine in mammalian cells and incorporated it into a nuclear localization signal to mask its endogenous activity.[62] Using temporally controlled photolysis of the masked nuclear localization tag, the authors were able to observe the kinetics of nuclear import in both a model system and for the tumor suppressor p53.

The same caged lysine was applied to create photoactivatable kinases.[62] This was demonstrated in a number of experiments on the MAPK pathway, in which the active site lysine of the MAPKK MEK1 was replaced by the photocaged lysine, disabling MEK1's catalytic activity. The resulting photoactivatable MEK1 was used to reveal several elements of the MAPK pathway, including the kinetics of phosphorylation-dependent ERK2 nuclear import. Other examples involving photoactive uAAs include the creation of a photoactivatable GFP,[62] site-specific

photocleavage of protein backbones,[63] and light-activated transcription through a photoactivatable RNA polymerase.[64] In short, genetically encoded photoactive uAAs constitute an important set of optogenetic tools for the spatiotemporal control of cellular processes.

2.3.2. Fluorescent Reporters

Complementary to photoreactive uAAs are ones that facilitate observation and imaging of cellular processes. Though this task has traditionally been the domain of GFP and its variants, genetically encoded fluorescent uAAs might offer several advantages over GFP. For example, fluorescent uAAs, owing to their small size, may be used to label proteins with minimal structural perturbation and may report on more intricate biochemical and biophysical events including protein conformational and local environmental changes. In addition, a wide range of fluorescence properties might be accessed through uAAs since there seems to be no hard limit on the types of small fluorophores that can be genetically encoded as uAAs.

Several groups have begun to capitalize on the opportunities afforded by genetically encoded fluorescent uAAs. For example, Lee, Chatterjee, Schultz et al. have added an environmentally sensitive fluorescent probe, 3-(6-acetylnaphthalen-2-ylamino)-2-aminopropanoic acid (Anap), to the genetic codes of yeast and mammalian cells and have used it both to detect conformational changes that occur during ligand binding by QBP (glutamine binding protein) and as a tool to visualize subcellular localization of proteins in living cells.[65,66] Others have added a coumarin-containing fluorescent uAA (CouAA) to the genetic code of E. coli and have used it to visualize contractile ring formation by the protein FtsZ during cytokinesis.[67,68] In the case of FtsZ and likely several other cytoskeletal proteins, fusion to GFP impairs cellular function, but the incorporation of the small CouAA into FtsZ did not disturb its wild-type function and allowed visualization of FtsZ subcellular localization.

Yet another genetically encoded fluorescent uAA for studying yeast and mammalian cell biology is 2-amino-3-(5-(dimethylamino)naphthalene-1-sulfonamido) propanoic acid (DanAla).[68] In an impressive study, Shen et al. demonstrated the genetic incorporation of DanAla into proteins expressed in neural stem cells that could subsequently differentiate into neural progenies.[69] Specifically, DanAla was incorporated into a voltage-sensitive domain (VSD) taken from a voltage-dependent membrane lipid phosphatase that undergoes conformational changes upon membrane polarization in neurons. Since the fluorescence of DanAla is highly sensitive to local polarity, DanAla acted as an optical reporter for conformational changes in VSD during membrane depolarization. The continued development of genetically encoded fluorescent amino acids for observing cellular processes in real time, especially at the resolution of conformational changes in single proteins, will lead to many new advances in cell biology.

2.3. Synthetic Biology

Synthetic biology applies the design principles of part modularity, orthogonality, independence, scalability, and portability to the composition of new biological functions.[70] Expanded genetic codes, which are achieved by engineering modular, orthogonal, and portable uAA-specific $codon_{BL}$/aaRS/tRNA sets that act independently in the context of natural codon/aaRS/tRNA sets, rely on similar principles. Therefore, there has been a dynamic exchange of ideas between the two fields in which synthetic biology principles have been applied to expanded genetic codes and vice-versa.

A perfect example of the former is represented in efforts to simultaneously encode multiple uAAs. This requires multiple independent $codon_{BL}$s, sets of mutually orthogonal aaRS/tRNA pairs, and might also benefit from engineered ribosomes that accommodate new $codon_{BL}$/tRNA recognition modes. Each of these areas has been addressed with scalable synthetic biology

strategies. For example, a number of labs have pursued the use of triplet codon$_{BL}$s beyond UAG as well as four-base codon$_{BL}$s to specify multiple uAAs, even using genome-wide engineering or synthesis efforts to remove codon redundancies from the natural genetic code (Figure 2).[16–19] Chin *et al.* have added to these efforts by developing orthogonal 30S subunit of ribosomes that recognize mRNAs with artificial ribosome binding sites (RBSs),[71] subsequently engineering one of these ribosomes to better accommodate four-base codon$_{BL}$s and their cognate tRNAs.[14] This allowed the efficient expression of proteins containing two different uAAs, incorporated by two mutually orthogonal aaRS/tRNA pairs, one of which recognizes a four-base codon$_{BL}$. To create a fully orthogonal ribosome-messenger system, Orelle *et al.* created a fully orthogonal ribosome by tethering the 30S and 50S subunits through rRNA rearrangement and linkage. This orthogonal ribosome, Ribo-T, potentially enables the engineering of the ribosomal-A site of the 50S subunit for uAA polymerization.[13] It might also be possible to borrow sense codons in specific contexts for uAA incorporation. For example, Bröcker *et al.* showed that the UGA codon, which specifies selenocysteine in a wide variety of organisms when present in a particular mRNA stem-loop structure, can be reassigned to 58 out of 64 of the sense anticodons.[72] Their work suggests the feasibility of reassigning sense codons for uAAs beyond selenocysteine.

In addition to new codon$_{BL}$s, the translational incorporation of multiple uAAs requires the establishment of mutually orthogonal aaRS/tRNA part sets, members of which can then be independently engineered to recognize distinct uAAs. This has been achieved either through finding disparate aaRS/tRNA pairs that naturally display mutual orthogonality[73,74] or through more general methods for engineering mutually orthogonal combinations of aaRS/tRNA pairs from parent pairs.[75,76] Taken together, these efforts represent both a significant example and goal of applied synthetic biology as they involve the creation of orthogonal part sets and their assembly

into custom genetic codes, ones that may eventually lead to the specification of synthetic polypeptides containing multiple or only uAAs in living systems.

The reverse direction, namely the application of expanded genetic codes to synthetic biology, has also been productive. For example, a major goal in synthetic biology is the predictable engineering of genetic switches that respond to user-defined inputs; these should facilitate the construction of custom cellular behaviors and allow synthetic and systems biologists to probe the principles of regulatory systems.[77] Towards this end, Liu *et al.* proposed the application of expanded genetic codes as a regulatory framework. This is based on the idea that codon$_{BLS}$ can be viewed as highly modular sensors for uAA inputs that the ribosome predictably integrates into the translation of a gene. Particular arrangements of codon$_{BLS}$ form different instructions for translational elongation that achieve the expression of output genes as a function of uAA inputs. To implement this framework, Liu *et al.* have engineered leader peptide elements, whose bacterial mechanisms couple translation of a short upstream open reading frame to transcription of regulated genes, into dedicated genetic switches that use codon$_{BLS}$ to specify custom regulatory functions.[20] These switches have been programmed into numerous uAA-induced ON and OFF behaviors, multi-input logic gates, and most recently, into higher-order regulatory functional forms with tunable sigmoidality and other complex properties (C. C. Liu, M. S. Samoilov, & A. P. Arkin, unpublished results). We believe that the success of these experiments will provide motivation to further explore the interface between expanded genetic codes and synthetic biology, both of which are rooted in a shared set of fundamental engineering principles.

2.4. Laboratory Evolution

The previous sections discuss the rational addition of uAAs into proteins and biological systems. An alternative and tantalizing possibility for genetically encoded uAAs is their integration

in evolution experiments, where the process of mutation, amplification, and selection determines how uAAs might contribute to new functions beyond those that can be engineered rationally (Figure 3). Several steps have been taken in this direction. For example, Liu *et al.* have developed a robust phage-based protein evolution system in which displayed protein libraries are generated in an expanded genetic code *E. coli* host that encodes 21 amino acids.[21] This system was used to demonstrate that expanded genetic codes can confer a selective advantage in protein evolution. In the process, several proteins whose functions depend on the novel chemistries of uAAs were evolved, including anti-gp120 antibodies that utilize sulfotyrosine for binding,[21,31] antibodies against a model glycan that use a boronate moiety to interact with diols,[78] zinc fingers that rely on iron instead of zinc for structure,[79] and metal-binding peptides that use uAAs with metal-chelating sidechains.[80] Ongoing studies include incorporating "chemical warheads" in the evolution of protease inhibitors, sulfotyrosine in the evolution of protein-protein interfaces, and metal-chelating uAAs in the evolution of new catalytic centers.

One may go beyond these directed evolution experiments and simply ask, in an open-ended manner, if uAAs can contribute to adaptation. Hammerling *et al.* studied exactly this idea, by passaging lines of T7 bacteriophage in the presence of an *E. coli* host with an expanded genetic code that incorporates 3-iodotyrosine (IodoY) at the amber stop codon.[22] After numerous serial transfers of the lytic phage, the frequency of mutants specifying IodoY in certain genes reached high levels. In one case, the IodoY mutation was beneficial: the Tyr39-to-IodoY mutation in T7 type II holin increased phage production over the wild-type version of the gene. The finding that phage adapt to expanded genetic code *E. coli* hosts by functionally using an uAA in their proteins suggests that uAAs may have many adaptive functions yet to be found. These, and the larger

implication that expanded genetic codes may be evolutionarily advantageous, are exciting avenues for experimental evolutionary biology.

2.5. Outlook

Through uAA protein mutagenesis, uAA probes of cellular function, the redesign and repurposing of genetic codes, and the incorporation of new chemistries into directed evolution platforms, the relatively young field of expanded genetic codes has already had an enormous impact on the control, design, and understanding of biology. Continued work in these areas and their extension into more organisms, including multicellular animals,[9,10] will solidify the position of expanded genetic codes as a foundational technology for the whole of biological sciences.

2.6. Acknowledgements

We would like to thank Prof. Dr. Peter Schultz for helpful comments and the University of California at Irvine for financial support.

2.7. References

1. Vetsigian, K., Woese, C. & Goldenfeld, N. Collective evolution and the genetic code. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 10696–701 (2006).

2. Lu, Y. & Freeland, S. On the evolution of the standard amino-acid alphabet. *Genome Biol* **7,** 102 (2006).

3. Weber, A. L. & Miller, S. L. Reasons for the occurrence of the twenty coded protein amino acids. *J. Mol. Evol.* **17,** 273–284 (1981).

4. Liu, C. C. & Schultz, P. G. Adding new chemistries to the genetic code. *Annu. Rev. Biochem.* **79,** 413–44 (2010).

5. Wang, L., Brock, a, Herberich, B. & Schultz, P. G. Expanding the genetic code of

Escherichia coli. *Science* **292,** 498–500 (2001).

6. Chin, J. W. *et al.* An expanded eukaryotic genetic code. *Science* **301,** 964–7 (2003).

7. Liu, W., Brock, A., Chen, S., Chen, S. & Schultz, P. G. Genetic incorporation of unnatural amino acids into proteins in mammalian cells. *Nat. Methods* **4,** 239–244 (2007).

8. Sakamoto, K. Site-specific incorporation of an unnatural amino acid into proteins in mammalian cells. *Nucleic Acids Res.* **30,** 4692–4699 (2002).

9. Greiss, S. & Chin, J. W. Expanding the genetic code of an animal. *J. Am. Chem. Soc.* **133,** 14196–14199 (2011).

10. Bianco, A., Townsley, F. M., Greiss, S., Lang, K. & Chin, J. W. Expanding the genetic code of Drosophila melanogaster. *Nat. Chem. Biol.* **8,** 748–50 (2012).

11. Parrish, A. R. *et al.* Expanding the genetic code of Caenorhabditis elegans using bacterial aminoacyl-tRNA synthetase/tRNA pairs. *ACS Chem. Biol.* **7,** 1292–1302 (2012).

12. Fried, S. D., Schmied, W. H., Uttamapinant, C. & Chin, J. W. Ribosome subunit stapling for orthogonal translation in E. coli. *Angew. Chemie - Int. Ed.* **54,** 12791–12794 (2015).

13. Orelle, C. *et al.* Protein synthesis by ribosomes with tethered subunits. *Nature* **524,** 119–124 (2015).

14. Neumann, H., Wang, K., Davis, L., Garcia-Alai, M. & Chin, J. W. Encoding multiple unnatural amino acids via evolution of a quadruplet-decoding ribosome. *Nature* **464,** 441–4 (2010).

15. Wang, K., Neumann, H., Peak-Chew, S. Y. & Chin, J. W. Evolved orthogonal ribosomes enhance the efficiency of synthetic genetic code expansion. *Nat. Biotechnol.* **25,** 770–7 (2007).

16. Ostrov, N. *et al.* Design, synthesis, and testing toward a 57-codon genome. *Science (80-.).* **353,** 819–822 (2016).

17. Isaacs, F., Carr, P., Wang, H. & Lajoie, M. Precise manipulation of chromosomes in vivo enables genome-wide codon replacement. *Science (80-.).* **333,** 348–353 (2011).

18. Lajoie, M. J. *et al.* Genomically Recoded Organisms Expand Biological Functions. *Science (80-.).* **342,** 357–360 (2013).

19. Lajoie, M. J. *et al.* Probing the Limits of Genetic Recoding in Essential Genes. *Science (80-.).* **342,** 361–363 (2013).

20. Liu, C. C., Qi, L., Yanofsky, C. & Arkin, A. P. Regulation of transcription by unnatural amino acids. *Nat. Biotechnol.* **29,** 164–8 (2011).

21. Liu, C. C. *et al.* Protein evolution with an expanded genetic code. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 17688–93 (2008).

22. Hammerling, M. J. *et al.* Bacteriophages use an expanded genetic code on evolutionary paths to higher fitness. *Nat. Chem. Biol.* 1–5 (2014). doi:10.1038/nchembio.1450

23. Liu, T. *et al.* Enhancing protein stability with extended disulfide bonds. *Proc. Natl. Acad. Sci.* **113,** 5910–5915 (2016).

24. Cho, H. *et al.* Optimized clinical performance of growth hormone with an expanded genetic code. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 9060–5 (2011).

25. Kim, C. H. *et al.* Bispecific small molecule – antibody conjugate targeting prostate cancer. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 17796–17801 (2013).

26. Lu, H. *et al.* Site-specific antibody-polymer conjugates for siRNA delivery. *J. Am. Chem. Soc.* **135,** 13885–13891 (2013).

27. Hutchins, B. M. *et al.* Site-specific coupling and sterically controlled formation of multimeric antibody fab fragments with unnatural amino acids. *J. Mol. Biol.* **406,** 595–603 (2011).

28. Kanan, Y. & Al-ubaidi, M. R. Tyrosine O Sulfation : An Overview. *JSM Biotechnol. Bioeng.* **1,** 1003–1007 (2013).

29. Liu, C. C. & Schultz, P. G. Recombinant expression of selectively sulfated proteins in Escherichia coli. *Nat. Biotechnol.* **24,** 1436–1440 (2006).

30. Liu, C. C., Brustad, E., Liu, W. & Schultz, P. G. Crystal structure of a biosynthetic sulfo-hirudin complexed to thrombin. *J. Am. Chem. Soc.* **129,** 10648–9 (2007).

31. Liu, C. C., Choe, H., Farzan, M., Smider, V. V & Schultz, P. G. Mutagenesis and evolution of sulfated antibodies using an expanded genetic code. *Biochemistry* **48,** 8891–8 (2009).

32. Neumann, H., Peak-Chew, S. Y. & Chin, J. W. Genetically encoding N(epsilon)-acetyllysine in recombinant proteins. *Nat. Chem. Biol.* **4,** 232–234 (2008).

33. Arbely, E. *et al.* Acetylation of lysine 120 of p53 endows DNA-binding specificity at effective physiological salt concentration. *Proc. Natl. Acad. Sci.* **108,** 8251–8256 (2011).

34. Lammers, M., Neumann, H., Chin, J. W. & James, L. C. Acetylation regulates cyclophilin A catalysis, immunosuppression and HIV isomerization. *Nat. Chem. Biol.* **6,** 331–337 (2010).

35. Neumann, H. *et al.* A Method for Genetically Installing Site-Specific Acetylation in Recombinant Histones Defines the Effects of H3 K56 Acetylation. *Mol. Cell* **36,** 153–163 (2009).

36. Oppikofer, M. *et al.* A dual role of H4K16 acetylation in the establishment of yeast silent chromatin. *EMBO J.* **30,** 2610–2621 (2011).

37. Elsässer, S. J., Ernst, R. J., Walker, O. S. & Chin, J. W. Genetic code expansion in stable cell lines enables encoded chromatin modification. *Nat. Methods* **advance on,** (2016).

38. Wang, Y.-S. *et al.* A genetically encoded photocaged Nε-methyl-l-lysine. *Mol. Biosyst.* **6,** 1557 (2010).

39. Groff, D., Chen, P. R., Peters, F. B. & Schultz, P. G. A genetically encoded e-N-Methyl lysine in mammalian cells. *ChemBioChem* **11,** 1066–1068 (2010).

40. Nguyen, D. P., Alai, M. M. G., Kapadnis, P. B., Neumann, H. & Chin, J. W. Genetically Encoding N E -Methyl- L -lysine in Recombinant Histones. *J. Am. Chem. Soc.* **2,** 14194–14195 (2009).

41. Ai, H., Lee, J. W. & Schultz, P. G. A method to site-specifically introduce methyllysine into proteins in E. coli. *Chem. Commun.* **46,** 5506 (2010).

42. Nguyen, D. P., Alai, M. M. G., Virdee, S. & Chin, J. W. Genetically directing ??-N, N-dimethyl-l-lysine in recombinant histones. *Chem. Biol.* **17,** 1072–1076 (2010).

43. Virdee, S. *et al.* Traceless and site-specific ubiquitination of recombinant proteins. *J. Am. Chem. Soc.* **133,** 10708–10711 (2011).

44. Virdee, S., Ye, Y., Nguyen, D. P., Komander, D. & Chin, J. W. Engineered diubiquitin synthesis reveals Lys29-isopeptide specificity of an OTU deubiquitinase. *Nat. Chem. Biol.* **6,** 750–7 (2010).

45. Chen, P. R. *et al.* A facile system for encoding unnatural amino acids in mammalian cells. *Angew. Chemie - Int. Ed.* **48,** 4052–4055 (2009).

46. Zhang, M. S. *et al.* Biosynthesis and genetic encoding of phosphothreonine through parallel selection and deep sequencing. *Nat. Methods* **14,** 729–736 (2017).

47. Rogerson, D. T. *et al.* Efficient genetic encoding of phosphoserine and its nonhydrolyzable analog. *Nat. Chem. Biol.* 1–11 (2015). doi:10.1038/nchembio.1823

48. Park, H.-S. *et al.* Expanding the genetic code of Escherichia coli with phosphoserine.

Science **333,** 1151–1154 (2011).

49. Chin, J. W., Martin, A. B., King, D. S., Wang, L. & Schultz, P. G. Addition of a photocrosslinking amino acid to the genetic code of Escherichia coli. *Proc. Natl. Acad. Sci.* **99,** 11020–11024 (2002).

50. Hino, N. *et al.* Protein photo-cross-linking in mammalian cells by site-specific incorporation of a photoreactive amino acid. *Nat. Methods* **2,** 201–206 (2005).

51. Hancock, S. M., Uprety, R., Deiters, A. & Chin, J. W. Expanding the genetic code of yeast for incorporation of diverse unnatural amino acids via a pyrrolysyl-tRNA synthetase/tRNA pair. *J. Am. Chem. Soc.* **132,** 14819–14824 (2010).

52. Ai, H. wang, Shen, W., Sagi, A., Chen, P. R. & Schultz, P. G. Probing Protein-Protein Interactions with a Genetically Encoded Photo-crosslinking Amino Acid. *ChemBioChem* **12,** 1854–1857 (2011).

53. Zhang, M. *et al.* A genetically incorporated crosslinker reveals chaperone cooperation in acid resistance. *Nat. Chem. Biol.* **7,** 671–677 (2011).

54. Chou, C., Uprety, R., Davis, L., Chin, J. W. & Deiters, A. Genetically encoding an aliphatic diazirine for protein photocrosslinking. *Chem. Sci.* **2,** 480–483 (2011).

55. Schlieker, C. *et al.* Substrate recognition by the AAA+ chaperone ClpB. *Nat. Struct. Mol. Biol.* **11,** 607–615 (2004).

56. Kimata, Y. *et al.* A mutual inhibition between APC/C and its substrate Mes1 required for meiotic progression in fission yeast. *Dev. Cell* **14,** 446–454 (2008).

57. Tagami, S. *et al.* Crystal structure of bacterial RNA polymerase bound with a transcription inhibitor protein. *Nature* **468,** 978–982 (2010).

58. Yap, M. N. & Bernstein, H. D. The Plasticity of a Translation Arrest Motif Yields

Insights into Nascent Polypeptide Recognition inside the Ribosome Tunnel. *Mol. Cell* **34,** 201–211 (2009).

59. Coin, I. *et al.* XGenetically encoded chemical probes in cells reveal the binding path of urocortin-i to CRF class B GPCR. *Cell* **155,** 1258–1269 (2013).

60. Chin, J. W. Reprogramming the genetic code. *Science (80-.).* **336,** 428–429 (2012).

61. Lemke, E. A., Summerer, D., Geierstanger, B. H., Brittain, S. M. & Schultz, P. G. Control of protein phosphorylation with a genetically encoded photocaged amino acid. *Nat. Chem. Biol.* **3,** 769–772 (2007).

62. Gautier, A. *et al.* Genetically encoded photocontrol of protein localization in mammalian cells - Supporting Information. *J. Am. Chem. Soc.* **16,** 1–19 (2010).

63. Peters, F. B., Brock, A., Wang, J. & Schultz, P. G. Photocleavage of the Polypeptide Backbone by 2-Nitrophenylalanine. *Chem. Biol.* **16,** 148–152 (2009).

64. Hemphill, J., Chou, C., Chin, J. W. & Deiters, A. Genetically encoded light-activated transcription for spatiotemporal control of gene expression and gene silencing in mammalian cells. *J. Am. Chem. Soc.* **135,** 13433–13439 (2013).

65. Chatterjee, A., Guo, J., Lee, H. S. & Schultz, P. G. A genetically encoded fluorescent probe in mammalian cells. *J. Am. Chem. Soc.* **135,** 12540–12543 (2013).

66. Lee, H. S., Guo, J., Lemke, E. A., Dimla, R. D. & Schultz, P. G. Genetic Incorporation of a Small, Environmentally Sensitive, Fluorescent Probe into Proteins in Saccharomyces cerevisiae. *J Am Chem Soc* **131,** 12921–12923 (2009).

67. Charbon, G. *et al.* Subcellular Protein Localization by Using a Genetically Encoded Fluorescent Amino Acid. *ChemBioChem* **12,** 1818–1821 (2011).

68. Wang, J., Xie, J. & Schultz, P. G. A genetically encoded fluorescent amino acid. *J. Am.*

Chem. Soc. **128,** 8738–8739 (2006).

69. Shen, B. *et al.* Genetically encoding unnatural amino acids in neural stem cells and optically reporting voltage-sensitive domain changes in differentiated neurons. *Stem Cells* **29,** 1231–1240 (2011).

70. Lucks, J. B., Qi, L., Whitaker, W. R. & Arkin, A. P. Toward scalable parts families for predictable design of biological circuits. *Curr. Opin. Microbiol.* **11,** 567–573 (2008).

71. Rackham, O. & Chin, J. W. Cellular logic with orthogonal ribosomes. *J. Am. Chem. Soc.* **127,** 17584–17585 (2005).

72. Bröcker, M., Ho, J. & Church, G. Recoding the Genetic Code with Selenocysteine. *Angew. Chemie ...* 319–323 (2014). doi:10.1002/anie.201308584

73. Wan, W. *et al.* A facile system for genetic incorporation of two different noncanonical amino acids into one protein in Escherichia coli. *Angew. Chemie - Int. Ed.* **49,** 3211–3214 (2010).

74. Anderson, J. C. *et al.* An expanded genetic code with a functional quadruplet codon. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 7566–71 (2004).

75. Chatterjee, A., Xiao, H. & Schultz, P. G. Evolution of multiple, mutually orthogonal prolyl-tRNA synthetase/tRNA pairs for unnatural amino acid mutagenesis in Escherichia coli. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 14841–6 (2012).

76. Neumann, H., Slusarczyk, A. L. & Chin, J. W. *De Novo* Generation of Mutually Orthogonal Aminoacyl-tRNA Synthetase/tRNA Pairs. *J. Am. Chem. Soc.* **132,** 2142–2144 (2010).

77. Lu, T. K., Khalil, A. S. & Collins, J. J. Next-generation synthetic gene networks. *Nat. Biotechnol.* **27,** 1139–1150 (2009).

78. Liu, C. C. *et al.* Evolution of Proteins with Genetically Encoded "Chemical Warheads" *J Am Chem Soc* 9616–9617 (2009). doi:10.1021/ja902985e

79. Kang, M. *et al.* Evolution of iron(II)-finger peptides by using a bipyridyl amino acid. *ChemBioChem* **15,** 822–825 (2014).

80. Day, J. W., Kim, C. H., Smider, V. V. & Schultz, P. G. Identification of metal ion binding peptides containing unnatural amino acids by phage display. *Bioorg. Med. Chem. Lett.* **23,** 2598–2600 (2013).

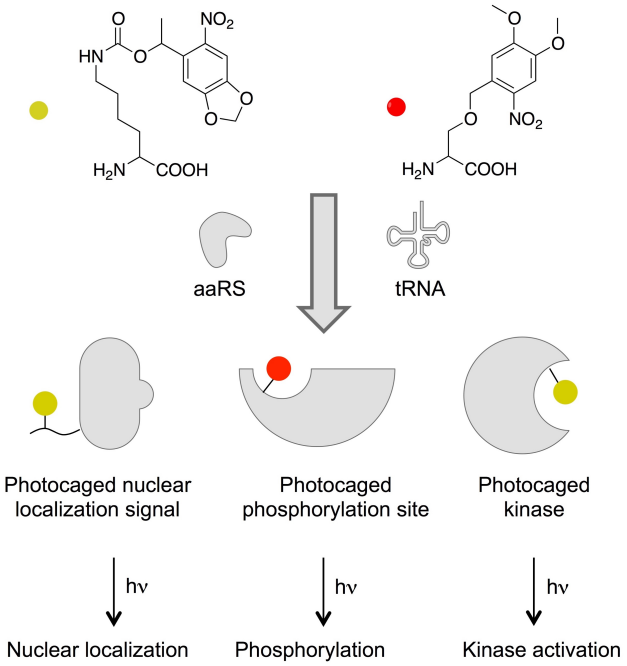

Figure 2.1. Light-activated control of biochemical pathways.

Using an expanded genetic code, photocaged serines and photocaged lysines are incorporated into target proteins. Under standard conditions, these proteins remain biologically inactive. UV-activation of photocaged proteins allows the controlled study of various biological pathways.

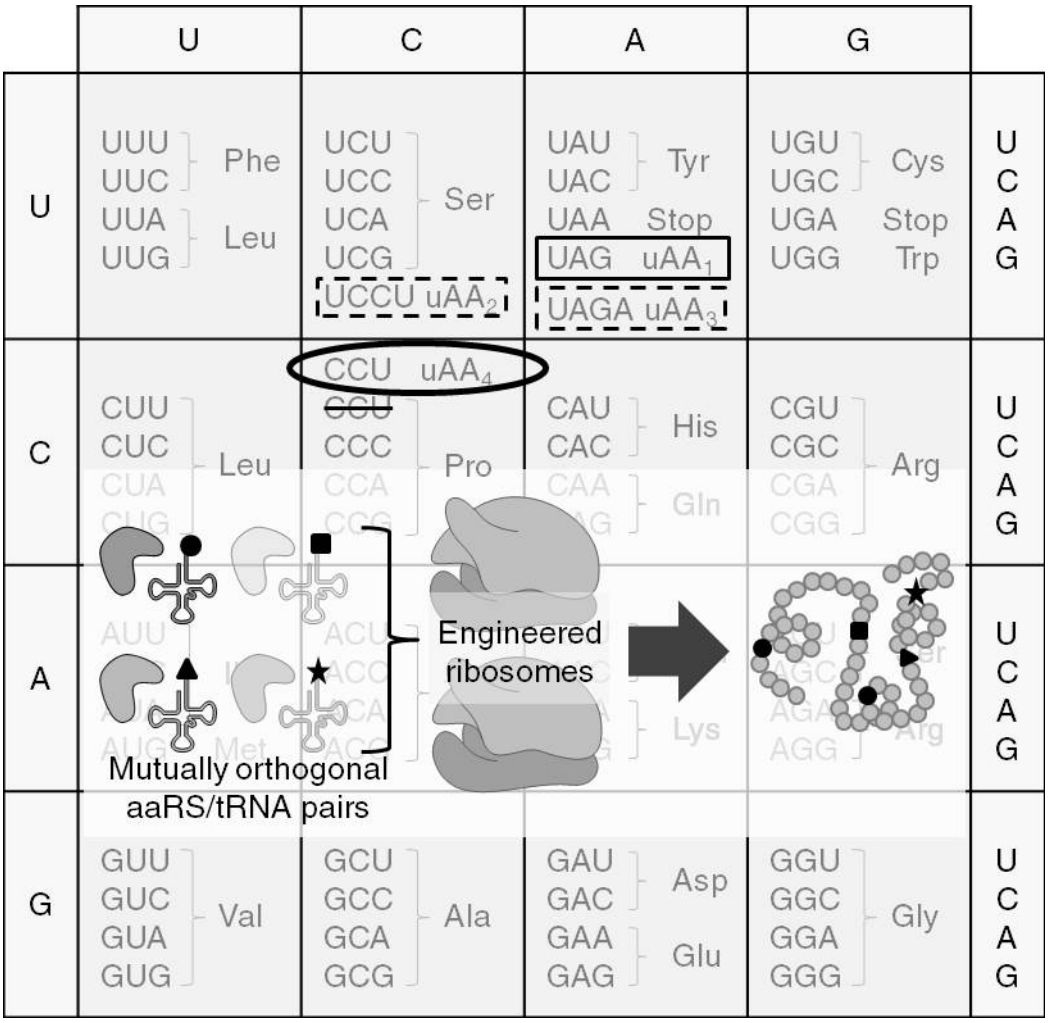

Figure 2.2. Synthetic expanded genetic code systems.

Evolved orthogonal 30S subunit ribosomes with high specificity for tRNAs that recognize three or four-base codons$_{BL}$ increase the efficiency of incorporating uAAs.[14,15] A wide range of mutually orthogonal aaRS/tRNA pairs have been engineered to incorporate specific uAAs at amber (box) and quadruplet (dashed box) codons.[73–76] Genome-wide removal of 13 redundant natural codons (e.g. circled), along with their respective tRNAs, will provide multiple codons$_{BL}$ for highly efficient uAA incorporation.[16–19]

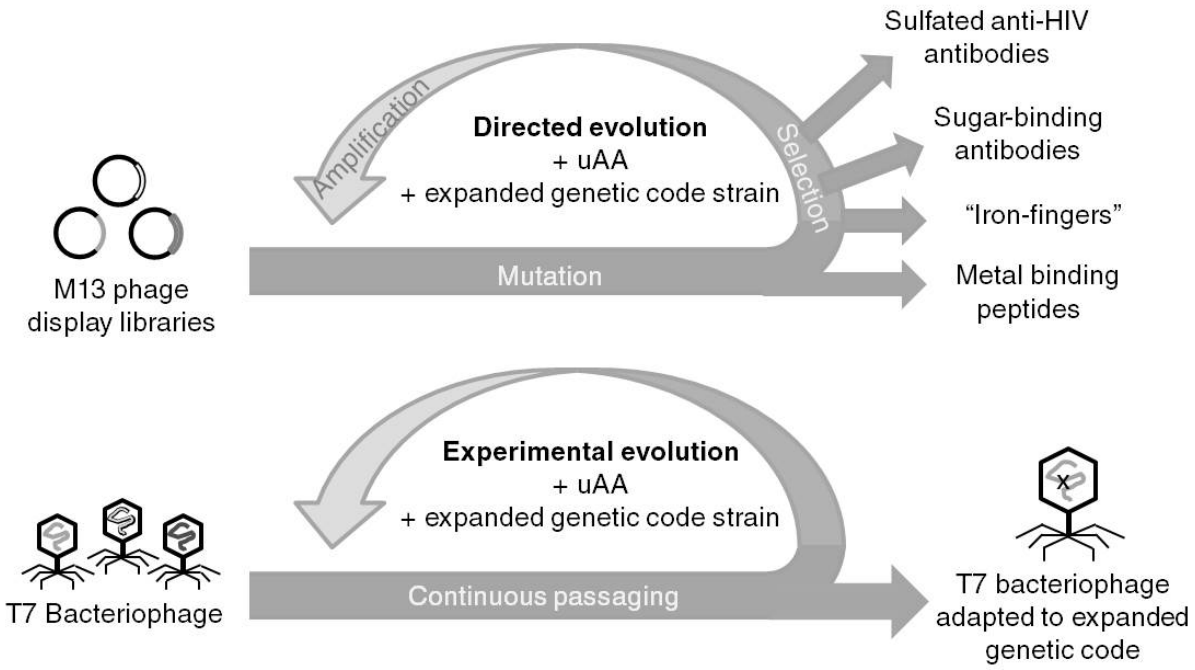

Figure 2.3. Laboratory evolution with expanded genetic codes.

