Spring 5-15-2016

Engineering Cre Recombinase for Genome Engineering

Chi Zhang
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology and Biomedical Sciences
Computational and Systems Biology

Dissertation Examination Committee:
James Havranek, Chair
Joseph Corbo
Robi Mitra
Gary Stormo
Ting Wang

Engineering Cre Recombinase for Genome Editing

by
Chi Zhang

A dissertation presented to the
Graduate School of Arts & Sciences
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

May 2016
St. Louis, Missouri

Table of contents

List of Figures

List of Tables

Acknowledgments

First and foremost, I would like to thank my thesis advisor and mentor, Dr. James Havranek. I was the first student at WashU to join his lab and he supported and guided me throughout these years. He provides expertise in both computational modeling and experimental techniques. He patiently accepted all of my failures and encouraged me to achieve success. He is an excellent mentor, and is not only a role model for my career but also for my life.

I would like to thank the members of my thesis committee who provided thoughtful advices to my research and supported my graduate career. In particular, I would like to thank Dr. Gary Stormo for serving as my committee chair and guiding my rotation project. I would also like to thank Dr. Joseph Corbo and Dr. Robi Mitra for their advices and facilitating with validation of asymmetric Cre recombinase in mammalian cells. I am grateful to Dr. Ting Wang for always being a positive mentor and a role model.

I would like to thank all the members of Havranek lab for their friendship, support and guidance over the years. Together, they make the lab a pleasant environment. I owe a great deal to Adam Joyce for his scientific insights and experimental suggestions. He helped me a lot with NGS library design and protocol. I would also like to thank Dr.Benjamin Borgo and Dr. Nicholas Bodmer for helpful discussions. Finally, I would like to thank our technician, Hollie Noia, for her support over the years.

I would like to thank my collaborators for their great work and comments on the manuscript. In particular, Connie Myers from Corbo lab helped with recombinase assays in mouse retinal explants and Zongtai Qi from Mitra lab helped with assays in mouse ES cell cultures. I am also grateful to Dr. Phil Bradley for his help writing the review article on modelling protein-DNA interactions and providing insights on my projects during RosettaCon.

I would like to thank everyone who have helped and supported me in the Computational Biology Program, the Genetics Department and the Biochemistry Department. My coordinators Melanie Relich, Jeanne Silvestrini and Melissa Torres have supported me and gone beyond their everyday work. I am grateful to Dr. Barak Cohen for serving as the chair of CSB program and giving me advices during my graduate study. I owe a lot to Jessica Hoisington-Lopez for helping with my NGS sequencing runs.

I am extremely fortunate to have a group of friends at Washington University. Shuang Chen and Meng Wu helped me with my career. Peichang Shi and Xiaochun Mou kindly provided me with transportation for a couple of years. Jie Zhang, GiNell Elliot and Mi Cai gave me a lot of advices. Xinhan Tao,Wan Shi, Lu Yu, Shuo Luo, Cong Li, Zhenghua He, Haowei Yuan, Peng Lin, Jiajing Xu, Xingxing Chen, Peng Li, and many more supported me one way or the other and gave me amazing memories.

Last by not least, I want to thank my parents, Yousu Zhang and Honggang Luo, my grandparents Laiwei Luo and Zhongyuan Liu, and Yunfei Shi, who is like a brother to me, for their unconditional love and support. They are always there for me whenever I need them.

Chi Zhang

Washington University in St. Louis

Martch 2016

ABSTRACT OF THE DISSERTATION

Engineering Cre Recombinase for Genome Editing

by

Chi Zhang

Doctor of Philosophy in Biology and Biomedical Sciences

Computational Biology

Washington University in St. Louis, 2016

Assistant Professor James Havranek, Chair

Cre recombinase recombines its DNA target, loxP sites, without help of accessory proteins or DNA repair systems. The simplicity of Cre-lox system has been widely utilized for genome editing, especially in mouse genetics. The goal of this dissertation is to construct Cre recombinase variants that will operate upon recombination target sites (RTs) present within the genome, instead of perturbing the genome by inserting wild-type RTs for subsequent genome engineering. In general, the desired RTs native to the genome are asymmetric. However, the loxP sequence is pseudo-palindromic, requiring a homotetrameric formation of Cre recombinase. As a first step, I broke the symmetry of Cre tetramer so that each Cre monomer could be arranged spatially to bind distinct RT half-sites. I designed an alternative protein-protein interface for Cre. Then, I separated the mutations into a pair of Cre monomers. I could then arrange the assembly of this pair of complementary Cre monomers to form a functional heterotetramer, even though neither monomer exhibits activity alone. When combined with other mutations that confer distinct DNA specificities, the monomers preferentially formed the desired complex and recombined asymmetric DNA sequences with greater fidelity. I've successfully found a pair of Cre monomers that do not work in isolation, but do when combined

together. This has been successfully demonstrated *in vivo* in E. coli, mouse ES cell cultures and mouse retinal explants. As the next step, I need to change the DNA specificity of Cre recombinase to recognize native genomic sites. Surprisingly, the DNA preferences of Cre recombinase have not been thoroughly characterized. The 34 bp RT site loxP contains two palindromic arm regions and an 8 bp spacer region. The arm region is recognized by Cre monomers while homology of the spacer region determines compatibility between RT sites. While the consensus sequence of loxP is known, I performed the first high-throughput studies to determine Cre's sequence specificity in the arm region. I broke the 13 bp arm region into 3 overlapping 5-6 bp small windows and used in *vitro* recombination and high-throughput sequencing data to generate logos for each window. I found that non-specific recombination can interfere with the analysis and careful selection of NaCl concentration is important for observing *in vitro* specificity. I have not only determined Cre's sequence preferences, but also used similar methods to determine CreC2#4 (a Cre mutant) and VCre (a Cre homolog). In contrast to zinc finger and TAL effector domains, no modular decomposition of DNA specificity exists for Cre recombinase homologs. As a result, the RT specificity of Cre has previously been modified using directed evolution, a laborious approach. To accelerate the process, I used sequence information from homologs of Cre. By searching across genomes of different bacteria species, I found hundreds of Cre homologs. Closely related homologs share similarity in both amino acid sequence and predicted RT DNA sequence. By comparing residues that differ between close homologs in the aligned regions where Cre contacts the switched base pairs, I found candidate possible mutations for a specificity switch. The change in specificity was validated by the high-throughput sequencing assay. This demonstrates the feasibility of leveraging sequence alignment

data to alter the specificity of Cre recombinase, reducing the amount of effort needed to generate mutants with novel RT preferences.

Chapter 1

Introduction:

Directed Evolution, Cre Recombinase and RMCE

Directed evolution of enzymes

Directed evolution is a method used in enzyme engineering that utilizes features of natural selection to evolve proteins to achieve a specific goal. It consists of two iterative steps: variant library creation/amplification and selection (1). It has proved to be a powerful and broadly applicable tool for altering the activities of enzymes. Strategies for directed evolution usually focus on both sequence diversifications for library construction and selection for altered activity.

The mutational sequence space of a full length protein is huge. Complete randomization of ten amino acids yields 10^{13} unique combinations. However, *in vivo* bacterial selections are limited by the transformation efficiency (at best 10^{9} - 10^{11}). Since exhaustive coverage of mutational space is impossible to achieve, directed evolution samples only a fraction of this large sequence space, with beneficial mutations accumulated successively in each round. Goeddel and co-workers first employed error-prone PCR to create random mutagenesis *in vitro* (2). By altering PCR conditions or using a proprietary mixtures of polymerases, mutation rates of 10^{-4} to 10^{-3} per base are achievable, and biases for different modes of mutation (e.g. transitions *versus* transversions) can be reduced (3) (4). When specific residues responsible for binding or catalysis are known, a more targeted strategy may be used. Synthetic DNA oligonucleotides containing degenerate codons targeted to the residues of interest can achieve exhaustive mutation of important residues (5). After several rounds of accumulating beneficial mutations, the technique of DNA shuffling can be applied to access combinations of these mutations. In this approach, first described by Stemmer, a family of genes is treated with DNase. Fragments of a desired length are then gel purified, and used as primers to reassemble a chimeric gene in a PCR reaction (6). To provide greater control over fragment size, recent protocols leverage misincorporation of

dUTP during PCR. Fragments of the amplicons are generated by treatment with uracil deglycosylases and apurinic/apyrimidinic lyases, and the size of the fragments can be controlled by the rate of dUTP misincorporation (7).

Effective selection strategies often link the desired enzymatic activity to the physical separation of the encoding DNA of the desired subpopulations of variants or to the survival of the organism expressing the enzyme. In binding affinity selections, protein library members are captured by an immobilized target together with their encoding DNA sequences. Phage display methods have broad applications in protein engineering and can serve as a successful example, where the protein library of interest is fused with phage coat proteins and expressed on the surface of the phage (8) (9). While non-binding phages are washed away, the bound phages then serve as the compartment to segregate the protein with the encoding DNA. The other strategy is to link the function of the enzyme to fitness survival of the organism. The most common practice is to evolve proteins by linking their activity to the expression of an antibiotic resistance gene. For example, Barbas and colleagues altered serine recombinases' DNA specificities by using the recombinase activity to reassemble a beta-lactamase gene (10). In the above strategies, cells or phages were used as compartments, where the library size was limited by the transformation efficiency. Purely *in vitro* selections, utilizing cell-free translation reactions, can bypass this limitation. One straight forward approach is to use mRNA display or ribosome display, where the mRNA is linked to the protein during translation (11) (12). In another *in vitro* approach, the encoding DNA and proteins are trapped in aqueous droplets of water-oil emulsions. This approach is particularly useful to evolve enzymes that directly operate on DNA. Reported examples include nucleases, RNA and DNA polymerases or enzymes with activities that can be linked to a polymerase (13) (14) (15).

The Cre-lox system

Cre recombinase is a 380 kDa protein originated from bacteriophage P1. It belongs to the λ integrase family, other examples of which include λ intergrase, Flp and more recently Vika, Scre, and VCre (16)-19). Cre recombines a specific 34 bp DNA recombination target (RT) site, the loxP sequence. LoxP site contains two palindromic 13 bp arm regions and an asymmetric 8 bp spacer region (Figure 1.1)(17). Each arm region is recognized by a Cre monomer. While there is no direct contact between Cre and the spacer region, the homology of the spacer is required for efficient recombination between two RT sites (18).

The recombination synaptic complex contains four Cre monomers, forming a homotetramer, and two loxP DNA sequances in an anti-parallel arrangement (Figure 1.2)(19). Cre cleaves and religates after position 14 of the top loxP strand and before position -14 of the bottom strand (Figure 1.1). During cleavage, residue Y324 is covalently linked to the 5' phosphate of the cleaved nucleotide. The four-fold protein symmetry imposes a pseudo-palindromic symmetry requirement on the RT site.

Altering Cre's DNA specificities

Several high resolution Cre-DNA crystals structures are available and can serve as a guide to rational redesign of DNA specificity. The protein-DNA interface is complex, with both direct interactions between the protein side chains and the DNA bases, and indirect interactions mediated by water molecules. In contrast to zinc finger or TAL effector domains, there is no modular decomposition for the interface. The absence of any simple scheme for understanding Cre's specificity for DNA makes the selection of specificity-altering mutations a significant challenge.

There have been several attempts to characterize the DNA sequence specificity of wil-type Cre recombinase. A sequencing study has been published for the specificity and compatibility of the loxP spacer region (20), but the specificity in the arm region was not characterized. In another study, a randomized library approach to identifying alternative lox sites showed some of Cre's sequence preferences *in vivo* (21). Because of extreme sequence promiscuity observed using their experimental conditions, they were unable to obtain detailed sequence preferences. Before attempting to change Cre's DNA specificity, it is important to fully characterize the protein-DNA interaction. One promising approach is to couple a functional screen for recombinase activity against a randomized library with the massive throughput capabilities of next-generation sequencing. With a detailed position weight matrix for Cre specificity in hand, we can improve our ability to find candidate target site in the genome by taking into account which mismatches with the loxP sequence have the most effect on Cre's recombinase activity.

Directed evolution is a powerful tool for altering Cre's DNA specificity. Buchholz and Stewart retargeted Cre to recombine a site from human chromosome 22, which they called the loxH site (22). In their approach, full length Cre recombinase was randomized by error-prone PCR. The loxH site has four changes in the arm region and a different spacer region. They used both positive and negative selection steps *in vivo* in bacterial cells. First, Cre mutants were evolved to recombine a loxP/loxH hybrid site for ten rounds. Then, Cre mutants were evolved for another ten rounds to efficiently recombine the full loxH site. However, the resulting mutants were still able to recombine loxP sites with significant activity. Therefore, they performed fifteen additional rounds of negative selection to evolve Cre against loxP sites, while retaining the ability to recombine loxH site. From the mutant library, a single clone (called Fre22) were

selected as the most specific mutant, with little activity for loxP and moderate activity for loxH. In a separate effort, Santoro and Schultz retargeted Cre to recombine loxM7 sites, which contain three changes from the loxP site (23). They used a focused mutational approach in which only five residues in the Cre protein (identified from the crystal structure) were randomized. They used GFP and RFP as reporter genes, and FACS sorting to separate functional clones. They also directly compared directed evolution using only positive selection with an approach that alternated rounds of positive and negative selection. They performed three rounds of positive selection alone on two libraries and five alternating rounds positive/negative selection on another two libraries. Mutants obtained from just positive selection were generally promiscuous in recombining both loxP and loxM7 sites, while mutants coming from alternating rounds of positive and negative selection exhibited much better specificity. A clone (named CreC2#4) was chosen as the most specific recombinase for loxM7 site. In a more recent approach, Buchholz and colleagues applied directed evolution to evolve Cre to target a sequence in the long-terminal repeat of the HIV-1 strain (named loxLTR) (24). This site has four changes in the left arm region, seven changes on the right, and has a different spacer region. The full length Cre sequence was randomized by error-prone PCR. They used an iterative positive selection strategy, sequentially targeting intermediate half-sites, separate symmetric half-sites, and the asymmetric full site. Surprisingly, although the two half sites have different sequences, they were able to obtain a Cre mutant that could recombine the asymmetric RT site after 126 rounds of directed evolution. The final Cre mutant, called Tre, successfully recombined the loxLTR sites *in vivo,* resulting in excision of the proviral HIV genome in cultured HeLa cells.

Symmetry of Cre recombinase tetramer

After directed evolution, Cre variants with altered DNA specificity that could target distinct RT sites can be obtained. However, the tetrameric formation of Cre complex puts a symmetric requirement on the RT sites they recombine, and the variants could only have limited use if only pseudo-palindromic sites are considered. Cre mutants with specificities towards the two half-sites of asymmetric RT site can be used to recombine these sites, but any combination of the two half-sites in the genome will be recombined, generating off-target recombination. Carmi and colleagues used a mixture of wild-type Cre and a Cre variant (CreC2#4) that recombines loxM7 site to recombine chimeric asymmetric sites containing both loxP and loxM7 half-sites (25). Their study not only demonstrated the promiscuity of mixture, it also showed application of wild type Cre or CreC2#4 alone could recombine these chimeric sites. Thus, by controlling the tetramer assembly, constructing separately mutatable Cre monomers that will function as obligate heterotetramer, but will be inactive in isolation is an attractive strategy for enhancing the specificity for RT site recognition. To address this issue, Baldwin and colleagues used a reciprocal small-to-large substitutions approach to create an alternative binding interface for Cre monomers (26). The redesigned pair of Cre monomers showed increased fidelity to recombine asymmetric sites, but one of the monomers showed reduced but clear activity in isolation. A pair of completely “orthogonal” Cre monomer is desirable for genomic applications.

Recombination mediated cassette exchange (RMCE)

There are several applications for Cre recombinase in genome editing. Recombination between two DNA molecules can drive either an insertion or a translocation event (Figure 2.1A). Recombination within one molecule generates either a deletion or an inversion event, depending on the relative orientations of the loxP sites (Figure 2.1B). One of the most common applications

of Cre recombinase is to generate conditional gene knockouts. By placing the expression of Cre recombinase under the control of promoters that are specific for particular developmental stages or tissues, a gene of interest that is flanked by loxP sites (floxed) can be excised from the genome (Figure 2.1A).

Insertional integration using native loxP sites is difficult because the final product contains two direct-repeat loxP sites, and is therefore susceptible to excision in the presence of the recombinases required to accomplish the integration (Figure 2.1A). One solution, first described in plants, is to utilize pairs of mutant RT sites (27). The crucial property for such a pair of sites is that each harbors a suboptimal half-site sequence on the left and right flanks, respectively. After recombination, one of the resulting RT sites contains both suboptimal half-sites, yielding a nonfunctional RT. Using this approach, the lox66/lox71 pair of mutant RT sites was later used to integrate into mouse embryonic stem cells (28). For the mutant RT sites, 16% of the insertions were targeted in their best conditions, while for wild type lox, the number was < 0.5%. This also implies that 84% of the insertions were random. The next advance came from Sauer and colleagues, which they called the method double-lox replacement, or reciprocal/segmental replacement, which was later named RMCE (29). The DNA fragment of interest (cassette) was flanked by incompatible heterologous RT sites. The idea leveraged off of the observation that lox site variants that differed in their spacer regions could not recombine with each other. By using heterologous lox sites, double crossovers could be accomplished with Cre that exchanged similarly bounded genetic intervals between different DNA molecules (Figure 1.3). They tested the introduction of a cassette harboring a selectable marker. Of the clones that become resistant to the selective drug, 75% had the insertion at the correct genome locus. This compares well to the 16% for the previous single integration effort. They also

analyzed the efficiency when no selectable markers were used. Cytometry indicated that 10% of the cells took up DNA. 96 of cells that took up DNA were analyzed, and two had the correct integration. Another group also reported results using double recombinations, introducing the term RMCE (30). When they tried RMCE without selection, ~1% of the cells that survived transfection had undergone RMCE. They got efficiencies of 4% and 16% for two different loci if only cells that took up DNA (as determined by cotransfection with a fluorescent reporter) were considered. When additional drug selection was performed, they got efficiencies of 10% and 50%.

This RMCE approach has several advantages over traditional Cre-mediated insertion. First, unlike Cre-mediated insertion, the recombination product is not susceptible to excision. Additional exogenous template vectors are added to push the reaction equilibrium towards gene conversion, so the efficiency of genome modification is higher with RMCE. Second, less cytotoxicity can be achieved because less Cre protein is required than for insertional integration. Third, the exchange boundaries are better defined in RMCE, making the integration more precise than Cre-mediated insertion.

Advantages of RMCE over double-stranded break (DSB) stimulated homology-directed repair (HDR)

DSB inducing agents, such as TALEN effectors, CRISPR/Cas, and zinc-finger nucleases (ZFNs) have emerged as attractive tools for genome engineering. The modular DNA binding domain of TALENs and ZFNs makes it easy to change their DNA specificities (31) (32). The Cas9 nuclease can be targeted to any site with a protospacer adjacent motif (PAM) sequence without the need to change the protein sequence (33). Loss of function mutants are generated

when the DSB is repaired by non-homologous end-joing (NHEJ) and gene conversion mutants are generated by HDR, when repair templates are exogenously provided (34).

DSB is efficient in generating local mutations or loss-of-function mutants, but not very efficient to operate on whole gene replacement. DSB stimulated gene conversion efficiency is highest close to the cleavage site. Efficiency for incorporation of altered bases falls to 25% of its peak value at distances of only 100 bp (35). As an example, $IL_2R\gamma$ is a target locus for gene therapy that causes X linked SCID (http://genome.nhgri.nih.gov/scid/index.shtml) (36). The cDNA is over 1 kb, and the gene itself is over 4 kb. Efficient targeting of all mutations for this locus would require dozens of double strand inducers. Each must be characterized and subjected to rigorous testing before application as a therapy. In contrast, 4 kbp is well within the efficient distance scale for Cre recombinase (30, 37, 38). Thus, a single combination of locus bracketing recombinases can be studied thoroughly and applied to all of the mutations, justifying the greater effort required to engineer a novel Cre recombinase compared to the inherently modular ZFNs, TALENs and CRESPER/Cas systems.

There are several other advantages of RMCE over DSB-induced HDR for genome modification. First, RMCE does not depend on the host cell's relative preference for HDR over NHEJ. Second, illegitimate integrations generated by undesired RMCE modifications can be identified by rapid inverse PCR techniques, while NHEJ healed DSBs can be difficult to identify. Finally, different technologies for genome modification (ZFNs, TALENs and RMCE) have been developed to different stages of maturity. Until all of these technologies are fully characterized and the relative merits determined, it is important to pursue all reasonable approaches.

Focus of the dissertation

The ultimate goal of this research is to engineer Cre recombinase to operate upon native genomic sequences. Chapter 2 describes a method to recombine asymmetric sites. Chapter 3 describes a method called 'Rec-Seq' to characterize DNA specificities of recombinases, and also discusses using bioinformatics to accelerate retargeting of recombinases.

The quaternary structure of the Cre complex creates a challenge for retargeting genomic RT sites. The four-fold symmetry of the synaptic complex imposes a pseudo-palindromic symmetry upon the RT site. In chapter 2, I redesigned the monomer-monomer interface for Cre recombinase using a combination of computational and rational design. I then assembled the complex to form an obligate heterotetramer. When combined with mutants with distinct DNA specificities, the heterotetramer recombined asymmetric sites with greater fidelity. The 'orthogonal' pair of Cre monomers were validated in both mammalian cell cultures and tissues.

As many recombinases originate from bacteriophages, the preferred RT sequence of some recombinases can be identified in the bacterial genome close to the encoding gene (39). However, just identifying the preferred RT sequence is not enough to fully characterize the DNA specificity of a recombinase. It is important to know the relative binding preferences for other sequences existing in the genome. In recent studies, the DNA binding specificities of transcription factors (TFs) have been determined *via* high-throughput sequencing of the bound subset of randomized DNA binding sites (40). In contrast to TFs, tyrosine recombinases exhibit their specificity in an enzymatic reaction in which two RT sites are cleaved and religated, making it a more challenging problem. In chapter 3, I have established a method called 'Rec-Seq' that utilizes recombination rather than binding as an enrichment method to characterize the DNA

specificity of recombinases. By analyzing the sequence alignment of Cre homologs, I have also managed to change the DNA specificity of Cre recombinase, using Rec-Seq as the confirmatory technology for the altered specificity.

Finally, chapter 4 concludes this dissertation and proposes future research in engineering recombinases.

References

1. Lutz,S. (2010) Beyond directed evolution--semi-rational protein engineering and design. *Curr Opin Biotechnol* **21**, 734-743.

2. Leung,D.W., Chen,E. and Goeddel,D.V. (1989) A method for random mutagenesis of a defined DNA segment using a modified polymerase chain reaction **1**, 11-15.

3. Cadwell,R.C. and Joyce,G.F. (1992) Randomization of genes by PCR mutagenesis. *PCR Methods Appl* **2**, 28-33.

4. Vanhercke,T., Ampe,C., Tirry,L. and Denolf,P. (2005) Reducing mutational bias in random protein libraries. *Anal Biochem* **339**, 9-14.

5. Wells,J.A., Vasser,M. and Powers,D.B. (1985) Cassette mutagenesis: an efficient method for generation of multiple mutations at defined sites. *Gene* **34**, 315-323.

6. Stemmer,W.P. (1994) Rapid evolution of a protein in vitro by DNA shuffling. *Nature* **370**, 389-391.

7. Müller,K.M., Stebel,S.C., Knall,S., Zipf,G., Bernauer,H.S. and Arndt,K.M. (2005) Nucleotide exchange and excision technology (NExT) DNA shuffling: a robust method for DNA fragmentation and directed evolution. *Nucleic Acids Res* **33**, e117.

8. Jespers,L.S., Roberts,A., Mahler,S.M., Winter,G. and Hoogenboom,H.R. (1994) Guiding the selection of human antibodies from phage display repertoires to a single epitope of an antigen. *Biotechnology (N Y)* **12**, 899-903.

9. Clackson,T., Hoogenboom,H.R., Griffiths,A.D. and Winter,G. (1991) Making antibody fragments using phage display libraries. *Nature* **352**, 624-628.

10. Gaj,T., Mercer,A.C., Gersbach,C.A., Gordley,R.M. and Barbas,C.F. (2011) Structure-guided reprogramming of serine recombinase DNA sequence specificity. *Proc Natl Acad Sci U S A* **108**, 498-503.

11. Hanes,J. and Plückthun,A. (1997) In vitro selection and evolution of functional proteins by using ribosome display. *Proc Natl Acad Sci U S A* **94**, 4937-4942.

12. Wilson,D.S., Keefe,A.D. and Szostak,J.W. (2001) The use of mRNA display to select high-affinity protein-binding peptides. *Proc Natl Acad Sci U S A* **98**, 3750-3755.

13. Takeuchi,R., Choi,M. and Stoddard,B.L. (2014) Redesign of extensive protein-DNA interfaces of meganucleases using iterative cycles of in vitro compartmentalization. *Proc Natl Acad Sci U S A* **111**, 4061-4066.

14. Ghadessy,F.J., Ong,J.L. and Holliger,P. (2001) Directed evolution of polymerase function by compartmentalized self-replication. *Proc Natl Acad Sci U S A* **98**, 4552-4557.

15. Ellefson,J.W., Meyer,A.J., Hughes,R.A., Cannon,J.R., Brodbelt,J.S. and Ellington,A.D. (2014) Directed evolution of genetic parts and circuits by compartmentalized partnered replication. *Nat Biotechnol* **32**, 97-101.

16. Landy,A. (1989) Dynamic, structural, and regulatory aspects of lambda site-specific recombination. *Annu Rev Biochem* **58**, 913-949.

17. Hoess,R.H. and Abremski,K. (1984) Interaction of the bacteriophage P1 recombinase Cre with the recombining site loxP. *Proc Natl Acad Sci U S A* **81**, 1026-1029.

18. Lee,G. and Saito,I. (1998) Role of nucleotide sequences of loxP spacer region in Cre-mediated recombination. *Gene* **216**, 55-65.

19. Guo,F., Gopaul,D.N. and van Duyne,G.D. (1997) Structure of Cre recombinase complexed with DNA in a site-specific recombination synapse. *Nature* **389**, 40-46.

20. Missirlis,P.I., Smailus,D.E. and Holt,R.A. (2006) A high-throughput screen identifying sequence and promiscuity characteristics of the loxP spacer region in Cre-mediated recombination. *BMC Genomics* **7**, 73.

21. Sheren,J., Langer,S.J. and Leinwand,L.A. (2007) A randomized library approach to identifying functional lox site domains for the Cre recombinase. *Nucleic Acids Res* **35**, 5464-5473.

22. Buchholz,F. and Stewart,A.F. (2001) Alteration of Cre recombinase site specificity by substrate-linked protein evolution *Nature biotechnology* **19**, 1047-1052.

23. Santoro,S.W. and Schultz,P.G. (2002) Directed evolution of the site specificity of Cre recombinase *Proc Natl Acad Sci USA* **99**, 4185-4190.

24. Sarkar,I., Hauber,I., Hauber,J. and Buchholz,F. (2007) HIV-1 proviral DNA excision using an evolved recombinase *Science* **316**, 1912.

25. Saraf-Levy,T., Santoro,S.W., Volpin,H., Kushnirsky,T., Eyal,Y., Schultz,P.G., Gidoni,D. and Carmi,N. (2006) Site-specific recombination of asymmetric lox sites mediated by a heterotetrameric Cre recombinase complex *Bioorganic & medicinal chemistry* **14**, 3081-3089.

26. Gelato,K.A., Martin,S.S., Liu,P.H., Saunders,A.A. and Baldwin,E.P. (2008) Spatially directed assembly of a heterotetrameric Cre-Lox synapse restricts recombination specificity *J Mol Biol* **378**, 653-665.

27. Albert,H., Dale,E.C., Lee,E. and Ow,D.W. (1995) Site-specific integration of DNA into wild-type and mutant lox sites placed in the plant genome *The Plant Journal* **7**, 649-659.

28. Araki,K., Araki,M. and Yamamura,K. (1997) Targeted integration of DNA using mutant lox sites in embryonic stem cells *Nucleic Acids Res.* **25**, 868.