Proteins selected from M13 phage display libraries exhibit unique biological activities and functions resulting from uAAs.[21,31,78–80] Likewise, the T7 bacteriophage has adapted to expanded genetic codes after many generations of experimental evolution.[22]

Chapter 3.Site-specific incorporation of sulfotyrosine using an expanded genetic code

Xiang Li[1] and Chang C. Liu[1,2,3]

1. Department of Biomedical Engineering, University of California, Irvine
2. Department of Chemistry, University of California, Irvine
3. Department of Molecular Biology and Chemistry, University of California, Irvine

Methods in Molecular Biology **1728**, 191-200 (2018).
DOI: 10.1007/978-1-4939-7574-7_12

Abstract

Tyrosine sulfation is an important posttranslational modification found in bacteria and higher eukaryotes. However, the chemical synthesis or expression of homogenously sulfated proteins is particularly difficult, limiting our study and application of tyrosine-sulfated proteins. With the recent development of genomically recoded organisms and orthogonal translation components, we can often treat otherwise posttranslationally-modified amino acids as unnatural amino acids (uAAs) encoded by an expanded genetic code. Here we describe methods for the cotranslational incorporation of one or multiple sulfotyrosines into proteins using standard or genomically recoded *Escherichia coli* stains, thereby achieving the direct expression of site-specifically tyrosine sulfated proteins *in vivo*.

3.1. Introduction

Sulfated proteins play a key role in chemotaxis[1,2], HIV and malarial infection[3,4], coagulation[5], and plant immunity[6,7]. In nature, proteins are sulfated by tyrosylprotein sulfotransferases (TPSTs), enzymes that post-translationally modify a target tyrosine with a sulfate group[8]. However, when TPSTs modify a protein that contains multiple target tyrosines in a short sequence window, the resulting products may contain heterogenous sulfation patterns[9,10] that are problematic for mass spectroscopy analysis and purification of minor sulfated products. In addition, TPSTs only modify specific target tyrosines, limiting our ability to dissect the role of non-native sulfation patterns or single sulfates in a stretch of multiple sulfated tyrosines.

We developed an expanded genetic code system to express homogenously sulfated proteins where the sulfated tyrosines can be installed at any location in a protein[11]. This system relies on the expression of an *Methanococcus jannaschii* (*Mj*) aminoacyl-tRNA synthetase (aaRS) and $tRNA_{CUA}$ pair that is engineered to incorporate sulfotyrosine (sY) at amber (UAG) stop codons. sY incorporation at a UAG codon is highly site-specific, as the engineered *Mj* aaRS and $tRNA_{CUA}$ do not cross-react with endogenous amino acids, aaRSs, or tRNAs[11]. While various homogenously sulfated proteins were successfully expressed using this expanded code for sY[12–14], our initial systems were limited to the incorporation of one or two sYs due to two main sources of inefficiency: suboptimal expression of the *Mj* aaRS for sY (STyrRS) and competition with the endogenous *E. coli* release factor 1 (RF-1) for UAG codons.

To improve the incorporation of multiple sYs, we implemented our expanded genetic code system for sY in a genomically recoded *E. coli* strain, C321.ΔA, whose genome lacks UAG codons and the gene that encodes RF-1[15]. The removal of RF-1 competition in C321.ΔA should improve the incorporation of uAAs at UAG codons (Figure 1A). In addition, we used an upgraded

expression vector, pULTRA[16], to express STyrRS and $tRNA_{CUA}$. Here, we outline protocols for incorporating one or multiple sYs into green fluorescent protein (GFP) in SS320 cells, a standard *E. coli* strain, or in C321.ΔA cells. STyrRS is encoded in the pULTRA-sY plasmid (Figure 2A). Wild-type GFP or GFP containing one or three permissive UAG codons is expressed under the AraBAD promoter in the pGLO expression vector (Figure 2B). After pULTRA-sY and pGLO are co-transformed into SS320 or C321.ΔA cells, the expression of STyrRS and GFP are co-induced by IPTG and L-arabinose, respectively. Fluorescence of cultures were analyzed using a fluorescent plate reader (Tecan Infinite M200 PRO). We found that the expression of GFP containing zero or one sY is higher in SS320 cells while the expression of GFP containing three sYs is higher in C321.ΔA cells (Figure 3). This should serve to guide any effort to heterologously produce any tyrosine-sulfated protein.

3.2. Materials

Solutions are made with sterilized DI water and stored at room temperature unless otherwise specified.

3.2.1. *E. coli* strains

1. SS320: purchased from Lucigen.
2. C321.ΔA: available on Addgene (ID # 48998).

3.2.2. Plasmids

1. pULTRA-sY: available on Addgene (ID # 82417).
2. pGLO-GFP-WT: purchased from Bio-Rad.
3. pGLO-GFP-1UAG: available on Addgene (ID # 82500).
4. pGLO-GFP-3UAG: available on Addgene (ID # 82501).
5. pUC19: purchased from NEB.

3.2.3. Stock solutions

1. 10x PBS (1 L, pH 7.4): dissolve 25.6 g $Na_2HPO_4 \cdot 7H_2O$, 80 g NaCl, 2 g KCl, and 2 g KH_2PO_4 in 900 mL of water. Adjust pH to 7.4. Fill the final volume to 1 L with water. Autoclave at 121°C for 45 min.
2. 20% D-glucose (1 L): dissolve 200 g of D-glucose in 900 mL of water. Fill the final volume to 1 L with water. Autoclave at 121°C for 15 min.
3. 10x Phosphate buffer: dissolve 23.1 g of KH_2PO_4 and 125.4 g of K_2HPO_4 in 900 mL of water. Fill the final volume to 1 L with water. Autoclave at 121°C for 45 min.
4. 80% Glycerol (1 L): add 800 mL of molecular biology grade Glycerol to 200 mL of water. Autoclave at 121°C for 45 min.
5. 5X M9 Salts (1 L): dissolve 64 g of $Na_2HPO_4 \cdot 7H_2O$, 15 g of KH_2PO_4, 2.5 g of NaCl, and 5 g of NH_4Cl in 900 mL of water. Fill the final volume to 1 L with water. Autoclave at 121°C for 45 min.
6. 2% L-arabinose (250 mL): dissolve 5 g of L-arabinose in 200 mL of water. Fill the final volume to 250 mL with water. Filter the solution using a 0.22 µm filter.
7. 100 mM IPTG (250 mL): dissolve 5 g of L-arabinose in 200 mL of water. Fill the final volume to 250 mL with water. Filter the solution using a 0.22 µm filter.
8. 1 M $MgSO_4$ (40 mL): dissolve 4.8 g of $MgSO_4$ in 40 mL of water. Mix with caution as the reaction will release heat. Filter the solution using a 0.22 µm filter.
9. 1 M $CaCl_2$ (40 mL): dissolve 4.4 g of $CaCl_2$ in 40 mL of water and vortex to mix. Filter the solution using a 0.22 µm filter.
10. 2 M $MgCl_2$ (40 mL): dissolve 3.8 g of $MgCl_2$ in 40 mL of water and vortex to mix. Filter the solution using a 0.22 µm filter.

11. 0.02% D-biotin (250 mL): dissolve 50 mg of D-biotin in 250 mL of water and vortex to mix. Filter the solution using a 0.22 μm filter.
12. Sulfotyrosine (250 mL): synthesize sY following a previously reported protocol[17]. Resuspend lyophilized sY in 250 mL of water. Filter the solution using a 0.22 μm filter. Stored at 4°C.

3.2.4. Media

1. SOB - Mg (1 L): dissolve 20 g Tryptone, 5 g Yeast extract, 0.584 g NaCl, 0.186 g KCl in 900 mL of water. Fill the final volume to 1 L with water. Autoclave at 121°C for 45 min.
2. M9T Medium (1 L): dissolve 10 g of tryptone and 5 g of NaCl in 200 mL of 5X M9 salts. Fill the final volume to 1 L with water. Autoclave at 121°C for 45 min. When the autoclaved M9T Medium reaches room temperature, add 20 mL of 20% D-glucose, 2 mL of 1 M $MgSO_4$, 100 μL of $CaCl_2$, and 1 mL of 0.02% D-biotin. Stir until all of the $CaCl_2$ precipitants have fully dissolved.
3. Terrific Broth (TB) Medium (1 L): dissolve 12 g Tryptone and 24 g Yeast extract in 800 mL of water. Add 5 mL of 80% Glycerol and then fill the final volume to 900 mL with water. Autoclave at 121°C for 45 min. Once the medium is cooled to less than 60°C, add 100 mL of 10x Phosphate buffer under sterile conditions.

3.2.5. Agar Plates

1. LB Agar with antibiotics (500 mL): dissolve 18.5 g of LB Agar Mix in water. Autoclave at 121°C for 20 min. Store the solution at 60°C. When ready, adjust the solution to ~45°C, add antibiotics of choice, swirl to mix, and pour plates. Store plates at 4°C.

3.2.6. Transformation Reagents

1. 15% Glycerol (1 L): add 150 mL of molecular biology grade glycerol to 850 mL of water. Autoclave at 121°C for 45 min. Store at 4°C.

2. SOC Medium (40 mL): vortex mix 200 µL of 2 M $MgCl_2$, 721 µL of 20% Glucose, and 39 mL of SOB - Mg under sterile conditions.

3.3. Methods

3.3.1. Construction of plasmids

pULTRA-sY can be obtained from Addgene. Alternatively, amplify STyrRS from pSup-STyrRS [11] using primers aaRStopUltraPCR1 (5'-CATCGCGGCCGCAT GGACGAATTTGAAATGATAAAGAGA-3') and aaRStopUltraPCR2 (5'- CATC GCGGCCGCTTATAATCTCTTTCTAATTGGCTCTAAAATC-3'), and subsequently insert STyrRS into pULTRA[16] using NotI cloning sites.

pGLO-GFP-1UAG can be obtained from Addgene. Alternatively, construct pGLO-GFP-1UAG via site-directed mutagenesis PCR from pGLO-GFP-WT using pGLO-GFP-1TAG-F (5'-CATATGGCTTAGAGCAAAGGAGAAGA ACTTTT-3') and pGLO-GFP-1TAG-R (5'-TCCTTTGCTCTAAGCCATATGTATAT CTCCTT-3').

pGLO-GFP-3UAG can be obtained from Addgene. Alternatively, construct pGLO-GFP-3UAG via site-directed mutagenesis PCR from pGLO-GFP-1UAG using pGLO-GFP-3TAG-F (5'-TCACACTAGGTATAGATCACGGCA GACAAACAAAA-3') and pGLO-GFP-3TAG-R (5'-GTGATCTATACCTAGTGTGA GTTATAGTTGTACTC-3').

3.3.2. Preparation of electrocompetent bacterial cells.

The following protocol is applicable to SS320 or C.321.ΔA cells.

Day 1:

1. Streak the frozen glycerol stock of SS320 cells onto an LB Agar plate or C.321.ΔA cells onto an LB Agar containing Zeocin plate.
2. Incubate the plate streaked with cells in a 37C incubator for 12-14 hours.

Day 2:

3. Pick a single colony into 10 mL of SOB - Mg media in a glass tube.
4. Incubate in a shaker at 37°C, 200 rpm for overnight.
5. Chill 15% glycerol, 250 mL plastic centrifuge bottles, 50 mL conical bottles, microcentrifuge tubes, and serological pipette tips in at 4°C for overnight.

Day 3:

6. Transfer 5 mL of the saturated culture into 500 mL of SOB - Mg media in a 2 L Erlenmeyer flask.
7. Incubate in a shaker at 37°C, 200 rpm for ~2.5 hours until the cells have reached mid-log phase (OD_{600} = 0.6).
8. Immediately chill the cells in an ice water bath for 10 minutes by constantly swirling the culture to evenly chill the cells.
9. The following steps should be performed at 4°C.
10. Evenly split the chilled cell cultures into four 250 mL plastic centrifuge bottles.
11. Centrifuge the cells down at 2500 g for 10 minutes.
12. Immediately discard supernatant and then place bottles on ice.
13. Resuspend the cells in chilled 250 mL 15% glycerol without inducing air bubbles.
14. Repeat steps 10-12.
15. Centrifuge the cells down at 2500 g for 10 minutes.
16. Discard as much supernatant as possible by dumping supernatant and then aspirating supernatant at the screw cap with a pipette. Immediately place bottles on ice.
17. Resuspend cells in the residual 15% glycerol and transfer to a chilled 50 mL conical bottle. The total volume should be ~ 1.5 mL.
18. Dilute 5 µL of the resuspended cells into 1 mL of ddH_2O.

19. Measure the OD_{600} of this diluted resuspension. If OD_{600} is between 0.85-1.0, then the cells are at the appropriate density for transformation. If OD_{600} is < 0.85, then centrifuge the cells down at 2500 g for 10 minutes and remove enough supernatant to achieve the right cell density.
20. Aliquot 50 µL of the resuspended cells into chilled microcentrifuge tubes.
21. Store the aliquots at -80°C.

3.3.3. Transformation via electroporation.

1. 1x Ampicilin is required to select for cells that were transformed with pGLO expression vectors, and 1x Spectinomycin is required to select for cells that were transformed with pULTRA-sY.
2. Thaw a 50 µL aliquot of electrocompetent cells on ice for 5 minutes.
3. Chill an electroporation cuvette on ice.
4. Add up to 5 µL of plasmids suspended in ddH_2O into the cells and incubate on ice for 5 minutes.
5. Transfer the cell suspension into a chilled electroporation cuvette without inducing air bubbles.
6. Electroporate using the Biorad MicroPulser at 1.6-1.8 kV.
7. Immediately rescue the cells in 1 mL of warm SOC. Carefully pipette up and down two times to mix without inducing air bubbles.
8. Incubate in a shaker at 37°C, 200 rpm for 1 hour.
9. Dilute the cells as appropriate and then spread 50-200 µL of the cells onto selective, warm LB Agar plates.
10. Incubate plates at 37°C for 12-16 hours.

3.3.4. Protein expression.

1. Pick a colony from step 9 into 1 mL of TB containing the appropriate antibiotics.
2. Incubate the culture in a shaker at 37°C, 200 rpm for 16 hours.
3. Add 20 µL of the saturated TB culture into 2 mL of M9T containing the appropriate antibiotics.
4. Incubate the culture in a shaker at 37°C, 200 rpm for 12 hours.
5. Add 10 µL of the saturated M9T culture into 1 mL of M9T containing the appropriate antibiotics with or without 20 mM sulfotyrosine.
6. Incubate the culture in a shaker at 37°C, 200 rpm for approximately 3 hours or until the culture has reached log phase (OD_{600} = 0.6-0.8).
7. Induce protein expression by adding 10 µL of 20% L-Arabinose and 10 µL of 100 mM IPTG to the culture.
8. Incubate the culture in a shaker at 30°C, 175 rpm for 18 hours.
9. Transfer 100 µL of the culture into a UV transparent 96-well plate.
10. Measure the absorbance at 600 nm and GFP fluorescence (excitation: 395 nm, emission: 509 nm).

3.4. Notes

1. Ideally, the concentration of sY stock solution should be higher than 100 mM so that the addition of sY to cell culture media will not significantly change the concentrations of media components.
2. Alternative to co-transformation, one can transform pGLO into chemically competent or electrocompetent cells that already contain pULTRA-sY. To prepare electrocompetent

cells containing pULTRA-sY, simply grow the cells in SOB - Mg supplemented with Spectinomycin and then follow the same preparation protocol in Subheading 3.2.

3. To achieve high transformation efficiency, add a volume of plasmids that is equal to or less than $1/10^{th}$ volume of the competent cells.
4. Avoid picking abnormally large colonies of cells that were transformed with pULTRA-sY from LB Agar plates, as large colonies are typically mutants that have disabled components for sY incorporation such as the tRNA.
5. C.321.ΔA cells have a slower doubling time than SS320 cells. pULTRA-sY slightly reduces the doubling time of SS320 cells grown in media containing sY. Take these factors into account when growing cells to their mid-log phase.
6. The volumes of cultures and solutions used for protein expression can be scaled up proportionally, as we routinely express sulfated proteins at 200 mL and 1 L culture volumes.
7. Incorporation of sY using low concentrations of sY in rich media has been difficult presumably due to transport of the negatively charged sY (Figure 1B). We and others found that the incorporation of one sY into proteins can be achieved at 3 mM in M9T [6,7]. Incorporation of multiple sYs will require higher sY concentration in the expression media even when using C321.ΔA cells. We have used concentrations of sY up to 25 mM.
8. C321.ΔA cells require D-biotin to grow in minimal media such as M9T. Before transfer to M9T, the initial growth of cells in rich media such as TB is recommended.
9. Expression of different proteins requires further optimization by altering sY concentration, expression temperature, induction culture OD, IPTG induction concentration, and protein

expression duration. As a rule of thumb, incorporation of three or more sYs is more efficient in C321.ΔA cells than in standard protein expression cells.

10. When a coding region contains a UAG codon, minimal protein should be produced without sY. The protein expression difference in the presence and absence of sY is a good quick check on sY being incorporated into the desired protein. To further ensure that sYs are properly incorporated into the proteins of interest, analyze the purified proteins via mass spectroscopy (MS). Unfortunately, MS positive ion mode, especially in MALDI, deprotonates sulfate modifications[7,18]. One can use selected reaction monitoring mass spectrometry to assess the relative loss of sulfates during MS based on the difference in retention times of the sulfated *versus* unsulfated tryptic peptides [7]. In our experience, the presence of tyrosine where sY is expected in mass spectra is almost always due to the loss of sulfate during mass spectrometry rather than the undesired incorporation of tyrosine during protein expression.
11. Do not use amber suppressor *E. coli* strains for incorporating sY to avoid incorporating natural amino acids at UAG codons.

3.5. References

1. Colvin, R. a, Campanella, G. S. V, Manice, L. a & Luster, A. D. CXCR3 requires tyrosine sulfation for ligand binding and a second extracellular loop arginine residue for ligand-induced chemotaxis. *Mol. Cell. Biol.* **26,** 5838–49 (2006).
2. Tan, J. H. Y. *et al.* Tyrosine sulfation of chemokine receptor CCR2 enhances interactions with both monomeric and dimeric forms of the chemokine monocyte chemoattractant protein-1 (MCP-1). *J. Biol. Chem.* **288,** 10024–34 (2013).
3. Choe, H. *et al.* Sulphated tyrosines mediate association of chemokines and Plasmodium

vivax Duffy binding protein with the Duffy antigen/receptor for chemokines (DARC). *Mol. Microbiol.* **55,** 1413–22 (2005).

4. Farzan, M. *et al.* Tyrosine Sulfation of the Amino Terminus of CCR5 Facilitates HIV-1 Entry. *Cell* **96,** 667–676 (1999).

5. Priestle, J. P., Rahuel, J., Rink, H., Tones, M. & Grutter, M. G. Changes in interactions in complexes of hirudin derivatives and human a-thrombin due to different crystal forms. 1630–1642 (1993).

6. Pruitt, R. N. *et al.* The rice immune receptor XA21 recognizes a tyrosine-sulfated protein from a Gram-negative bacterium. *Sci. Adv.* **1,** e1500245–e1500245 (2015).

7. Schwessinger, B. *et al.* A second-generation expression system for tyrosine-sulfated proteins and its application in crop protection. *Integr. Biol.* 2006–2009 (2016). doi:10.1039/C5IB00232J

8. Choe, H. & Farzan, M. Chapter 7. Tyrosine sulfation of HIV-1 coreceptors and other chemokine receptors. *Methods Enzymol.* **461,** 147–70 (2009).

9. Mikkelsen, J., Thomsen, J. & Ezban, M. Heterogeneity in the tyrosine sulfation of Chinese hamster ovary cell produced recombinant FVIII. *Biochemistry* **30,** 1533–7 (1991).

10. Seibert, C. *et al.* Sequential tyrosine sulfation of CXCR4 by tyrosylprotein sulfotransferases. *Biochemistry* **47,** 11251–62 (2008).

11. Liu, C. C. & Schultz, P. G. Recombinant expression of selectively sulfated proteins in Escherichia coli. *Nat. Biotechnol.* **24,** 1436–1440 (2006).

12. Liu, C. C., Brustad, E., Liu, W. & Schultz, P. G. Crystal structure of a biosynthetic sulfo-hirudin complexed to thrombin. *J. Am. Chem. Soc.* **129,** 10648–9 (2007).

13. Liu, C. C., Choe, H., Farzan, M., Smider, V. V & Schultz, P. G. Mutagenesis and

evolution of sulfated antibodies using an expanded genetic code. *Biochemistry* **48,** 8891–8 (2009).

14. Liu, C. C. *et al.* Protein evolution with an expanded genetic code. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 17688–93 (2008).

15. Lajoie, M. J. *et al.* Genomically Recoded Organisms Expand Biological Functions. *Science (80-.).* **342,** 357–360 (2013).

16. Chatterjee, A., Sun, S. B., Furman, J. L., Xiao, H. & Schultz, P. G. A Versatile Platform for Single- and Multiple-Unnatural Amino Acid Mutagenesis in Escherichia coli. *Biochemistry* (2013). doi:10.1021/bi4000244

17. Liu, C. C., Cellitti, S. E., Geierstanger, B. H. & Schultz, P. G. Efficient expression of tyrosine-sulfated proteins in E. coli using an expanded genetic code. *Nat. Protoc.* **4,** 1784–9 (2009).

18. Hartmann-Fatu, C. & Bayer, P. Determinants of tyrosylprotein sulfation coding and substrate specificity of tyrosylprotein sulfotransferases in metazoans. *Chem. Biol. Interact.* **259,** 17–22 (2016).

3.6. Figure log

Figure No	Copyright holder	Copyright holder's required wording
3	The Royal Society of Chemistry	Adapted from Ref. 7 with permission from The Royal Society of Chemistry.

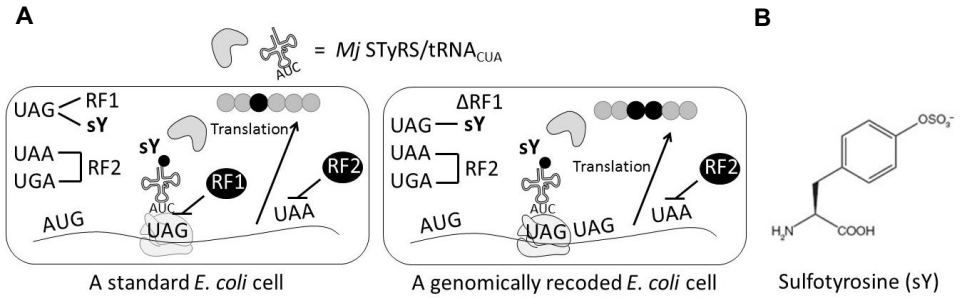

Figure 3.1. Expanded genetic code for sY in a standard or a genomically recoded E. coli cell.

(A) Incorporation of sY is more efficient in a genomically recoded *E. coli* cell such as C.321.ΔA. Genome-wide reassignment of UAG to UAA codons and the deletion of RF-1 in C.321.ΔA cells enable maximum incorporation of sY at UAG codons via *Mj* STyrRS/tRNA$_{CUA}$. (B) Chemical structure of sY.

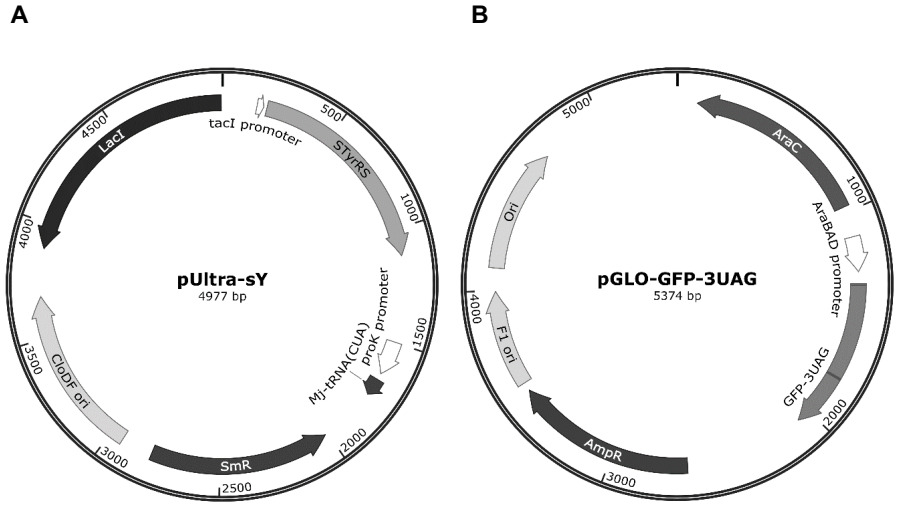

Figure 3.2. Plasmid Maps.

(A) pULTRA-sY, the expression plasmid for the orthogonal STyrRS/$tRNA_{CUA}$ pair. Expression of STyrRS is driven by an IPTG-inducible tacI promoter and *Mj* $tRNA_{CUA}$ is constitutively expressed from the proK promoter. pULTRA-sY contains the CloDF origin and spectinomycin resistance gene (SmR), both of which are compatible with common expression vectors. (B) pGLO, the expression plasmid for GFP variants. pGLO-GFP-3UAG encodes GFP with three UAG codons at amino acid positions 3, 151, and 153. pGLO-GFP-1UAG and pGLO-GFP-WT (plasmid maps are not shown) encode GFP with one UAG codon at amino acid position 3 and the wild-type GFP gene without UAG codons, respectively. Expression of GFP variants is driven by an L-arabinose-inducible AraBAD promoter. Plasmid maps were created using SnapGene.

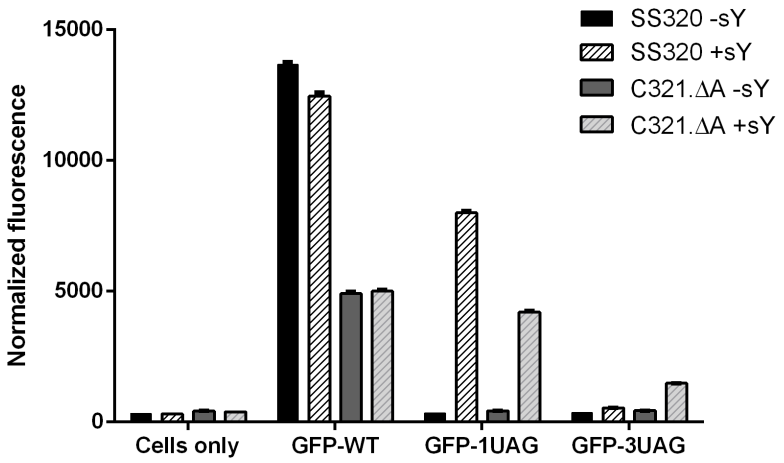

Figure 3.3. Incorporation of sY into GFP.

Fluorescence of wild-type GFP (GFP-WT), GFP with one or three UAG codons (GFP-1UAG or GFP-3UAG) expressed in SS320 or C321.ΔA cells without or with 20 mM sY in the growth media. Only cells that express GFP-1UAG or GFP-3UAG contain pULTRA-sY. 'Cells only' are cells that do not contain expression vectors. Fluorescence signals were normalized by dividing by OD_{600}. Error bars represent ± s.d. of triplicate samples. Data was adapted from [7] with permission from The Royal Society of Chemistry. Graph was made using Graphpad Prism.

Chapter 4.A second-generation expression system for tyrosine sulfated proteins and its application in crop protection

Benjamin Schwessinger,[1,2,3] Xiang Li,[4] Thomas L. Ellinghaus,[2,5] Leanne Jade G. Chan,[2] Tong Wei,[1,2] Anna Joe,[1,2] Nicholas Thomas,[1] Rory Pruitt,[1,2] Paul D. Adams,[2,5] Maw Sheng Chern,[1,2] Christopher J. Petzold,[2] Chang C. Liu,[4] Pamela C. Ronald[1,2]

1. Department of Plant Pathology and the Genome Center, University of California, Davis
2. Joint BioEnergy Institute and Physical Biosciences Division, Lawrence Berkeley National Laboratory
3. The Australian National University, Research School of Biology
4. Department of Biomedical Engineering, University of California, Irvine
5. Physical Biosciences Division, Lawrence Berkeley National Laboratory

Integrative Biology **8**, 542-545 (2016).
DOI: 10.1039/C5IB00232J

Abstract

Posttranslational modification (PTM) of proteins and peptides is important for diverse biological processes in plants and animals. The paucity of heterologous expression systems for PTMs and the technical challenges associated with chemical synthesis of these modified proteins has limited detailed molecular characterization and therapeutic applications. Here we describe an optimized system for expression of tyrosine-sulfated proteins in *Escherichia coli* and its application in a bio-based crop protection strategy in rice.

4.1. Introduction

Tyrosine sulfation is an important posttranslational modification involved in diverse biological processes including immunity and development.[1] For example, in humans, tyrosine sulfation of cellular co-receptors mediates their interaction with the gp120 (glycoprotein 120) coat protein of HIV (Human Immunodeficiency Virus).[2] The therapeutic potential of tyrosine-sulfated proteins is reflected in a recent report of an engineered sulfated immunoglobulin that neutralized 100% of a diverse panel of neutralization-resistant HIV isolates.[3] Historically, tyrosine sulfation was thought to be restricted to eukaryotic biology. However, we recently demonstrated that the plant bacterial pathogen *Xanthomonas oryzae* pv. *oryzae* (*Xoo*) harbors a functional tyrosine sulfotransferase, called RaxST (required for activation of XA21-mediated immunity sulfotranserase), that sulfates the 60 amino acid protein RaxX (required for activation of XA21-mediated immunity X), on its central tyrosine (Y41).[4,5] Tyrosine sulfated RaxX, but not unsulfated RaxX, triggers an immune response in rice plants carrying the receptor XA21.[6] This growing body of research solidifies the role of tyrosine sulfation as a mediator of protein-protein interactions and immune recognition[1,6] and demonstrates the relevance of this PTM to human and plant health.

The therapeutic application of posttranslationally modified peptides or proteins requires efficient synthesis. Peptide synthesis of tyrosine-sulfated proteins is technically challenging especially for longer peptides and proteins.[1] An alternative approach is to express recombinant proteins together with the corresponding sulfotransferase in *E. coli* or perform *in vitro* sulfation assays.[1,7] However these strategies often result in heterogeneous sulfation of the target, limiting applications.

In 2006, we described a first-generation system to express sulfated proteins in *E. coli* that overcomes these drawbacks.[7,8] This system relies on an expanded genetic code that enables *E. coli* cells to direct the incorporation of sulfotyrosine (sY) at UAG (amber) codons during translation. Highly specific incorporation of sY is achieved through a specially-engineered tRNA/aminoacyl-tRNA synthetase (tRNA/aaRS) pair that recognizes sY and the amber codon without cross reacting with endogenous aaRSs and tRNAs.[7] This expanded genetic code system has enabled a number of applications including the recombinant production of the therapeutic anticoagulant sulfo-hirudin[7,9] and phage display evolution studies on sulfated anti-gp120 antibodies[10,11]. Several advances in expanded genetic code technology have taken place since these studies were carried out.[12,13] Here, we incorporate two main advances into an improved second-generation system for the recombinant expression of tyrosine-sulfated proteins. We demonstrate the high-quantity expression and characterization of highly purified, sulfated RaxX proteins and show that these proteins can induce immunity in rice plants carrying the XA21 immune receptor. This bio-based strategy provides a new avenue for protecting crops from disease.