29. Sadowski,P.D. (1995) The Flp recombinase of the 2-microns plasmid of Saccharomyces cerevisiae. *Prog Nucleic Acid Res Mol Biol* **51**, 53-91.

30. Feng,Y.Q., Seibler,J., Alami,R., Eisen,A., Westerman,K.A., Leboulch,P., Fiering,S. and Bouhassira,E.E. (1999) Site-specific chromosomal integration in mammalian cells: highly efficient CRE recombinase-mediated cassette exchange1 *Journal of molecular biology* **292**, 779-785.

31. Hockemeyer,D., Wang,H., Kiani,S., Lai,C.S., Gao,Q., Cassady,J.P., Cost,G.J., Zhang,L., Santiago,Y., Miller,J.C., Zeitler,B., Cherone,J.M., Meng,X., Hinkley,S.J., Rebar,E.J., Gregory,P.D., Urnov,F.D. and Jaenisch,R. (2011) Genetic engineering of human pluripotent cells using TALE nucleases. *Nat Biotechnol* **29**, 731-734.

32. Miller,J.C., Holmes,M.C., Wang,J., Guschin,D.Y., Lee,Y.L., Rupniewski,I., Beausejour,C.M., Waite,A.J., Wang,N.S., Kim,K.A., Gregory,P.D., Pabo,C.O. and Rebar,E.J. (2007) An improved zinc-finger nuclease architecture for highly specific genome editing *Nat Biotechnol* **25**, 778-785.

33. Cong,L., Ran,F.A., Cox,D., Lin,S., Barretto,R., Habib,N., Hsu,P.D., Wu,X., Jiang,W., Marraffini,L.A. and Zhang,F. (2013) Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819-823.

34. Hsu,P.D., Lander,E.S. and Zhang,F. (2014) Development and applications of CRISPR-Cas9 for genome engineering. *Cell* **157**, 1262-1278.

35. Elliott,B., Richardson,C., Winderbaum,J., Nickoloff,J.A. and Jasin,M. (1998) Gene conversion tracts from double-strand break repair in mammalian cells. *Mol Cell Biol* **18**, 93-101.

36. Urnov,F.D., Miller,J.C., Lee,Y.L., Beausejour,C.M., Rock,J.M., Augustus,S., Jamieson,A.C., Porteus,M.H., Gregory,P.D. and Holmes,M.C. (2005) Highly efficient endogenous human gene correction using designed zinc-finger nucleases. *Nature* **435**, 646-651.

37. Bethke,B. and Sauer,B. (1997) Segmental genomic replacement by Cre-mediated recombination: genotoxic stress activation of the p53 promoter in single-copy transformants. *Nucleic Acids Res* **25**, 2828-2834.

38. Lauth,M., Moerl,K., Barski,J.J. and Meyer,M. (2000) Characterization of Cre-mediated cassette exchange after plasmid microinjection in fertilized mouse oocytes. *Genesis* **27**, 153-158.

39. Surendranath,V., Chusainow,J., Hauber,J., Buchholz,F. and Habermann,B.H. (2010) SeLOX--a locus of recombination site search tool for the detection and directed evolution of site-specific recombination systems. *Nucleic Acids Res* **38**, W293-8.

40. Liu,J. and Stormo,G.D. (2005) Combining SELEX with quantitative assays to rapidly obtain accurate models of protein-DNA interactions. *Nucleic Acids Res* **33**, e141.

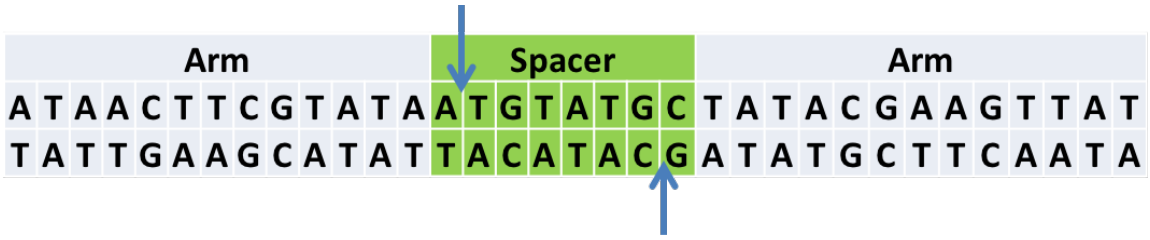


Figure 1.1 34 bp loxP DNA sequence

The pseudo-palindromic 34 bp loxP DNA sequence contains two symmetrical 13 bp arm (colored in grey) region and an asymmetrical 8 bp spacer region (colored in green). The blue arrows indicate the cleavage point introduced by Cre recombinase.

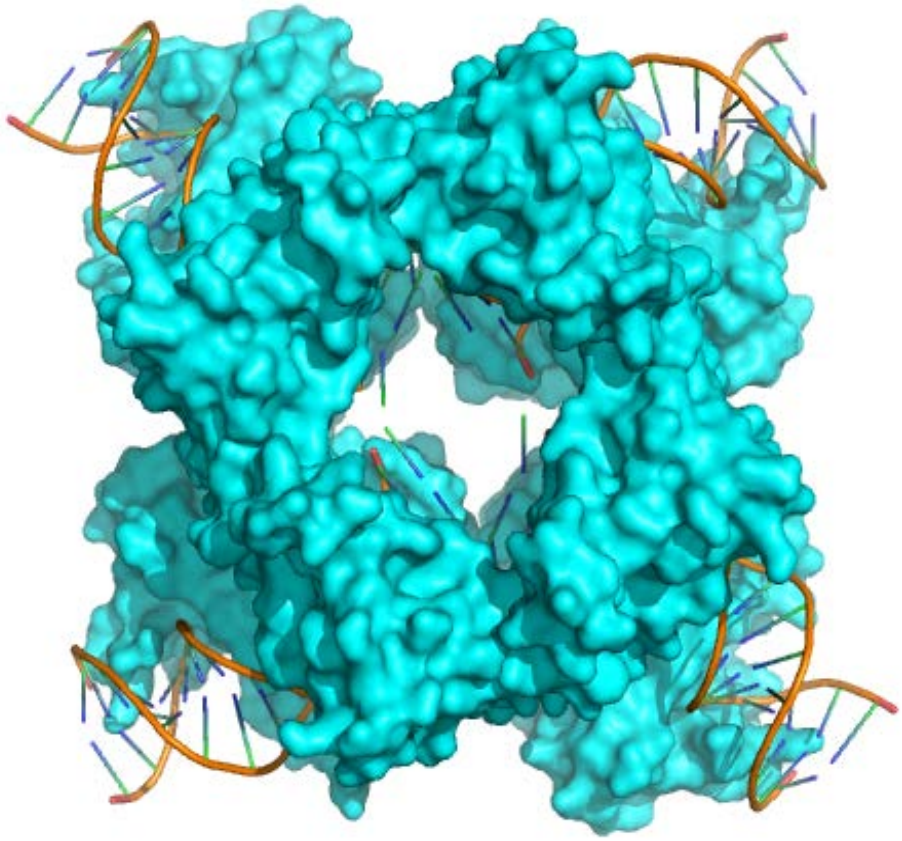


Figure 1.2 Synaptic complex of Cre homotetramer

A structural view of the Cre tetramer in complex of two loxP DNA sequence taken from PDB structure 1KBU. Cre monomers are colored in cyan and DNA is colored in orange.

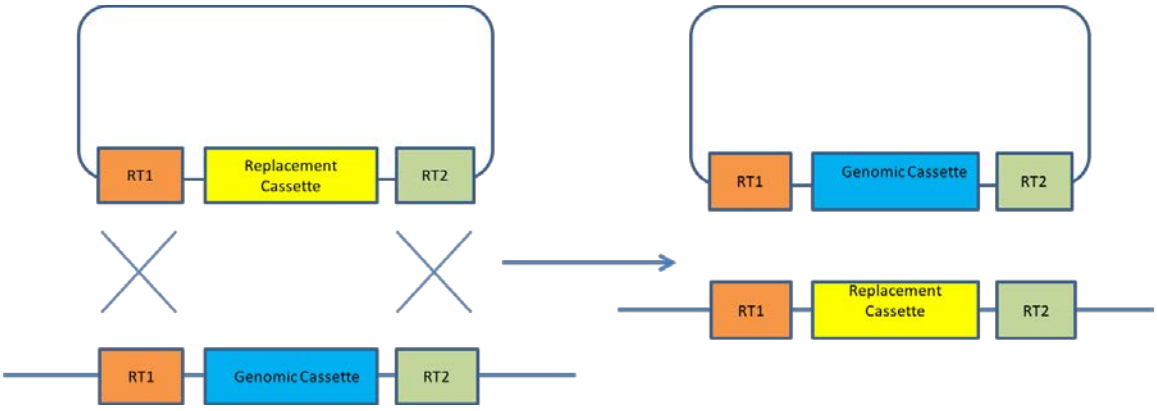


Figure 1.3 Schemes of RMCE

RMCE utilizes two incompatible RT sites (colored in orange and green) to define the boundaries of the cassette of interest in the genome (colored in blue) and replace it with the exogenously provided template (colored in yellow). Unlike traditional Cre induced integration, the resulting exchange product is not susceptible to excision.

Chapter 2

Redesign of the Monomer-monomer Interface of Cre Recombinase Yields an Obligate Heterotetrameric Complex

This Chapter has been published:

Zhang, C, CA Myers, Z Qi, RD Mitra, JC Corbo, and JJ Havranek. 2015. Redesign of the monomer-monomer interface of Cre recombinase yields an obligate heterotetrameric complex. *Nucleic Acids Res* 43, no. 18: 9076-9085.

The author contributions are the following: James Havranek and Chi Zhang did the protein design. Chi Zhang cloned, purified and tested the mutants *in vitro*. Connie Myers and Joseph Corbo designed and performed the mouse retinal explant assay. Zongtai Qi and Robi Mitra designed and performed the mouse ES cell culture assay. The manuscript was written and revised primarily by Chi Zhang and James Havranek.

ABSTRACT

Cre recombinase catalyzes the cleavage and religation of DNA at loxP sites. The enzyme is a homotetramer in its functional state, and the symmetry of the protein complex enforces a pseudo-palindromic symmetry upon the loxP sequence. The Cre-lox system is a powerful tool for many researchers, particularly those working in mouse genetics. However, broader application of the system is limited by the fixed sequence preferences of Cre, which are determined by both the direct DNA contacts and the homotetrameric arrangement of the Cre monomers. As a first step towards achieving recombination at arbitrary asymmetric target sites, we have broken the symmetry of the Cre tetramer assembly. Using a combination of computational and rational protein design, we have engineered an alternative interface between Cre monomers that is functional yet incompatible with the wild-type interface. Wild-type and engineered interface halves can be mixed to create two distinct Cre mutants, neither of which are functional in isolation, but which can form an active heterotetramer when combined. When these distinct mutants possess different DNA specificities, control over complex assembly directly discourages recombination at unwanted half-site combinations, enhancing the specificity of asymmetric site recombination. The engineered Cre mutants exhibit this assembly pattern in a variety of contexts, including mammalian cells. The availability of obligate heterotetrameric mutants opens the door to the controlled assembly of Cre monomers whose DNA specificities may be altered independently.

INTRODUCTION

Cre recombinase forms a tetrameric complex that splices DNA molecules containing the 34-bp recombination target (RT) site loxP (1), recombining two DNA molecules in *trans* to accomplish an insertion or translocation event, or in *cis* to achieve either gene excision or

inversion, depending on the relative orientation of the loxP sites (Figure 2.1). Cre recombinase has been used to generate conditional gene knockouts, where a gene of interest is flanked by loxP sites ('floxed') (2). Expression of Cre recombinase under the control of promoters that are specific for particular tissues or developmental stages abrogates gene function by physical excision from the genome. The utility of this system depends on the functional autonomy of Cre recombinase: the enzyme requires no other factors to splice DNA, and is capable of modifying genomes in non-replicating cells, where the efficacy of gene conversion via double-strand break (DSB) induced homologous recombination is expected to be low (3, 4).

Another application for Cre recombinase is recombination-mediated cassette exchange (RMCE) (5), also known as double-reciprocal crossover (6, 7) or double-lox replacement (8, 9). In this approach, (reviewed in ref. (10)) recombination between DNA molecules that share two neighboring heterologous RT sites accomplishes the exchange of the bounded genetic interval (the cassette) between the sites (Figure 2.1C). This has been demonstrated using both Flp and Cre recombinase with heterologous RT variants (5, 8), as well as simultaneously with Cre and the Flp recombinases (11). Although RMCE has so far only been demonstrated with wild-type recombinase proteins and RT sites, the approach has many attractive features as a tool for genome engineering. First, it has a higher efficiency for gene conversion than does Cre-mediated insertion, as it does not require survival of insertional events that are susceptible to reversal by excision (8). Second, the cassettes that are exchanged are precisely demarcated, yielding truly 'scarless' genomic surgery. Third, the process requires less Cre protein than recombinational insertion, resulting in less cytotoxicity(8). Finally, the autonomy of Cre as a recombinase suggests that RMCE could prove to be effective in terminally differentiated cells, in contrast to strategies for gene conversion that rely upon homology directed repair.

One impediment to broader use of Cre recombinase is the inflexibility of the binding site specificity. In contrast to DNA binding proteins whose specificity derives from the assembly of small recognition modules such as zinc finger or TAL effector domains, Cre recombinase interacts with DNA through large interfaces that defy a modular decomposition. Nevertheless, altered RT specificities have been elicited in mutant Cre recombinases using directed evolution (12-14).

The quaternary structure of the Cre complex creates a second challenge for engineering novel RT specificities. The four-fold symmetry in the functional protein complex imposes a pseudo-palindromic symmetry upon the RT site. The loxP site consists of two 13 bp palindromic half-sites separated by an asymmetric 8 bp spacer that gives loxP its direction. The utility of targeting Cre mutants to altered RT sites is severely compromised if only pseudo-palindromic sites may be considered. This limitation has been addressed by using directed evolution to generate mutant homotetrameric complexes that can operate on asymmetric sites (14, 15). However, requiring a single Cre mutant to operate on two different half-sites is likely to result in promiscuous enzymes. Separate Cre mutants with specificities towards the two half-sites of an asymmetric RT site may be able to recombine these sites, but the lack of control over assembly of the complex allows for any combination of these half-sites as potential sites for recombination (16). Some of these combinations will be undesired, generating off-target recombination events and exacerbating the cytotoxicity of Cre recombinase (17).

A similar technical challenge has been overcome in the design of zinc finger nucleases (ZFNs). ZFNs are DSB agents that achieve their sequence specificity by concatenating multiple zinc finger modules, each of which recognizes 3-4 base pairs. The cleavage activity is provided by the dimeric *Fok*I nuclease. *Fok*I monomers are genetically fused to zinc finger arrays, and two

such constructs that converge upon a DNA site reconstitute a functional nuclease dimer, inducing a DSB. The development of obligate heterodimer *Fok*I mutants has increased target specificity and reduced cytotoxicity in this system (18). Under this approach, the ZFNs that co-locate on desired cleavage sites must contribute two distinct *Fok*I monomers; misassembly of two copies of the same ZFN at an off-target site cannot reconstitute a functional nuclease. Constructing a functional Cre complex from distinguishable and separately mutatable monomers is an attractive strategy for enhancing the specificity of RT site recognition. An earlier effort to generate heterotetramer Cre mutants succeeded in forming a novel functional interface, but one of the two mutants retained significant activity as a homotetramer (19).

In this manuscript we describe the engineering of Cre mutants that are inactive in isolation, but are functional as a (ABAB) heterotetramer when both mutants are present. We use a combination of computational and rational design to select mutations that are predicted to form a novel interface between Cre monomers that is functional, but whose halves are incompatible with their wild-type counterparts. We show that the negative engineering goal (incompatibility with wild-type) is more difficult to achieve than the positive goal (full functionality), requiring three iterations of mutation. The obligate heterotetrameric assembly of the pair of mutants is demonstrated *in vitro* and *in vivo*, notably in mammalian cells. We hope that the availability of these mutants enables the specific and reliable targeting of Cre to asymmetric RT sites.

MATERIALS AND METHODS

Computational modeling and design

We selected the 2.2 Å crystal structure of a Cre-loxP Holliday junction as a template for computational design (PDB code: 1KBU (20)). The protein design capabilities of Rosetta3 (21) were used to select amino acids to form an alternative interface between Cre monomers. Amino acid positions 25, 29, 32, 33, 35 from chain A and 69, 72, 76, 119, 123 from chain B were chosen by eye for redesign because they form multiple interactions across the largest region of contact between monomers, but do not participate in the protein-DNA interface (Figure 2.2). At each of these positions, the calculation permitted mutation to a subset of amino acids including positive, negative or non-polar amino acids (AVMLDERK). The redesign calculation used the standard RosettaDesign fixed backbone algorithm. Sidechain rotamers were built using a backbone-dependent rotamer library (22). Extra rotamers sampling additional values for the χ_1 and χ_2 side chain torsion angles were included in the design calculation (command line options –ex1, -ex2 in Rosetta). A scoring function using a softened form of the Lennard-Jones potential (soft_rep_design) was used (23) to evaluate the interactions between the rotamers and the fixed backbone, and between rotamers at different positions. The combinatorial search through conformational space was accomplished using a Monte Carlo method with Metropolis acceptance criteria.

Gene construction and protein expression

A gene encoding wild-type Cre recombinase with an N-terminal Met-His_7 tag was constructed from 100bp overlapping oligonucleotides ordered from Integrated DNA technologies

(IDT) and cloned into the pET42a vector. Cre mutants were generated by site-directed mutagenesis. Proteins were expressed in BL21(DE3) star cells at 30°C using the autoinduction protocol of Studier (24). Details of the purification strategy are given in the Supplemental Materials.

In vitro recombinase activity assay

Two direct loxP repeats or other variants of loxP/M7 sites separated by a ~0.5 kb spacer were cloned between the XbaI and SphI sites of the pBAD33 plasmid. The 0.7 kb DNA substrate for *in vitro* recombination assays was generated by PCR amplification with pBAD-forward and pBAD-reverse primers. 1µg of the DNA substrate was incubated with 1 µM Cre in 10 µL of 50 mM Tris-HCl, pH 7.8, 50 mM NaCl and 10 mM $MgCl_2$ for 12 hours at 37° C. Total amount of protein used is the same across all *in vitro* assays. Reactions were stopped by incubation at 98° C for 20 minutes. Reactions were analyzed on 2% agarose gels and visualized by staining with GelGreen nucleic acid stain (Biotium).

Cell culture and transfection

The plasmid pGL4.23 (GenBank accession number: DQ904455) containing a multiple cloning site (MCS) for insertion of a response element of interest upstream of a minimal promoter and a gene encoding luc2 was purchased from Promega. The original minimal promoter in pGL4.23 was replaced with the haemoglobin beta (HBB) gene minimal promoter 144bp upstream of the HBB transcription start site. The HBB minimal promoter has only the basic components for transcription (i.e. TATA box and GC box) and was amplified by PCR from mouse genomic DNA. The coding sequence of luc2 in PGL4.23 gene was replaced with different

mutants of Cre recombinase using Gibson assembly. The enhancer candidates (CMV and SP1 enhancers) were then cloned into the MCS upstream of the minimal promoter. The engineered plasmids were isolated using standard molecular biology techniques and were confirmed by Sanger sequencing.

Ai14 mouse embryonic stem (ES) cells were engineered by targeted insertion of a construct containing the CAG promoter, followed by a floxed stop cassette-controlled red fluorescent marker gene (tdTomato) (Figure 2.4A) (25). The Ai14 mouse ES cells were cultured in complete media consisting of Dulbecco's modified eagle media (DMEM; Gibco) supplemented with 10% new born calf serum, 10% fetal bovine serum (FBS; Gibco), and 0.3 mM of each of the following nucleosides: adenosine, guanosine, cytosine, thymidine, and uridine (Sigma-Aldrich). To maintain their undifferentiated state, the cells were also cultured in flasks coated with a 0.1% gelatin solution (Sigma-Aldrich) in the presence of 1000 U/mL leukemia inhibitory factor (LIF; Chemicon) and 20 mM β-mercaptoethanol (BME; Invitrogen).

Plasmids used for transfection of cells were prepared using EndoFree Plasmid Maxi Kits (Qiagen). About 2 x 10^5 Ai14 ESCs were plated in one well of a six-well plate one day prior to transfection with complete medium plus LIF in feeder free conditions. The cells were then transfected at 70% confluence by Lipofectamine 2000 (Invitrogen). For each transfection experiment, a total of 1 μg of plasmid DNA and 8 μL of Lipofectamine 2000 reagent were mixed following the manufacturer’s protocol, and incubated at room temperature for 5 minutes before adding to the culture medium. The medium was replaced with fresh ESC medium plus LIF the following day and cells were cultured for another day before harvested for fluorescence activated cell sorting (FACS).

Flow cytometry

Upon reaching approximately 100% confluence, the cells were trypsinized from the plate and were suspended in Hank's Balanced Salt Solution (HBSS) supplemented with 2 mM EDTA, washed once with PBS, and resuspended in 500 µl PBS. Cellular fluorescence was analyzed on an iCyt Reflection HAPS2 cell sorter at the Washington University Siteman Flow Cytometry Core. Cells were treated with propidium iodide (2 µg/ml) prior to sorting to counter-select dead cells. The gate was set relative to cells transfected with plasmids lacking red fluorescent protein genes (negative controls) to eliminate nonspecific background reporting. A minimum of 7000 total live single cells was analyzed from each FACS and post-sort analysis was performed with FlowJo software to obtain the percentage of RFP positive cells.

Recombinase assay in mouse retinal explants

Electroporations and explant cultures were performed as previously described (26). Retinal explants were electroporated in a chamber containing 0.5 µg/mL each of supercoiled DNA encoding a gene for *Nrl*-eGFP as a control for electroporation efficiency, a reporter construct for Cre activity comprised of DsRed preceded by a floxed stop codon, and a gene encoding either wild-type or engineered Cre under control of the *Nrl* promoter (27). Three replicates were performed for each electroporation. Quantification of fluorescence in retinal explants was accomplished using the ImageJ program (http://rsbweb.nih.gov/ij/) using a previously described protocol (28).

RESULTS

Computational redesign of a non-native but functional protein-protein interface between Cre recombinase monomers

We desired an engineered protein interface between Cre recombinase monomers that could form a functional complex, yet be incompatible with the wild-type interface. The two sides of such an interface could then be mixed with the other sides of the wild-type interface to yield two distinct Cre mutants. These mutants, by virtue of possessing incompatible interfaces, could not form functional homotetrameric complexes, but could be combined to form a functional heterotetramer (Figure 2.1D). We selected the 2.2 Å crystal structure of a Cre-loxP Holliday junction (PDB code: 1KBU) (20) as our template for computational design. We then selected the largest monomer-monomer interface patch for redesign, focusing on residues that did not participate in any contacts with DNA (cyan oval on left side of Figure 2.2A). We used the Rosetta molecular modeling program to redesign five residues on each side of the interface (see Methods), although in some cases (two of ten) the wild-type amino acid was retained by the design calculation. The set of mutations that constitute the redesigned interface are the combined A1 and B1 mutations given in Table 2.1. When evaluated with the Rosetta full atom scoring function, the redesigned interface is 2.8 Rosetta units (RU) worse than the wild-type interface. Although the redesigned sequence was selected without regard for destabilizing mixed engineered/wild-type interfaces, models for the Cre-A1 and Cre-B1 homotetramer interfaces were predicted to be 9.8 and 20.49 RU worse than wild-type, respectively. Inspection of individual terms shows that the Cre-A1 and Cre-B1 models contain interface residues with side chains in strained conformations, presumably due to lack of favorable interactions.

We tested the redesigned interface by generating pairs of Cre mutants such that each mutant possesses one side of the interface, with the other side fixed as wild-type. We assayed members of each pair for recombinase activity *in vitro* both individually and in combination (Figure 2.3). While the combined pair of redesigned mutants was active (Cre-A1+Cre-B1 in Figure 2.3B; see Table 2.1 for mutations), one of the mutants (Cre-A1) was active individually, indicating that this hybrid redesign/wild-type interface was functionally compatible, in violation of our negative engineering goal (Figure 2.1D).

Iterative rounds of rational design enhance the formation of (ABAB) complexes

We attempted to find another region of contact between monomers in the Cre complex that we could mutate in an attempt to further destabilize homotetrameric Cre-A1 complexes. Visual inspection of the Cre crystal structure revealed a salt bridge between Glu308 and Arg337 (Figure 2.2C) that we hypothesized could be inverted to obtain additional specificity for the heterotetramic complex (Figure 2.2E). We therefore further mutated Cre-A1 (adding R337E) to yield Cre-A2, and mutated Cre-B1 (adding E308R) to yield Cre-B2. Thus, homotetrameric complexes of Cre-A2 would place two glutamate residues at 308 and 337 in close proximity, and Cre-B2 would likewise pair two arginine residues, yielding unfavorable electrostatic repulsion in either case. Our *in vitro* recombinase assay showed that the Cre-A2+Cre-B2 combination exhibited robust recombinase activity. However, while its activity is reduced relative to Cre-A1, the Cre-A2 monomer was still capable of forming a functional homotetrameric complex (Figure 2.3B).

We selected a polar interaction between monomers as the final site for mutagenesis. We hypothesized that a replacement interaction consisting of hydrophobic residues would be

incompatible with the pre-existing polar interaction. Structural modeling suggested that the mutation E123L and Q35L could create a tight packing interaction between leucine residues across the monomer-monomer interface, but that interfaces combining a polar residue from the wild-type interface with either leucine from the engineered interface would be energetically unfavorable.

In vitro assays indicated that the E123L mutation did indeed penalize formation of functional homotetrameric complexes, but that the Q35L mutation unexpectedly facilitated homotetramer formation in the previously inactive B2 mutant (data not shown). Consequently, we applied the E123L mutation to Cre-A2 to create Cre-A3. This mutation successfully disrupted formation of Cre-A2 homotetramers while preserving activity in the Cre-A3+Cre-B2 heterotetramer (Figure 2.3B). The improvement in specificity appears to come from selective destabilization of the Cre-A3 homotetramer with limited destabilization of the heterotetramer. The reactivation of Cre-B1 by the Q35L mutation is especially puzzling, as the distance between the alpha carbon of position 35 and that of the nearest mutated position from Cre-B1 (position 76) is roughly 13 Å, suggesting that this mutation cannot directly alleviate the interface deficiencies introduced by the Cre-B1 mutations. This mutation may allow for a subtle rearrangement of the charged side chains at the interface that yields a functional complex through a mechanism that requires introduction of modes of relaxation that are not captured by our model.

To test whether our round 1 mutations are essential to enforce heterotetramer formation, we generated Cre mutants with only round 2 and round 3 mutations. The salt-bridge swap from round 2 alone yields two Cre mutants with reduced but clear activity (data not shown). We combined round 2 and round 3 mutations to create Cre-E123L-E308R and Cre-E123L-R337E. *In vitro* assays indicated that these mutants do not form an obligate heterotetrameric pair

(Supplemental Figure 2.1A). We conclude that the combined effects of mutations from all three rounds are necessary to achieve our design goal.

Heterotetrameric mutations can be combined with DNA specificity altering mutations to enhance target site specificity

We hypothesized that the ability to control the assembly of functional Cre complexes would lead to higher fidelity recognition of asymmetric RT sites if used in combination with recombinases with different DNA specificities. Directed evolution has already been exploited to generate mutants of Cre recombinase that can utilize altered RT sites. A mutant (termed Cre-C2#4) with five amino acid mutations relative to wild-type has been shown to recombine an alternate RT site termed loxM7 (13). The monomer-monomer interface mutations from Cre-A3 and Cre-B2 were applied separately to the Cre-C2#4 mutant. If the proteins with different DNA specificities exhibit the expected ABAB heterotetrameric pattern assembly, they should only recombine DNA half-sites with a specific spatial arrangement, yielding enhanced target specificity.