4.2. Results and discussion

The Schultz group recently developed the pULTRA system for efficient unnatural amino acid incorporation through expanded genetic codes. pULTRA encodes an optimized amber suppressor $tRNA_{CUA}$/aaRS pair that increases the level of tRNA aminoacylation, minimizes tRNA toxicity, and optimizes aaRS expression levels.[14] To host $tRNA_{CUA}$/aaRS pairs, the Church and Isaacs groups created a recoded *E. coli* strain, C321.ΔA.exp, that has all genomic amber stop codons replaced with the alternate stop codon UAA. This strain carries a deletion of release factor 1 (RF1), eliminating active termination at amber stop

codons.[13] To improve sulfated protein expression, we first cloned our sY-specific aaRS into the pULTRA system. The resulting plasmid, pULTRA-sY (Figure 3), was tested for its ability to insert one or three sYs into GFP (Green Fluorescence Protein) specified by either one or three UAG codons (Figure 3). Comparison of pULTRA-sY with a previous plasmid system for sY incorporation, pEVOL-sY[15], revealed that pULTRA-sY achieved substantially higher incorporation of a single sY into GFP in standard SS320 *E. coli* cells. This advantage was lost when we attempted to incorporate three sYs into GFP, presumably because RF1 competition for UAG codons increases when multiple sY incorporation events are required. However, when we used pULTRA-sY in *E. coli* strain C321.ΔA.exp, high levels of sY incorporation at single or multiple UAG codons were achieved (Figure 1).

We next used our improved second-generation expression system for tyrosine-sulfated proteins to produce full-length sulfated RaxX, a protein relevant to crop protection.[16] We designed a C-terminal his-tagged RaxX fused to maltose binding protein (MBP) at the N-terminus followed by a 3C protease cleavage site (MBP-3C-RaxX60-His). Because RaxX is normally sulfated at position 41 in *Xoo,* we specified tyrosine sulfate using an UAG codon at the corresponding position. We expressed MBP-3C-RaxX60-His under the control of the araBAD promoter and induced expression by the addition of arabinose during growth in minimal media. We obtained similar expression levels for RaxX60-Y-His (unsulfated) and RaxX60-sY-His (containing a sulfated tyrosine at position 41) in the presence of 5 mM sY. We obtained up to 4 mg of RaxX60-Y/sY per liter of culture after a three-step purification process including the removal of the N-terminal fusion tag by treatment with 3C protease. The resulting peptides were of a purity > 90%, as

estimated by standard LC-MS/MS analysis of trypsin digested protein preparations (Figure 4, Table 1).

Next we assessed the sulfation status of RaxX60-sY by selected reaction monitoring mass spectrometry (SRM-MS). Because of the different physiochemical properties of the sulfated and unsulfated peptide variants, trypsin peptides covering the central tyrosine (Y41) eluted with markedly different retention times (ΔRT ~ 1.1 min). Sulfate modifications of tyrosines are inherently unstable during electrospray ionization in the positive ion mode and easily lost making quantification of the sulfation status challenging.[17] To account for this, we used the observed difference in retention time of sulfated *versus* unsulfated tryptic peptides to estimate the relative loss of SO_4^{2-} from Y41 under our ionization conditions to be approximately 11%. This estimate enabled us to calculate the relative sulfation status of RaxX60-sY produced using our improved sulfated protein expression system to be over 99.5%, which is well above the sulfation status of three shorter commercially synthesized RaxX-sY peptide variants (Table 2, see supplemental information for experimental details). We were unable to detect any sY-containing peptides derived from RaxX60-Y samples. These results indicate that our optimized second-generation recombinant expression system produces tyrosine-sulfated proteins that are superior to standard commercially synthesized sY peptides.

The impact of tyrosine sulfation on protein structure is unpredictable. In some cases it has been shown that sY modifications stabilize intermolecular interactions.[9,18] We therefore investigated the structural effect of RaxX60 tyrosine sulfation. Using our highly purified RaxX60-Y and RaxX60-sY, we performed circular dichroism (CD) spectroscopy of RaxX60. CD spectroscopy is an excellent technique to estimate protein secondary

structure and the impact of post-translation modification on protein folding as a function of temperature. [19] Standard CD analysis at 20ºC indicated that tyrosine sulfation had little impact on the overall fold of RaxX60 with both protein versions being largely disordered with a 20-30% β-strand content, most likely corresponding to a single β-sheet (Figure 5, Table 3, see supplemental information for experimental details).[20] Next we tested if sulfation has an impact on thermal protein stability and recorded CD spectroscopy-melting curves. Indeed sulfation had a significant impact on RaxX60 thermal stability with RaxX60-sY keeping its overall fold at much higher temperatures when compared to RaxX60-Y (Figure 2A, see supplemental information for experimental details). This finding suggests that tyrosine sulfation of RaxX60 positively impacts protein stability.

We previously demonstrated that sulfated RaxX variants of different lengths, including a 21 amino acid fragment called RaxX21-sY are immunogenic on rice plants carrying the immune receptor XA21.[16] We tested if recombinantly produced sulfated RaxX60-sY is able to induce a similar set of responses. We found that at a concentration of 500 nM, RaxX60-sY activates the rice immune response in an XA21-dependent manner as measured by immune gene expression over time (Figure 2B). RaxX60-sY induced the expression of two well-established rice marker genes (*Os04g10010* and *PR10b*) as early as 3 hours post treatment with a prolonged transcriptional activation observed over 24 hours.

We found that at a concentration of 500 nM, RaxX60-sY activates the rice immune response in an XA21-dependent manner as measured by immune gene expression over time (Figure 2B). RaxX60-sY induced the expression of two well-established rice marker genes (*Os04g10010* and *PR10b*) as early as 3 hours post treatment with a prolonged transcriptional activation observed over 24 hours.

In rice production, resistance to *Xoo* is agronomically important. Given our observed induction of immunity by RaxX60-sY, we aimed to test whether exogenous application of RaxX60-sY on XA21 rice plants could confer immunity to infectious strains of *Xoo*. To explore this possibility, we used a variation of a recently established assay[21]. In this experiment we inoculated rice plants with a virulent strain of *Xoo* (PXO99AΔ*raxX*) that is not recognized by XA21.[16] Two days post-infection, we treated rice leaves with a 1 μM solution of RaxX60-Y or RaxX60-sY. We found that post-treatment with RaxX60-sY dramatically slowed disease progression and symptom development (Figure 2C and D).

4.3. Conclusion

We developed an improved second-generation high efficiency expression system for tyrosine sulfated proteins in *E. coli* and demonstrated its application in immune receptor-mediated crop protection using the model activator-receptor pair RaxX-sY-XA21 in rice. The application of recombinantly produced RaxX60-sY in real field conditions could in principal lead to a reduction of disease progression even after initial infection.

4.4. Author contribution

B.S., X.L., C.C.L., and P.C.R. wrote manuscript. B.S., X.L., T.L.E., L.J.G.C., C.J.P., C.C.L., and P.C.R. designed experiments. B.S., X.L., T.L.E., L.J.G.C., T.W., A.J., N.T., R.P., P.D.A., M.S.C., C.J.P., C.C.L., and P.C.R. analyzed data. B.S., X.L., T.L.E., L.J.G.C., T.W., A.J., N.T., R.P., and C.J.P. performed experiments. B.S., P.D.A., C.C.L., and P.C.R. contributed financial support.

All authors read the manuscript and agreed on it's submission.

4.5. Competing financial interests

R.N.P., B.S., A.J., and P.C.R. have filed a patent describing the isolation and expression of RaxX and its application for engineering resistance in plants (U.S. provisional patent application no. 62/013,709).

4.6. Acknowledgements

Funded by NIH GM59962 to P.C.R. and start-up funds from UC Irvine to C.C.L. This work was also conducted in part by the Joint BioEnergy Institute and was supported by the Office of Science, Office of Biological and Environmental Research, of the U.S. Department of Energy under contract no. DE-AC02-05CH11231. B.S. was supported by a Human Frontiers Science Program long-term postdoctoral fellowship (LT000674/2012) and a Discovery Early Career Award (DE150101897).

4.7. References

1. Seibert, C. & Sakmar, T. P. Toward a framework for sulfoproteomics: Synthesis and characterization of sulfotyrosine-containing peptides. *Biopolymers* **90,** 459–77 (2008).

2. Stone, M. J., Chuang, S., Hou, X., Shoham, M. & Zhu, J. Z. Tyrosine sulfation: an increasingly recognised post-translational modification of secreted proteins. *N. Biotechnol.* **25,** 299–317 (2009).

3. Gardner, M. R. *et al.* AAV-expressed eCD4-Ig provides durable protection from multiple SHIV challenges. *Nature* **519,** 87–91 (2015).

4. Han, S.-W. *et al.* Tyrosine sulfation in a Gram-negative bacterium. *Nat. Commun.* **3,** 1153 (2012).

5. Pruitt, R. N. *et al.* The rice immune receptor XA21 recognizes a tyrosine-sulfated protein from a Gram-negative bacterium. *Sci. Adv.* **1,** e1500245–e1500245 (2015).

6. Ludeman, J. P. & Stone, M. J. The structural role of receptor tyrosine sulfation in chemokine recognition. *Br. J. Pharmacol.* **171,** 1167–1179 (2014).

7. Liu, C. C. & Schultz, P. G. Recombinant expression of selectively sulfated proteins in Escherichia coli. *Nat. Biotechnol.* **24,** 1436–1440 (2006).

8. Liu, C. C., Cellitti, S. E., Geierstanger, B. H. & Schultz, P. G. Efficient expression of tyrosine-sulfated proteins in E. coli using an expanded genetic code. *Nat. Protoc.* **4,** 1784–9 (2009).

9. Liu, C. C., Brustad, E., Liu, W. & Schultz, P. G. Crystal structure of a biosynthetic sulfo-hirudin complexed to thrombin. *J. Am. Chem. Soc.* **129,** 10648–9 (2007).

10. Liu, C. C. *et al.* Protein evolution with an expanded genetic code. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 17688–93 (2008).

11. Liu, C. C., Choe, H., Farzan, M., Smider, V. V & Schultz, P. G. Mutagenesis and evolution of sulfated antibodies using an expanded genetic code. *Biochemistry* **48,** 8891–8 (2009).

12. Liu, C. C. & Schultz, P. G. Adding new chemistries to the genetic code. *Annu. Rev. Biochem.* **79,** 413–44 (2010).

13. Lajoie, M. J. *et al.* Genomically Recoded Organisms Expand Biological Functions. *Science (80-.).* **342,** 357–360 (2013).

14. Chatterjee, A., Sun, S. B., Furman, J. L., Xiao, H. & Schultz, P. G. A Versatile Platform for Single- and Multiple-Unnatural Amino Acid Mutagenesis in Escherichia coli. *Biochemistry* (2013). doi:10.1021/bi4000244

15. Young, T. S., Ahmad, I., Yin, J. A. & Schultz, P. G. An Enhanced System for Unnatural Amino Acid Mutagenesis in E. coli. *J. Mol. Biol.* **395,** 361–374 (2010).

16. Pruitt, R. N. *et al.* The rice immune receptor XA21 recognizes a tyrosine-sulfated protein from a Gram-negative bacterium. *Sci. Adv.* **1,** e1500245 (2015).

17. Monigatti, F., Hekking, B. & Steen, H. Protein sulfation analysis—A primer. *Biochim. Biophys. Acta - Proteins Proteomics* **1764,** 1904–1913 (2006).

18. Huang, C.-C. *et al.* Structures of the CCR5 N terminus and of a tyrosine-sulfated antibody with HIV-1 gp120 and CD4. *Science* **317,** 1930–4 (2007).

19. Greenfield, N. J. Using circular dichroism spectra to estimate protein secondary structure. *Nat. Protoc.* **1,** 2876–2890 (2007).

20. Gellman, S. H. Minimal model systems for β-sheet secondary structure in proteins. *Curr. Opin. Chem. Biol.* **2,** 717–725 (1998).

21. Wei, T., Chen, M., Thomas, N. & Ronald, P. C. Induction of XA21-mediated immunity by in planta application of RaxX.

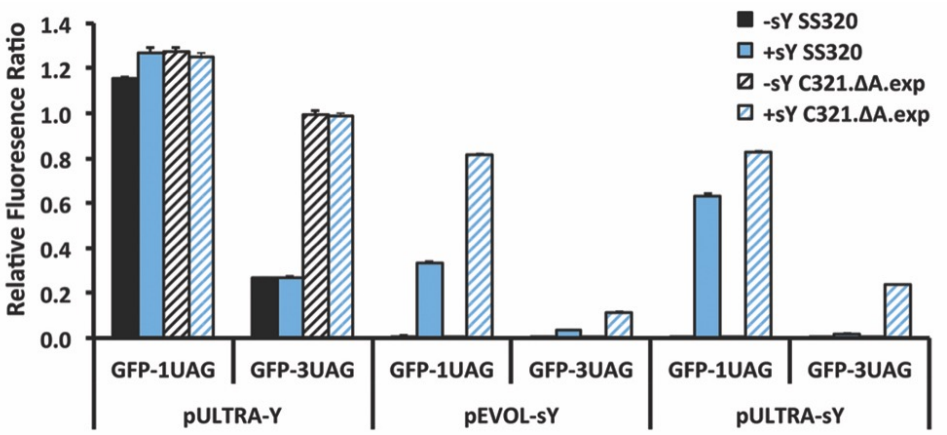

Figure 4.1. Highly efficient sY protein production using a second-generation sY protein expression system.

Relative fluorescence ratio, with respect to wild-type GFP expression, of GFP with one (GFP-1UAG) or three (GFP-3UAG) amber codons. GFPs were expressed in the presence of a control plasmid or plasmids encoding sY incorporation systems (pULTRA-Y and pEVOL-sY or pULTRA-sY) in two *E. coli* cell lines (SS320 or C321.ΔA.exp). –sY and +sY indicate the absence and presence of 20 mM sY in the growth media. Bars indicate the mean ± SD (n = 3).

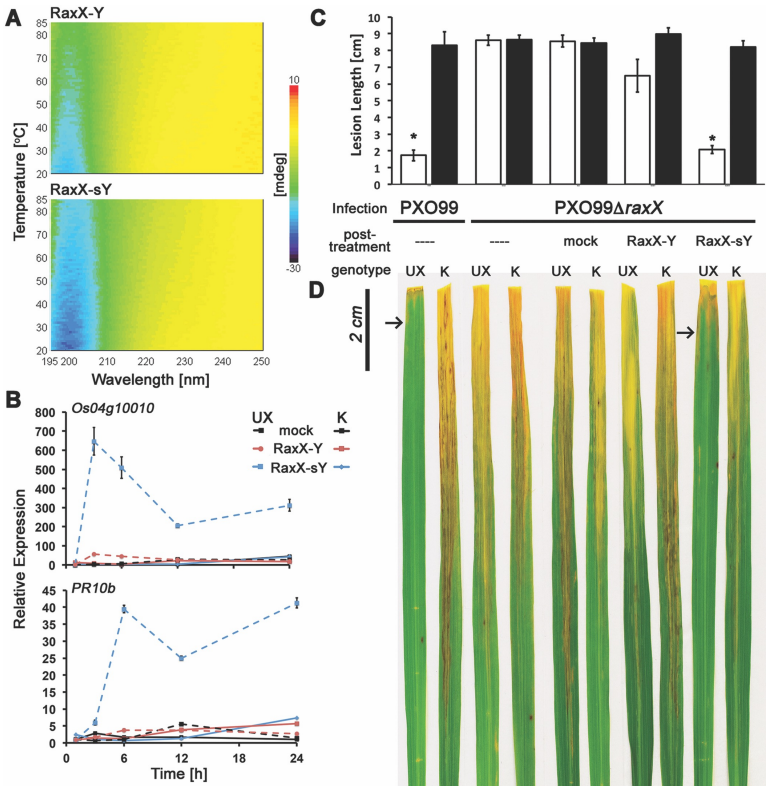

Figure 4.2. Recombinantly produced tyrosine-sulfated RaxX60-sY induces rice immunity.

(A) Tyrosine 41 sulfation leads to increased thermal stability of the native fold of RaxX60 as measured by CD spectroscopy. (B) Application of RaxX60-sY (500 nM) leads to temporal immune gene activation (*Os04g10010* and *PR10b*) in detached rice leaves in an immune receptor kinase dependent manner (XA21). All data points depict means ± SE measured 13 days after infection ($n \geq 5$). The "*" indicates statistically significant difference from no-treatment control (- - -) using Dunnett's test ($\alpha = 0.01$). (D) Rice leaves displaying water-soaked lesions 13 days after infection. Ony rice leaves carrying XA21 (UX) treated with RaxX60-sY and the non-infectious control display significantly reduced disease symptoms as highlighted by arrows. UX and K indicates the plant genotypes Ubi::MYC::XA21 and wild-type Kitaake control, respectively.

4.8. Supplemental Information

4.8.1. Experimental Details

4.8.1.1. Construction of second-generation sY protein expression vectors

pULTRA-sY and pEVOL-sY were cloned by replacing the *Methanocaldococcus janaschii* (*Mj*) tyrosyl-tRNA synthetase (*Mj*-TyRS) in pULTRA[1] and pEVOL[2] with STyrRS, , the *Mj* aminoacyl-tRNA synthetase (aaRS) evolved to specifically charge *Mj*-$tRNA_{CUA}$ with sY. *Mj*-$tRNA_{CUA}$ suppresses UAG codons.[3] To clone pULTRA-sY, STyrRS was amplified from pCDF-sY[4] using primers aaRStopULTRAPCR1 and aaRStopULTRAPCR2, and subsequently inserted into pULTRA using the NotI cloning site. To clone pEVOL-sY, two pEVOL backbone fragments were amplified from pEVOL using primer pairs pEVOL-BB-F/pEVOL-BB-R and pEVOL-BB2-F/pEVOL-BB2-R, respectively. Two STyrRS fragments were amplified from pULTRA using primer pairs aaRS-F/aaRS-R and aaRS2-F/aaRS2-R, respectively. pEVOL was constructed by assembling the pEVOL and STyrRS fragments using the Gibson method.[5]

4.8.1.2. Cloning, expression, and analyses of GFP

The pGLO expression vector was used to express wild-type GFP (pGLO-GFP-WT) or GFP with one or three UAG codons. pGLO-GFP-1UAG was cloned via site-directed mutagenesis PCR from pGLO-GFP-WT using primers pGLO-GFP-1TAG-F and pGLO-GFP-1TAG-R. pGLO-GFP-3UAG was cloned via site-directed mutagenesis from pGLO-GFP-1UAG using primers pGLO-GFP-3TAG-F and pGLO-GFP-3TAG-R.

To compare sY incorporation in GFP in different strains and systems, SS320 (Lucigen) or C321.ΔA.exp[6] *E. coli* cells were cotransformed with 1) a plasmid expressing *Mj* $tRNA_{CUA}$ and *Mj*TyRS (pULTRA-Y) or STyrRS (pULTRA-sY or pEVOL-sY), and 2) a plasmid expressing GFP with one or three UAG codons (pGLO-GFP-1UAG or pGLO-GFP-3UAG). Transformed cells were grown in M9T medium (1X M9 salts, 10 g/L tryptone, 5 g/L NaCl, 1 mM $MgSO_4$, and 80 μM biotin) with or without 20 mM of sY, synthesized as described previously[7]. Once cultures

reached mid-log phase (OD_{600} of 0.6-0.8), the expression of GFP and aaRS were co-induced with 0.02% L-Arabinose (for pEVOL-sY) or 1 mM IPTG and 0.02% L-Arabinose (for pULTRA-Y and pULTRA-sY). The cultures were incubated for 18 hours at 30°C. 100 μL of each cell culture were transferred into 96-well black clear-bottom plates (Corning), and GFP fluorescence (excitation/emission, 395/509 nm) was measured using an Infinite 200 PRO plate reader (Tecan). Fluorescence signals were normalized by dividing by OD_{600}. Autofluorescence of cells not expressing GFP and aaRS was subtracted, and the resulting fluorescence for each sample was divided by fluorescence of *E. coli* cells expressing wild-type GFP to obtain the relative fluorescence ratio.

4.8.1.3. Cloning, Expression and Purification of RaxX60

Full length RaxX60 from *Xoo* PXO99 was expressed as a MBP-3C-RaxX-His fusion protein in *E. coli* C321.ΔA.exp[6]. For this purpose RaxX60 genomic DNA was amplified using the two primers RaxX-FL_EcoRI-3C_forward and RaxX-His-stop-HindIII-reverse and cloned into pMAL-c4X (NEB) (Table S4). A sequence verified clone was used as template for amplification using the two primers pMAL-start_E_NcoI and RaxX-His-stop-HindIII-reverse and the product inserted into pBAD_A_Myc (Invitrogen). The resulting plasmid was sequence verified and named pBAD/MBP-3C-RaxX-His. The UAG codon was introduced at the corresponding position of Y41 in RaxX60 using point mutagenesis on pBAD/MBP-3C-RaxX-His with the primer pair RaxX_amber_TAG_F and RaxX_amber_TAG_R. The resulting plasmid was sequence verified and named pBAD/MBP-3C-RaxX-His (Amber). *E. coli* C321.ΔA.exp was cotransformed with pULTRA-sY and pBAD/MBP-3C-RaxX-His (Amber) to generate sulfated RaxX60 or transformed with pBAD/MBP-3C-RaxX-His to generate unsulfated RaxX60. Transformed bacteria were grown in 0.5 L of M9T media in 2L baffled Erlenmeyer flasks. For expression of sulfated RaxX, sY was added to a final concentration of 5 mM. Cultures were grown at 37°C with shaking at 300

rpm. Expression was induced at an $O.D._{600}$ of 0.7-0.9 by addition of 1 mM IPTG and 0.25% (w/vol) L-Arabinose for 5 h. MBP-3C-RaxX-His was purified from intracellular total protein extracts in extraction buffer A (25mM Tris pH 8, 500 mM NaCl, 40 mM Imidazol, fresh 1mM PMSF, 0.1 µl/ml Nuclease mix (GE Healthcare), 2.5 mM $MgCl_2$, 1 mM DTT) over a 5 mL Ni-NTA column on an FPLC (Akta, GE Healthcare) and eluted with a constant linear gradient of buffer A containing 500 mM Imidazol. Fractions containing MBP-3C-RaxX-His were concentrated using Amicon Ultra-15 centrifugal filter units (10 kDa) and resuspended in buffer B (20 mM Hepes pH 8) to a final concentration of 1-5 mg/ml. The MBP-tag was removed by overnight digestion with 3C protease (Thermo Scientific) at a concentration of 1:200 (w/w) at 4°C with head over mixing. The MBP was separated from full length RaxX-His over a 1 mL SP-XL column, a strong cation exchanger, on an FPLC (Akta, GE Healthcare) with a constant gradient of buffer B containing 1 M NaCl. Fractions containing MBP-3C-RaxX-His were concentrated using Amicon Ultra-15 centrifugal filter units (3 kDa). In a third step RaxX60 was further purified by size exclusion chromatography on a Superdex 75 10/300 in buffer C (20mM Tris pH 8, 50 mM NaCl). During some experiments over 50% of the sample was lost at the final third purification step. It is recommended to avoid this last purification step if not required. Unsulfated RaxX was expressed and purified in the same way using pBAD/MBP-3C-RaxX-His in the absence of sY in the media. The yield for sulfated and unsulfated highly purified RaxX-His was up to 4 mg protein per 1L of culture.

4.8.1.4. Gene expression assay in rice

Rice defense gene expression assays were performed as described previously.[8]

4.8.1.5. Statistical analysis

Statistical analyses were performed using JMP software (ASAS Institute Inc., Cary, NC, USA).

4.8.1.6. Proteomic Analysis

RaxX peptides were digested with 1/10 (w/w) of trypsin at 37ºC o/n in a buffer containing 25 mM Tris pH 8, 50 mM NaCl, and 5 mM DTT. Peptides were purified over C18 bench top spin column (Harvard Apparatus).

4.8.1.7. Shotgun proteomics

Samples were analyzed on an Agilent 1290 liquid chromatography system coupled to an Agilent 6550 iFunnel Q-TOF mass spectrometer (Agilent Technologies; Santa Clara, CA). Peptide samples (4 µg) were loaded onto an Ascentis Express ES-C18 column (10 cm length x 2.1 mm, 2.7 µm particle size) (Sigma Aldrich; St. Louis, MO) operating at 60 ºC at a flow rate of 400 µL/min. A 13.5-minute method with the following separation gradient was used: 95% Buffer A (0.1% formic acid) and 5% Buffer B (99.9% acetonitrile, 0.1% formic acid) were held for 1 minute, then B was increased to 35% over 5.5 minutes, followed increasing to 80% B over 1 minute where it was held at 600 µL/min for 3.5 minutes. Buffer B was ramped back down to 5% over 0.5 minutes and the system was re-equilibrated for 2 minutes. Peptides were introduced into the mass spectrometer from the LC by using a Dual Jet Stream Electrospray Ionization source (Agilent Technologies) operating in positive-ion mode. Source parameters used include: gas temperature (250°C), drying gas (14 L/min), nebulizer (35 psig), sheath gas temp (250°C), sheath gas flow (11 L/min), VCap (5,000 V), fragmentor (180 V), and OCT 1 RF Vpp (750 V). The data were acquired with Agilent MassHunter Workstation Software, LC/MS Data Acquisition B.05.01 (Build 5.01.5125.2) and peak lists were extracted via the MassHunter Qualitative Analysis MGF file export option. Protein and peptide identifications from MS/MS data were accomplished with Mascot (version 2.3.02; Matrix Science), filtered, and validated using Scaffold (version 4.4.5; Proteome Software).

4.8.1.8. Targeted proteomics

Samples were analyzed using an Agilent 1290 liquid chromatography system coupled to an Agilent 6460 mass spectrometer (Agilent Technologies). Peptide samples (0.5 µg) were loaded and separated on an Ascentis Express Peptide C18 column (15 cm length x 2.1 mm ID, 2.0 µm particle size; Sigma Aldrich) operating at 60 ºC in normal flow at 400 µL/min. A 43-minute method with the following gradient was used: 98% Buffer A (0.1% formic acid) and 2% Buffer B (99.9% acetonitrile, 0.1% formic acid) were held for 12 seconds, then B was increased to 35% in 35.5 minutes, followed by an increase to 90% B in 18 seconds where it was held for 2 minutes. Buffer B was brought down to 5% over 30 seconds, then further decreased to 2% B over 2.5 minutes where it was held for 2 minutes to re-equilibrate the column to the starting conditions. Peptides were introduced into the mass spectrometer from the LC, using an Jet Stream Electrospray Ionization source (Agilent Technologies) operating in positive-ion mode. Source parameters used include: gas temperature (250°C), gas flow (13 L/min), nebulizer (35 psi), sheath gas temp (250°C), sheath gas flow (11 L/min), and VCap (3,500 V). The data were acquired with Agilent MassHunter Workstation Software, LC/MS Data Acquisition B.06.00 (Build 6.0.6025.3 SP3). Targeted proteomic methods were developed and data were analyzed with Skyline software (MacCoss Lab, University of Washington). Data can be accessed at http://tinyurl.com/2nd-gen-sY-RaxX (E-mail: jbeireviewer@gmail.com and Password: jbeireview1).

4.8.1.9. Calculations of Relative sY status

It is well described that sulfate groups are abundantly lost from tyrosine during electrospray ionization (ESI) in the positive ion mode.[9,10] In order to adequately measure relative sulfation status we therefore first estimated the loss of SO_4^{2-} from the tryptic peptide covering RaxX Y41 [K.HVGGGDsYPPPGANPK.H] during ESI. Because of their distinct physiochemical properties, unsulfated [K.HVGGGDYPPPGANPK.H] and sulfated [K.HVGGGDsYPPPGANPK.H] tryptic

peptides were eluted at markedly different retention times (RT) of 6.7 and 7.8 min, respectively, without any overlap. This led to the observation that from RaxX60-Y derived actual unsulfated $[K.HVGGGDYPPPGANPK.H]_{RT=6.7}$ peptides eluted at RT of 6.7 min and from RaxX60-sY derived originally sulfated peptides that lost their SO_4^{2-} during ESI $[K.HVGGGDYPPPGANPK.H]_{RT=7.8}$ peptides eluted at RT of 7.8 min. The relative neutral loss of SO_4^{2-} from [K.HVGGGDsYPPPGANPK.H] $\%SO_4^{2-}loss$ was estimated by dividing the normalized intensity of $[K.HVGGGDYPPPGANPK.H]^{3+}_{RT=7.8}$ derived from RaxX60-sY by the normalized intensity (I) of $[K.HVGGGDYPPPGANPK.H]^{3+}_{RT=6.7}$ derived from RaxX60-Y. Equal loading (0.5 µg) and division by the intensity of an unmodified tryptic peptide $[K.GAASLQRPAGAK.G]^{2+}_{RT=4.7}$ of the same protein preparation achieved normalization (equation 1).

$$\%SO_4^{2-}loss = \frac{I_{RaxX60-sY}\,[\text{K.HVGGGDYPPPGANPK.H}]^{3+}_{\text{RT=7.8}}}{I_{RaxX60-sY}\,I\,[\text{K.GAASLQRPAGAK.G}]^{2+}_{\text{RT=4.7}}} \;/\; \frac{I_{RaxX60-Y}\,[\text{K.HVGGGDYPPPGANPK.H}]^{3+}_{\text{RT=6.7}}}{I_{RaxX60-Y}\,[\text{K.GAASLQRPAGAK.G}]^{2+}_{\text{RT=4.7}}}$$

[Equation 1]

Based on two different protein preparations and three technical replicate runs each we estimated an approximate loss of SO_4^{2-} from [K.HVGGGDsYPPPGANPK.H] of 11% ($\%SO_4^{2-}loss = 11$). This enabled us to calculate the relative sulfation status ($\%sY$) of RaxX-sY as shown in equation 2 and Table S2.

$$\%sY = (100 - \%SO_4^{2-}loss) + \%SO_4^{2-}loss * \frac{I\,[\text{K.HVGGGDYPPPGANPK.H}]^{3+}_{\text{RT=7.8}}}{I\,[\text{K.HVGGGDYPPPGANPK.H}]^{3+}_{\text{RT=6.7}} + I\,[\text{K.HVGGGDYPPPGANPK.H}]^{3+}_{\text{RT=7.8}}}$$

[Equation 2]

4.8.1.10. Circular Dichroism

RaxX60-Y and RaxX60-sY melting curves from 20°C to 85°C were recorded with a JASCO J-815 CD spectrometer. The experimental concentrations were 0.1 mg/mL protein, 5 mM NaCl and 2 mM Tris (pH 8). Two accumulations were taken at a wavelength range from 195 nm to 250 nm applying 0.2 nm data pitches. The 20°C spectra represent the cross sections at this temperature point. Data processing of the 20°C CD spectra involved the programs CRDATA, CDSSTR, SELCON3 and CONTIN/LL of CDPro (http://lamar.colostate.edu/~sreeram/CDPro/).[11,12] The protein reference datasets 1, 7 and 10 were chosen to cover the largest wavelength range, the largest dataset of soluble proteins, or the overall largest dataset, respectively.

4.8.1.11. Post-treatment experiment

Kitaake, and a Kitaake transgenic line expressing Myc-XA21 under the maize ubiquitin promoter (UX) were used for the post-treatment assay.[13] Rice seeds were surface-sterilized using 15% bleach, rinsed with water and germinated in distilled water at 28°C for 1 week. Well-grown seedlings were transplanted into trays filled with A-OK Starter Plugs (Grodan) and watered twice a week with Hoagland's solution. For the post-treatment assay, 4-week-old plants were inoculated using the scissors clipping method, and treated with 1 μM RaxX60-Y, RaxX60-sY or mock solution (with 0.02% Tween20) at 2 days post-infection (dpi). The lesions were measured at 10 to 14 dpi depending on lesion progression.

4.8.1.12. Commercial peptides

Sulfated (sY) and unsulfated (Y) versions of RaxX39 (KGRPEPLDQRLWKHVGGGDYPPPGANPKHDPPPRNPGHH), RaxX24 (LWKHVGGGDYPPPGANPKHDPPPR), RaxX21 (HVGGGDYPPPGANPKHDPPPR), RaxX18 (LWKHVGGGDYPPPGANPK) were ordered from Pacific Immunology. In one case the supplier was unable to provide high quality peptides for RaxX39-sY. Mass-spectrometry analysis

by the supplier indicated a K^+ adduct, which was indirectly supported by in house SRM analysis. In addition we ordered RaxX21-sY peptides also from one additional commercial vendor, however this vendor was not able to provide us with peptides meeting in commercial quality standards. The peptides were assessed for purity by HPLC and MS analysis by the vendor. All peptides were resuspended in ddH_2O. No other solvents were required to dissolve the peptide.