To this end, we designed DNA substrates harboring direct repeats of six different loxP/M7 hybrid RT sites as a rigorous test of specificity (Figure 2.3C). We expect that a mixture of wild-type Cre and Cre-C2#4 (both of which lack our obligate heterotetrameric mutations) could recombine all of the six RT sites, as the individual monomers can combine in any manner dictated by the sequences of the RT half-sites. In contrast, a combination of the designed Cre-A3-C2#4 and Cre-B2 recombinases, or similarly the Cre-A3 and Cre-B2-C2#4 recombinases, would specifically recombine the LM-LM site, but not the other five RT sites (Figure 2.3C). This

would imply that heterotetrameric Cre mutants will have less off-target activity when used for genome editing.

In vitro assays confirmed that the heterotetrameric Cre is more specific in recombining different arrangement of loxP/M7 sites (Figure 2.3D). Cre-C2#4 is slightly promiscuous, and can recombine loxP sites when incubated with DNA substrate for a long period of time ((13), Supplemental Figure 2.2B). The observed partial activity of the two designed pairs on LL-ML site (lane 2 in the middle and right gels of Figure 2.3D) is most likely the result of promiscuity of Cre-C2#4's DNA specificity. It is also interesting to note that, because the four Cre monomers work cooperatively to recombine the DNA target, wild-type Cre and Cre-C2#4 homotetramers recombined most of the loxP/M7 hybrid sites on their own ((29), Supplemental Figure 2.1B). The specificity shown here by the two designed pairs provides strong evidence that our mutant recombinases indeed form an ABAB heterotetrameric complex.

Obligate heterotetramer formation is preserved in mammalian cells

We envision RMCE in mammalian cells as the target application for our heterotetramer-forming Cre mutants. We employed two reporter systems to determine whether the engineered proteins satisfy our design goals in mammalian cells. First, we assayed the recombinase activity of the Cre mutants in a mouse ES cell reporter line by flow cytometry. We inserted a gene for the tandem dimer tomato (tdTomato) fluorescent protein downstream of a floxed stop codon at the *rosa26* locus (Figure 2.4A). Constructs encoding genes for the Cre mutants driven by the haemoglobin beta (HBB) minimal gene promoter, either alone or in combination with one of two enhancers (see Methods), were transfected into the reporter line,

and the cells expressing tdTomato were quantified by flow cytometry (Figure 2.4B, Supplemental Table 2.1).

Similar to the *in vitro* results, we observed the Cre-A2+Cre-B2 combination to be functional, while the Cre-A2 mutant retains significant activity as a homotetramer. Combining Cre-A3 with Cre-B2 yielded a suitable obligate heterotetrameric pair, retaining roughly 40% of wild type Cre activity. Neither the Cre-A3 nor the Cre-B2 mutants exhibited appreciable activity alone.

We also evaluated the activity of the Cre mutants in mouse retinal explants. Dissected newborn mouse retinas were electroporated with a construct expressing GFP under the control of the rod photoreceptor-specific *Nrl* promoter (27) (as a loading control), Cre mutants under the control of the same *Nrl* promoter, and a floxed tdTomato reporter construct. After eight days in explant culture, the retinas were harvested, and imaged. The appearance of the flat-mounted retinas under epifluorescent illumination is shown in Figure 2.5. GFP fluorescence indicates areas of successful electroporation, and red fluorescence reports recombinase activity. Wild-type Cre shows robust activity, with all green cells also exhibiting red fluorescence (Figure 2.5A). The Cre-A3 and Cre-B2 mutants alone show very little activity (Figure 2.5B,C), while combining the two restores robust activity (Figure 2.5D). Quantification confirms that Cre-A3 and Cre-B2 form an obligate heterotetrameric pair in photoreceptor cells (Figure 2.5E).

DISCUSSION

We sought to engineer a pair of mutants of Cre recombinase that form an obligate ABAB heterotetrameric complex. The Cre-A3 and Cre-B2 mutants are the result of an iterative process of computational and rational protein engineering. We have shown that the two mutants are

inactive in isolation, but are functional when combined. Furthermore, we have shown that when additional mutations are used to confer an altered DNA specificity upon either one of the mutants, the arrangements of half-sites that are recombined are consistent with the formation of an ABAB complex. Although our attempts to confirm the composition of the functional complex directly via crystallography were unsuccessful, our data are strongly suggestive that we have succeeded in our goal.

Engineering a novel interface for Cre recombinase monomers that is incompatible with the wild-type interface involves two distinct requirements, one positive and one negative. The positive requirement is that the novel interface must give rise to a functional tetrameric complex. The negative requirement is that any combination of wild-type and engineered monomer surfaces must be functionally incompatible. We found that the negative engineering goal was more difficult to achieve. We were able to generate a novel functional interface using straightforward computational protein design. However, the mutations on one side of the interface (the Cre-A1 mutations) were still compatible with the wild-type residues on the other side. We found that additional rounds of rational design were required to reduce the residual activity of homotetrameric complexes.

A previous effort to create a heterotetrameric Cre complex identified concerted small-to-large and large-to-small hydrophobic mutations in an expression library that combinatorially mutagenized three tightly coupled residues (19). The engineered interface was functional, but one of the mutant surfaces retained significant activity in complex with the complementary wild-type surface. Perhaps unsurprisingly, a small-to-large mutation was incompatible with the wild-type surface, presumably due to steric clash. However, large-to-small mutations exhibited reduced activity relative to wild-type, likely a consequence of creating a

cavity that destabilized, but did not destroy, the integrity of the interface. Although our mutations were selected to drive heterotetramer formation primarily based upon electrostatics rather that sterics, we ascribe our elimination of homotetramer activity to the increased number of residues and contact regions that we altered rather than the nature of the interactions that were altered or introduced.

Researchers have successfully demonstrated a different strategy for partitioning Cre recombinase into two variants that are only active when combined. It has been shown that Cre recombinase can be split into N- and C-terminal fragments (split-Cre) that can reconstitute a functional complex when co-expressed *in vivo* by virtue of coiled coil dimerization tags appended to each fragment (30). The motivation for this approach was to place the split-Cre fragments under different promoters, yielding enhanced control over the cell types in which functional Cre complexes are present and resulting in highly specific conditional gene regulation. However, this approach to splitting Cre is not suitable for our purpose of combining monomers with different DNA specificities. Each split-Cre complex retains specificity for the loxP RT site. Even if specificities of the DNA-contacting regions are altered, the assembly of N and C-terminal fragments is uncontrolled, allowing for multiple combinations of half-site RT site specificities (16), and making this decomposition unsuitable for targeting asymmetric sites with high specificity.

CRISPR-based systems have emerged as an attractive tool for genome engineering due to the ease with which the Cas9 nuclease can be redirected to arbitrary targets (31-33). CRISPR/Cas technology represents the logical conclusion of modular DSB inducing agents, as the Cas9 nuclease can be targeted to any site that contains a protospacer adjacent motif (PAM) sequence (typically 3-5 bases in length) without mutating the protein itself. In cell culture, this

activity can drive the efficient generation of loss-of-function mutants when the DSB is repaired by non-homologous end-joining, or gene conversion when homology-directed repair occurs in the presence of an exogenously provided repair template (34). Given these features of CRISPR/Cas systems, what role can mutants of Cre recombinase play in genome engineering applications?

Gene conversion by RMCE possesses advantages over DSB-induced gene conversion that are unique to enzymatically autonomous recombinases. A crucial advantage is that no other cofactors or endogenous cellular machinery are necessary. In particular, this avenue for genome editing does not rely upon the homology-directed DNA repair (HDR) system. The balance between DNA repair *via* HDR and *via* non-homologous end-joining (NHEJ) is highly dependent on cell type, and HDR itself is not a significant route for DNA repair in cells that are not replicating (3, 4). Thus, RMCE approaches may prove to be the only effective route to gene conversion for postmitotic cells, where DSB-induced HDR performs poorly. Furthermore, DSB-stimulated gene conversion is efficient over a relatively short range (~100 bp) (35). In contrast, cassette-mediated exchange is capable of correcting any mutation that falls within the RT site boundaries. Using RMCE, genetic intervals of >100kb of DNA have been exchanged, with the size of the interval limited by the size of the donor construct, and not by the method itself (36).

The disadvantage of targeting mutant recombinases to endogenous sites in a genome is the difficulty with which recombinase DNA specificity is altered. While directed evolution has proven to be successful in generating novel RT specificities, the compatibility of DNA specificity altering mutations with our interface mutations is a concern. Our results show that in at least one case (loxM7) the mutations that alter DNA specificity are compatible with our

mutations for controlling tetramer assembly. With respect to obtaining novel specificity Cre mutants, there is no realistic hope for any retargeting strategy that can rival the speed and ease of retargeting in CRISPR/Cas systems. We anticipate that endogenous site RMCE will be useful when a particular genomic locus is of sufficient interest to merit the effort required to obtain mutant recombinases whose RT specificities bracket the locus, or when there is a need to repeatedly exchange the DNA within the genetic interval. This may be the case when a locus harbors a large number of disease-associated polymorphisms that span several kb, or when a 'promoter bashing' experimental approach is desired in an endogenous context.

We have presented an obligate heterotetrameric pair of Cre recombinase mutants. We have demonstrated that this pair can be used to form functional complexes that can recognize asymmetric RT sites. However, to realize the RMCE approach with maximal control over Cre complex formation, we will require a second pair of recombinase monomers to target the second asymmetric RT site that brackets the genetic cassette. This may be accomplished by engineering two additional Cre monomers that form a second obligate heterotetramer that is incompatible with the mutants we have described here. As this involves a large number of positive and negative constraints on monomer association, we suggest that an easier approach will be to use the knowledge of interacting residues we have identified in this study to direct rational redesign of the interface of a Cre homolog (37-39). Although no crystal structures are available for close homologs of Cre, sequence homology between recombinases has been recognized that could assist in generating obligate heterotetrameric mutants (37, 40). We are currently investigating the feasibility of this approach.

FUNDING

Research reported in this publication was supported by the National Institute Of General Medical Sciences of the National Institutes of Health under Award Number R01GM101602 (to J.J.H.) and by the National Eye Institute under Award Number R01EY018826 (to J.C.C.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

REFERENCES

1. Sternberg,N. and Hamilton,D. (1981) Bacteriophage P1 site-specific recombination. I. Recombination between loxP sites. *J Mol Biol* **150**, 467-486.
2. Gu,H., Zou,Y.R. and Rajewsky,K. (1993) Independent control of immunoglobulin switch recombination at individual switch regions evidenced through Cre-loxP-mediated gene targeting. *Cell* **73**, 1155-1164.
3. Saleh-Gohari,N. and Helleday,T. (2004) Conservative homologous recombination preferentially repairs DNA double-strand breaks in the S phase of the cell cycle in human cells. *Nucleic Acids Res* **32**, 3683-3688.
4. Rothkamm,K., Kruger,I., Thompson,L.H. and Lobrich,M. (2003) Pathways of DNA double-strand break repair during the mammalian cell cycle. *Mol Cell Biol* **23**, 5706-5715.
5. Bouhassira,E.E., Westerman,K. and Leboulch,P. (1997) Transcriptional behavior of LCR enhancer elements integrated at the same chromosomal locus by recombinase-mediated cassette exchange *Blood* **90**, 3332-3344.
6. Schlake,T. and Bode,J. (1994) Use of mutated FLP recognition target (FRT) sites for the exchange of expression cassettes at defined chromosomal loci *Biochemistry* **33**, 12746-12751.
7. Seibler,J. and Bode,J. (1997) Double-reciprocal crossover mediated by FLP-recombinase: a concept and an assay *Biochemistry* **36**, 1740-1747.
8. Bethke,B. and Sauer,B. (1997) Segmental genomic replacement by Cre-mediated recombination: genotoxic stress activation of the p53 promoter in single-copy transformants. *Nucleic Acids Res* **25**, 2828-2834.
9. Soukharev,S., Miller,J.L. and Sauer,B. (1999) Segmental genomic replacement in

embryonic stem cells by double lox targeting. *Nucleic Acids Res* **27**, e21.

10. Turan,S., Zehe,C., Kuehle,J., Qiao,J. and Bode,J. (2013) Recombinase-mediated cassette exchange (RMCE) - a rapidly-expanding toolbox for targeted genomic modifications. *Gene* **515**, 1-27.
11. Anderson,R.P., Voziyanova,E. and Voziyanov,Y. (2012) Flp and Cre expressed from Flp-2A-Cre and Flp-IRES-Cre transcription units mediate the highest level of dual recombinase-mediated cassette exchange. *Nucleic Acids Res* **40**, e62.
12. Buchholz,F. and Stewart,A.F. (2001) Alteration of Cre recombinase site specificity by substrate-linked protein evolution *Nature Biotechnol* **19**, 1047-1052.
13. Santoro,S.W. and Schultz,P.G. (2002) Directed evolution of the site specificity of Cre recombinase *Proc Natl Acad Sci USA* **99**, 4185-4190.
14. Sarkar,I., Hauber,I., Hauber,J. and Buchholz,F. (2007) HIV-1 proviral DNA excision using an evolved recombinase *Science* **316**, 1912.
15. Bolusani,S., Ma,C.H., Paek,A., Konieczka,J.H., Jayaram,M. and Voziyanov,Y. (2006) Evolution of variants of yeast site-specific recombinase Flp that utilize native genomic sequences as recombination target sites *Nucleic Acids Res* **34**, 5259.
16. Saraf-Levy,T., Santoro,S.W., Volpin,H., Kushnirsky,T., Eyal,Y., Schultz,P.G., Gidoni,D. and Carmi,N. (2006) Site-specific recombination of asymmetric lox sites mediated by a heterotetrameric Cre recombinase complex *Bioorg Med Chem* **14**, 3081-3089.
17. Loonstra,A., Vooijs,M., Beverloo,H.B., Allak,B.A., van Drunen,E., Kanaar,R., Berns,A. and Jonkers,J. (2001) Growth inhibition and DNA damage induced by Cre recombinase in mammalian cells. *Proc Natl Acad Sci U S A* **98**, 9209-9214.
18. Szczepek,M., Brondani,V., Buchel,J., Serrano,L., Segal,D.J. and Cathomen,T. (2007)

Structure-based redesign of the dimerization interface reduces the toxicity of zinc-finger nucleases. *Nat Biotechnol* **25**, 786-793.

19. Gelato,K.A., Martin,S.S., Liu,P.H., Saunders,A.A. and Baldwin,E.P. (2008) Spatially directed assembly of a heterotetrameric Cre-Lox synapse restricts recombination specificity *J Mol Biol* **378**, 653-665.
20. Martin,S.S., Pulido,E., Chu,V.C., Lechner,T.S. and Baldwin,E.P. (2002) The order of strand exchanges in Cre-LoxP recombination and its basis suggested by the crystal structure of a Cre-LoxP Holliday junction complex. *J Mol Biol* **319**, 107-127.
21. Andrew Leaver-Fay, Michael Tyka, Steven M. Lewis, Oliver F. Lange, James Thompson, Ron Jacak, Kristian Kaufman, P. Douglas Renfrew, Colin A. Smith, Will Sheffler, Ian W. Davis, Seth Cooper, Adrien Treuille, Daniel J. Mandell, Florian Richter, Yih-En Andrew Ban, Sarel J. Fleishman, Jacob E. Corn, David E. Kim, Sergey Lyskov, Monica Berrondo, Stuart Mentzer, Zoran Popović, James J. Havranek, John Karanicolas, Rhiju Das, Jens Meiler, Tanja Kortemme, Jeffrey J. Gray, Brian Kuhlman, David Baker, and Philip Bradley (2011) ROSETTA3: An Object-Oriented Software Suite for the Simulation and Design of Macromolecules *Methods Enzymol* **487**, 545.
22. Shapovalov,M.V. and Dunbrack,R.L. (2011) A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure* **19**, 844-858.
23. Dantas,G., Corrent,C., Reichow,S.L., Havranek,J.J., Eletr,Z.M., Isern,N.G., Kuhlman,B., Varani,G., Merritt,E.A. and Baker,D. (2007) High-resolution structural and thermodynamic analysis of extreme stabilization of human procarboxypeptidase by computational protein design. *J Mol Biol* **366**, 1209-1221.

24. Studier,F.W. (2005) Protein production by auto-induction in high-density shaking cultures *Protein Expr Purif* **41**, 207-234.

25. Madisen,L., Zwingman,T.A., Sunkin,S.M., Oh,S.W., Zariwala,H.A., Gu,H., Ng,L.L., Palmiter,R.D., Hawrylycz,M.J., Jones,A.R., Lein,E.S. and Zeng,H. (2010) A robust and high-throughput Cre reporting and characterization system for the whole mouse brain. *Nat Neurosci* **13**, 133-140.

26. Hsiau,T.H., Diaconu,C., Myers,C.A., Lee,J., Cepko,C.L. and Corbo,J.C. (2007) The cis-regulatory logic of the mammalian photoreceptor transcriptional network. *PLoS ONE* **2**, e643.

27. Akimoto,M., Cheng,H., Zhu,D., Brzezinski,J.A., Khanna,R., Filippova,E., Oh,E.C., Jing,Y., Linares,J.L., Brooks,M., Zareparsi,S., Mears,A.J., Hero,A., Glaser,T. and Swaroop,A. (2006) Targeting of GFP to newborn rods by Nrl promoter and temporal expression profiling of flow-sorted photoreceptors. *Proc Natl Acad Sci U S A* **103**, 3890-3895.

28. Montana,C.L., Myers,C.A. and Corbo,J.C. (2013) Quantifying the activity of cis-regulatory elements in the mouse retina by explant electroporation. *Methods Mol Biol* **935**, 329-340.

29. Sheren,J., Langer,S.J. and Leinwand,L.A. (2007) A randomized library approach to identifying functional lox site domains for the Cre recombinase. *Nucleic Acids Res* **35**, 5464-5473.

30. Hirrlinger,J., Scheller,A., Hirrlinger,P.G., Kellert,B., Tang,W., Wehr,M.C., Goebbels,S., Reichenbach,A., Sprengel,R., Rossner,M.J., and Frank Kirchhoff (2009) Split-cre complementation indicates coincident activity of different genes in vivo *PLoS ONE* **4**, e4286.

31. Cong,L., Ran,F.A., Cox,D., Lin,S., Barretto,R., Habib,N., Hsu,P.D., Wu,X., Jiang,W.,

Marraffini,L.A. and Zhang,F. (2013) Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819-823.

32. Jinek,M., Chylinski,K., Fonfara,I., Hauer,M., Doudna,J.A. and Charpentier,E. (2012) A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science*

33. Mali,P., Yang,L., Esvelt,K.M., Aach,J., Guell,M., DiCarlo,J.E., Norville,J.E. and Church,G.M. (2013) RNA-guided human genome engineering via Cas9. *Science* **339**, 823-826.

34. Hsu,P.D., Lander,E.S. and Zhang,F. (2014) Development and applications of CRISPR-Cas9 for genome engineering. *Cell* **157**, 1262-1278.

35. Elliott,B., Richardson,C., Winderbaum,J., Nickoloff,J.A. and Jasin,M. (1998) Gene conversion tracts from double-strand break repair in mammalian cells. *Mol Cell Biol* **18**, 93-101.

36. Wallace,H.A., Marques-Kranc,F., Richardson,M., Luna-Crespo,F., Sharpe,J.A., Hughes,J., Wood,W.G., Higgs,D.R. and Smith,A.J.H. (2007) Manipulating the mouse genome to engineer precise functional syntenic replacements with human sequence. *Cell* **128**, 197-209.

37. Karimova,M., Abi-Ghanem,J., Berger,N., Surendranath,V., Pisabarro,M.T. and Buchholz,F. (2013) Vika/vox, a novel efficient and specific Cre/loxP-like site-specific recombination system. *Nucleic Acids Res* **41**, e37.

38. Sauer,B. and McDermott,J. (2004) DNA recombination with a heterospecific Cre homolog identified from comparison of the pac-c1 regions of P1-related phages *Nucleic Acids Res* **32**, 6086-6095.

39. Suzuki,E. and Nakayama,M. (2011) VCre/VloxP and SCre/SloxP: new site-specific

recombination systems for genome engineering *Nucleic Acids Res* **39**, e49-e49.

40. Nunes-Duby,S.E., Kwon,H.J., Tirumalai,R.S., Ellenberger,T. and Landy,A. (1998) Similarities and differences among 105 members of the Int family of site-specific recombinases. *Nucleic Acids Res* **26**, 391-406.

Cre mutant	Mutations
Cre-A1	K25R, D29R, R32E, D33L, Q35R
Cre-B1	E69D, R72K, L76E
Cre-A2	Cre-A1 + R337E
Cre-B2	Cre-B1 + E308R
Cre-A3	Cre-A2 + E123L
Cre-B3	Cre-B2 + E123L

Table 2.1. Mutations for each round of protein engineering.

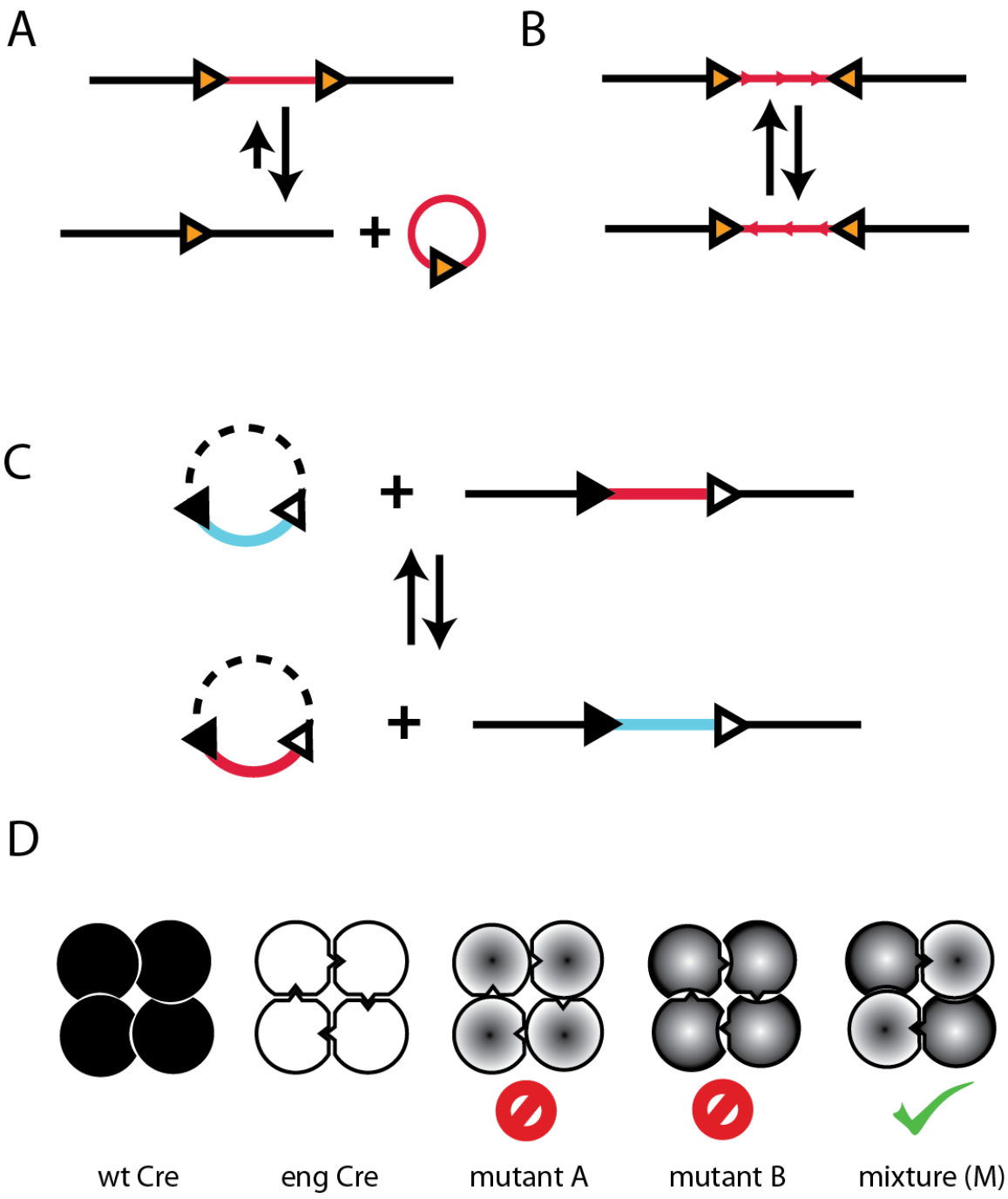


Figure 2.1. Genomic applications of Cre recombinase.

Depending on the number and relative orientation of the loxP sites, Cre recombinase can perform deletion, inversion, insertion or exchange of genetic content. (A) Direct repeats of the loxP site can be recombined to excise the intervening genetic interval (downward arrow). This reaction is

also catalyzed in the reverse direction, yielding a genetic insertion (upward arrow). For thermodynamic reasons, the excision reaction is favored, and insertion events occur with low frequency. (B) Inverted loxP repeats can be recombined to yield an inversion of the bracketed DNA. (C) Recombination at pairs of distinct RT sites gives rise to exchange of the intervening genetic ‘cassette’. (D) Cre recombinase is a homotetramer in its functional complex (wt Cre), imparting a preference for a symmetric RT as a consequence. As a first step to achieving recombination at asymmetric sites, we desire an orthogonal engineered interface between Cre monomers (eng Cre). We seek to construct a novel homotetramer Cre mutant with monomer-monomer interfaces that, while functional, are incompatible with the wild-type protein. Combining wild-type and engineered half-interfaces gives rise to two distinct mutants that cannot form functional complexes (mutants A and B). Combining the two mutants (denoted by ‘M’) can reconstitute a functional heterotetrameric complex, which contains two wild-type and two engineered interfaces.

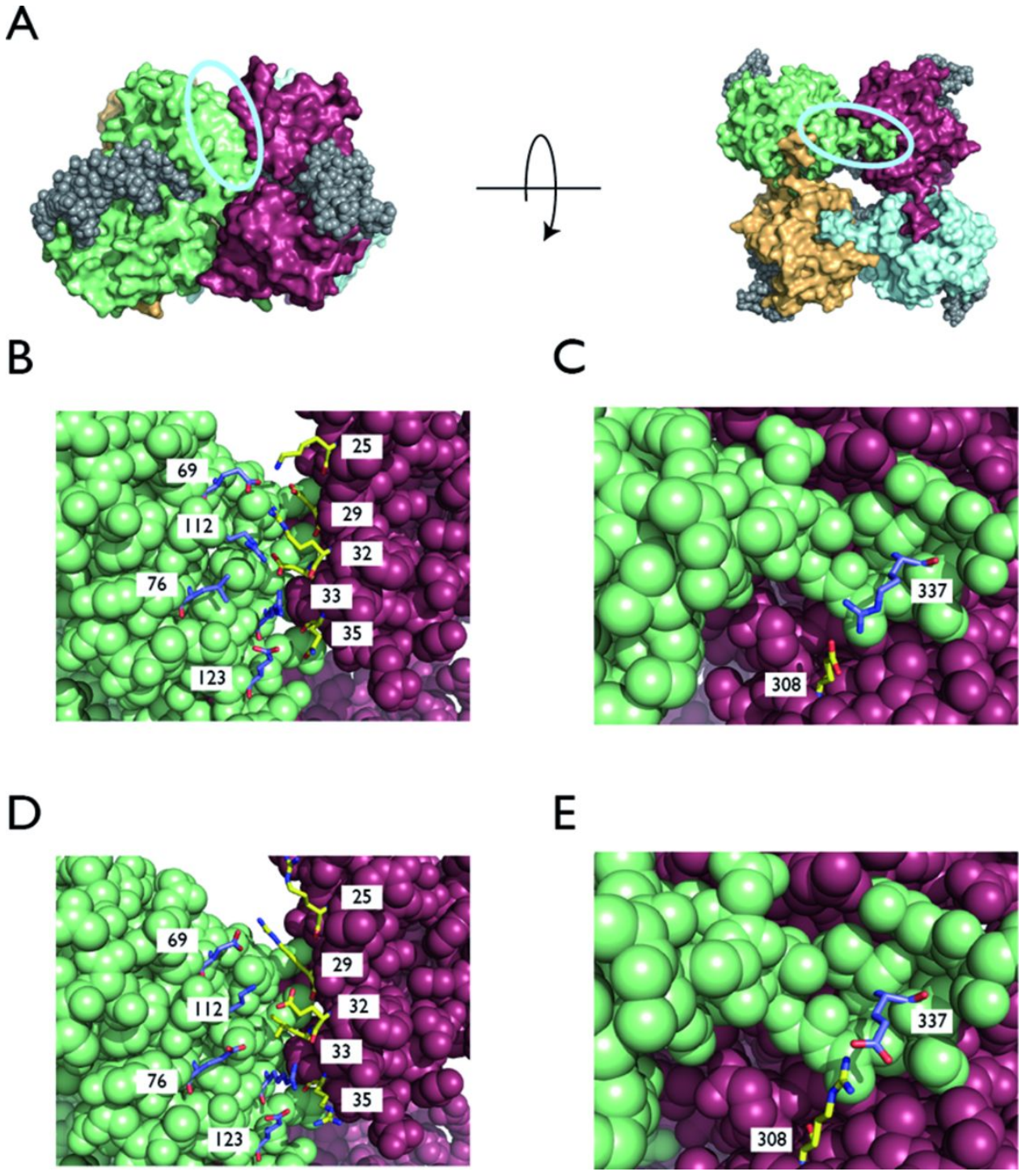


Figure 2.2. Mutated positions in the monomer-monomer interface.