4.8.1.13. Sequences of GFP-1UAG

MA*SKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKLPVPWP

TLVTTFSYGVQCFSRYPDHMKRHDFFKSAMPEGYVQERTISFKDDGNYKTRAEVKFEGD

TLVNRIELKGIDFKEDGNILGHKLEYNYNSHNVYITADKQKNGIKANFKIRHNIEDGSVQ

LADHYQQNTPIGDGPVLLPDNHYLSTQSALSKDPNEKRDHMVLLEFVTAAGITHGMDE

LYK

"*" indicates a UAG amber codon

4.8.1.14. Sequences of GFP-3UAG

MA*SKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKLPVPWP

TLVTTFSYGVQCFSRYPDHMKRHDFFKSAMPEGYVQERTISFKDDGNYKTRAEVKFEGD

TLVNRIELKGIDFKEDGNILGHKLEYNYNSH*V*ITADKQKNGIKANFKIRHNIEDGSVQL

ADHYQQNTPIGDGPVLLPDNHYLSTQSALSKDPNEKRDHMVLLEFVTAAGITHGMDEL

YK

"*" indicates a UAG amber codon

4.8.2. Supplemental figures

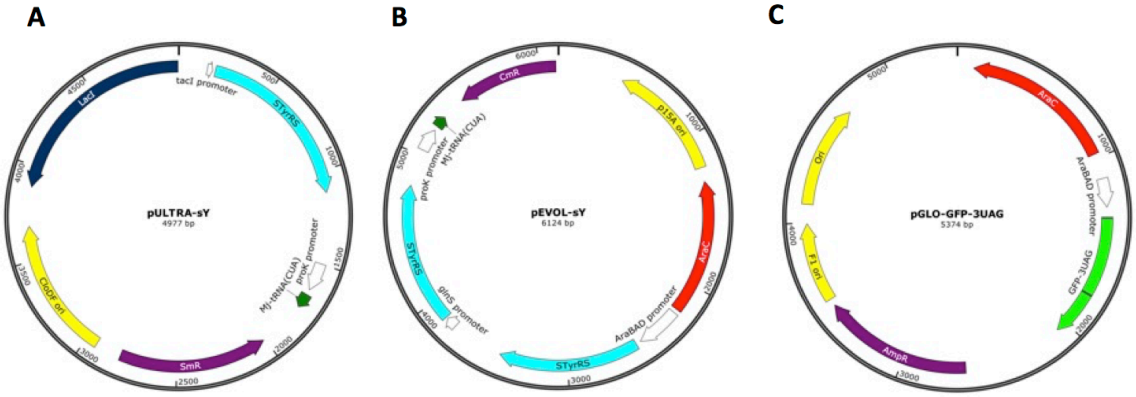

Figure 4.3. Vector Maps of pULTRA-sY, pEVOL-sY, and pGLO-GFP-3UAG.

Plasmid maps of (A) pULTRA-sY, encoding an inducible STyrRS (the sY-specific *Mj* aaRS) controlled by the *tacI* promoter, an optimized *Mj*-$tRNA_{CUA}$, an antibiotic resistance marker for spectinomycin (SmR), and origin (CloDF) for compatibility with common plasmids, (B) pEVOL-sY, an older generation plasmid for sY incorporation, and (C) pGLO-GFP-3UAG, the plasmid for expressing GFP with 3 UAG codons at previously reported permissive sites.[1]

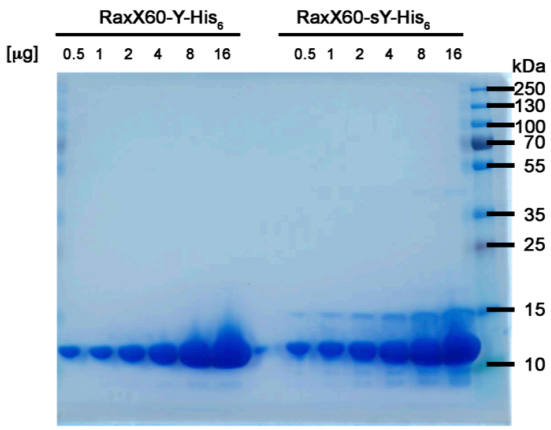

Figure 4.4. High purity isolation of recombinantly produced RaxX60-Y and RaxX60-sY.

8-20% SDS-PAGE gradient gel of different amounts of three-step purified RaxX60-Y and RaxX60-sY from *E. coli* stained with SimplyBlue safe stain. Both proteins migrate at around 12kDa slightly above their predicted molecular weight of 7.2kDa.

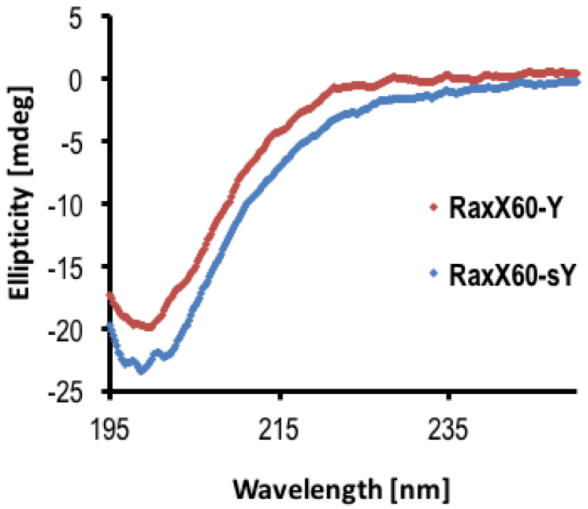

Figure 4.5. RaxX60-Y and RaxX60-sY are largely disordered and potentially form a single β-sheet.

Circular dichroism (CD) spectra of 0.1mg/ml RaxX60-Y and RaxX60-sY at 20^{o}C.

4.8.3. Supplemental tables

Identified Proteins	Accession Number	Molecular Weight	Normalized total spectral count	
			RaxX60-Y	RaxX60-sY
RaxX60-His_6	RS05990_PXO	7 kDa	76 (~97%)	71 (~92%)
Cyclic di-GMP phosphodiesterase YfgF	YFGF_ECOLI	86 kDa	2	0
50S ribosomal protein L2	RL2_ECOLI	30 kDa	0	6

Table 4.1. High purity isolation of RaxX60-Y and RaxX60-sY.

Normalized total spectral count of tryptic peptides identified by HPLC-ESI-MS/MS analysis of three-step purified recombinantly produced RaxX60-Y and RaxX60-sY. Proteins are identified by at least one peptide with a Mascot Ion score ≥ 25.0. Values in brackets indicate approximate relative purity of RaxX60-Y and RaxX60-sY, respectively.

RaxX variant	Purity	Relative sulfation status [%sY]
RaxX21-Y[a]	>80%[c]	N.A.
RaxX24-Y[a]	>80%[c]	N.A.
RaxX39-Y[a]	>95%[c]	N.A.
RaxX60-Y[b]	>90%[d]	N.A.
RaxX21-sY[a]	>80%[c]	95%
RaxX24-sY[a]	>80%[c]	97%
RaxX39-sY[a]	>95%[c]	97%
RaxX60-sY[b]	>90%[d]	>99.5%

Table 4.2. High quality production of sulfated RaxX60-sY in the second-generation sY E. coli expression system.

Summary of the approximated relative sulfation status as measured by SRM-MS analysis of different RaxX variants either produced by standard commercial chemical synthesis ([a]) or recombinant expression in *E. coli* ([b]). Purity is given as provided by the commercial supplier ([c]) or measured by MS/MS analysis ([d]) in Supplementary table 1. "N.A." indicates not applicable.

		RaxX60-Y				RaxX60-sY			
		α-helix [%]	β-strand [%]	turns [%]	unordered [%]	α-helix [%]	β-strand [%]	turns [%]	unordered [%]
CDPro	overall	1.4 – 7.0	18 – 34	11 – 25	39 – 69	3.7 – 10.7	10 – 36	11 – 29	31 – 73
	1	1.4 – 5.8	24 – 34	23 – 25	41 – 48	3.7 – 8.7	23 – 36	24 – 29	31 – 43
	7	2.6 – 4.3	18 – 21	11 – 14	61 – 69	5.4 – 9.0	10 – 18	11 – 14	65 – 73
	10	3.0 – 7.0	29 – 34	23 – 25	39 – 42	5.5 – 10.7	25 – 31	24 – 26	37 – 40
	CDSSTR	1.4 – 3.3	19 – 34	13 – 25	40 – 64	3.7 – 6.2	17 – 33	11 – 26	37 – 65
	CONTIN/LL	4.3 – 7.0	21 – 31	13 – 23	40 – 62	5.1 – 6.7	16 – 36	14 – 29	31 – 65
	SELCON3	3.7 – 5.8	18 – 34	11 – 24	39 – 69	8.7 – 10.7	10 – 25	11 – 26	40 – 73

Table 4.3. RaxX60-Y and RaxX60-sY are largely disordered and potentially form a single β-sheet.

Summary of RaxX60-Y and RaxX60-sY relative secondary structure contents as determined by circular dichroism (CD) spectroscopy. The top row summarizes the overall output ranges generated from the data. The following rows split up the data into output subsets specifying the relationship between the outputs and the CDPro reference protein datasets (1, 7 or 10) or programs (CDSSTR, CONTIN/LL or SELCON3) used to generate them, respectively.

Name	Sequence
aaRStopULTRAPCR1	catcgcggccgcATGGACGAATTTGAAATGATAAAGAGA
aaRStopULTRAPCR2	catcgcggccgcTTATAATCTCTTTCTAATTGGCTCTAAAATC
pEVOL-BB-F	GTCGACCATCATCATCATCA
pEVOL-BB-R	AGATCTAATTCCTCCTGTTA
pEVOL-BB2-F	gagattataaCTGCAGTTTCAAACGCTAAA
pEVOL-BB2-R	attcgtccatATGGGATTCCTCAAAGCGTA
aaRS-F	taacaggaggaattagatctATGGACGAATTTGAAATGAT
aaRS-R	tgatgatgatgatggtcgacTTATAATCTCTTTCTAATTGGC
aaRS2-F	ggaatcccatATGGACGAATTTGAAATGAT
aaRS2-R	gaaactgcagTTATAATCTCTTTCTAATTGGCTC
pGLO-GFP-1TAG-F	catatggcttagAGCAAAGGAGAAGAACTTTT
pGLO-GFP-1TAG-R	tcctttgctctaAGCCATATGTATATCTCCTT
pGLO-GFP-3TAG-F	tcacactaggtatagATCACGGCAGACAAACAAAA
pGLO-GFP-3TAG-R	gtgatctatacctaGTGTGAGTTATAGTTGTACTC
RaxX-FL_EcoRI-3C_forward	AAATGGAATTCGGTCTGGAAGTTCTGTTTCAGGGCCCGAACCACTCGAAAAAATCG
RaxX-His-stop-HindIII-reverse	AAATGTAAGCTTTCAGTGATGGTGGTGATGGTGATGGTGCCCGGGGTTGCG
pMAL-start_E_NcoI	AAATGTCCATGGAAATCGAAGAAGGTAAACTG
RaxX_amber_TAG_F	GGCGGTGGGGACTAGCCCCCGCCGGGCGCGAAT
RaxX_amber_TAG_R	ATTCGCGCCCGGCGGGGGCTAGTCCCCACCGCC

Table 4.4. Primers used in this study.

4.8.4. Supplemental references

1. Chatterjee, A., Sun, S. B., Furman, J. L., Xiao, H. & Schultz, P. G. A Versatile Platform for Single- and Multiple-Unnatural Amino Acid Mutagenesis in Escherichia coli. *Biochemistry* (2013). doi:10.1021/bi4000244

2. Young, T. S., Ahmad, I., Yin, J. A. & Schultz, P. G. An Enhanced System for Unnatural Amino Acid Mutagenesis in E. coli. *J. Mol. Biol.* **395,** 361–374 (2010).

3. Liu, C. C. & Schultz, P. G. Adding new chemistries to the genetic code. *Annu. Rev. Biochem.* **79,** 413–44 (2010).

4. Liu, C. C. & Schultz, P. G. Recombinant expression of selectively sulfated proteins in Escherichia coli. *Nat. Biotechnol.* **24,** 1436–1440 (2006).

5. Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **6,** 343–345 (2009).

6. Lajoie, M. J. *et al.* Genomically Recoded Organisms Expand Biological Functions. *Science (80-.).* **342,** 357–360 (2013).

7. Liu, C. C., Cellitti, S. E., Geierstanger, B. H. & Schultz, P. G. Efficient expression of tyrosine-sulfated proteins in E. coli using an expanded genetic code. *Nat. Protoc.* **4,** 1784–9 (2009).

8. Pruitt, R. N. *et al.* The rice immune receptor XA21 recognizes a tyrosine-sulfated protein from a Gram-negative bacterium. *Sci. Adv.* **1,** e1500245 (2015).

9. Seibert, C. & Sakmar, T. P. Toward a framework for sulfoproteomics: Synthesis and characterization of sulfotyrosine-containing peptides. *Biopolymers* **90,** 459–77 (2008).

10. Monigatti, F., Hekking, B. & Steen, H. Protein sulfation analysis—A primer. *Biochim. Biophys. Acta - Proteins Proteomics* **1764,** 1904–1913 (2006).

11. Sreerama, N. & Woody, R. W. Estimation of Protein Secondary Structure from Circular

Dichroism Spectra: Comparison of CONTIN, SELCON, and CDSSTR Methods with an Expanded Reference Set. *Anal. Biochem.* **287,** 252–260 (2000).

12. Sreerama, N. & Woody, R. W. On the analysis of membrane protein circular dichroism spectra. *Protein Sci.* **13,** 100–112 (2004).

13. Park, C.-J. *et al.* Ectopic expression of rice Xa21 overcomes developmentally controlled resistance to Xanthomonas oryzae pv. oryzae. *Plant Sci.* **179,** 466–471 (2010).

Chapter 5.Characterization of a sulfated anti-HIV antibody using an expanded genetic code

Xiang Li[1], Justin Hitomi[1], and Chang C. Liu*[1,2,3]

1. Department of Biomedical Engineering, University of California, Irvine
2. Department of Chemistry, University of California, Irvine
3. Department of Molecular Biology and Chemistry, University of California, Irvine

Abstract

Tyrosine sulfation is a crucial post-translational modification for sulfated anti-HIV antibodies to neutralize HIV. One of the most neutralizing sulfated anti-HIV antibodies, E51, contains a region in its V_HCDR3 loop with five tyrosines, which are hypothesized to be partially or fully sulfated to bind to HIV's gp120 coat protein. However, the gp120-binding contribution of each sulfate is unclear. In addition, natural sulfation of tyrosine-rich loops usually yields a mixture of multiply sulfated products, complicating attempts to dissect the function of individual E51 sulfoforms with unique sulfation patterns. Here, we use an upgraded expanded genetic code for sulfotyrosine to express homogenous E51 sulfoforms containing up to five sulfates. Through systematic characterization of the 32 possible sulfoforms of E51, we show that only a subset of E51 sulfoforms with two, three, or four sulfated tyrosines bind to gp120 with similar potency as naturally sulfated E51, which we find is a mixture of sulfoforms. We show that sulfation of Tyr100i is necessary for gp120-binding whereas sulfation of Tyr100n is detrimental to binding. These results reveal that gp120-binding by E51 requires very specific sulfation patterns and should aid in the further design of sulfated E51-based peptides and immunoadhesins against HIV.

5.1. Introduction

Tyrosine sulfation is a post-translational modification (PTM) that plays a major role in mediating extracellular protein-protein interactions involved in chemotaxis[1,2], leukocyte trafficking[3,4], anti-coagulation[5], plant immunity[6,7], and viral infection[8,9]. Among these, perhaps the most extensively studied tyrosine-sulfate-dependent process is how HIV hijacks a tyrosine-sulfated co-receptor, either chemokine receptor CCR5 or CXCR4, during infection. Not only does the interaction between CCR5 with HIV's gp120 coat protein depend on tyrosine sulfation[8], a class of anti-HIV antibodies has been discovered to also require tyrosine sulfation to inhibit HIV infection. These CD4-induced (CD4i) antibodies use tyrosine sulfation to mimic the tyrosine-sulfated region of CCR5 in order to compete for gp120 binding[10].

Among tyrosine-sulfated anti-HIV antibodies, the most neutralizing is the E51 antibody, which prevents a range of HIV isolates from infecting either CCR5 or CXCR4 cells *in vitro*[12]. Recently, an engineered immunoadhesin consisting of a peptide derived from the tyrosine-sulfated region of E51, located in the third complementary determining region of the heavy chain (V_HCDR3), and two domains of CD4 was shown to effectively neutralize a broad panel of different HIV isolates[12]. When delivered as a gene therapy, this immunoadhesin also prevented multiple challenges of SIV in macaques, making it a possible alternative to an HIV vaccine.

Determining which sulfated tyrosine of E51 interact with gp120 may provide biochemical insights to improve sulfated anti-HIV antibodies, peptides, or immunoadherins. However, dissecting the sulfation patterns of E51 and the contribution of each sulfate is a difficult task. Tyrosine sulfation is mediated by tyrosylprotein sulfotransferases (TPSTs), which enzymatically install sulfate groups onto target tyrosines within sulfation sequence motifs. The V_HCDR3 of E51 contains five tyrosines within its predicted sulfation sequence motif (Figure A), and it is

hypothesized that most or all of these tyrosines can be sulfated[13–15]. However, tyrosine sulfation often yields heterogeneously sulfated proteins due to incomplete sulfation or sequence-dependent enzymatic activities of different TPST isoforms[16–18]. One method to dissect the sulfation patterns of multiply-sulfated proteins utilizes advanced liquid chromatography to isolate sulfated peptides with unique sulfation patterns from mixtures[17,19,20], but this method cannot separate large intact proteins, which are often required for functional assays. Alternatively, one may mutate candidate sulfated tyrosines, but this is complicated by the possibility that the unsulfated tyrosine may also contribute to function and the fact that sulfated tyrosines can promote nearby tyrosines to become sulfated[21]. Thus far, neither the crystal structure nor biochemical studies of E51 have fully elucidated which sulfated tyrosines interact with gp120[13,15].

We have shown that an *Escherichia coli* strain with an expanded genetic code system for sulfotyrosine (sY) can express homogenously, site-specifically tyrosine-sulfated proteins without reliance on any post-translational modification processes[22–24]. In this system, *E. coli* encodes an orthogonal sY-specific aminoacyl-tRNA synthetase (STyrRS)/$tRNA_{CUA}$ pair to incorporate sY at amber (UAG) codons (Figure 1B). The incorporation of sY specifically occurs at UAG codons, as STyrRS does not aminoacylate endogenous *E. coli* tRNAs nor do endogenous *E. coli* synthetases aminoacylate $tRNA_{CUA}$[24]. Here, we use an upgraded version of this system to express all 32 possible sulfoforms of E51 (Table 1) including one that contains five sYs. We find that several pure E51 sulfoforms bind to gp120 with high potency, matching the activity of a mixture of E51 sulfoforms obtained through natural sulfation by TPSTs in human cells. We show that among the five tyrosines, sulfation of Y100i is the most critical for gp120-binding whereas sulfation of Y100n is detrimental to gp120. Our systematic characterization of all 32 possible sulfoforms of E51 demonstrates the capabilities of expanded genetic code systems to recapitulate exact

posttranslational modification patterns, elucidates essential sulfation events in E51 important for gp120-binding, and highlights how the gp120-binding potencies of multiply-tyrosine-sulfated E51s are not simply the sum of the contribution of individual sulfates.

5.2. Methods

5.2.1. Construction of expression plasmids for E51 Fabs and small hairpin RNAs.

For expression in bacterial cells, E51 Fab cassette for sulfoform **1** was synthesized (IDT) and cloned into the pGLO expression vector *via* Gibson Assembly. The remaining E51 sulfoforms with different numbers of UAG codons were sequentially cloned *via* inverse PCR using primers listed in Table 4. The resulting pGLO-E51 plasmids were co-transformed with pULTRA-sY (Addgene #82417) into C321.ΔA.exp cells (Addgene #49018).

For expression in mammalian cells, the variable and constant region of E51 heavy chain fused to a C-terminal HA tag for sulfoform **1** was cloned into pCMV expression vector from pGLO-E51 via Gibson Assembly while the variable and constant region of E51 light chain for sulfoform **1** was cloned into a separate pCMV following the same process. DNA fragments corresponding to small hairpin RNAs (shRNAs) that target tyrosylprotein sulfotransferase-1 (TPST-1) (5'-GGGCCATGCTGGACGCACATC-3') or tyrosylprotein sulfotransferase-2 (TPST-2) (5'-GGGACAGCAGGTGCTAGAGTG-3') were cloned into a custom mammalian expression vector via inverse PCR.

5.2.2. Expression of E51 Fabs in bacterial cells with an expanded genetic code for sY.

32 E51 sulfoforms were expressed in C321.ΔA.exp cells in M9T minimal media. Briefly, single colonies of C321.ΔA.exp cells with pULTRA-sY and pGLO-E51 were picked into Terrific Broth (TB). After growing the cells at 37°C and 200 rpm for 16 hours, the saturated cultures in TB were diluted into M9T (1X M9 salts, 10 g/L tryptone, 5 g/L NaCl, 1 mM $MgSO_4$, and 80 μM biotin) at a ratio of 1:100 and grown to saturation at 37°C and 200 rpm for 12 hours. 2 mL of the

saturated M9T culture was added to 200 mL of M9T with 0, 3, 5, or 10 mM of sY for expressing E51 sulfoforms with 0, 1-3, 4, or 5 UAG codons, respectively, and grown to mid-log phase (OD_{600} = 0.6-0.8) at 37°C and 200 rpm. The expression of STyrRS and E51 Fabs was induced by adding 0.2% of L-arabinose and 1 mM of IPTG and incubating the cultures at 22°C and 200 rpm for 24 hours.

Fabs were extracted from the cultures using either of two methods. The first method involves extracting Fabs from the periplasm through cold osmotic shock. Cell pellets were resuspsended in 16 mL of cold periplasmic lysis buffer (20% sucrose, 20 mM Tris, 2 mM EDTA, pH 8.0) with protease inhibitor cocktail (Sigma) and incubated at 22°C and 200 rpm for 4 hours. The supernatant of the lysis consisting of the first periplasmic fraction was stored at 4°C, and the cell pellet was resuspended in 16 mL of cold periplasmic lysis buffer with protease inhibitor cocktail and incubated at 4°C for overnight. The second supernatant of the lysis consisting of the second periplasmic extraction was combined with the first periplasmic extraction. The combined periplasmic fractions were filtered using a 0.22 μm filter, concentrated using a 10 kDa molecular weight cut off (MWCO) spin-filter column (MilliporeSigma), and then purified using Protein G columns (Pierce) according to the manufacturer's protocols. The second method involves concentrating Fabs from the supernatant of the culture via ammonium sulfate precipitation. After the supernatant of Fab cultures was filtered using a 0.22 μm filter, 0.2 M of ammonium sulfate was gradually added to the supernatant with stirring, and the solution was incubated to fully mix at 22°C for 1 hour. The precipitated Fabs were pelleted through centrifugation at 4°C and 10,000 g for 15 min, resuspended in PBS, concentrated using a 10 kDa molecular weight cut off (MWCO) spin-filter column (MilliporeSigma), and then purified using Protein G columns (Pierce) according to the manufacturer's protocols. The final concentrations of Fabs were measured on a microvolume

spectrophotometer (NanoDrop) using the appropriate molecular weight and ε value calculated on ExPASy ProtParam.

5.2.3. Expression of E51 Fabs in mammalian cells.

E51 Fabs sulfated by TPSTs were expressed in HEK293T cells. To increase the sulfation of E51, plasmids encoding the variable and constant region of E51 light or heavy chain were co-transfected into HEK293T cells in serum media at ~90% confluency with plasmids encoding TPST-1 or TPST-2 (Addgene #11252 and 11253 respectively). To decrease the sulfation of E51, plasmids encoding the variable and constant region of E51 light or heavy chain were co-transformed into HEK293T cells in serum media at ~90% confluency with plasmids encoding shRNAs that target TPST-1 or TPST-2. After 12 hours, the media was swapped with serum-free media and subsequently incubated at 28°C for 5 days for protein expression. The supernatant of the culture was collected, concentrated using a 10 kDa MWCO spin-filter column, and purified using protein G columns.

5.2.4. Analysis of gp120-binding using ELISA.

0.2 μg of soluble gp120 (ADA, Clade B) or gp120(Q422L) (Immune-tech) in 100 μL of PBS was coated onto the surface of a polystyrene high binding microplate well at 4°C for overnight. After blocking the wells with 200 μL of 2% Milk and then washing 3 times with 300 μL of PBST (PBS + 0.05% Tween-20), equal amounts of Fab samples with or without the addition of 0.2 μg of sCD4 (Progenics) in 100 μL of 2% Milk PBST were added and incubated at 37°C for 2 hours. The wells were washed 5 times with 200 μL of PBST, incubated with 100 μL of a rabbit anti-HA polyclonal antibody (ThermoFisher Scientific) at 37°C for 1 hour, washed 5 times with 200 μL of PBST, and then incubated with 100 μL of an HRP-conjugated anti-rabbit polyclonal antibody (Sigma) at 37°C for 1 hour, washed 5 times with 200 μL of PBST, and incubated with 100 μL of QuantaBlu solution (ThermoFisher Scientific) at 22°C for 30 minutes. Fluorescence (ex: 325 nm,

em: 420 nm) was measured using a microplate reader (TECAN). ELISA binding curves were fitted using a 4PL model fit and the concentration at half of the maximum binding signal (EC_{50}) was determined by interpolation. ELISA binding curves and EC_{50} graphs were made using Graphpad Prism 5.0.

5.2.5. Mass spectrometry of Fabs.

Intact Fab samples were analyzed for intact mass by UPLC-MS. E51 Fabs were dialyzed in 0.1% formic acid and diluted to a concentration of 100 ng/μL. 2 μL of the diluted samples were injected into a Xevo G2-XS QTof mass spectrometer equipped with a lockspray ion source (Waters). Intact Fabs were separated with on a AQUITY UPLC protein column (300Å, 2.1 x 100 mm, BEHC C4 1.7 μm) using a 10 minute linear gradient at a flow rate of 0.2 mL/min. The instrument was operated in positive-ion mode with a capillary voltage of 3.0 kV. Solvent A was composed of 0.1% formic acid in water, and solvent B was composed of 0.1% formic acid in acetonitrile. The samples were calibrated with lock mass composed of 50 ng/mL of Leucine encephalin, and the instrument was calibrated with sodium iodide. MS1 scan was performed at 0.5 seconds across the range of 400-4000 m/z. The desolvation gas and ion source temperatures were set to 300 C and 100C, respectively.

The extracted charge state ladders of the intact Fabs were deconvolved using MaxEnt 1 on the Applied Waters Markerlynx XS software. The deconvolution was done with the following settings: mass range of 45,000–50,000 Da, resolution of 0.45 Da/channel, damage model set to Uniform Gaussian mode with width at half height of 0.8 Da, minimum intensity ratios at 33% each, and maximum 12 iterations.

5.3. Results

5.3.1. Expression of homogenous sulfoforms of E51 Fabs using an upgraded expanded genetic code for sY.

Our upgraded expanded genetic code for sY[7] can incorporate up to five sYs in a single protein, enabling the expression of all 32 sulfoforms of E51. To express all 32 sulfoforms of E51 (either Tyr or sY at positions 100f, 100i, 100m, 100n, and 100o, corresponding to the possible sites of tyrosine sulfation in E51), we transformed pULTRA-sY, an improved expression vector for optimized expression of STyrRS and the corresponding $tRNA_{UAG}$[25], alongside a plasmid encoding an E51 Fab (Figure 7) into a genomically recoded *E. coli* strain, C321.ΔA.exp[26], that lacks genomic amber stop codons and release factor 1. We initially obtained good yields for 30 out of 32 sulfoforms of E51 in the Fab format with the addition of 3 mM sY to the growth media (Figure 8). While the expression of unsulfated E51 Fab, **1**, in C321.ΔA.exp cells (0.245 mg/L) was lower than Fab expression yields in standard protein expression strains (~1-5 mg/L), the incorporation of one to four sYs was relatively efficient. However, the expression levels of **28**, one of the E51 Fabs with four sYs, and of **32**, the E51 Fab with five sYs, were low. We hypothesized that the permeability of the negatively charged sY into cells limited the incorporation of multiple sYs in these cases. Increasing the concentration of sY to 5 mM for **28** and 10 mM for **32** in the growth media improved the yields of both Fabs (Figure 9). Mass spectrometry analysis of each of the 32 sulfoforms of E51 Fabs showed a single major peak corresponding to the predicted mass (Table 2), confirming that sY-incorporation into E51 Fabs results in homogenously sulfated products.

5.3.2. Expression of heterogenously sulfated E51 Fabs in mammalian cells capable of tyrosine sulfation.

To compare E51 sulfoforms with naturally sulfated E51, we expressed E51 sulfated by TPSTs (E51-TPSTs) in mammalian cells and determined that E51-TPSTs is a mixture of E51 isoforms primarily with two, three, and four sulfates. For the expression of E51-TPSTs, we co-transfected HEK293T cells with plasmids encoding E51 Fab and plasmids encoding human TPST-

1 or TPST-2 to elevate tyrosine sulfation. Mass spectra of E51-TPSTs showed multiple peaks corresponding to E51 with different number of sulfates: three major peaks corresponding E51 with two, three, and four sulfates, and two minor peaks corresponding E51 with one and five sulfates (Figure 2A). This result indicates that naturally sulfated E51 is a mixture composed of different sulfoforms. We also obtained E51 Fab with reduced tyrosine sulfation by TPSTs (E51-shRNAs) by co-transfecting HEK293T cells with the E51 Fab expression plasmids and plasmids encoding shRNAs capable of interfering the message of endogenous TPST-1 or TPST-2. Mass spectra of E51-shRNAs showed two major peaks that correspond to E51 with zero and one sulfate (Figure 2B) presumably due to the incomplete knock-down of endogenous human TPSTs.

5.3.3. gp120-binding by E51 sulfoforms.

Although only a subset of the 32 possible E51 sulfoforms showed detectable binding to gp120, a few sulfoforms exhibited similar gp120-binding potency and CD4-induced binding as naturally sulfated E51. We used ELISA to test the gp120-binding of 32 sulfoforms of E51 as well as E51-TPSTs and E51-shRNAs. Since E51 is a CD4i antibody, we performed the binding assays in the presence or absence of soluble CD4 (sCD4). We used a range of the Fab concentrations in our ELISAs in order to determine the half-maximum gp120-binding concentrations (EC_{50}). Most of the E51 sulfoforms did not exhibit detectable ELISA binding signals (Figure 10). However, 12 out of 32 E51 sulfoforms along with E51-TPSTs and E51-shRNAs bound to gp120 in the presence of sCD4 with EC_{50} values in the nanomolar range. In particular, several E51 sulfoforms with two, three, or four sYs (**7**, **17**, and **28**) bound to gp120 strongly as E51-TPSTs with EC_{50} of ~0.9 nM (Figure 3A and Table 3). Six of the 12 E51 sulfoforms and E51-TPSTs showed weak binding to gp120 in the absence of sCD4. Since E51 and sulfated immunoadhesions neutralize HIV infection in the presence of CD4[11–13], we used EC_{50}s from ELISA conducted with sCD4 for further analysis.

In addition to their similar binding profiles, E51 sulfoforms and E51-TPSTs expectedly targeted the same CCR5-binding epitope of gp120. This was determined by assaying our E51 Fabs against mutant gp120(Q422L). The Q422L mutation in the CCR5-binding epitope of gp120 has been shown to disrupt the binding of CCR5 and E51[27]. We observed no binding for E51 sulfoforms and E51-TPSTs to gp120(Q422L) in contrast to a control anti-gp120 Fab that binds a different epitope of gp120 (Figure 3B,C). The similar binding profiles between certain E51 sulfoforms and E51-TPSTs, their strong sCD4 induction, and the binding epitope consistency suggest that E51 sulfoforms expressed using an expanded genetic code accurately recapitulate the activity of the sulfoforms found in E51 mixtures resulting from posttranslational modification by TPSTs.

5.3.4. Key E51 determinants for binding to gp120.

Analysis of the EC_{50} values of E51 sulfoforms shows that sulfation of Tyr100i is the most crucial E51 residue for gp120 binding. Every E51 sulfoform with detectable gp120 binding contains sulfated Tyr100i, and sulfation of Tyr100i alone (*i.e.* sulfoform **3**) enables E51 to bind to gp120 (EC_{50} = 1.854 nM), although the combination of sulfated Tyr100i and sulfated Tyr100f provides the strongest interaction with gp120 (EC_{50} = 0.883 nM). On the other hand, sulfation of Tyr100n is deleterious for binding to gp120 (Figure 4). E51 sulfoforms that contain sulfated Tyr100n (**18**, **27**, **29**, and **32**) have higher EC_{50} than that of their E51 sulfoform counterparts that contain unsulfated Tyr100n (**7**, **17**, **19**, and **28**). The negative impact of 100i is particularly noticeable for **7**. While sulfation of Tyr100m and Tyr100o of **7** (i.e. sulfoform **17** and **19**) minimally changes its EC_{50}, sulfation of Tyr100n of **7** (i.e. sulfoform **18**) decreases its gp120-binding EC_{50} by 6-fold. Moreover, there are four E51 sulfoforms that contain sulfated Tyr100i but did not detectably bind to gp120, likely due to the sulfation of Tyr100n. The fact that the sulfation of Tyr100n would be detrimental to binding to gp120 is surprising, because we do find a

component in E51-TPSTs with all five tyrosines sulfated, necessitating sulfation at Tyr100n, although this is a minor peak in the mass spec analysis.

Additional mutagenic experiment suggests that Tyr100n in the absence of sulfate strongly contributes to gp120 binding. Since our expanded genetic code enables the expression of any sulfoform without risking a change in sulfation pattern due to mutations in the sulfation motif, we were able to assay whether unsulfated Tyr100n is superior to other amino acids at position 100n in the context of the highest-gp120-affinity E51 sulfoform, **7**. We first tested Phe and Asp mutations because phenylalanine is structurally similar to tyrosine and aspartic acid mimics the negative charge of sY and is also commonly found in the binding site of unsulfated CD4i anti-HIV antibodies[15]. Aspartic acid at 100n reduced gp120-binding whereas phenylalanine at 100n did not (Figure 5A). Further site-saturation mutagenesis of 100n with all other amino acids allowed us to conclude that an aromatic amino acid without a sulfate modification is optimal at position 100n (Figure 5B). We also tested whether Phe or Asp mutations to Tyr100m and Tyr100o of **7** could alter gp120-binding, as sulfation of these tyrosines only marginally contribute to gp120-binding. In the presence of sCD4, every Phe mutant bound to gp120 with similar binding profiles as unmutated **7** (Figure 5A). The Tyr-100m-Asp mutant and the Tyr-100n-Asp mutant, however, showed reduced gp120-binding. In the case of Tyr-100n-Asp, the aspartic acid mutation completely eliminated its ability to bind to gp120, suggesting that Tyr100n strongly contributes to gp120-binding.

5.4. Discussion

Tyrosine sulfation plays an important role for E51 and immunoadherins derived from E51 to neutralize HIV infection[11,12]. However, the contribution of critical sulfated tyrosines and sulfation patterns responsible for E51's binding to gp120 are not well-studied. We have

demonstrated that an upgraded expanded genetic code for sY can site-specifically incorporate up to five sYs into E51 for the expression of all 32 homogenous sulfoforms of E51. Our systematic characterization of E51 sulfoforms reveals not only key determinants for binding gp120 but also the possible sulfation patterns of naturally sulfated E51 based on highly functional sulfoforms and their comparison to the activity of a mixture of E51s produced through sulfation by TPSTs in mammalian cells.

It was previously assumed that E51 uses at least three sYs to prevent gp120 from binding to multiply-sulfated receptors such as CCR5[13–15], but our results indicate that one of the best binding E51 sulfoforms, **7**, only requires the sulfation of two tyrosines: Tyr100f and Tyr100i. There are two other high-affinity sulfoforms with EC_{50} values approximately the same as **7** that contain three and four sYs, but the presence of those one or two sYs makes little impact. In addition, E51 with 5 sYs actually weakly associates with gp120. This is due to the negative contribution of sulfation at Tyr100n. These results suggest that the specific patterns of sYs are important for interacting with the CCR5-binding epitope of gp120 and it is not enough to consider just the number of sYs localized in the V_HCDR3. In fact, the sulfation pattern of the doubly-sulfated **7** closely resembles that of a well-characterized, doubly-sulfated anti-HIV antibody, 412d, where sulfated tyrosine 100c buries itself into the CCR5-binding epitope of gp120 while sulfated tyrosine 100 forms peripheral electrostatic interactions with gp120[10]. Sequence alignment of the V_HCDR3 of 412d and E51 shows homology between Tyr100i of E51 and Y100c of 412d (Figure 5). Interestingly, we find that sulfation of Tyr100i of E51 contributes the most to gp120 binding by E51 much as sulfation of Y100c contributes the most to gp120 binding by 412d[28]. The similarity in which sulfates contribute most to gp120 binding between E51 and 412d suggests that the gp120-

binding mechanism between E51 and gp120 is also similar and that one can model the E51-gp120-CD4 interaction based on the crystal structure of 412d in complex with gp120 and CD4[10].

Given that sulfation of Tyr100f and Tyr100i are important E51 determinants for binding to gp120 and that sulfoforms **7**, **17** and **28**, all of which contains sulfation of Tyr100f and Tyr100i, bind gp120 equally well as E51-TPSTs (Figure 3A), we postulate that the major components in the mixture of E51-TPSTs are also these sulfoforms. The sulfation patterns of sulfoforms **7**, **17** and **28** (containing two, three and four sYs respectively) are consistent with the mass spec analysis of E51-TPSTs, which shows major peaks corresponding to two, three, and four sulfates. However, mass spec analysis of E51-TPSTs also shows a peak, albeit minor, corresponding to five sulfates. Given that this requires the sulfation of Tyr100n, which is detrimental to binding in all contexts, the E51-TPSTs mixture would be improved if this component were absent. Indeed, one may ask why nature chooses to have Tyr100n in the first place, risking its sulfation that decreases activity, if our mutagenesis data shows that Tyr100n can be replaced with Phe without any functional penalty. We hypothesize that this is due to the sequence motif requirements of posttranslational sulfation. Since sulfated tyrosines themselves can induce further sulfation of nearby tyrosines, it may be the case that sulfation of Tyr100n increases the efficiency of TPSTs in sulfating Tyr100f or Tyr100i whose sulfation is important for gp120 binding. For instance, a mutagenesis study on E51 peptides sulfated by TPSTs showed that Phe mutants (YYFYY, YYYFY, and YYYYF) did not bind to gp120,13 whereas we showed in this study that certain Phe mutants don't reduce gp120-binding when key sulfates are left intact. This suggests that for posttranslationally sulfated E51, Phe mutation in YYFYY, YYYFY and YYYYF abrogates binding to gp120 because it prevents nearby tyrosines from being sulfated, and not because the mutation per se reduces activity. Similarly, the sulfation of the N-terminus of CCR5 (CCR5 1-18) occurs on Y3 before Y10, Y14,

and Y15[29], even though only the sulfation of the latter three tyrosines are essential for binding to gp120[30]. In this case, sulfation of Y3 increases the overall sulfation of CCR5 1-18, ensuring the sulfation of the key tyrosines, Y10, Y14, and Y15. In short, dependency of tyrosine sulfation on a sulfation motif means that the sulfation of one tyrosine cannot always be decoupled from the sulfation of another, creating necessary tradeoffs at the functional level.

These tradeoffs can be broken with an expanded genetic code that directly incorporates sY at any location in a protein, with important implications. First, we can now dissect the contribution of tyrosine sulfates down to the resolution of an individual sulfated tyrosine. Second, we can now access sulfation patterns and sulfated mutants that natural TPSTs cannot. Third, we can now to produce homogenously sulfated proteins in cases where natural TPSTs would yield a mixture. As we have shown in this study, these three implications of using an expanded genetic code to produce sulfated E51s enabled us to fully survey the functional landscape of E51's sulfated tyrosines in gp120 binding. In future work, we should be able to discover more active variants of E51 that are not accessible through sulfation by TPSTs, produce homogenous preparations of only the most active sulfoforms of E51 and E51-derived peptides and immunoadherins for therapeutic applications, and carry out similar studies on other multiply-sulfated proteins such as chemokine receptors.

5.5. Acknowledgements

The following reagent was obtained through the NIH AIDS Reagent Program, Division of AIDS, NIAID, NIH: Human Soluble CD4 Recombinant Protein (sCD4) from Progenics. We would like to thank the University of California at Irvine for financial support.

5.6. References

(1) Tan, J. H. Y.; Ludeman, J. P.; Wedderburn, J.; Canals, M.; Hall, P.; Butler, S. J.; Taleski, D.; Christopoulos, A.; Hickey, M. J.; Payne, R. J.; Stone, M. J. *J. Biol. Chem.* **2013**, *288* (14), 10024–10034.

(2) Colvin, R. a; Campanella, G. S. V; Manice, L. a; Luster, A. D. *Mol. Cell. Biol.* **2006**, *26* (15), 5838–5849.

(3) Sako, D.; Comess, K. M.; Barone, K. M.; Camphausen, R. T.; Cumming, D. A.; Shaw, G. D. *Cell* **1995**, *83* (2), 323–331.

(4) Wilkins, P. P.; Moore, K. L.; Mcever, R. P.; Cummings, R. D. *J. Biol. Chem.* **1995**, *50* (21), 22677–22681.

(5) Priestle, J. P.; Rahuel, J.; Rink, H.; Tones, M.; Grutter, M. G. **1993**, 1630–1642.

(6) Pruitt, R. N.; Schwessinger, B.; Joe, a.; Thomas, N.; Liu, F.; Albert, M.; Robinson, M. R.; Chan, L. J. G.; Luu, D. D.; Chen, H.; Bahar, O.; Daudi, a.; De Vleesschauwer, D.; Caddell, D.; Zhang, W.; Zhao, X.; Li, X.; Heazlewood, J. L.; Ruan, D.; Majumder, D.; Chern, M.; Kalbacher, H.; Midha, S.; Patil, P. B.; Sonti, R. V.; Petzold, C. J.; Liu, C. C.; Brodbelt, J. S.; Felix, G.; Ronald, P. C. *Sci. Adv.* **2015**, *1* (6), e1500245–e1500245.

(7) Schwessinger, B.; Li, X.; Ellinghaus, T. L.; Chan, L. J. G.; Wei, T.; Joe, A.; Thomas, N.; Pruitt, R.; Adams, P. D.; Chern, M. S.; Petzold, C. J.; Liu, C. C.; Ronald, P. C. *Integr. Biol.* **2016**, 2006–2009.

(8) Farzan, M.; Mirzabekov, T.; Kolchinsky, P.; Wyatt, R.; Cayabyab, M.; Gerard, N. P.; Gerard, C.; Sodroski, J.; Choe, H. *Cell* **1999**, *96*, 667–676.

(9) Choe, H.; Moore, M. J.; Owens, C. M.; Wright, P. L.; Vasilieva, N.; Li, W.; Singh, A. P.; Shakri, R.; Chitnis, C. E.; Farzan, M. *Mol. Microbiol.* **2005**, *55* (5), 1413–1422.

(10) Huang, C.-C.; Lam, S. N.; Acharya, P.; Tang, M.; Xiang, S.-H.; Hussan, S. S.-U.;

Stanfield, R. L.; Robinson, J.; Sodroski, J.; Wilson, I. a; Wyatt, R.; Bewley, C. a; Kwong, P. D. *Science* **2007**, *317* (5846), 1930–1934.

(11) Choe, H.; Li, W.; Wright, P. L.; Vasilieva, N.; Venturi, M.; Huang, C.-C.; Grundner, C.; Dorfman, T.; Zwick, M. B.; Wang, L.; Rosenberg, E. S.; Kwong, P. D.; Burton, D. R.; Robinson, J. E.; Sodroski, J. G.; Farzan, M. *Cell* **2003**, *114* (2), 161–170.

(12) Gardner, M. R.; Kattenhorn, L. M.; Kondur, H. R.; von Schaewen, M.; Dorfman, T.; Chiang, J. J.; Haworth, K. G.; Decker, J. M.; Alpert, M. D.; Bailey, C. C.; Neale, E. S.; Fellinger, C. H.; Joshi, V. R.; Fuchs, S. P.; Martinez-Navio, J. M.; Quinlan, B. D.; Yao, A. Y.; Mouquet, H.; Gorman, J.; Zhang, B.; Poignard, P.; Nussenzweig, M. C.; Burton, D. R.; Kwong, P. D.; Piatak, M.; Lifson, J. D.; Gao, G.; Desrosiers, R. C.; Evans, D. T.; Hahn, B. H.; Ploss, A.; Cannon, P. M.; Seaman, M. S.; Farzan, M. *Nature* **2015**, *519* (7541), 87–91.

(13) Dorfman, T.; Moore, M. J.; Guth, A. C.; Choe, H.; Farzan, M. *J Biol Chem* **2006**, *281* (39), 28529–28535.

(14) Chiang, J. J.; Gardner, M. R.; Quinlan, B. D.; Dorfman, T.; Choe, H.; Farzan, M. *J. Virol.* **2012**, *86* (22), 12417–12421.

(15) Huang, C.; Venturi, M.; Majeed, S.; Moore, M. J.; Phogat, S.; Zhang, M.-Y.; Dimitrov, D. S.; Hendrickson, W. a; Robinson, J.; Sodroski, J.; Wyatt, R.; Choe, H.; Farzan, M.; Kwong, P. D. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101* (9), 2706–2711.

(16) Mikkelsen, J.; Thomsen, J.; Ezban, M. *Biochemistry* **1991**, *30* (6), 1533–1537.

(17) Seibert, C.; Veldkamp, C. T.; Peterson, F. C.; Chait, B. T.; Volkman, B. F.; Sakmar, T. P. *Biochemistry* **2008**, *47* (43), 11251–11262.

(18) Hartmann-fatu, C.; Bayer, P. *Chem. Biol. Interact.* **2016**, *259*, 17–22.

(19) Yu, Y.; Hoffhines, A. J.; Moore, K. L.; Leary, J. A. *Nat. Methods* **2007**, *4* (7), 583–588.

(20) Seibert, C.; Cadene, M.; Sanfiz, A.; Chait, B. T.; Sakmar, T. P. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99* (17), 11031–11036.

(21) Tanaka, S.; Nishiyori, T.; Kojo, H.; Otsubo, R.; Tsuruta, M.; Kurogi, K.; Liu, M.-C.; Suiko, M.; Sakakibara, Y.; Kakuta, Y. *Sci. Rep.* **2017**, *7* (1), 8776.

(22) Liu, C. C.; Cellitti, S. E.; Geierstanger, B. H.; Schultz, P. G. *Nat. Protoc.* **2009**, *4* (12), 1784–1789.

(23) Liu, C. C.; Mack, A. V; Tsao, M.-L.; Mills, J. H.; Lee, H. S.; Choe, H.; Farzan, M.; Schultz, P. G.; Smider, V. V. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105* (46), 17688–17693.

(24) Liu, C. C.; Schultz, P. G. *Nat. Biotechnol.* **2006**, *24* (11), 1436–1440.

(25) Chatterjee, A.; Sun, S. B.; Furman, J. L.; Xiao, H.; Schultz, P. G. *Biochemistry* **2013**.

(26) Lajoie, M. J.; Rovner, a. J.; Goodman, D. B.; Aerni, H.-R.; Haimovich, a. D.; Kuznetsov, G.; Mercer, J. a.; Wang, H. H.; Carr, P. a.; Mosberg, J. a.; Rohland, N.; Schultz, P. G.; Jacobson, J. M.; Rinehart, J.; Church, G. M.; Isaacs, F. J. *Science (80-.).* **2013**, *342* (6156), 357–360.

(27) Xiang, S.-H.; Wang, L.; Abreu, M.; Huang, C.-C.; Kwong, P. D.; Rosenberg, E.; Robinson, J. E.; Sodroski, J. *Virology* **2003**, *315* (1), 124–134.

(28) Liu, C. C.; Choe, H.; Farzan, M.; Smider, V. V; Schultz, P. G. *Biochemistry* **2009**, *48* (37), 8891–8898.

(29) Jen, C. H.; Moore, K. L.; Leary, J. A. *Biochemistry* **2009**, *48* (23), 5332–5338.

(30) Liu, X.; Malins, L. R.; Roche, M.; Sterjovski, J.; Duncan, R.; Garcia, M. L.; Barnes, N. C.; Anderson, D. A.; Stone, M. J.; Gorry, P. R.; Payne, R. J. *ACS Chem. Biol.* **2014**, *9* (9), 2074–2081.

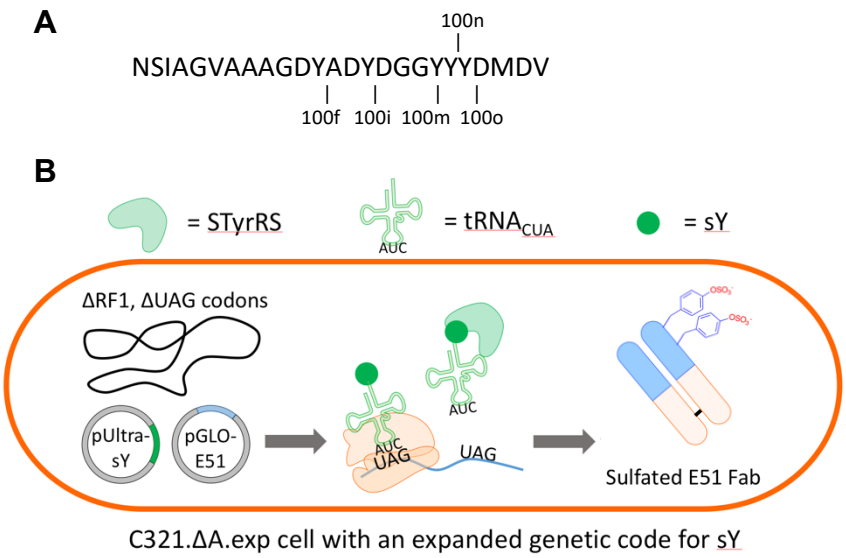

Figure 5.1. Expression of sulfated E51 Fab using an expanded genetic code for sY. (A) Protein sequence of E51 V_HCDR3 and candidate sulfated tyrosines at positions 100f, 100i, 100m, 100n, and 100o (Kabat numbering). (B) Depiction of the expression of sulfated E51 Fab in a C321.ΔA.exp cell capable of sY-incorporation at UAG codons. The lack of UAG stop codons and release factor 1 (RF1) in the C321.ΔA.exp genome eliminates RF1 competition in the cell, maximizing sY-incorporation at UAG codons, mediated by STyrRS and $tRNA_{CUA}$ expressed from pULTRA-sY, in E51 Fab transcripts expressed from pGLO-E51. Highly efficient sY-incorporation yields homogenously sulfated E51 Fabs containing up to five sYs.

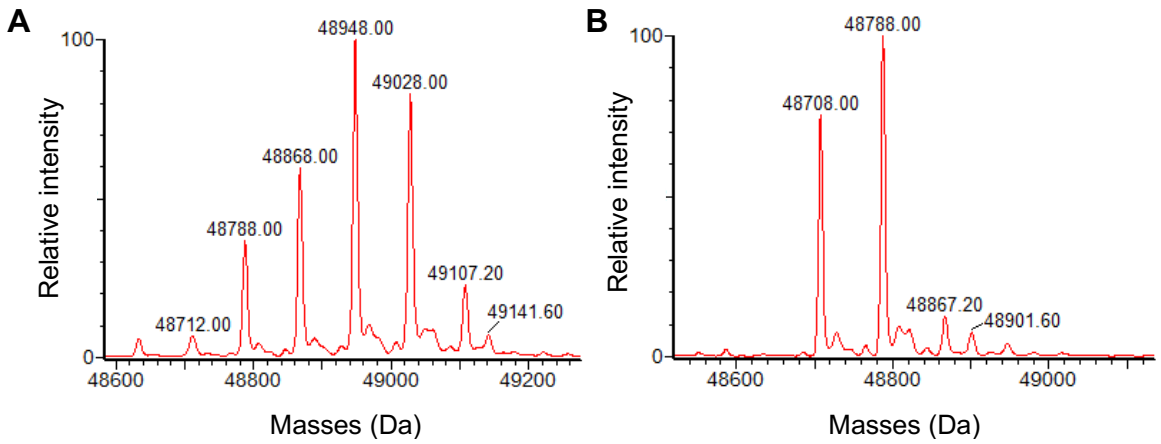

Figure 5.2. Mass spectra of naturally sulfated E51 show heterogenous sulfation by TPSTs.

(A) Deconvoluted mass spectrum of intact E51-TPSTs. 48,712 Da = E51 with 0 sY, 48,788 Da = E51 with 1 sY, 48,866 Da = E51 with 2 sYs, 48,948 Da = E51 with 3 sYs, 49,208 Da = E51 with 4 sYs, 49,107 Da = E51 with 5 sYs, 49,141.60 Da = E51 with 5 sYs that did not undergo cyclization of N-terminal glutamines in both heavy and light chain. (B) Deconvoluted mass spectrum of intact E51-shRNAs. 48,708 Da = unsulfated E51, 48,788 Da = E51 with 1 sY, 48,867.20 Da = E51 with 2 sYs, 48,901.60 Da = E51 with 2 sYs that did not undergo cyclization of N-terminal glutamines in both heavy and light chain.

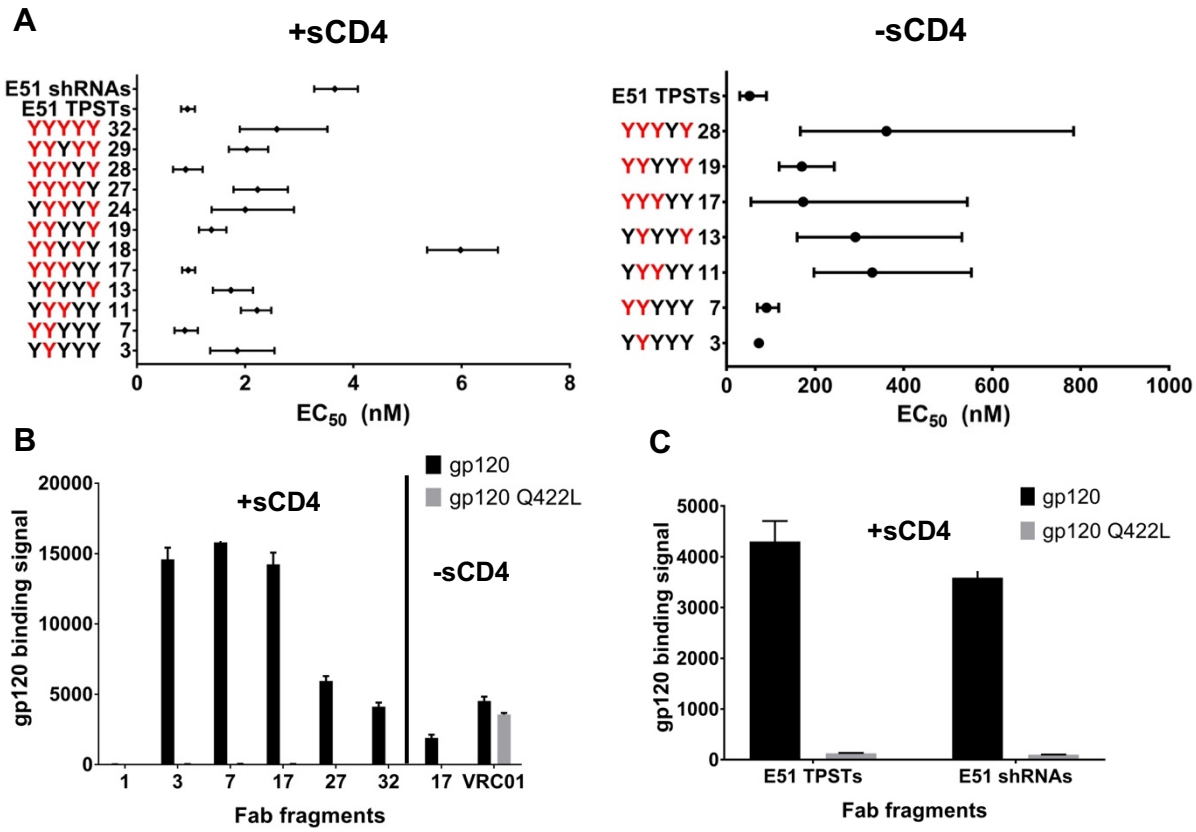

Figure 5.3. gp120-binding profiles of E51 sulfoforms and naturally sulfated E51.

(A) gp120-binding EC_{50}s of E51 sulfoforms, E51-TPSTs, and E51-shRNAs assayed against gp120 with or without sCD4 in an ELISA. Red 'Y' denotes sY. Error bars represent 95% confidence intervals. EC_{50} values are listed in Supplemental Table 3. (B) ELISA on 1 μg of E51 sulfoforms and VRC01, an anti-HIV antibody that targets the CD4-binding epitope of gp120, against gp120 or gp120(Q422L) with or without sCD4 (n = 3). (C) ELISA on 1 μg of E51-TPSTs and E51-shRNAs with sCD4 against gp120 or gp120(Q422L) (n = 3).

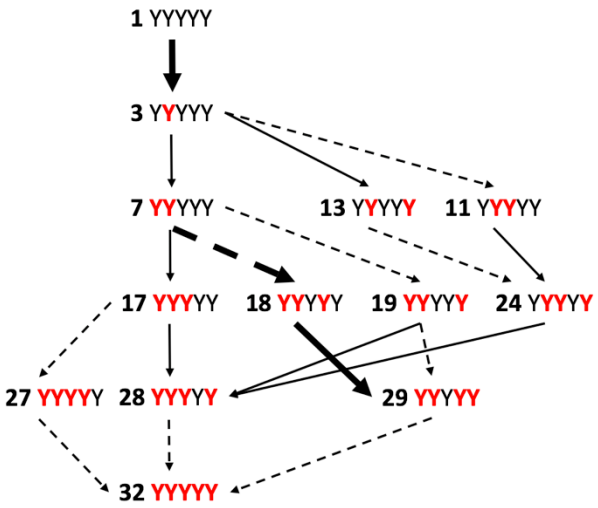

Figure 5.4. Functional diagram of E51 sulfoforms.

Based on EC_{50}s from ELISA conducted in the presence of sCD4, solid arrows indicate no change to or decrease in EC_{50} while dashed arrows indicate increase in EC_{50}. Bold arrows represent marked changes in EC_{50} values (> 3 nM). Red "Y" denotes sY.

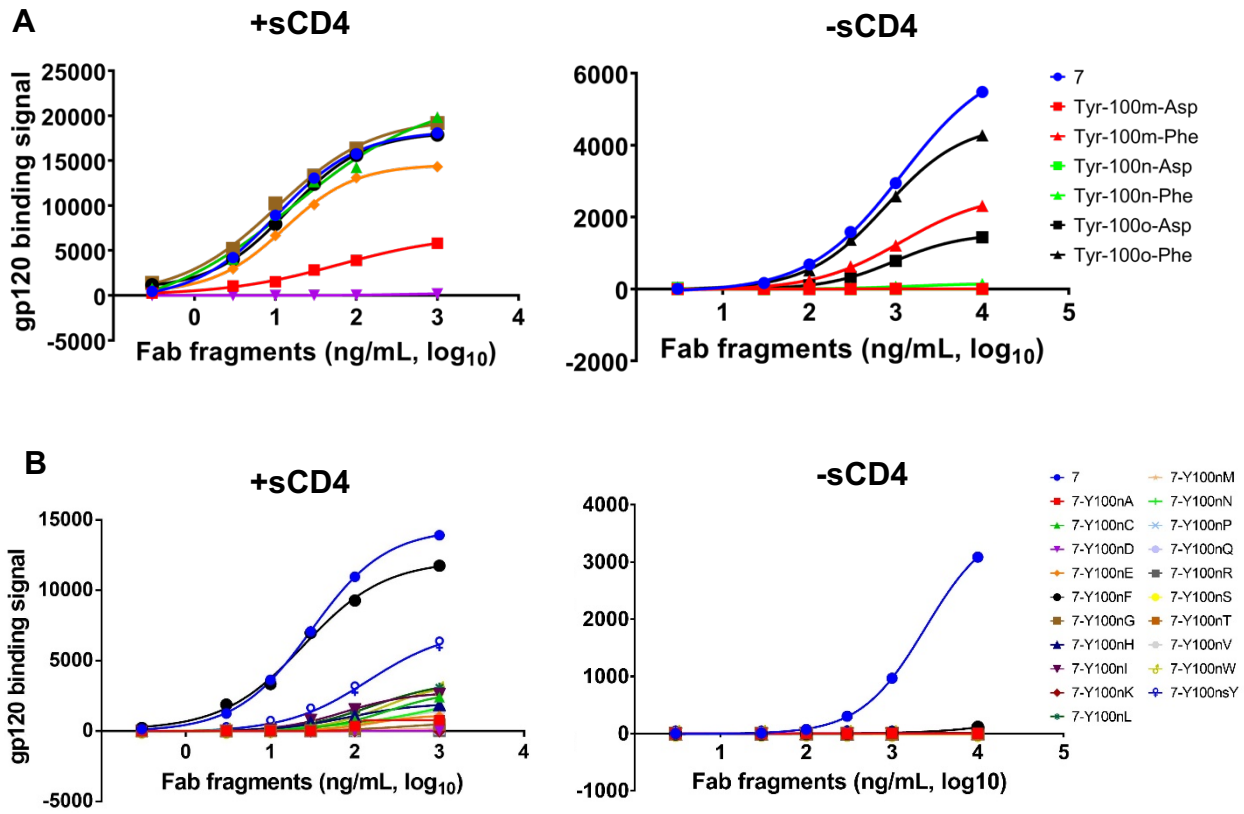

Figure 5.5. Contribution of individual tyrosines to binding to gp120.

(A) gp120-binding of E51 sulfoform **7** mutants with Phe or Asp mutations at 100m, 100n, or 100o in an ELISA with or without sCD4. (B) gp120-binding of **7** mutants with saturated site-directed mutagenesis at 100n in an ELISA with or without sCD4.

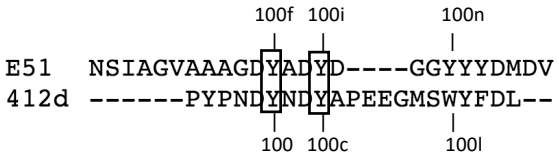

Figure 5.6. Sequence alignment of the VHCDR3 of E51 and 412d anti-HIV antibodies show homology between tyrosine-sulfates (boxed) and tyrosines important for binding gp120.

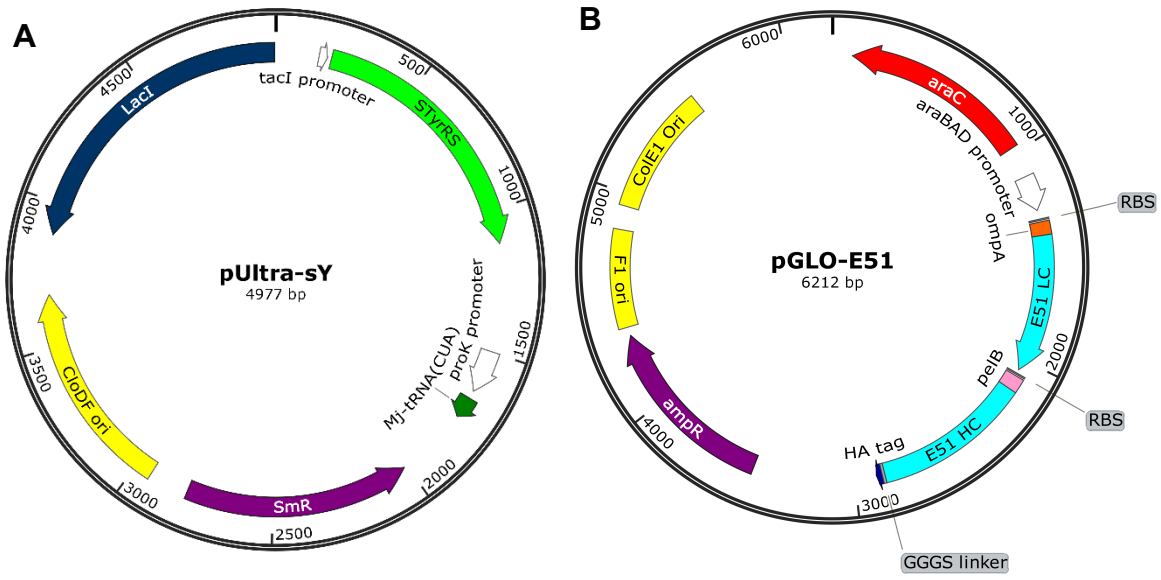

Figure 5.7. Plasmid maps.

(A) pUltra-sY constitutively expresses *Methanococcus jannaschii* tRNA engineered to decode UAG codons (Mj-tRNA$_{CUA}$) and expresses *M. jannaschii* aminoacyl tRNA synthetase engineered to charge sulfotyrosine (STyrRS) under a tacI promoter inducible by IPTG. (B) pGLO-E51 contains a E51 Fab expression cassette under an araBAD promoter inducible by L-arabinose. The Fab cassette encodes for E51 light chain (LC) with a N-terminal OmpA signaling peptide and E51 heavy chain (HC) with a N-terminal pelB signaling peptide and a C-terminal HA tag. Up to five tyrosine codons in the V$_H$CDR3 were replaced with UAG stop codons in each E51 sulfoform.