(A) The arrangement of Cre monomers on a loxP Holliday junction. The nucleic acid is shown as grey spheres, and each Cre monomer is rendered in a separate color. The largest area of contact

is indicated with a cyan oval on a side view of the complex (left side), and likewise the salt bridge that was inverted, shown in a bottom view (right side). (B) A set of interacting residues across the monomer-monomer interface was selected by eye for computational redesign (positions 25, 29, 32, 33, 35, 69, 72, 76, 119, and 123). The experimentally determined conformations of the sidechains at these positions are shown (PDB code: 1KBU) (20)). In a third round and rational design, positions 35 and 123 were mutated to hydrophobic residues. (C) A putative salt bridge between a glutamate at position 308 and an arginine at position 337 is observed in the wild-type crystal structure. (D) The predicted model of the monomer-monomer interface after computational redesign is shown. The amino acids at positions 29 and 32 switch their electrostatic charge relative to wild-type, position 33 switches from charged to hydrophobic, and positions 76 and 35 switch from uncharged to charged amino acids. (E) A putative model for the charge swap at positions 308 and 337 preserves a salt bridge, but with a change in polarity.

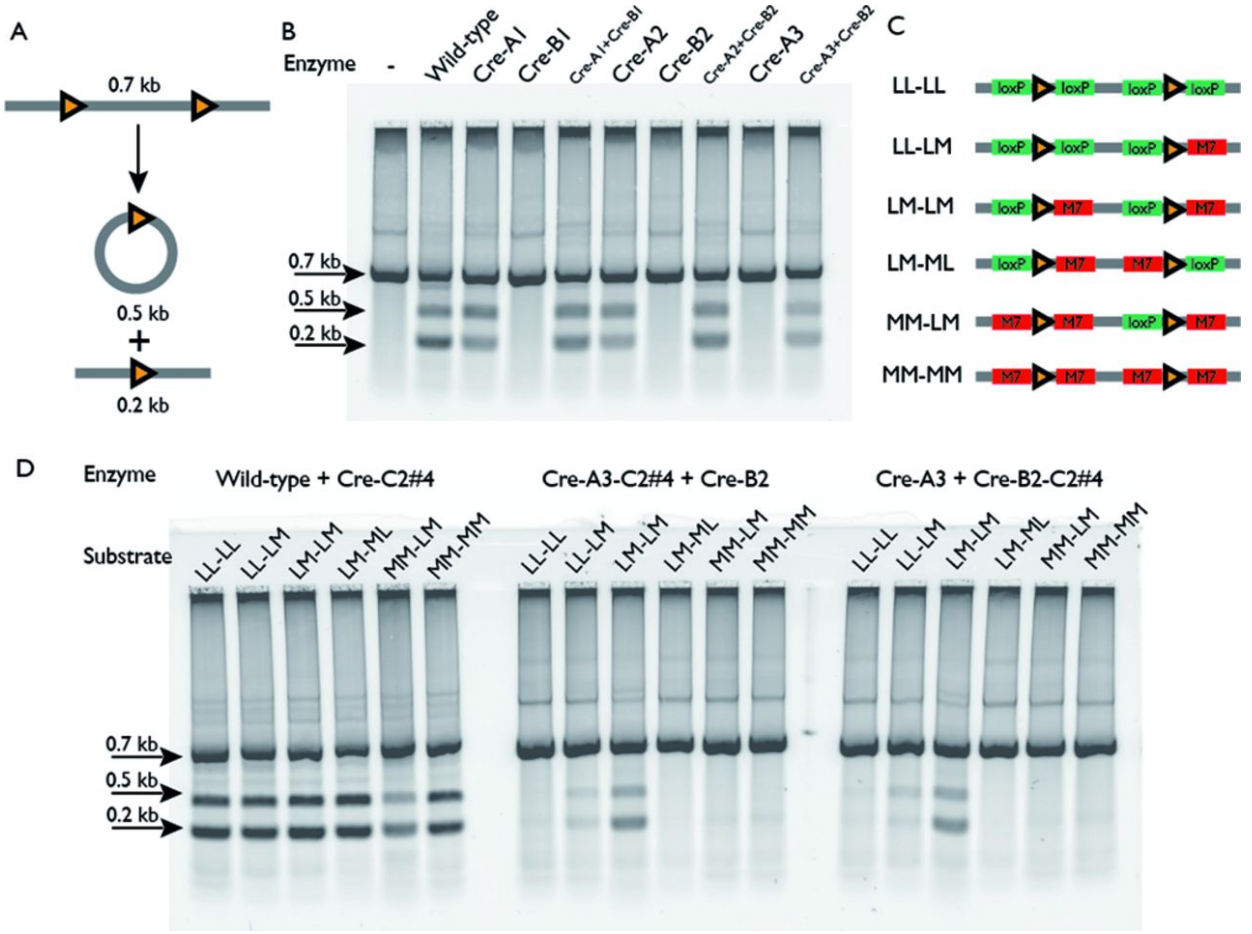
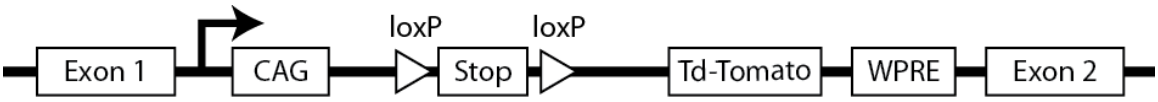


Figure 2.3. *In vitro* recombination assay of Cre mutants.

(A) *In vitro* recombinase assay. A 0.7 kb linear DNA substrate with direct repeats of the loxP site (orange triangles) is incubated with wild-type or mutant Cre recombinase. The activity of functional Cre complexes results in production of a 0.5 kb circular product and a 0.2 kb linear product through intra-molecular excision. (B) *In vitro* assay results. Lane 1: DNA substrate alone; lane 2: wild-type Cre; lanes 3-5: 1st round redesigned Cre mutants Cre-A1, Cre-B1 and a mixture of the two (CreA1+CreB1); lanes 6-8: 2nd round redesigned Cre mutants Cre-A2, Cre-B2 and a mixture of the two (Cre-A2+Cre-B2); lanes 9-10: 3rd round mutant Cre-A3 and a mixture of Cre-A3+Cre-B2. All Cre-B mutants are inactive in isolation. Cre-A mutants progressively lose homotetramer activity through the three rounds of design. (C) *In vitro* substrates for asymmetric

recombination target site experiments. RT half-sites in the linear DNA substrate described in panel (A) were systematically varied to incorporate the M7 sequence (13). LoxP and M7 half-sites are rendered as green and red boxes, and abbreviated by the letters L and M, respectively. Combinations ranged from entirely loxP (LL-LL, the same as in panel (A)) to entirely M7 (MM-MM), including hybrid RT sites situated as both direct (LM-LM) and inverted (LM-ML) repeats. (D) The effect of controlled assembly of heterotetrameric Cre complexes. Each of the mixed loxP/M7 substrates was incubated with a pair of recombinases, one with mutations that recognize the M7 RT halfsite (Cre-C2#4) and the other with preference for the loxP half-site. In the left panel, the two proteins have no additional mutations to control complex formation. In the middle and right panels, recombinases with different RT specificities are combined with the Cre-A3 and Cre-B2 mutations, with both possible combinations tested. The restriction of permissible substrates by the Cre-A3 and Cre-B2 mutations are consistent with a requirement for an (ABAB) heterotetramer to achieve recombinase activity.

A



B

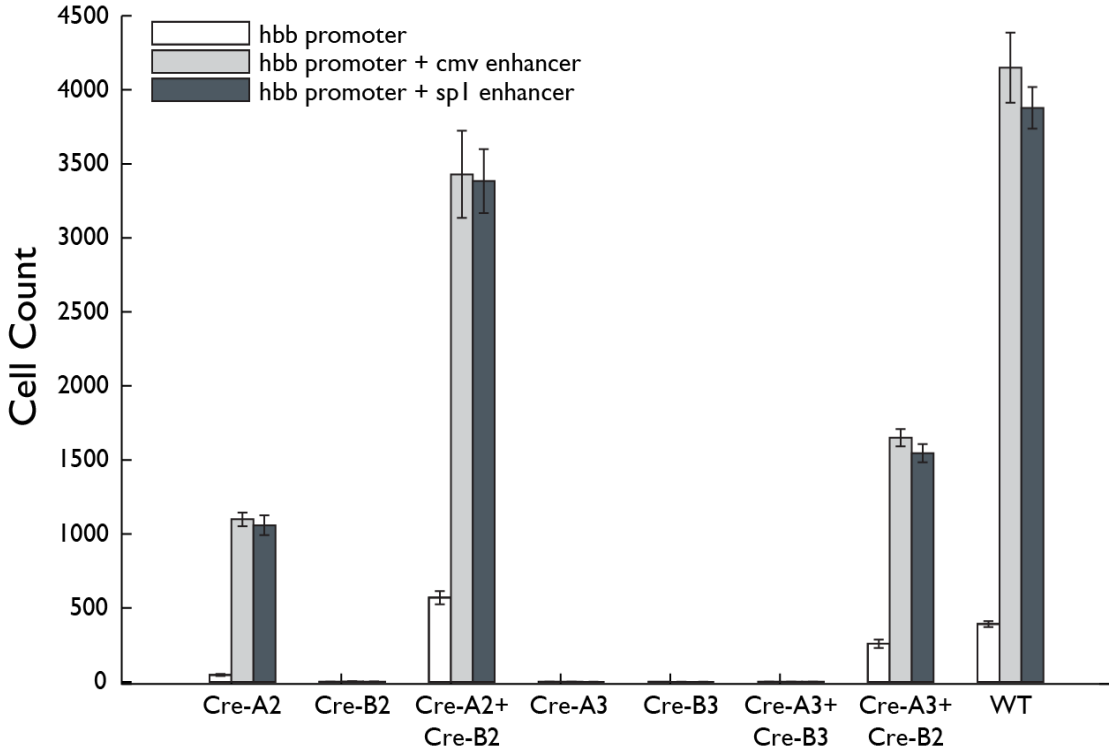
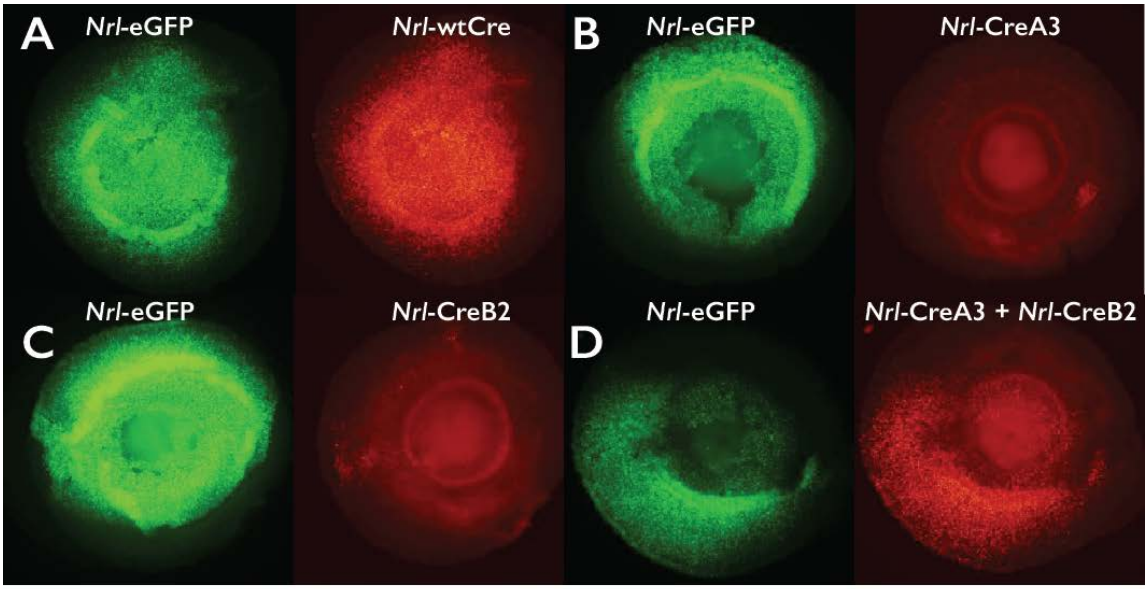


Figure 2.4. Cre mutant pair recapitulates requirement for heterotetramer formation in mouse ES cell cultures.

(A) Diagram of the Cre-reporter cell line. The Cre-reporter cassette was inserted into the Rosa26 locus, in the intron between endogenous exons 1 and 2. In the cassette, a red-fluorescent protein (RFP) is preceded by a floxed stop codon and followed by the woodchuck post-transcriptional regulatory element (WPRE). (B) Plasmids with the hbb minimal promoter driving expression of

different Cre variants either alone or augmented with the CMV or SP1 enhancers were co-transfected into Ai14 mouse embryonic stem (ES) cells containing a fluorescent reporter cassette. The same total amount of DNA was used for all transfections, and 3 independent transfections were performed for each Cre variant. The percent of RFP positive cells was measured by flow cytometry. A total of 7000 cells were sorted after each transfection. The average number of RFP positive cells for each Cre variant or combination of variants is shown. For Cre-B2, Cre-A3, and Cre-A3, cell counts were less than five (out of 7000) for all promoter constructs (Supplemental Table 2.1).