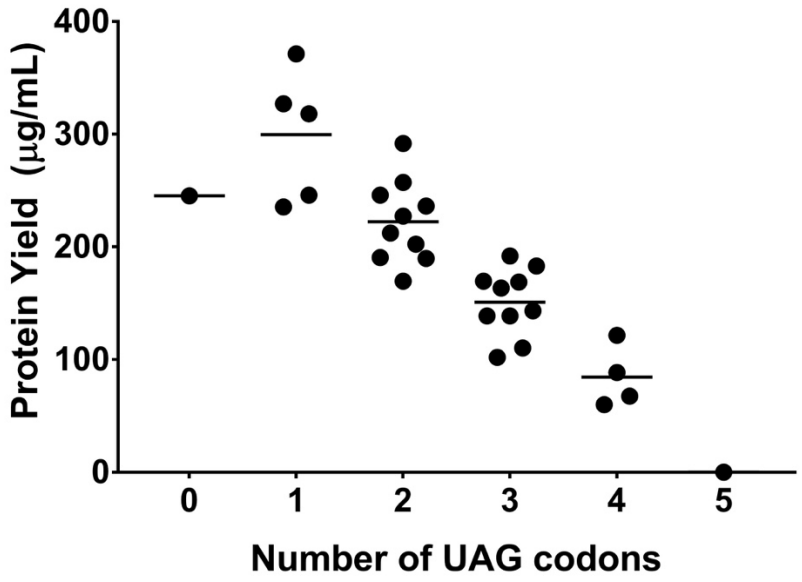

Figure 5.8. Yields of E51 Fab fragments incorporated with up to five sYs expressed in C321.ΔA.exp cells with an EGC for sY.

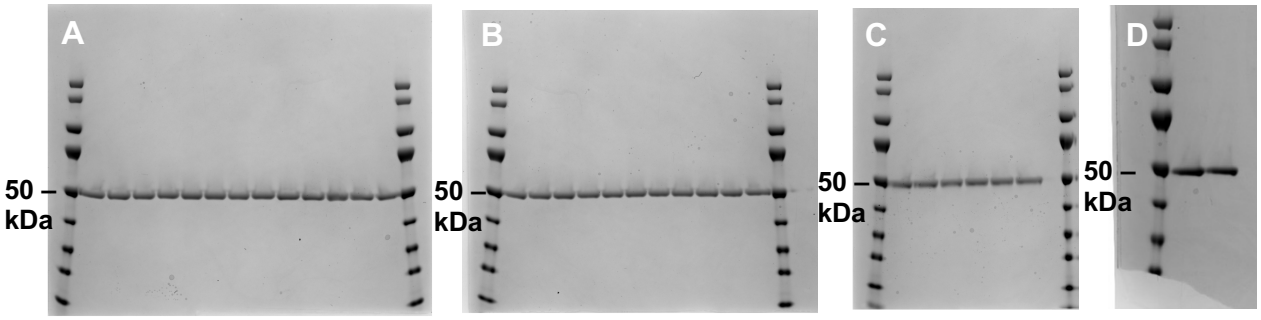

Figure 5.9. Visualization of 32 E51 sulfoform Fabs on SDS PAGE gels.

(A) Sulfoforms **1**-**13**. (B) Sulfoforms **14**-**24**. (C) Sulfoforms **25**-**27** and **29**-**31**. (D) Sulfoforms **28** and **32**.

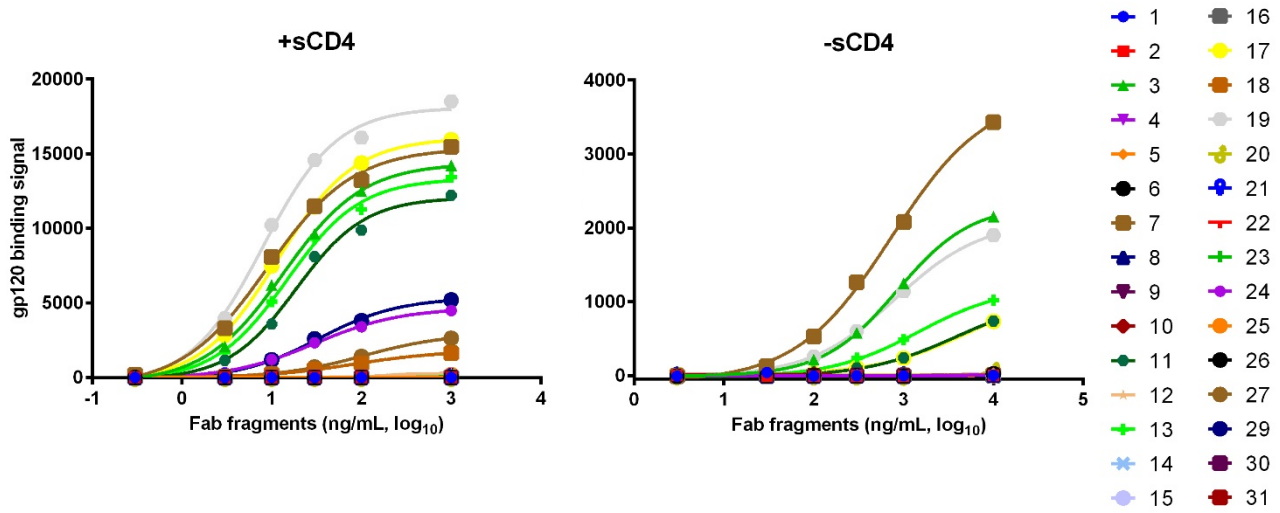

Figure 5.10. ELISA on 30 out of 32 E51 sulfoforms against gp120 with or without sCD4.

E51 sulfoforms that did not exhibit binding signals were not tested in Supplemental Figure 5.

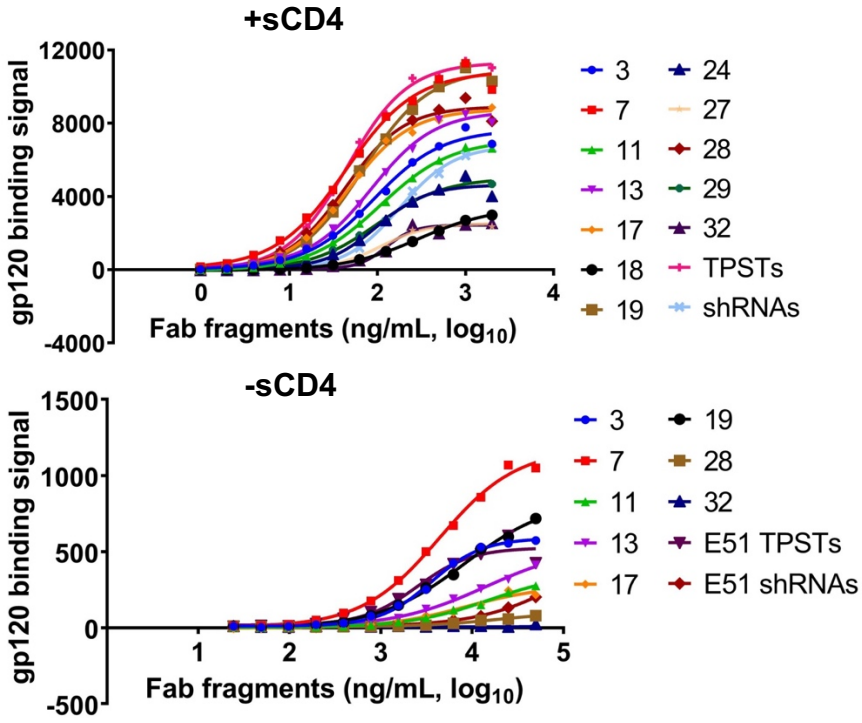

Figure 5.11. ELISA on gp120-binding E51 sulfoforms, E51-TPSTs, and E51-shRNAs against gp120 with or without sCD4

0 sY	1 sY	2 sYs	3 sYs	4 sYs	5 sYs
1 YYYYY	**2** YYYYY	**7** YYYYY	**17** YYYYY	**27** YYYYY	**32** YYYYY
		8 YYYYY	**18** YYYYY		
	3 YYYYY	**9** YYYYY	**19** YYYYY	**28** YYYYY	
		10 YYYYY	**20** YYYYY		
	4 YYYYY	**11** YYYYY	**21** YYYYY	**29** YYYYY	
		12 YYYYY	**22** YYYYY		
	5 YYYYY	**13** YYYYY	**23** YYYYY	**30** YYYYY	
		14 YYYYY	**24** YYYYY		
	6 YYYYY	**15** YYYYY	**25** YYYYY	**31** YYYYY	
		16 YYYYY	**26** YYYYY		

Table 5.1. Sulfation patterns of the 32 E51 sulfoforms.

Simplified sequences of E51 V_HCDR3 contain Tyr (black) or sY (red) at positions 100f, 100i, 100m, 100n, and 100o, corresponding to the possible sites of tyrosine sulfation in E51. Bold numbers refer to the names of E51 sulfoforms.

Number of sYs	Name	Predicted mass (Da)	Observed mass (Da)
0	**1**	48742	48742
1	**2**	48822	48822
	3	48822	48822
	4	48822	48822
	5	48822	48822
	6	48822	48823
2	**7**	48902	48902
	8	48902	48903
	9	48902	48905
	10	48902	48903
	11	48902	48903
	12	48902	48902
	13	48902	48901
	14	48902	48903
	15	48902	48902
	16	48902	48902
3	**17**	48982	48983
	18	48982	48981
	19	48982	48982
	20	48982	48980
	21	48982	48982
	22	48982	48982
	23	48982	48984
	24	48982	48983
	25	48982	48982
	26	48982	48983
4	**27**	49062	49063
	28	49062	49064
	29	49062	49064
	30	49062	49064
	31	49062	49063
5	**32**	49142	49143
1	E51-TPSTs	48788	48788
2		48868	48868
3		48948	48948
4		49028	49028
5		49108	49108
0	E51-shRNAs	48708	48708
1		48788	48788
2		48868	48867

Table 5.2. ESI-MS of 32 E51 sulfoforms and naturally sulfated E51.

The major peak in the mass spectrum each of the 32 E51 sulfoforms correlates to the predicted mass of the sulfoform. Multiple major and minor peaks of E51 TPSTs and E51 shRNAs correlate to predicted masses of E51 sulfoforms with up to five sYs. The mass shifts of -34 Da in E51 TPSTs and E51 shRNAs correspond to the conversion of N-terminal glutamine to pyroglutamate in both the heavy and light chain of E51 expressed in HEK293T cells.

	Sequence	Name	EC_{50} (nM)	C.I. (nM)
+sCD4	YYYYY	**3**	1.854	1.352-2.543
	YYYYY	**7**	0.883	0.694-1.124
	YYYYY	**11**	2.218	1.956-2.516
	YYYYY	**13**	1.735	1.403-2.144
	YYYYY	**17**	0.947	0.834-1.075
	YYYYY	**18**	5.980	5.362-6.669
	YYYYY	**19**	1.375	1.145-1.651
	YYYYY	**24**	2.000	1.379-2.902
	YYYYY	**27**	2.231	1.783-2.789
	YYYYY	**28**	0.899	0.666-1.212
	YYYYY	**29**	2.209	1.699-2.426
	YYYYY	**32**	2.584	1.898-3.517
		E51-TPSTs	0.934	0.815-1.071
		E51-shRNAs	3.657	3.275-4.082
-sCD4	YYYYY	**3**	72.809	64.560-82.105
	YYYYY	**7**	90.191	69.136-117.628
	YYYYY	**11**	320.111	188.303-544.141
	YYYYY	**13**	290.971	159.204-531.828
	YYYYY	**17**	172.912	54.956-543.977
	YYYYY	**19**	169.854	118.654-243.115
	YYYYY	**28**	361.235	166.366-784.322
		E51-TPSTs	51.693	29.756-89.842

Table 5.3. EC_{50} of E51 sulfoforms, E51-TPSTs, and E51-shRNAs based on ELISA binding curves from Figure 11.

C.I. = confidence interval. Red "Y" denote sY.

Name	Sequence
E51sY1-F	cggagactagGCTGACTACGATGGGGGCTA
E51sY1-R	gtagtcagccTAGTCTCCGGCAGCTGCTAC
E51sY2-F	cgctgactagGATGGGGGCTACTACTACGA
E51sY2-R	gcccccatccTAGTCAGCGTAGTCTCCGGC
E51sY3-F	tgggggctagTACTACGATATGGATGTCTG
E51sY3-R	atcgtagtacTAGCCCCCATCGTAGTCAGC
E51sY4-F	gggctactagTACGATATGGATGTCTGGGG
E51sY4-R	catatcgtacTAGTAGCCCCCATCGTAGTC
E51sY5-F	ctactactagGATATGGATGTCTGGGGCCA
E51sY5-R	atccatatccTAGTAGTAGCCCCCATCGTA
E51sY2sY3-F	ctaggatgggggctagTACTACGATATGGATGTCTG
E51sY2sY3-R	gtactagcccccatccTAGTCAGCGTAGTCTCCGGC
E51sY1sY2sY3-F	cggagactagGCTGACTAGGATGGGGGCTA
E51sY1sY2sY3-R	ctagtcagccTAGTCTCCGGCAGCTGCTAC
E51sY1sY2sY3sY4sY5-F	tgactaggatgggggctagtagtaggATATGGATGTCTGGGGCCAA
E51sY1sY2sY3sY4sY5-R	actagcccccatcctagtcagcctagTCTCCGGCAGCTGCTACTCC
E51sY2sY4-R	ccatatcgtacTAGTAGCCCCCATCCTAGTC
E51sY2sY5-R	catccatatccTAGTAGTAGCCCCCATCCTA
E51sY3sY4-R	ccatatcgtacTACTAGCCCCCATCGTAGTC
E51sY3sY5-R	catccatatccTAGTACTAGCCCCCATCGTA
E51sY4sY5-F	ctactagtagGATATGGATGTCTGGGGCCA
E51sY4sY5-R	catccatatccTACTAGTAGCCCCCATCGTA
E51sY2sY4sY5-R	catccatatccTACTAGTAGCCCCCATCCTA
E51sY3sY4sY5-F	ctagtagtagGATATGGATGTCTGGGGCCA
E51sY3sY4sY5-R	catccatatccTACTACTAGCCCCCATCGTA
E51sY2sY3sY4sY5-F	cgctgactagGATGGGGGCTAGTAGTAGGA
XL-TPST1shRNAvecF	TTAATGCGTCTCggcccTTTTTTAAGCTTGGGCCGCTCGAGG
XL-TPST1shRNAvecR	ATTAATCGTCTCggcccGGTGTTTCGTCCTTTCCACAAG
XL-TPST1shRNAupper	GGGCCATGCTGGACGCACATCTTCGATAGAGATGTGCGTCCAGCATG
XL-TPST1shRNAlower	GGGCCATGCTGGACGCACATCTCTATCGAAGATGTGCGTCCAGCATG
XL-TPST2shRNAvecF	TTAATGCGTCTCgtcccTTTTTTAAGCTTGGGCCGCTCGAGG
XL-TPST2shRNAvecR	ATTAATCGTCTCgtcccGGTGTTTCGTCCTTTCCACAAG

Table 5.4. Primers used to clone the 32 E51 sulfoforms and shRNAs.

Chapter 6.Directed evolution of tyrosine sulfated anti-HIV antibodies

Abstract

Tyrosine sulfation, a post-translational modification, plays a critical role for a class of sulfated CD4-induced (CD4i) antibodies to target HIV's coat protein gp120. Sulfates are only installed onto target tyrosines surrounded by specific sequence contexts known as sulfation motifs. Although important for tyrosine sulfation, these sulfation motifs may not be optimal for sulfated CD4i antibodies to bind to gp120. We hypothesize that repurposing multiple sulfation motifs with residues that bind to gp120 while maintaining tyrosine sulfation will improve the HIV neutralizing potency of one of the best sulfated CD4i antibodies, E51. To test our hypothesis, we use phage display with an expanded genetic code system to express sulfated antibody phage libraries with randomized residues replacing E51's sulfation motifs and screen these phages against gp120. We show that selections without the addition of soluble CD4 (sCD4) did not enrich for CD4i antibodies. Moreover, we show that phage selections with the addition of sCD4 enriched for CD4i antibodies that contain sulfation motifs and exhibit similar gp120-binding as E51. Therefore, we conclude that E51's sulfation motifs are optimal at binding gp120.

6.1. Introduction

A myriad of different antibodies have been generated by the human immune system to target HIV. These antibodies neutralize HIV through the interference with quaternary structural changes of HIV coat proteins or blockage of HIV coat proteins from interacting with surface proteins of immune cells[1]. Some of the most effective antibodies prevent HIV infections through competitive association with glycan-dependent[2], CD4-binding[3], or chemokine receptor CCR5- or CXCR4-binding[4] epitopes of the HIV coat protein, gp120.

Recently, Gardner *et al.* developed a highly potent anti-HIV immunoadhesion[5] composed of the binding domains of CD4 and a peptide derived from a sulfated CD4-induced (CD4i) antibody, E51. This engineered immunoadhesion, eCD4-Ig, can penetrate through the glycan shell of gp120 and simultaneously block both the CD4- and CCR5-binding epitopes of gp120. This dual blockage is so potent that eCD4-Ig can, at IC_{50}s lower than any known broadly neutralizing antibodies, prevent cellular entry of neutralization-resistant HIV isolates, including ones that hijack chemokine receptor CXCR4 or both CCR5 and CXCR4 for infection. Even more impressively, the delivery of rhesus eCD4-Ig *via* adeno-associated virus vectors into macaques protected their immune system from multiple injections of SHIV over more than 40 weeks. The potencies of eCD4-Ig and its derivatives make them an attractive alternative to HIV vaccine.

We aim to improve the neutralizing potency of eCD4-Ig and other E51-based therapeutics through enhancing the gp120-binding affinity of E51 antibody. eCD4-Ig can be improved through engineering E51. Chiang et al. showed that the gp120-binding of E51-based peptide can be improved through an Ala-100g-Tyr mutation in the V_HCDR3 of E51[6]. When implemented into eCD4-Ig, this Ala-100g-Tyr mutation lowered the IC_{50}s of eCD4-Ig for neutralizing several challenging isolates of HIV. On the contrary, efforts to engineer the CD4-binding domains of

eCD4-Ig remain unfruitful. For example, VRC01, a CD4-mimetic antibody that binds to the CD4-binding site of gp120, was less efficient than CD4-Ig at promoting the binding of E51 to HIV's Env trimer[7], because only CD4 can enhance E51's binding to gp120; this suggests that a hypothetical E51-VRC01-Ig immunoadhesion would exhibit less dual-receptor synergistic binding to gp120 than that of eCD4-Ig. Improvements to the gp120-binding affinity of E51 can potentially improve the neutralizing ability of eCD4-Ig *in vitro* or even *in vivo*.

Tyrosine sulfation is perhaps the most important element for E51 to effectively neutralize HIV infection[4,8]. E51 is sulfated in the *trans* Golgi apparatus of mammalian cells by tyrosylprotein sulfotransferase (TPST) I and II, which enzymatically transfer a sulfate group from a 3'-phosphoadenosine-5'-phosphosulfate onto an acceptor tyrosine. To modify acceptor tyrosines at the correct locations, TPSTs recognize specific sequence contexts within seven amino acids of the candidate tyrosine; these sequence contexts, termed sulfation motifs, include acidic residues (e.g. aspartic acid and glutamic acid), small residues (e.g. glycine and alanine), and previously sulfated tyrosines that promote tyrosine sulfation of additional nearby tyrosines[9]. The long V_HCDR3 loop of E51 contains multiple aspartic acids, glycines, and alanines that recruit TPSTs in order to sulfate up to five tyrosines in the third complementary determining regions of the heavy chain (V_HCDR3) of E51 for binding gp120.

Although necessary for tyrosine sulfation, sulfation motifs may not be optimal at binding gp120. Minimal interactions between sulfation motifs and gp120 is exemplified in the crystal structure of an extensively characterized sulfated anti-HIV antibody, 412d, in complex with gp120 and CD4[10]. The crystal structure shows that sulfation motifs of 412d only form peripheral interactions with gp120 residues outside of the CCR5-binding epitope while sulfated Tyr100c of 412d is deeply buried within the binding pocket. Replacement of only one of the E51's sulfation

motifs, Ala100g, with a tyrosine actually improved the binding between E51 and gp120[6], presumably because the single mutation minimally altered the overall tyrosine sulfation of the E51 variant.

We hypothesize that repurposing multiple sulfation motifs to bind gp120, without affecting tyrosine sulfation of E51, will drastically improve E51 at neutralizing HIV infections. To sulfate E51 without sulfation motifs, we plan to directly incorporate sulfotyrosine (sY) into E51 in a cell with an expanded genetic code for sY[11]. The expanded genetic code system for sY utilizes an orthogonal aminoacyl-tRNA synthetase for sulfotyrosine (STyrRS)/$tRNA_{CUA}$ pair to site-specifically incorporate sY at amber (UAG) stop codons. STyrRS and $tRNA_{CUA}$ are engineered such that they do not react with any endogenous *E. coli* tRNA or aminoacryl-tRNA synthetases. By replacing tyrosine codons with amber stop codons in E51's V_HCDR3, this system can express tyrosine sulfated E51 regardless of the sequence context surrounding amber stop codons, enabling mutagenesis of multiple sulfation motifs. Moreover, an expanded genetic code for sY combined with phage display, a standard method for protein evolution, has been shown to discover high gp120-binding sulfated antibodies that do not contain sulfation motifs[12].

Here, we use the expanded genetic code phage display system to express and screen sulfated E51 antibody libraries against gp120. In phage display, bacteriophage serves as a physical vessel to link the genotype and phenotype of proteins of interest, enabling the enrichment of protein sequences that exhibit strong binding affinities for a target protein through several iterative rounds of panning, washing, and amplification. Using this system, M13 bacteriophage displaying sulfated E51 single chain fragment variables (scFvs) were expressed in a sY-dependent manner. We show that our phage display selections enrich for sulfated antibodies that specifically bind to the CCR5-binding epitope of gp120. Our initial gp120 selection enriched for a non-CD4i antibody with a

higher gp120-binding than E51 in the absence of sCD4 but a lower gp120-binding than E51 in the absence of soluble CD4 (sCD4). After several gp120 selections on several different libraries with the addition of sCD4, we only find sulfated CD4i antibodies, some of which resemble sulfation motifs, that bind to gp120 with similar or lower affinities than E51. For one of the selections, we show that NGS analysis can track the enrichment of individual sequences from each round of selection and is compatible with a sulfation prediction algorithm to assess if those sequences can be sulfated by TPSTs. Taking together our selection and gp120-binding results, we conclude that the sulfation motifs of E51 are in fact optimal at binding gp120.

6.2. Materials and methods

6.2.1. Cloning of E51 scFvs displayed on phage, phage libraries, and E51 Fabs.

Codon-optimized E51 scFv with no UAG codons (**1**) was synthesized (IDTDNA) and cloned into pSEX81 phagemid (Progen) *via* Gibson Assembly. E51 scFv variants with one, two, or five UAG codons were cloned *via* inverse PCR. E51 scFv-phage variants were transformed into SS320 cells containing pULTRA-sY.

PCR transcripts of sulfated antibody libraries for phage display were generated *via* inverse PCR using 5'-phosphorylated primers and phagemid expressing **17** as the template for L1 and L2, phagemid expressing **3** as the template for L3, and phagemid expressing **2** as the template for L4 (Table 1). Linear library DNA transcripts were then blunt-ligated using T4 DNA ligase at 4°C for 12 hours, and ethanol precipitated with carrier yeast tRNA and 3M sodium acetate at -20°C for 2 hours. For L1 and L2, the precipitate DNA was transformed into electrocompetent Top10 cells. The transformed cells were added to 2YT 5 times the volume of the rescued culture, supplemented with 1% Glucose and 1X Ampicillin, and grown to saturation at 37°C for 8-10 hours. Library DNA were purified from 50 mL of the transformed cells and subsequently transformed into

electrocompetent SS320 cells containing pULTRA-sY. For L3 and L4, the precipitated DNA was directly transformed into electrocompetent SS320 cells containing pULTRA-sY.

The V_HCDR3 of sulfated antibody and E51 variants from pSEX were amplified and then cloned into pGLO expression vector via Gibson Assembly. The resulting DNA were transformed into C321.ΔA.exp cells containing pULTRA-sY.

6.2.2. Expression and purification of phage.

A starter culture was grown by picking a colony of SS320 cells into 2YT with 1% Glucose, incubating the culture at 37°C, 200 rpm for 12-16 hours, and diluting in fresh 2YT with 1% Glucose at a ratio of 1:100. Once the culture has reached mid-log phase (OD600 = 0.6-0.8), cells from 20 mL of the culture was centrifuged, washed once with 40 mL of 2YT to remove residual glucose, and resuspended in 2YT with hyperphages (Progen) at multiplicity of infection (MOI) of 20. After 1 hour of infection at 37°C, 100 rpm, the cells were resuspended in 50 mL of 2YT with 20 mM of sY, and this phage expression culture was incubated at 30°C, 175 rpm for 24 hours. The volumes of cultures and media for phage expression can be scaled by the same factor. Phages were precipitated from the culture supernatant by adding $1/5^{th}$ volume of cold phage precipitation solution (20% PEG-8000, 2.5 M NaCl) to the supernatant, incubating at 4°C for 1 hour, and centrifuging at 15,000 g, 4°C for 30 minutes. Phage pellets were resuspended in PBS.

For the helper phage system, the 1:100 dilution culture was grown to early mid-log phase (OD600 = 0.4). Cells from 20 mL of the culture was centrifuged, washed once with 40 mL of 2YT, and resuspended in 2YT with M13K07 Helper Phage (NEB) at MOI of 4.6. After 1 hour of infection at 37°C, 200 rpm, the cells were resuspended in 50 mL of 2YT with 20 mM of sY, and this phage expression culture was incubated at 30°C, 175 rpm for 18 hours. Phage precipitation method is the same as the hyperphage system.

To titer phage, scFvs were cleaved from the pIII of phages by adding 5 μL of phages to 195 μL of 4 μg/mL of Trypsin in PBS and incubating the sample at 37°C for 20 minutes. After diluting the digested phage samples by 400-fold in PBS, 5 μL of phages were added to 195 μL of mid-log phase (OD600 = 0.6-0.8) SS320 cells for phage infection at 37°C, 200 rpm for 1 hour. Phage concentration was calculated based on the number of infected SS320 cells.

6.2.3. ELISAs against gp120 or gp120(Q422L).

0.2 μg of gp120 or gp120(Q422L) (Clade-B ADA, Immune Technology) in 100 μL of PBS was coated each well of a high-binding microplate (Corning) at 4°C for 12 hours. After the wells were blocked with 2% blocking-grade milk in 200 μL of PBS and washed 5 times with 300 μL of PBST (PBS with 0.05% Tween 20), the wells were incubated with equal amounts of phages or soluble Fab fragments at 37°C for 2 hours, washed 5 times with 300 μL of PBST, incubated with 100 μL of anti-M13 Rabbit antibody (Sigma-Aldrich) at 37°C for 1 hour, washed 5 times with 200 μL of PBST, incubated with 100 μL of HRP-conjugated anti-Rabbit antibody (Sigma-Aldrich), at 37°C for 1 hour, and washed 5 times with 200 μL of PBST. To develop binding signal, 100 μL of QuantaBlu fluorogenic peroxidase substrate solution (Thermo Fisher Scientific) was incubated in the each of the wells for 30 minutes at room temperature. Fluorescence signals (em: 325 nm, ex: 420 nm) was measured on a plate reader (TECAN).

6.2.4. Screening sulfated antibody libraries *via* phage display.

For the first round, 12 μg of gp120 in 100 μL of PBS was coated in 12 wells of a high-binding microplate at 4°C for 12 hours. After the wells were blocked with 2% blocking-grade milk in 200 μL of PBS and washed 5 times with 300 μL of PBST, at least 10^{12} phages with or without 12 μg of sCD4 (Progenics Pharmaceuticals) were panned against gp120 at 37°C for 20 minutes (K_{on} condition) or 4 hours (K_{off} condition). Unbound phages were washed with 200 μL of PBST

at different conditions (Table 2). Bound phages were eluted with 150 μL of 4 μg/mL of Trypsin in PBS at 37°C for 20 minutes and used to infect mid-log phase (OD600 = 0.6-0.8) SS320 cells containing pULTRA-sY at 37°C, 100 rpm for 1 hour. A small portion of the cells infected with eluted phage was used to titer phage and the rest was added to 25x volume of 2YT with 1% glucose and incubated at 37°C, 200 rpm for 10-14 hours. DNA of the resulting saturated culture was extracted for sequencing analysis. In subsequent rounds of phage display, the volume of phages expressed, gp120 coated in microplate wells, and phages and sCD4 panned were reduced (Table 2).

6.2.5. Next-generation sequencing preparation and analysis of phage libraries.

DNA fragments of each round of phage display that contain V_HCDR3 of the scFv were amplified using recommended Illumina primers. PCR products were gel purified, normalized based on the intensities of DNA bands on an electrophoresis agarose gel, and pooled at a ratio of 5 to 1 for round 1 amplicons to round 2+ amplicons. The pooled DNA were submitted to Applied biomedical for 250 bp pair-end sequencing using Illumina MiSeq. NGS sequencing results were analyzed using a python script that excludes reads containing nucleotides with a Phred score below 30. To predict whether the scFv sequence can be sulfated by TPSTs, full length VH-CDR3 of the scFv hits was analyzed using GPS-TSP 1.0[13].

6.2.6. Expression and purification of Fab fragments.

A single colony of C321.ΔA.exp cells containing pULTRA-sY and pGLO-E51 were picked into Terrific Broth (TB). After growing the cells at 37°C and 200 rpm for 16 hours, the saturated cultures in TB were diluted into M9T containing antibiotics at a ratio of 1:100. The cells were grown in M9T to saturation at 37°C and 200 rpm for overnight. 2 mL of cells in M9T were added to 200 mL of M9T containing antibiotics and 3 or 10 of mM sY and grown to mid-log phase (OD_{600} = 0.6-0.8) at 37°C and 200 rpm. Fab and STyrRS expression were induced by adding 0.2%

of L-arabinose and 1 mM of IPTG and incubating the cultures at 22°C and 200 rpm for 24 hours. Cell pellets were resuspended in 16 mL of cold periplasmic lysis buffer (20% sucrose, 20 mM Tris, 2 mM EDTA, pH 8.0) with protease inhibitor cocktail (Sigma) and incubated at 22°C and 200 rpm for 4 hours. The supernatant of the lysis consisting of the first periplasmic fraction was stored at 4°C, and the cell pellet was again resuspended in 16 mL of cold periplasmic lysis buffer with protease inhibitor cocktail and incubated at 4°C for overnight. The second supernatant of the lysis was combined with the first supernatant. The combined supernatant was filtered using a 0.22 μm filter, concentrated, and then purified using Protein G columns (Pierce) by following the manufacturer's instruction. The final concentrations of Fab fragments were measured on a microvolume spectrophotometer (NanoDrop) using the appropriate molecular weight and ε value calculated on ExPASy ProtParam. Fab fragments were stored in 50% glycerol at -80°C for up to 3 months. Positive mode ESI-TOF was used to confirm the incorporation of sYs in E51 sulfoforms and sulfation of naturally sulfated E51 Fab fragments.

6.3. Results

6.3.1. sY-dependent expression of sulfated E51 scFv displayed on phage.

We designed our phage expression system to ensure phage production is dependent on the incorporation of sY. We fused E51 scFv to the N-terminus of pIII, the bacteriophage protein responsible for bacterial infection, in a phagemid vector, pSEX. We replaced tyrosine codons with amber stop codons in the corresponding E51 scFv variants. This way, sulfated E51 scFv-pIII will only express when sY is incorporated at amber stop codon. To express phage, we adopted the hyperphage system[14]. The genome of hyperphage has a mutated origin of replication for weak replication and lacks the entire pIII gene. Since pIII is essential for the assembly of phage, hyperphage by itself cannot replicate, and it can only produce phage when an external source of

pIII is provided. In our phage expression system, the only external source of pIII comes from the expression of E51 scFv-pIII fusion. Thus, our design ensures that phage is produced only if sY is incorporated at the amber stop codons in E51 scFv.

High amounts of phage displaying sulfated E51 scFv was expressed in SS320 cells with an expanded genetic code for sY only when the growth media was supplemented with sY. In the presence of sY, we obtained ~ 10^{10} or 10^{9} phage/mL of E51 scFv-phage variants with one (**2**-**6**), two (**11**), and three (**17**) or five (**32**) UAG codons, respectively (Figure 1A). The titers of E51 scFv-phage variants with one UAG codon expressed with sY were at least 10 times higher than those expressed without sY. For E51 scFv-phage variants with multiple UAG codons (**11**, **17**, and **32**), the titer of phage expressed without sY was below the range of detection ($<10^{7}$ phage). The absence of sY in the growth media presumably caused early termination of protein translation at UAG codons, suggesting that natural amino acids are scarcely misincorporated at UAG codons. This sY-dependent phage expression indicates that sYs are site-specifically incorporated into E51 scFvs displayed on phage.

6.3.2. Sulfated E51 scFv displayed on phage binds to gp120.

Binding assays on E51 scFv-phage variants showed that the association of sulfated E51 scFv with gp120 is induced by CD4 and is specific to the CCR5-binding epitope. The binding profiles of E51 scFv-phage variants are similar to those of E51 Fab fragments in Chapter 5: **3** exhibits the highest gp120-binding signal among the singly-sulfated variants and **17** strongly binds to gp120 (Figure 1B). However, E51 with 5 sYs, **32**, exhibits equal binding signal as **3** possibly due to avidity effects from the multivalent display of scFvs on phage. Phage alone or phage displaying unsulfated E51, **1**, did not bind to gp120. Since **3** and **17** were among the highest gp120-binding E51 variants as scFvs-phage and soluble Fab fragments (Chapter 5), their sequences were the template for constructing our sulfated antibody libraries.

6.3.3. First selection for anti-gp120 antibodies.

Our first and selection selection enriched for a non-CD4i sulfated antibody that does not resemble the conventional sulfation motifs. For our selection, we constructed two sulfated antibody libraries, L1 and L2, in which six different sulfation motifs were randomized (Table 3) to generate a theoretical diversity of 10^9 unique DNA sequences. The phage display selections involved two schemes to select for antibodies (Table 2): high Kon and low Koff (short panning, long washing) or antibodies with low Koff (long panning, long washing). After six rounds of selection against gp120, we sequenced and expressed phage from individual colonies. ELISA on several scFv-phage clones showed marked improvement in binding to gp120 (Figure 2A). When we converted these antibodies into soluble Fab fragments, however, only XLS92 retained its enhanced binding to the CCR5-binding epitope of gp120 (Figure 2B, C, Figure 3A), yet XLS92 bound to gp120 with much lower affinity than **17** in the presence of soluble CD4 (sCD4) (Figure 3). The binding affinity of XLS92 for gp120 is the same with or without sCD4, and unlike E51 and 412d, XLS92 does not enhance the ability of sCD4 to bind to gp120 (data not shown). Therefore, we conclude that XLS92 is not a CD4-induced antibody. However, XLS92 contains several interesting features. The VH-CDR3 of XLS92 is three amino acids shorter than our intended design of L2 presumably due to oligonucleotide synthesis or cloning errors that occurred starting at the first NNK. As a result, XLS contains only two sYs that loosely resembles the general features of 412d antibody but the lacks the surrounding E51 amino acids including sulfation motifs (Table 3).

6.3.4. Second selection for antibodies that target the CCR5-binding epitope of gp120.

We wanted to ensure our phage display selection was enriching for sulfated antibodies that specifically target the CCR5-binding epitope of gp120. To this end, we implemented an internal negative selection within each round of phage display by incubating a mutant gp120, gp120

Q422L, to phage libraries before the panning step of each selection. gp120 Q422L contains a key mutation in the CCR5-binding epitope of gp120 that prevents gp120 from binding to both CCR5 and E51 in the presence or absence of sCD4[15]. A few antibodies enriched from our first selection also exhibited activity for gp120(Q422L) (data not shown), possibly because they bind to a different epitope of gp120. The addition of gp120 (Q422L) to the phage library before selection should decrease off-target antibodies from undergoing enrichment in the selection.

Interestingly, XLS92 was also enriched from this selection even though most of the other enriched sequences showed low to moderate improvement in binding gp120. Following the same selection conditions as the first selection, we observed the enrichment of XLS92 in beginning in the fourth round of both the Kon and Koff selections. The enrichment of other sequences, however, occurred only in Kon or Koff selections. Aside from XLS92, no other antibodies from the first selection was enriched in the second selection, possibly due to the addition of gp120(Q422L) that eliminated variants with high off-target binding. ELISA on the enriched scFv-phage from the second selection showed only moderate (up to 3-fold) gp120-binding improvements (Figure 2D), all of which except that of XLS92 did not translate in the form of soluble Fab fragments against gp120 (Figure 2E,F). Obtaining a non CD4-induced antibody like XLS92 from our first and second selection is expected, for we did not introduce sCD4 in these phage selections; this is because previous studies showed that CD4 induction minimally impacts the gp120-binding of E51[4,15] in contrast to our findings in Chapter 5.

6.3.5. Third selection for CD4i antibodies.

Although the addition of sCD4 to the selection enriched for CD4-induced antibodies, these antibodies contain sequences that resemble sulfation motifs and bind to gp120 with similar binding affinities as sulfated E51. We conducted five rounds of phage display on L1 and L2 in parallel under similar selection conditions as the first selection with the addition of sCD4 to phage libraries

prior to panning. Since sCD4 binds to gp120(Q422L), we did not concurrently add gp120(Q422L) with sCD4 to phage libraries to ensure maximum selection for CD4i antibodies. Sangar sequencing of 24 colonies from round 5 of each selection showed high enrichment of sulfation motifs (Figure 4). Several highly-enriched variants that are computationally predicted to be sulfated by TPSTs (Table 3), because most of variants contain Asp100h and Asp100j, which are common sulfation motifs. This result indicates that our gp120 selection with sCD4 enriched for CD4i antibodies with sulfation motifs, although we did not detect E51 enrichment. In addition, this selection did not enrich for XLS92, presumably due to subpar binding compared to CD4i antibodies in the presence of sCD4. Binding assays of enriched antibody variants in Fab format against gp120 with sCD4 showed similar binding profiles as **17** (Figure 5), indicating that the evolved residues in these variants exhibit similar gp120-binding as sulfation motifs of E51.

6.3.6. Fourth selection on a smaller library with negative control experiments and NGS analysis.

Instead of conducting phage display on large libraries, we attempted selection on a smaller library to ensure complete coverage of each library member and enable sequence analysis using next-generation sequencing (NGS). One major bottleneck of our previous selections was that we lost a lot of our library members due to limited transformation competency of SS320 cells with pULTRA-sY. The theoretical diversity of L1 and L2 were ~10^9 based on six NNKs, yet the experimental diversity of our libraries obtained from our transformation titers, excluding high amounts of random truncations, insertions, mutations, and sequence biases in the library, was at best 3 x 10^9. A large library of six randomized residues allowed us to explore a wide protein sequence space, but not every combination of amino acid sequences was screened in a library with low coverage. Contrarily, screening a library with fewer randomized residues will navigate a much smaller protein landscape that may not contain high binding variants. However, one major

advantage of a smaller library is full site-saturated mutagenesis at each designated position through complete sequence coverage. Furthermore, a smaller sequence diversity allows one to analyze selection DNA sequences using NGS. Traditional directed evolution experiments involves Sangar sequencing of ~100 individual clones, which is costly and may inaccurately portray the composition of a library population or the frequency of enriched variants. An affordable NGS run on one sample provides 10^6-10^5 sequences, which should cover the diversity of a smaller library to provide accurate sequence analysis.

We constructed a small library L3 (theoretical diversity ~ 3 x 10^4) based on our interpretation of the gp120-binding data on 32 E51 sulfoforms, which showed that sY100i and Tyr100n are important for binding to gp120 in the presence of sCD4 while the sulfation of the other three tyrosines (Tyr100f, Tyr100m and Tyr100o) moderately altered gp120-binding. L3 contains an UAG stop codon at position 100i for the incorporation of sY and NNKs replacing codons for Tyr100f, Tyr100m and Tyr100o. As opposed to our previous cloning strategy, in which we first transformed our libraries into a standard cloning strain prior to transformation into SS320 with pULTRA-sY, we directly transformed L3 into SS320 with pULTRA-sY and after a short growth period to select for transformed cells, proceeded to produce phage for the first round selection.

Prior to first round of selection, NGS on L3 showed complete sequence diversity coverage and minimal sequence biases from library construction. We sequenced DNA extracted from cells used to produce phage for the first round selection to account for additional sequence biases caused by cell growth in liquid culture. The NGS data shows a greater number of unique DNA sequences than L3's theoretical diversity due to additional sequence variants generated from point mutations (Table 5), confirming the complete coverage of every antibody variant in L3. L3 also contains

minimal sequence biases, as NGS analyses show even distributions of nucleotides at NNKs (Figure 6) and no overrepresentation of any individual DNA or protein sequence (Table 5).

We designed additional negative control experiments for this phage selection to select for CD4i antibodies by identifying undesired sequence enrichment as a result of fast cell growth and phage production or non-specific binding. The pipeline of the selection experiments is shown in Figure 7. We split the phage produced at the beginning of each round for three different experiments in parallel: selection for antibodies against gp120 with sCD4, selection for non-specific binding antibodies against gp120(Q422L) with sCD4, and phage passaging through direct phage infection into new SS320 cells with pULTRA-sY. We extracted DNA from cells at the beginning of every round of selection for NGS to identify highly frequent sequences. While the aim of the selection against gp120 with sCD4 is to enrich superior antibodies against the CCR5-binding epitope, the selection actually enriches for antibodies that bind to every epitope of gp120 as well as other components in the selection such as the blocking reagent and plastic surface of a micro-well. Therefore, we performed a parallel negative selection against gp120(Q422L) for each round and subsequently used NGS to help eliminate highly enriched antibodies with high non-specific affinities. In addition, phage display of peptide libraries has been shown to enrich for parasitic antibody sequences that enable fast cell growth and high phage production but have low binding affinities for the antigen of interest[16]. To identify potential parasitic sequences, we passaged phage from each round without selection by directly infecting new cells and subsequently analyzed sequence enrichment using NGS.

Although the selection against gp120 highly enriched for specific sequences that were not identified in the negative control experiments, they did not bind to gp120 with higher affinity than E51 sulfoform **3**. We performed three rounds of selection against gp120 under equal selection

pressure. From these three rounds, we observed minor increases in the percentages of eluted phage (Table 4) and no improvements in the gp120-binding, albeit a decrease the gp120(Q422L)-binding, of the total phage populations (Figure 9A). Even though these results indicate minimal gp120-improvement from the gp120 selection, antibodies containing Trp100f was heavily enriched only in the gp120 selection (Figure 8). Since the only difference between the gp120 selection and the gp120(Q422L) selection is the defective Q422L mutation in the CCR5-binding epitope of gp120, the high enrichment of these Trp100f variants in the gp120 selection (Figure 10) seems to indicate that they exhibit improved epitope-specific gp120-binding. However, several of these antibodies assayed against gp120 with sCD4 showed similar or lower gp120-binding and elevated non-specific binding than E51 sulfoform **3** (Figure 9B). Since Trp is common in motifs that bind to plastic[17], under long duration of washing during the gp120 selection perhaps the binding of Trp100f variants to the CCR5-binding epitope of gp120 assisted local non-specific binding to plastic. The enrichment of Trp100f variants could also be attributed to binding to multiple gp120 epitopes, as ELISA shows higher Trp100f variants' binding to gp120(Q422L) than E51 sulfoform **7**. As a result, E51 sulfoforms with minimal non-specific binding were enriched by less than 80-fold than the most frequent variant after three rounds of selection. Lastly, given the small sequence diversity of L3, it is highly possible that an antibody with higher gp120-binding did not exist in L3.

6.3.6. Fifth selection using the helper phage system.

Alternatively, we tested whether the helper phage system can improve our selection for superior anti-HIV antibodies. The helper phage system uses wild type M13 phage containing the pIII gene to initiate phage production, meaning that phage assembly in the helper phage system will randomly use wild type pIII or pIII-scFv. This results in the production of phage without displaying scFv, which can be problematic to the selection if the production of pIII-scFv is low.

However, the advantage of the helper phage system is that it can produce monovalent display of scFv-phage, thereby eliminating avidity effects associated with polyvalent scFv-phage produced in the hyperphage system. The improved sY-incorporation from pULTRA-sY should produce high amounts of sulfated scFv fused to pIII to enable phage display using helper phage. Indeed, the amount of functional E51 **7** scFv displayed on phage expressed using helper phage is comparable to that of E51 **7** scFv-phage expressed using hyperphage based on their gp120-binding (Figure 11A).

The helper phage system could not enrich for sulfated CD4i antibodies using our phage display with an expanded genetic code for sY. To test the feasibility of using helper phage for phage display with an expanded genetic code, we observed sequence enrichment in samples of a mock gp120 selection containing doubly-sulfated E51, **7**, and unsulfated E51, **1**, at ratios of 1:1, 1:100, 1:100, and 1:1000 (Figure 11B). Since **7** binds to gp120 with much higher affinity than **1**, successful selection is demonstrated by the enrichment of TAG codons at Tyr100f and Tyr100i (i.e. enrichment for **7**). Under stringent washing conditions (wash 20 times with PBSTs over 3 hours) after one round of gp120 selection, the 1:1000 sample only modestly enriched for **7** as demonstrated by ~10% of TAC to TAG conversion in the sequencing results. In contrast, the gp120 selection with standard washing (wash 5 times with PBST over 10 minutes) markedly enriched for **7** in the 1:1000 sample expressed using the hyper phage system. Furthermore, we randomized Tyr100i, Tyr100m, and Tyr100o in library L4 to compared the enrichment for the essential sY, sulfated Tyr100i, in a gp120 selection on L4 phage expressed using helper phage or hyperphage. We sequenced 12 colonies after one round of each of the selections. 11 out of 12 colonies sequenced from the helper phage selection correspond to an aberrant truncated sequence containing a frameshift that would mistranslate scFv-pIII while 7 out of 12 colonies sequenced

from hyperphage selection enriched for sequences containing sulfated Tyr100i. Taken together, these results show that helper phage used for phage display with an expanded genetic code for sY is not suited to enrich for sulfated CD4i antibodies against gp120.

6.4. Discussion

While E51 requires tyrosine sulfation to bind to gp120 and neutralize HIV infections, we hypothesized that the E51 sulfation motifs are suboptimal at binding gp120 and can be repurposed to residues that bind to gp120 with high affinities. However, our gp120 selections performed in the presence of sCD4 in the third selection enriched for CD4i antibodies that showed similar or lower gp120 binding. Some of these enriched antibodies also contain sulfation motifs that are different from E51, and a sulfation prediction algorithm predicts that these motifs can recruit TPSTs to sulfate adjacent tyrosines. Moreover, XLS92, a non-CD4i sulfated antibody without sulfation motif that showed higher gp120-binding than E51 only in the absence of sCD4, was enriched only in gp120 selections performed without sCD4, suggesting other residues can replace sulfaiton motifs for superior non-CD4i binding. However, the EC_{50} of the non-CD4i gp120-binding of XLS92 is more than 20-fold higher than the CD4i gp120-binding of E51. Taken together, these results indicate that sulfation motifs are optimal for binding to gp120 in the presence of CD4. This conclusion is not entirely surprising, as other non-sulfated proteins are involved with binding the CCR5-binding epitope of gp120. For example, CXCR4, unlike CCR5, does not require tyrosine sulfation for HIV to specifically bind to the CCR5-binding epitope of gp120 for HIV infection[18,19], yet the deletion of N-terminus of CXCR4, which contains sulfation motifs, reduces HIV infection. Additionally, not only do sulfated CD4i antibodies mimic CCR5 but non-sulfated CD4i antibodies also leverage sulfation motifs to block the CCR5-binding epitope of gp120 in order to neutralize HIV infection.

Given that we modified a few selections to specifically select for CD4i antibodies and to attempt to eliminate non-specific bindings, we propose ways to overcome additional deficiencies in our phage display selections. First, immobilization of gp120 antigens on plastic surfaces poses several issues for binding selection. Adsorption of target antigens on plastic surfaces through hydrogen bond and non-polar interactions has been shown to partially or fully denature the protein, possibly disabling the CCR5- or CD4-binding epitope of gp120 and exposing off-target epitopes. Antigen coating on surfaces is also non-uniform; this yields patches of densely immobilized antigens that trap local dissociated binders, creating undesired avidity effects and possibly promoting non-specific binding to plastic as evident by the enrichment of Trp variants from our fourth selection. One way to circumvent these issues is to conduct phage display using streptavidin-coated magnetic microbeads. In this selection scheme, beads covalently attached to biotinylated antigens are incubated with phages in solution, enabling dissociated binders to fully separate from previously bound antigens. Second, our selection for CD4i antibodies could had been limited by the CD4-gp120 rate of association ($6.27 \times 10^4\ S^{-1}\ M^{-1}$)[20], which is slower than the K_{on} of sulfated 412d binding to gp120 ($1.06 \times 10^5\ S^{-1}\ M^{-1}$)[12]. Instead of externally adding sCD4 to the panning step of the selection, physical linkage of CD4 and gp120 through the expression of gp120-CD4 fusion protein will increase the number of gp120 bound to CD4 and should improve our selection for CD4i antibodies. Third, perhaps superior gp120-binding residues were not present in our phage libraries with large sequence diversities (e.g. L1 and L2) due to subpar sequence coverage, which can be improved through reduction in codon redundancy, optimization of cloning processes, and increase in the uptake of DNA of SS320 with pULTRA-sY for higher yield of DNA transformation.

Cost-effective bulk sequencing using NGS offers invaluable sequence analysis not only for phage display but also for other experimental evolution experiments with an expanded genetic code system. By splitting a single Illumina MiSeq run, we obtained 10^5-10^6 sequence reads, of which ≥ 74% has Phred quality scores above 30, enough to completely cover the maximum sequence diversity of L3 with three NNKs (10^4 theoretical diversity). For a standard phage display library with six NNKs (10^9 theoretical diversity), the current standard NGS will only cover ~1.5% of the library sequences, but we anticipate that future NGS will provide higher number of reads at similar costs. These large amount of reads allows for analysis of library sequence bias and determination of highly enriched variants, including ones that contain additional sY. More generally, monitoring variants with additional sYs or other non-canonical amino acids (ncAAs) using NGS could identify proteins with novel functions for protein evolution as well as reveal strains with ncAA-dependency for experimental evolution that tests for expanded genetic codes adaptations.

6.5. References

1. Burton, D. R. *et al.* A Blueprint for HIV Vaccine Discovery. *Cell Host Microbe* **12,** 396–407 (2012).

2. Pejchal, R. *et al.* A Potent and Broad Neutralizing Antibody Recognizes and Penetrates the HIV Glycan Shield. *Science (80-.).* **334,** 1097–1103 (2011).

3. Zhou, T. *et al.* Structural Basis for Broad and Potent Neutralization of HIV-1 by Antibody VRC01. *Science (80-.).* **120,** 811–817 (2010).

4. Choe, H. *et al.* Tyrosine sulfation of human antibodies contributes to recognition of the CCR5 binding region of HIV-1 gp120. *Cell* **114,** 161–70 (2003).

5. Gardner, M. R. *et al.* AAV-expressed eCD4-Ig provides durable protection from multiple

SHIV challenges. *Nature* **519,** 87–91 (2015).

6. Chiang, J. J. *et al.* Enhanced recognition and neutralization of HIV-1 by antibody-derived CCR5-mimetic peptide variants. *J. Virol.* **86,** 12417–21 (2012).

7. Gardner, M. R. *et al.* CD4-induced antibodies promote association of the HIV-1 envelope glycoprotein with CD4-binding site antibodies. *J. Virol.* **90,** JVI.00803-16 (2016).

8. Dorfman, T., Moore, M. J., Guth, A. C., Choe, H. & Farzan, M. A tyrosine-sulfated peptide derived from the heavy-chain CDR3 region of an HIV-1-neutralizing antibody binds gp120 and inhibits HIV-1 infection. *J Biol Chem* **281,** 28529–28535 (2006).

9. Stone, M. J., Chuang, S., Hou, X., Shoham, M. & Zhu, J. Z. Tyrosine sulfation: an increasingly recognised post-translational modification of secreted proteins. *N. Biotechnol.* **25,** 299–317 (2009).

10. Huang, C.-C. *et al.* Structures of the CCR5 N terminus and of a tyrosine-sulfated antibody with HIV-1 gp120 and CD4. *Science* **317,** 1930–4 (2007).

11. Liu, C. C. & Schultz, P. G. Recombinant expression of selectively sulfated proteins in Escherichia coli. *Nat. Biotechnol.* **24,** 1436–1440 (2006).

12. Liu, C. C., Choe, H., Farzan, M., Smider, V. V & Schultz, P. G. Mutagenesis and evolution of sulfated antibodies using an expanded genetic code. *Biochemistry* **48,** 8891–8 (2009).

13. Pan, Z. *et al.* Systematic analysis of the in situ crosstalk of tyrosine modifications reveals no additional natural selection on multiply modified residues. *Sci. Rep.* **4,** (2014).

14. Rondot, S., Koch, J., Breitling, F. & Dübel, S. A helper phage to improve single-chain antibody presentation in phage display. *Nat. Biotechnol.* 75–78 (2001).

15. Xiang, S.-H. *et al.* Epitope mapping and characterization of a novel CD4-induced human

monoclonal antibody capable of neutralizing primary HIV-1 strains. *Virology* **315,** 124–134 (2003).

16. Matochko, W. L., Cory Li, S., Tang, S. K. Y. & Derda, R. Prospective identification of parasitic sequences in phage display screens. *Nucleic Acids Res.* **42,** 1784–98 (2014).

17. Adey, N. B., Mataragnon, A. H., Rider, J. E., Carter, J. M. & Kay, B. K. Characterization of phage that bind plastic from phage-displayed random peptide libraries. *Gene* **156,** 27–31 (1995).

18. Basmaciogullari, S., Babcock, G. J., Van Ryk, D., Wojtowicz, W. & Sodroski, J. Identification of conserved and variable structures in the human immunodeficiency virus gp120 glycoprotein of importance for CXCR4 binding. *J. Virol.* **76,** 10791–800 (2002).

19. Farzan, M. *et al.* The role of post-translational modifications of the CXCR4 amino terminus in stromal-derived factor 1 alpha association and HIV-1 entry. *J. Biol. Chem.* **277,** 29484–9 (2002).

20. Myszka, D. G. *et al.* Energetics of the HIV gp120-CD4 binding reaction. *Proc. Natl. Acad. Sci. U. S. A.* **97,** 9026–9031 (2000).

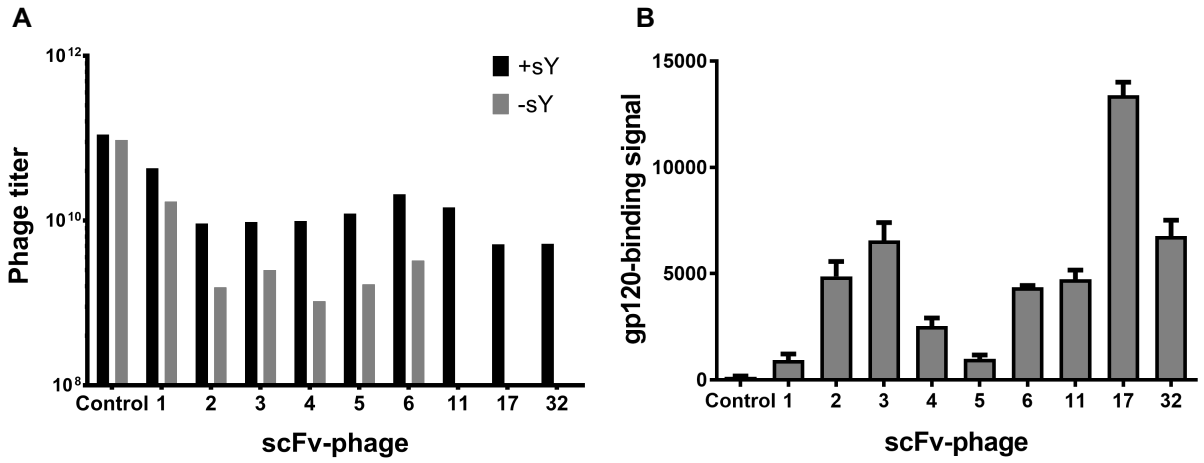

Figure 6.1. sY-incorporated E51 scFvs displayed on phage bind gp120.

(A) Phage titers of E51 scFv displayed on M13 bacteriophage. Phage was expressed from *E. coli* grown in media with (black bars) or without (gray bars) sY. (B) ELISA on 5 x 10^9 phage displaying E51 scFv sulfoforms against gp120. Control = phage without scFv displayed. E51 scFv number names are listed in Table 1.

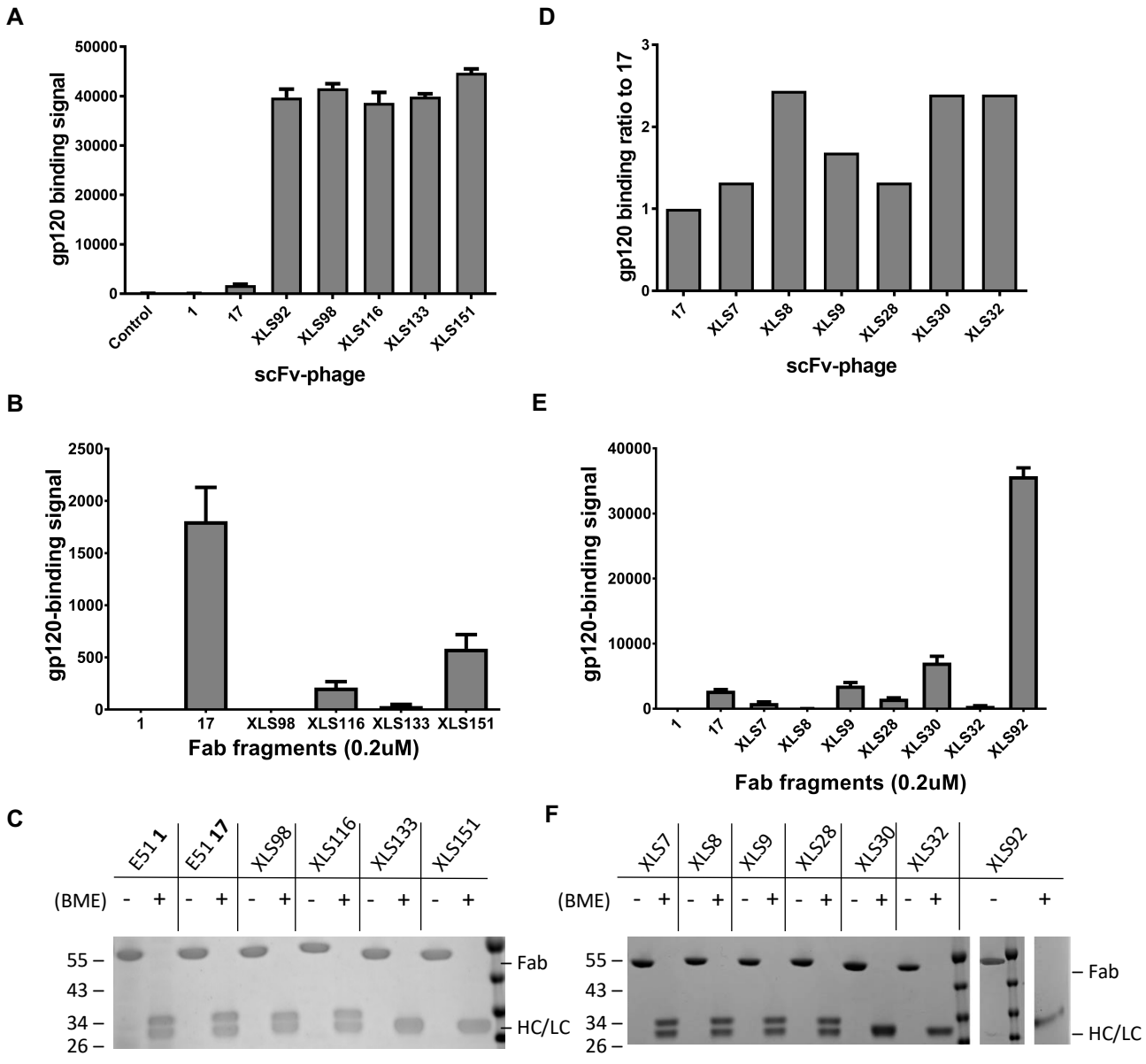

Figure 6.2. E51 library hits from the first and second selection.

ELISA against 1 μg of gp120 on scFv library hits from (A) the first selection and (B) the second selection, in which the ELISA binding signals are normalized to **17** binding signal. ELISA against 1 μg of gp120 on 0.2 μM of Fab fragments from (C) the first selection and (D) the second selection. (E, F) Coomassie gels on 0.8 μg of Fab fragments used in (B) and (C), respectively. BME = β-mercaptoethanol.

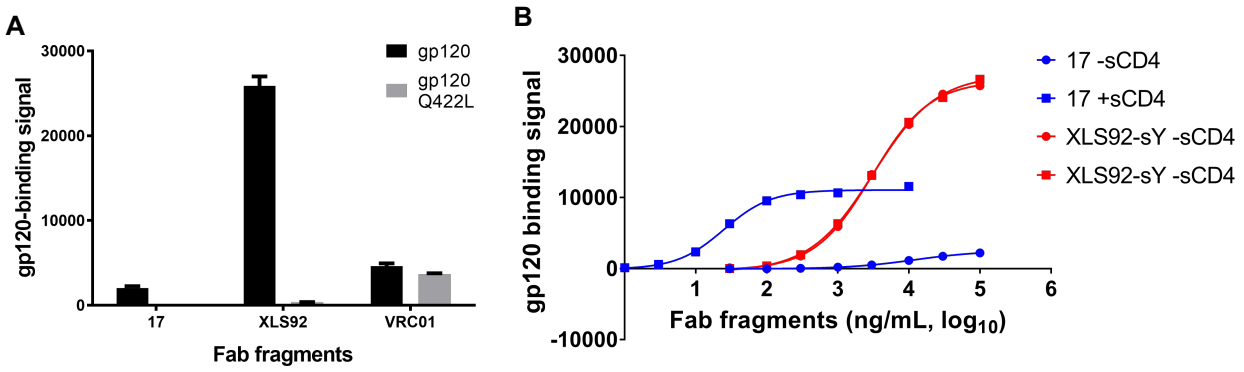

Figure 6.3. XLS92 is not a CD4i antibody.

(A) ELISA on 1 μg of Fabs against gp120 or gp120(Q422L). VRC01 is an anti-HIV antibody that targets the CD4-binding epitope of gp120. (B) ELISA on **17** and XLS92 Fabs at various concentrations against gp120 with or without the addition of sCD4.

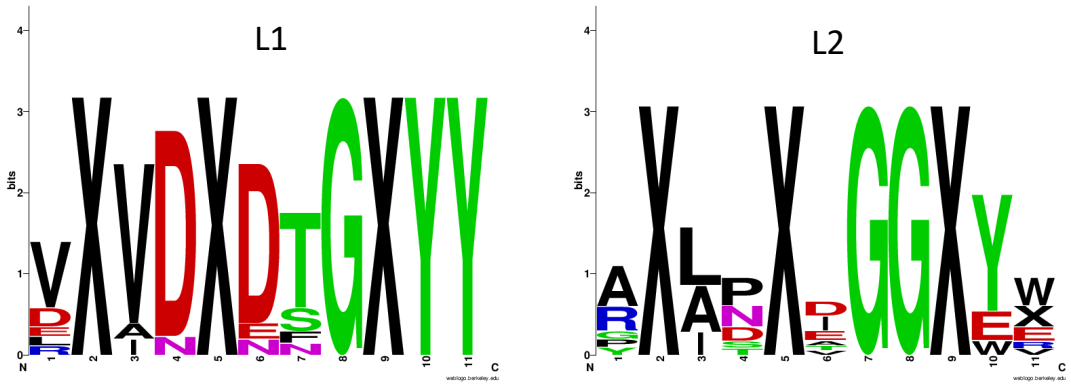

Figure 6.4. VHCDR3 amino acid sequence logos of library L1 or L2 gp120-binding variants from the third selection.

The generation of sequence logos was based on Sangar sequencing results of 24 randomly selected colonies at the end of round 5 of the third selection. A 'X' denotes a sY.

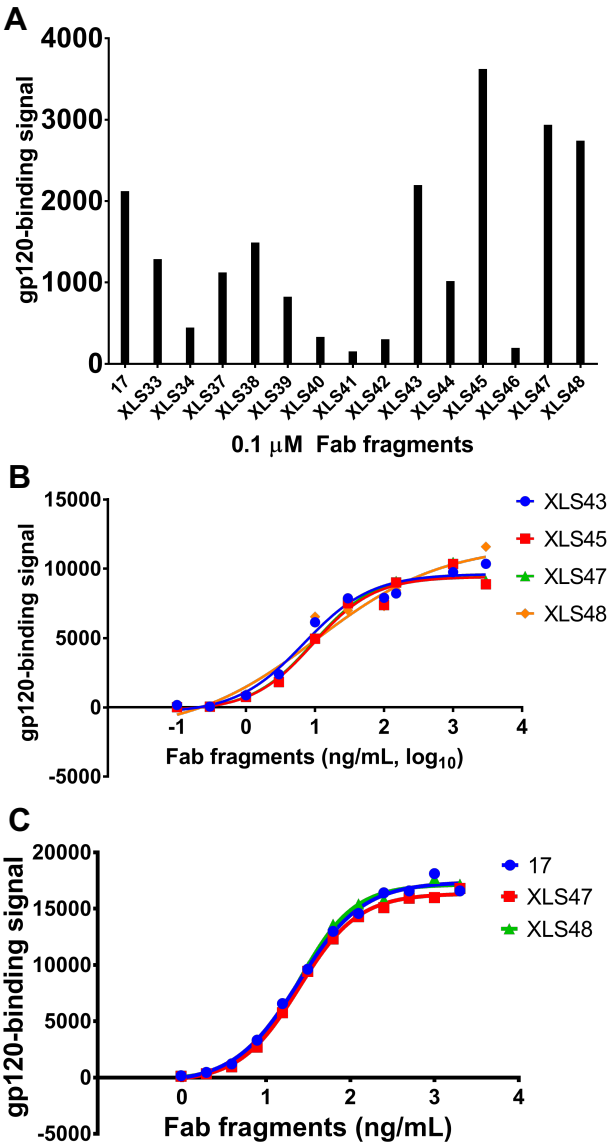

Figure 6.5. gp120-binding profiles of antibodies from the third selection.

(A) ELISA on Fabs against gp120 in the presence of sCD4. (B, C) ELISA on 17 and high-binding XLS Fabs at various concentrations against gp120 in the presence of sCD4.

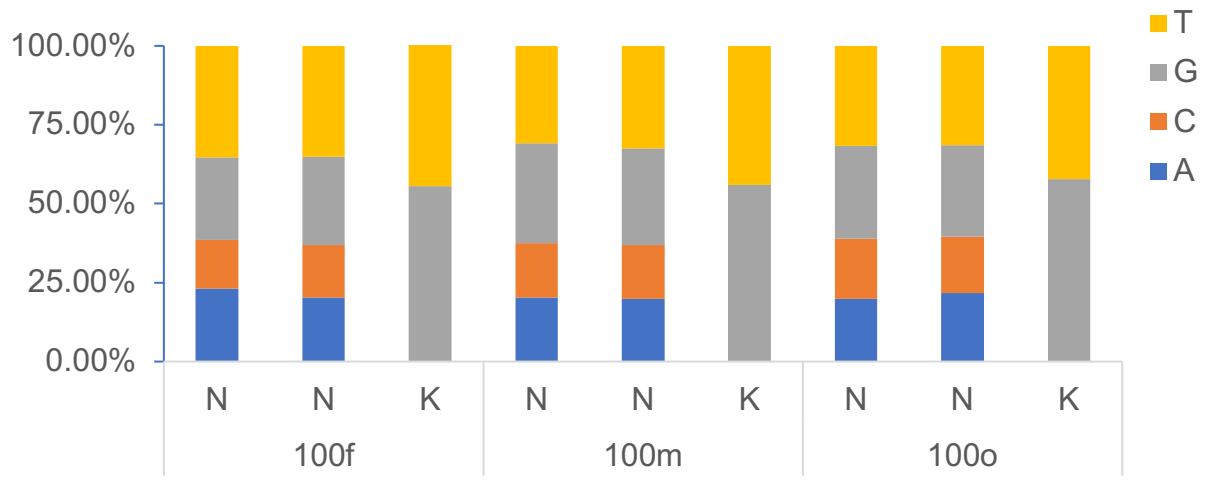

Figure 6.6. Distribution of nucleotides in NNK of L4.

Frequencies of nucleotides at every base positions of NNKs in L4 were counted from L4 Total DNA reads (Table 5). 100f, 100m, and 100o refer to the amino acid positions randomized in the E51 V_HCDR3.

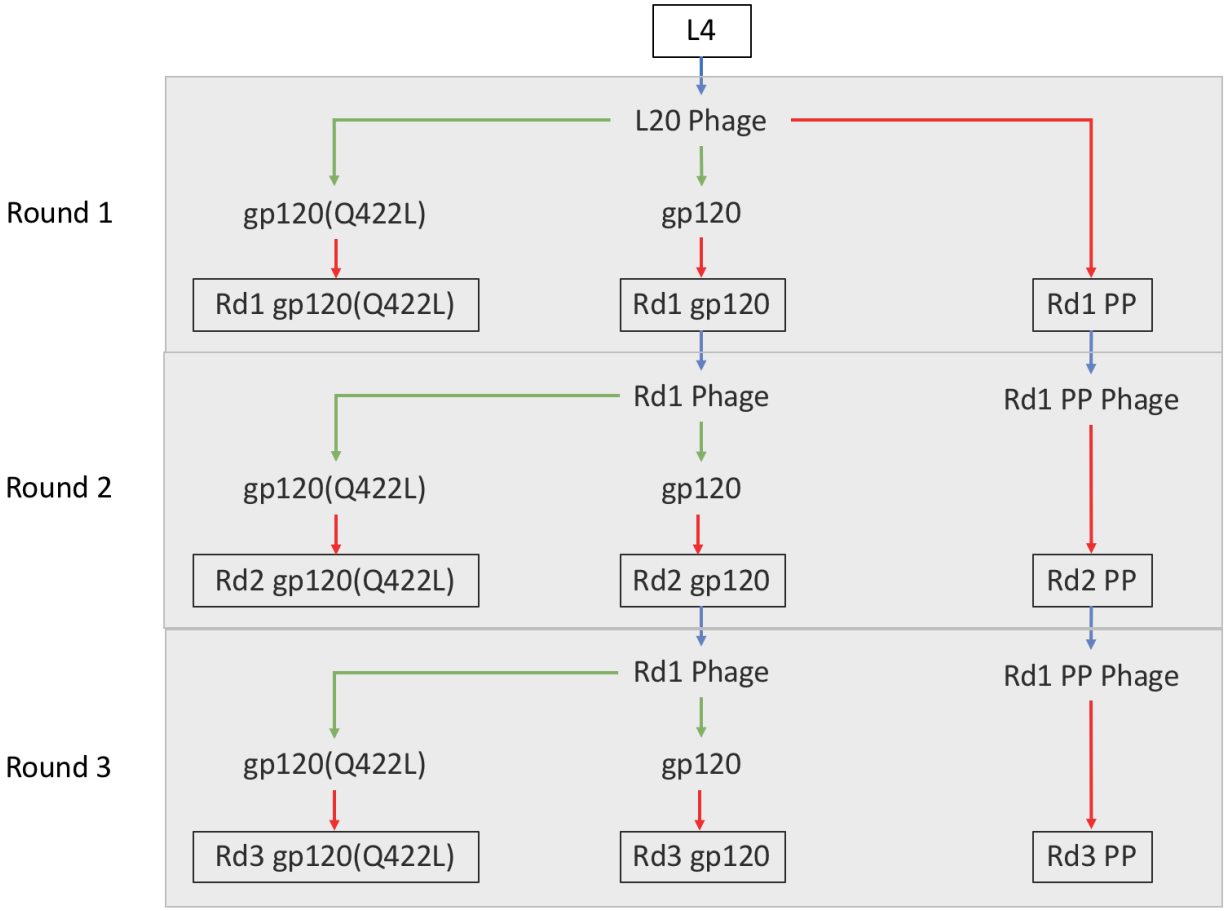

Figure 6.7. Diagram of the fourth phage display selections.

Phage expressed from L4 and amplified cultures from round 1-3 selections against gp120 (Rd1, 2, or 3 gp120) were evenly split for the next round of selection against gp120 or gp120(Q422L) or phage passaging (PP) experiments. Blue arrows indicate phage expression and precipitation. Green arrows indicate selection panning. Red arrows indicate phage infection into SS320 cells with pULTRA-sY. Boxes indicate that DNA extracted from those cultures after phage infection were submitted for NGS.

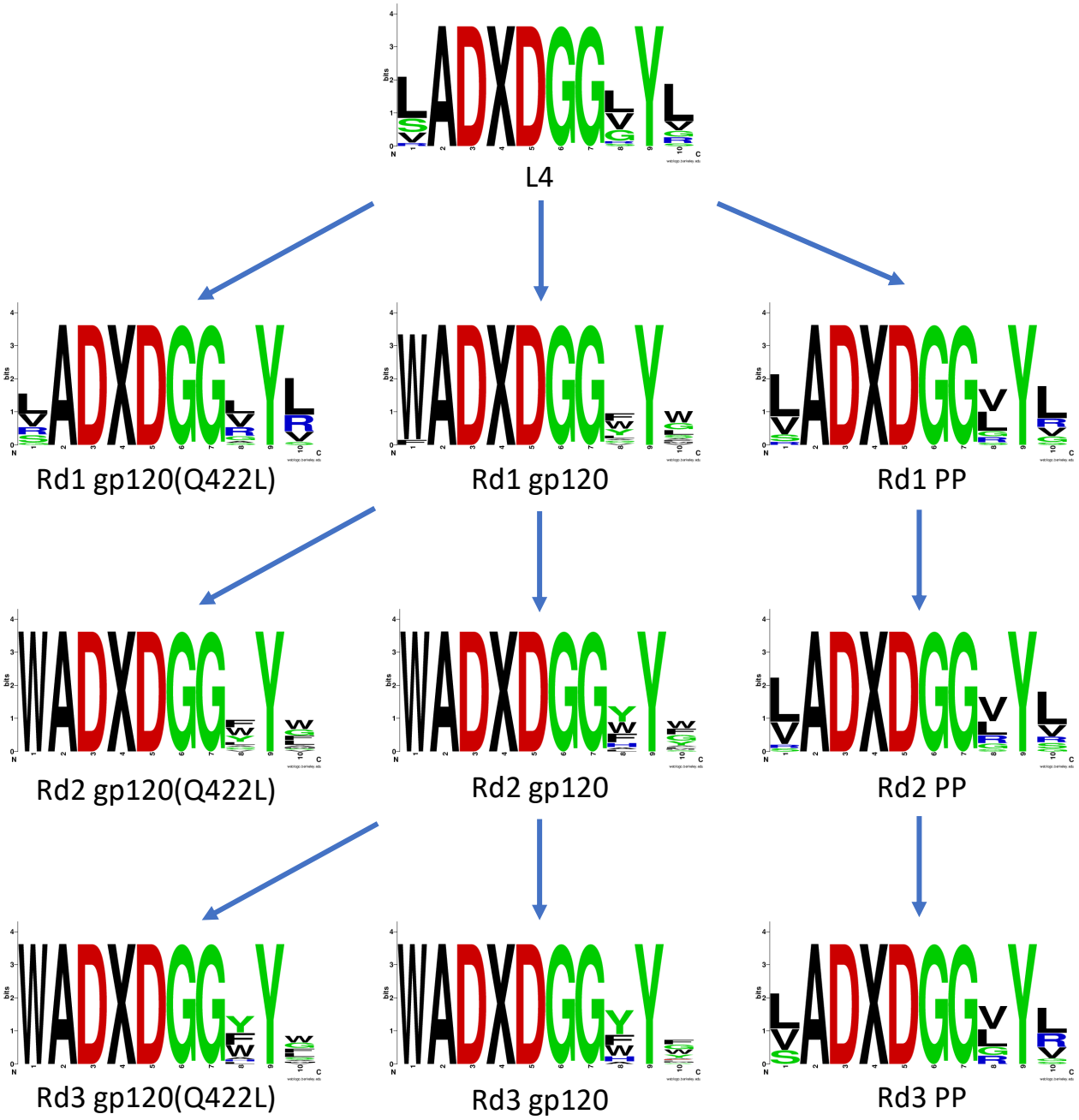

Figure 6.8. Sequence logos of 20 of the most frequent sequences from the fourth selection.

These sequences were identified from NGS results of cultures indicated in Figure 7 (Boxed). A 'X' denotes a sY.

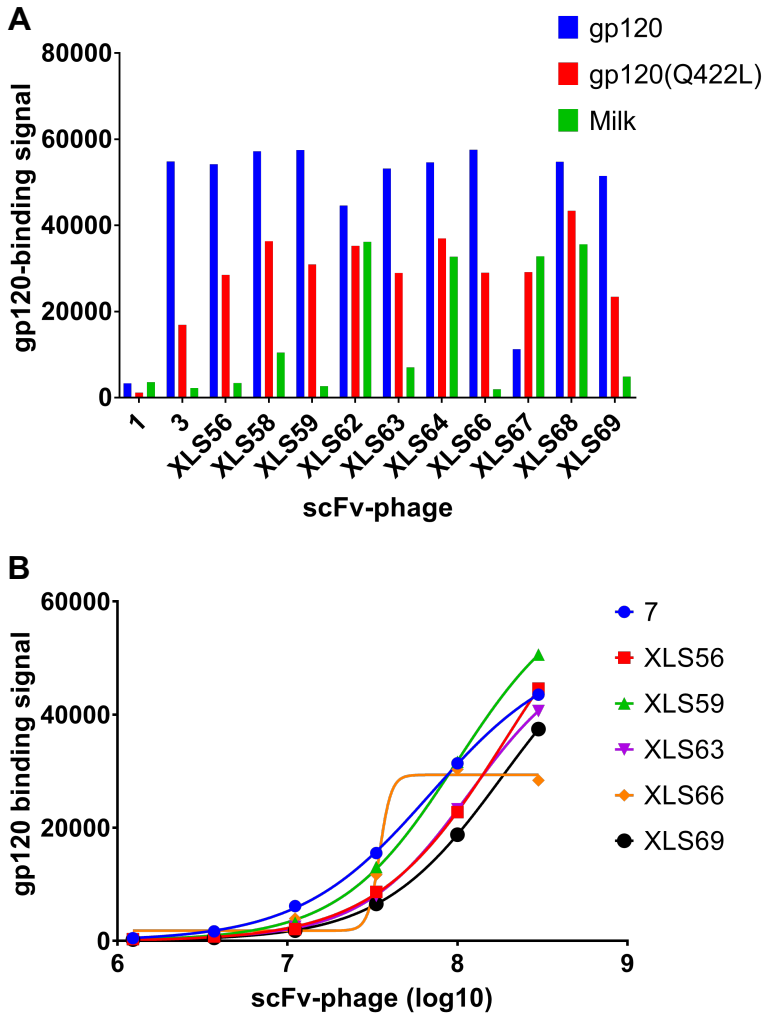

Figure 6.9. gp120-binding profiles of antibodies from the fourth selection.

(A) ELISA on scFv-phage of E51 sulfoforms 1 and 7, and antibodies the top 20 most frequent sequences from Rd3 gp120 and Rd3 gp120(Q422L) against gp120, gp120(Q422L), or 2% Milk in the presence of sCD4. (B) ELISA on various concentrations of scFv-phage variants that showed low non-specific binding against gp120 with sCD4.

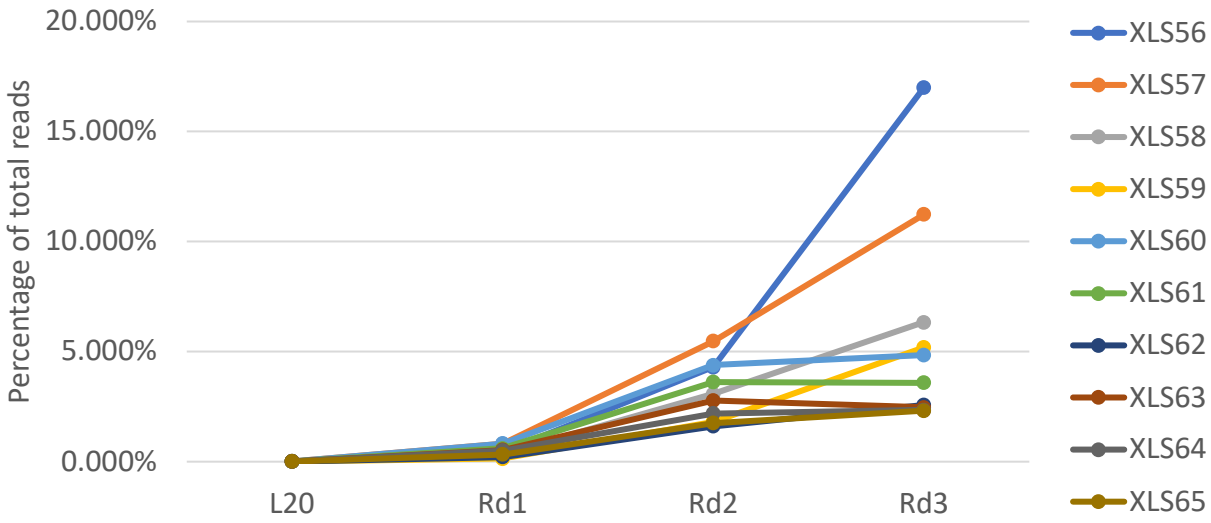

Figure 6.10. Sequence enrichment of 10 most frequent sequences from round 3 of the fourth selection.

Fraction percentages were calculated by dividing the DNA frequency of each variant by the total DNA sequences with high Phred Scores (Good QS DNA from Table 5). The sum of these percentages for L20, Rd1, Rd2, and Rd3 are 0.07%, 4.80%, 31.03%, and 57.92%, respectively. V_HCDR3 sequences of these most frequent variants are listed in Table 6.

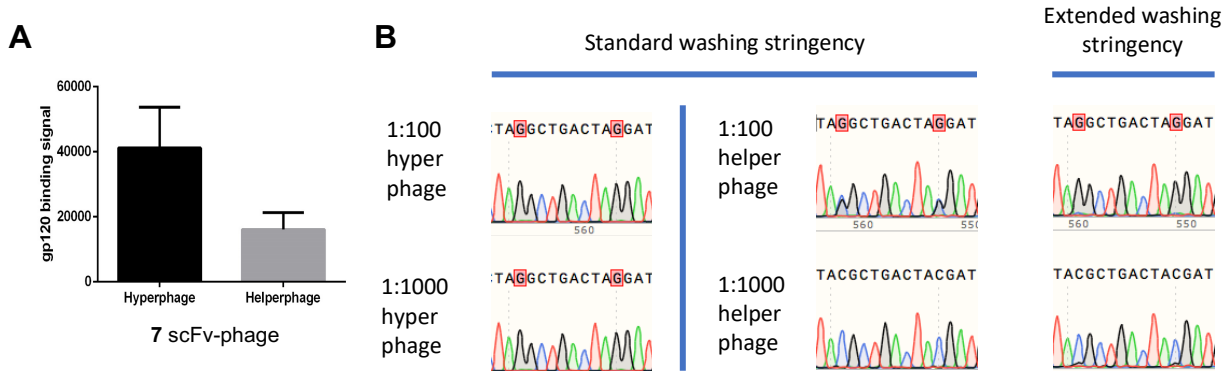

Figure 6.11. Phage enrichment comparison between helperphage and hyperphage in mock selections.

(A) ELISA on (B) One round of mock selection was performed on different dilutions, 1:100 or 1:1000, of the doubly-sulfated E51 sulfoform **7** scFv-phage to unsulfated E51 sulfoform **1** scFv-phage, expressed using either helperphage or hyperphage, against gp120 with sCD4. The conversion from TAC to TAG indicates enrichment for the CD4i gp120-binding sulfoform, **7**. Standard washing stringency refers to washing selection wells after panning 5 times with PBST for 5 minutes. Extended washing stringency refers to washing selection wells after panning 20 times with BPST for 3 hours.

Name	Sequence
1	DYADYDGGYYY
2	DYADYDGGYYY
3	DYADYDGGYYY
4	DYADYDGGYYY
5	DYADYDGGYYY
6	DYADYDGGYYY
7	DYADYDGGYYY
11	DYADYDGGYYY
17	DYADYDGGYYY
32	DYADYDGGYYY

Table 6.1. List of VHCDR3 amino acid sequences of E51 sulfoforms displayed as scFv on phage or as Fabs.

A red 'Y' denotes a sY.

Selection	Round #	gp120 coated in wells (μg/mL)	sCD4 added (μg/mL)	Panning duration	Washing stringency
1 Kon	1*	125	0	4 hours	15 min, 5 times
	2	25	0	20 min	3 hours, 10 times
	3	25	0	20 min	10 hours, 10 times
	4	25	0	20 min	10 hours, 10 times
	5	5	0	20 min	26 hours, 20 times
	6	5	0	20 min	48 hours, 20 times
1 Koff	1	125	0	4 hours	15 min, 5 times
	2	25	0	4 hours	3 hours, 10 times
	3	25	0	4 hours	10 hours, 10 times
	4	25	0	4 hours	10 hours, 10 times
	5	5	0	4 hours	26 hours, 20 times
	6	5	0	4 hours	48 hours, 20 times
2 Kon	1*	125	50	4 hours*	15 min, 5 times
	2	25	2.5	20 min	1 hour, 10 times
	3	25	2.5	20 min	3 hours, 10 times
	4	5	1	20 min	8 hours, 10 times
	5	5	1	20 min	24 hours, 20 times
2 Koff	1	125	50	4 hours	15 min, 5 times
	2	25	5	4 hours	1 hour, 10 times
	3	25	5	4 hours	3 hours, 10 times
	4	5	1	4 hours	8 hours, 10 times
	5	5	1	4 hours	24 hours, 20 times
3	1	125	50	2 hours	15 min, 5 times
	2	25	5	2 hours	1 hour, 10 times
	3	25	5	2 hours	3 hours, 10 times
	4	5	1	2 hours	8 hours, 10 times
	5	5	1	2 hours	24 hours, 20 times
4	1	10	10	2 hours	3 hours, 20 times
	2	10	10	2 hours	3 hours, 20 times
	3	10	10	2 hours	3 hours, 20 times
	4	10	10	2 hours	3 hours, 20 times

Table 6.2. Phage display selection conditions.

Selection "Kon" refers to selecting for antibodies with high binding association while "Koff" refers to selecting for antibodies with low binding disassociation. gp120(Q422L) or sCD4 was added to phage prior to panning. * refers to the fact that Round 1 of first and third selection was performed under only the "Koff" panning condition.

Name	Library	Sequence	Selection	Tyrosine sulfation by TPSTs prediction
E51	None	DYADYGDDYYY	None	DYADYGDDYYY
L1	L1	XYXXYXXXYYY	First, second, and third	
XLS98	L1	DYYVYWTRYYY	First	DYYVYWTRYYY
XLS133	L1	DYIVYWTRYYY	First	DYIVYWTRYYY
XLS151	L1	PYILYWTRYYY	First	PYILYWTRYYY
XLS7	L1	PYADYYRGYYY	Second	PYADYYRGYYY
XLS8	L1	AYNTYEDGYYY	Second	AYNTYEDGYYY
XLS9	L1	PYLDYNGGYYY	Second	PYLDYNGGYYY
XLS33	L1	VYVDYDTGYYY	Third	VYVDYDTGYYY
XLS39	L1	FYLSYEEGYYY	Third	FYLSYEEGYYY
XLS40	L1	PYTNYEQGYYY	Third	PYTNYEQGYYY
XLS44	L1	WYTQYEGGYYY	Third	WYTQYEGGYYY
XLS47	L1	DYVQYDMGYYY	Third	DYVQYDMGYYY
XLS48	L1	DYWDYESGYYY	Third	DYWDYESGYYY
L2	L2	XYXXYXGGYXX	First, second, and third	
XLS92	L2	QYMWYFSA	First & second	QYMWYFSA
XLS116	L2	PYPWYYGGYEG	First	PYPWYYGGYEG
XLS28	L2	LYLDYDAGYYY	Second	LYLDYDAGYYY
XLS30	L2	LYYFSG	Second	LYYFSG
XLS32	L2	PYVLYWTRYYY	Second	PYVLYWTRYYY
XLS34	L2	RYLNYIGGYYY	Third	RYLNYIGGYYY
XLS37	L2	AYSPYTGGYWE	Third	AYSPYTGGYWE
XLS38	L2	PYIDYTGGYYE	Third	PYIDYTGGYYE
XLS41	L2	FYTDYMGGYVG	Third	FYTDYMGGYVG
XLS42	L2	PYQPYDGGYFW	Third	PYQPYDGGYFW
XLS43	L2	YYIDYEGGYYW	Third	YYIDYEGGYYW
XLS45	L2	DYVDYNNGYYY	Third	DYVDYNNGYYY
XLS46	L2	WYTDYFGGYEG	Third	WYTDYFGGYEG
L3	L3	DXADYDGGXYX	Fourth	
XLS56	L3	DWADYDGGYYW	Fourth, gp120	DWADYDGGYYW
XLS58	L3	DWADYDGGYYF	Fourth, gp120	DWADYDGGYYF
XLS59	L3	DWADYDGGYYY	Fourth, gp120	DWADYDGGYYY
XLS63	L3	DWADYDGGYYA	Fourth, gp120	DWADYDGGYYA
XLS64	L3	DWADYDGGWYF	Fourth, gp120	DWADYDGGWYG
XLS66	L3	DWADYDGGFYY	Fourth, gp120	DWADYDGGFYY
XLS67	L3	DWADYDGGWYG	Fourth, gp120	DWADYDGGYYW
XLS62	L3	DWADYDGGHYF	Fourth, gp120(Q422L)	DWADYDGGHYF
XLS68	L3	DWADYDGGYYE	Fourth, gp120(Q422L)	DWADYDGGYYE
XLS69	L3	DWADYDGGYYS	Fourth, gp120(Q422L)	DWADYDGGYYS
L4	L4	DYADXDGGXYX	Fifth	

Table 6.3. VHCDR3 sequences of libraries and selection hits tested for gp120-binding.

A blue 'X' denotes a randomized residue that can be either of the 20 natural amino acids or sY. A red 'Y' denotes a sY incorporated by an expanded genetic code. A green 'Y' denotes a Tyr computationally predicted for tyrosine sulfation using GPS-TSP.

Selection	Round number	Kon or Koff	Eluted/ panned phage	Selection	Round number	Kon or Koff	Eluted/ panned phage
Library 1				Library 2			
First	1	Kon	1.77×10^{-3}*	First	1	Kon	2.66×10^{-4}*
	2		9.35×10^{-6}		2		2.04×10^{-5}
	3		3.85×10^{-6}		3		4.15×10^{-6}
	4		2.27×10^{-4}		4		1.18×10^{-4}
	5		4.20×10^{-5}		5		1.08×10^{-3}
	6		NA		6		NA
	1	Koff	1.77×10^{-3}*		1	Koff	2.66×10^{-4}*
	2		4.61×10^{-6}		2		1.95×10^{-5}
	3		1.81×10^{-5}		3		1.32×10^{-6}
	4		3.82×10^{-4}		4		3.00×10^{-4}
	5		4.17×10^{-3}		5		2.42×10^{-3}
	6		NA		6		NA
Second	1	Kon	2.45×10^{-3} *	Second	1	Kon	4.93×10^{-3} *
	2		1.62×10^{-4}		2		2.24×10^{4}
	3		2.41×10^{-5}		3		1.80×10^{4}
	4		3.98×10^{-4}		4		1.38×10^{3}
	5		2.46×10^{-5}		5		NA
	1	Koff	2.45×10^{-3} *		1	Koff	4.93×10^{-3} *
	2		7.52×10^{-4}		2		4.46×10^{-4}
	3		1.96×10^{-5}		3		4.81×10^{-4}
	4		2.89×10^{-3}		4		3.62×10^{-5}
	5		3.17×10^{-5}		5		3.39×10^{-5}
Library 3				Library 4			
Third	1	Koff	4.56×10^{-7}	Fourth	1	Koff	1.45×10^{-5}
	2		3.55×10^{-6}		2		4.82×10^{-4}
	3		2.23×10^{-8}		3		4.21×10^{-4}
	4		8.59×10^{-5}		4		NA
	5		6.98×10^{-7}		5		NA

Table 6.4. Phage enrichment in multiple selections of phage display.

Overall enrichment of phages is indicated by an increase of eluted/panned phage ratio as long as those rounds were subjected to the same selection stringency, which is not necessarily the case for the first three selections. * denotes the same round 1 phage enrichments shared between the first and second selection, because the round 2 starting populations are the same for the first and second selection.

	L4	Rd1 gp120	Rd2 gp120	Rd3 gp120	Rd1 gp120 Q422L	Rd2 gp120 Q422L	Rd3 gp120 Q422L
Total DNA reads	1085522	179516	189789	255685	215480	190694	191213
Good QS DNA[a]	807396, 74.0%	145591, 81.0%	159231, 84.0%	214140, 84.0%	168582, 78.0%	154347, 81.0%	157923, 83.0%
Exact design DNA[b]	652271, 81.0%	142913, 98.0%	157719, 99.0%	212561, 99.0%	150963, 90.0%	152155, 99.0%	156353, 99.0%
Unique DNA[c]	35983, 109.8%	6197, 18.9%	1901, 5.8%	1810, 5.5%	16378, 50.0%	6361, 19.4%	3322, 10.1%
Most redundant DNA	165	1206	8722	36383	208	1476	4664
Unique Proteins[d]	11720, 126.6%	3894, 42.0%	1300, 14.0%	1279, 13.8%	7279, 78.6%	4005, 43.2%	2177, 23.5%
Most redundant proteins	1039	1449	8725	36409	332	1693	4666
Proteins with additional sYs	2571	470	23	72	200	78	138

Table 6.5. Next-generation sequencing analysis of the fourth selection.

Total DNA reads represents the number of raw reads obtained from splitting a single Illumina MiSeq run. NGS was performed on the V_HCDR3 of L4, Rd1-3 selections against gp120, Rd1-3 selections against gp120(Q422L), and Rd1-3 phage passaging (PP) experiments.

a. DNA sequences from total raw reads with Phred scores above 30 at every nucleotide. Percentages are with respect to Total DNA reads.
b. DNA sequences without extra insertions, deletions, and frameshift mutations to the V_HCDR3. Percentages are with respect to Good QS DNA.
c. Non-redundant DNA sequences. Percentages are with respect to the theoretical DNA diversity of L4 (32,768).
d. Non-redundant protein sequences. Percentages are with respect to the theoretical protein diversity of L4 including sY (9261).

	Rd1 PP	Rd2 PP	Rd3 PP
Total DNA reads	293206	175497	189789
Good QS DNA[a]	220236, 75.0%	134296, 77.0%	139592, 73.6%
Exact design DNA[b]	206414, 94.0%	127260, 95.0%	132434, 99.0%
Unique DNA[c]	32486, 99.1%	28491, 86.9%	28666, 87.5%
Most redundant DNA	58	41	53
Unique Protein[d]	10135, 109.4%	9047, 97.7%	9088, 98.1%
Most redundant protein	320, 0.0015%	216, 0.0017%	226, 0.0017%
Proteins with additional sYs	183	20	30

Table 5 (continued).

Name	Sequence	L20	Rd1	Rd2	Rd3	Rd1 Q422L	Rd1 PP	Rd2 PP	Rd3 PP
XLS56	WADYDGGYYW	38	641	6857	36409	39	11	4	13
XLS57	WADYDGGFYW	101	1184	8725	24053	25	31	28	21
XLS58	WADYDGGYYF	39	517	4912	13580	32	17	5	7
XLS59	WADYDGGYYY	36	214	2895	11118	28	7	7	5
XLS60	WADYDGGWYW	90	1207	6976	10374	31	40	27	30
XLS61	WADYDGGYYG	84	883	5758	7696	38	25	19	20
XLS62	WADYDGGHYF	24	312	2571	5514	24	7	10	9
XLS63	WADYDGGYYA	56	771	4424	5269	29	22	11	13
XLS64	WADYDGGWYF	77	740	3485	5045	28	23	24	9
XLS65	WADYDGGWYY	45	474	2802	4964	26	16	5	17
3	YADYDGGYYY	19	34	68	58	0	9	6	2
7	YADYDGGYYY	12	70	112	227	2	5	0	0
17	YADYDGGYYY	25	13	0	0	0	0	0	0
32	YADYDGGYYY	23	0	0	0	0	0	0	0

Table 6.6. Selection enrichment of 10 most frequent antibodies from round 3 of fourth selection and E51 sulfoforms in the fourth selection.

Numbers represent the frequency of the corresponding amino acid sequences in each NGS result. A red 'Y' denotes UAG stop codon in the DNA sequence and sY in the amino acid sequence.