A
Nrl-eGFP
Nrl-wtCre
B
Nrl-eGFP
Nrl-CreA3
C
Nrl-eGFP
Nrl-CreB2
D
Nrl-eGFP
Nrl-CreA3 + Nrl-CreB2

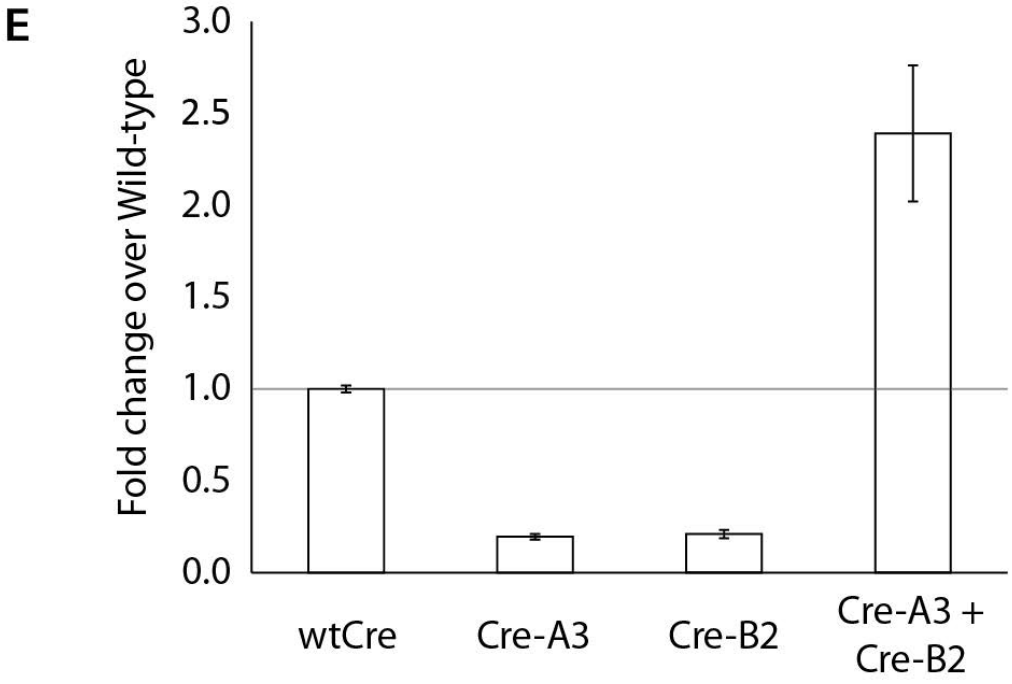

E
Fold change over Wild-type
3.0
2.5
2.0
1.5
1.0
0.5
0.0
wtCre
Cre-A3
Cre-B2
Cre-A3 +
Cre-B2

Figure 2.5. Engineered Cre mutants retain preference for heterotetrameric complex in mouse retinal cells.

Dissected newborn mouse retinas (with lens in place) were electroporated with constructs encoding: (1) *Nrl*-eGFP as a control for electroporation efficiency, (2) a reporter construct for Cre activity comprised of DsRed preceded by a floxed stop codon, and (3) a gene encoding either wild-type (panel A) or engineered Cre (Panels B-D) under control of the *Nrl* promoter. The left side of each panel shows the fluorescence from the green channel, which indicates cells that were successfully electroporated. Fluorescence from the red channel results from removal of the floxed stop codon, indicating Cre activity. The lens shows some autofluorescence which is apparent as a central circular region of red fluorescence in B, C, and D. (E) Quantification of activity of electroporated constructs relative to wild-type Cre.

Chapter 3

Rec-Seq: A High-Throughput Specificity Assay for Recombinases

This Chapter is adapted from a manuscript in preparation for publication.

ABSTRACT

Cre recombinase is a member of the integrase family of site-specific DNA recombinases. The recognition site for Cre is the 34 base pair *loxP* sequence. However, the specificity of Cre is similar to a transcription factor in that it is tolerant of modest substitutions relative to the *loxP* sequence. We have developed a high-throughput assay (Rec-Seq) for characterizing the DNA specificity of recombinase activity. The assay employs a library of synthetic DNA substrates in which one of the four participating half sites has been fully or partially randomized. The library is incubated with purified recombinase and functionally tolerated members of the library are isolated by their altered size. We demonstrate the utility of the assay by characterizing the specificity of the wild-type Cre and VCre recombinases and their relative tolerance to base substitutions across their recognition sequences. We next apply the assay to evaluate the change in DNA specificity resulting from the directed evolution of Cre recombinase towards an altered DNA site. Finally, we utilize sequence analysis of recombinase homologs to predict the change in DNA site specificity due to an amino acid mutation, and use Rec-Seq to validate the predicted specificity determinant.

INTRODUCTION

Cre recombinase recognizes the 34 bp *loxP* recombination target (RT) sequence, and catalyzes a recombination reaction between two copies of *loxP*. The *loxP* sequence possesses pseudo-palindromic symmetry: inverted repeats of a 13 bp arm region are separated by an asymmetric 8 bp spacer region (1). Cre's site specificity is predominantly exhibited in the recognition of the *loxP* arm regions. The enzyme is tolerant of many spacer region sequences, although recombination of two RT sites is dependent on the strong similarity, if not identity, between their spacer regions (2).

The Cre-*loxP* system requires no other cellular cofactors for robust activity, and has been employed for a number of genome engineering tasks. The most common is the creation of a conditional gene knock-out, in which a gene of interest is flanked by *loxP* sites, leading to loss of the gene in cells expressing Cre (3). The reversibility of this excision reaction can be exploited to achieve gene insertions, albeit at low efficiency (4). Pairs of *loxP* variants have been identified that could accomplish this insertion more efficiently (5). A more recent use of recombinases for genome manipulation is recombinase mediated cassette exchange (RMCE). In the RMCE approach, the genetic interval between heterologous RT sites can be exchanged from a donor DNA molecule to an acceptor. This has been demonstrated using several different recombinase-RT sequence pairs (6, 7). Recently, there has been increased interest in obtaining recombinase mutants with altered DNA specificity capable of recognizing sequences already present in human genomes (8-10).

The discovery or engineering of recombinases with new RT specificities is of significant interest as a way to expand the utility of recombinase-based applications. Other than the Cre-*loxP* system, the most studied recombinase-RT pair is Flp-FRT, derived from the 2 μ plasmid of yeast. More recently, a number of Cre homologs have been reported, including the VCre-*VloxP*, SCre-*SloxP*, and Vika-*vox* systems. Many more homologs can be readily identified through searches of sequence databases, and in some cases their RT sequences can be predicted by analysis of the genomic context surrounding the recombinase gene body (11). Each newly identified RT sequence increases the likelihood that a given genetic locus harbors a subsequence similar to the DNA specificity of a known recombinase.

Directed evolution is a powerful tool for altering Cre's DNA specificity. Buchholz and Stewart retargeted Cre to recombine a site from human chromosome 22 called *loxH* (9). The *loxH* site has four changes in the arm region and a different spacer region. They first randomized full length Cre recombinase by error-prone PCR. They used both positive and negative selection steps *in vivo* in bacterial cells. First, Cre mutants were evolved to recombine a *loxP*/loxH hybrid site for ten rounds. Then, Cre mutants were evolved for another ten rounds to efficiently recombine loxH full site. However, the resulting mutants were still able to recombine *loxP* sites with significant activity. Therefore, they performed fifteen additional rounds of negative selection to discourage Cre activity against *loxP* sites, while retaining the ability to recombine *loxH* sites. From the mutant library, a single clone called Fre22 was selected as the most specific mutant, with no activity for *loxP* and moderate activity for *loxH*. Another group used similar techniques to evolve the Flp recombinase towards a site in the human genome (8). Buchholz and colleagues also used directed evolution to evolve Cre to target a sequence in the long-terminal repeat of the HIV-1 strain (named *loxLTR*) (10). This asymmetric site has four changes in the left arm region and seven changes in the right, and has a different spacer region. The full length Cre sequence was randomized by error-prone PCR. They used an iterative positive selection strategy, sequentially targeting intermediate half-sites, separate symmetric half-sites, and finally the full asymmetric site. Surprisingly, although the two half-sites have different sequences, they were able to obtain a Cre mutant that could recombine the asymmetric RT site after 126 rounds of directed evolution. The final Cre mutant (called Tre) successfully recombined the loxLTR sites *in vivo*, resulting in excision of the proviral HIV genome in cultured cells (10).

Using a different strategy, Santoro and Schultz retargeted Cre to recombine *loxM7* sites, which contain three substitutions to each arm region of the *loxP* site (12). They used a focused

mutational approach in which only five residues in the Cre protein (identified from the crystal structure) were randomized. They used GFP and RFP as reporter genes, and FACS sorting to separate functional clones.. They also directly compared directed evolution using only positive selection with an approach that incorporated negative selections in the later rounds. They performed three rounds of positive selection on two libraries and five alternating rounds positive/negative selection on another two libraries. Mutants obtained from just positive selection were generally promiscuous in recombining both *loxP* and *loxM7* sites, while mutant coming from alternating rounds of positive and negative selection exhibited much better specificity. A clone (named CreC2#4) was chosen as the most specific recombinase for the *loxM7* site.

A current weakness of directed evolution applied to selection of novel recombinase specificities is that there is no assay for assessing the specificity of the resultant mutants across a full RT half-site. The selections or screens used to identify recombinases with activity towards a desired novel RT site only report on the activity of the recombinase against a single site. Often, a negative selection is applied by screening the mutant library for activity against the relevant wild-type RT site for the recombinase used as the starting point for evolution. In one directed evolution effort, it was subsequently discovered that although a mutant with activity against the desired novel RT site had indeed been obtained, the mutant actually had a stronger preference for an entirely different RT site (8). This underscores the necessity to take into account not just the desired activity, but also activity against all other possible substrates when assessing the specificity of recombinases. This is a daunting task in the context of mammalian cells.

A sequencing study had been published for *loxP* spacer region specificity and compatibility (13), but the specificity in the arm region have not been thoroughly characterized. A previous randomized library approach to identifying alternative lox sites showed some Cre's

sequence preferences *in vivo* (14), but did not yield highly specific data because Cre recombinase was highly promiscuous in their experimental settings. In this manuscript, we overcame the promiscuity issue by adjusting salt concentrations *in vitro*, and used high-throughput sequencing to determine Cre's DNA specificity.

In this manuscript we describe a high-throughput assay for the specificity of DNA recombinases that we have named Rec-Seq. Using this assay we have characterized the specificity of the arm region of Cre recombinase. We also present the experimentally determined specificity for the VCre recombinase, a homolog of Cre with a different recombination target site sequence. We furthermore assess the specificity of a Cre mutant that has been obtained by directed evolution to exhibit altered DNA specificity in a subset of the arm region (12), giving a clearer evaluation of the nature of the specificity change. Finally, we demonstrate how the simple consideration of amino acid-base correspondences between members of the tyrosine recombinase family can uncover specificity determinants in protein alignments. We validate the utility of this method for identifying specificity-altering mutations by characterizing the changes in specificity that result from the identified amino acid mutation.

MATERIALS AND METHODS

Molecular cloning, expression, and purification of recombinases

Genes encoding N-terminal His^7-tagged recombinase were cloned into a pET42a vector using standard molecular biology techniques. Proteins were expressed in ArcticExpress cells (Agilent Genomics) at 20°C for 48 hours using the autoinduction protocol of Studier (15). Recombinases were purified using a HisTrap™ HP column (Amersham) following protocols

described previously (16). Protein concentrations were determined by UV absorbance using an extinction coefficient at 280nm of 49 $mM^{-1}cm^{-1}$ for Cre and 51 $mM^{-1}cm^{-1}$ for VCre. After purification, the proteins were stored at 4° C.

Library preparation for Illumina sequencing

Primers and library sequences are described in supplemental materials and Supplemental Table 3.1. The substrate libraries were generated by PCR amplification of pBAD33 vector's multiple cloning site, with the Library-forward and Library-reverse primers. 700ng of the substrate libraries were incubated with 3 µM recombinase protein in 10 µL of 50 mM Tris-HCl, pH 7.8, and 10 mM $MgCl_2$ for 20 hours at 37° C. NaCl conditions for each assay are described in Supplemental Table 3.1. The reactions were stopped at 98° C for 10 minutes, and then diluted to 50 µL. The mixtures were digested the XbaI restriction enzyme (NEB) to prevent PCR amplification of the 824 bp substrate in the next step, ensuring that only the 97 bp recombination product can be amplified (Figure 3.1). The selection sequencing libraries were generated by PCR with primers pair Selected-sequencing-forward and Sequencing-reverse, using 5 µL of the above digested mixture as template. The background sequencing libraries were generated with primers pair Unselected-sequencing-forward and Sequencing- reverse, using 1 µL substrate libraries as template. Both libraries were amplified by Q5® DNA polymerase (New England Biolabs) using two-step protocol, 30 cycles, 15 s extension time.

High-throughput sequencing data analysis

For each individual sequence, the enrichment over background is calculated as:

$$Enrichment(seq) = \frac{N_{sel}^{seq}}{N_{bg}^{seq}} \frac{N_{bg}^{total}}{N_{sel}^{total}}$$

where N_{sel}^{seq} and N_{bg}^{seq} are the number of occurrences of the individual sequence in the selected and background libraries, respectively, and N_{sel}^{total} and N_{bg}^{total} are the total number of sequence counts in the selected and background libraries. For the 4N, 5N and 6N libraries, sequences with enrichment of less than 2.5 were discarded. For the 9N libraries, sequences with enrichment of less than 8 were discarded. Sequence logos were generated by Weblogo3 (17) using enrichment data as input.

Identification of Cre homologs and prediction of their RT sites

We identified homologs of Cre recombinase by performing a PSI-BLAST search of the wild-type Cre sequence against the NCBI set of non-redundant protein sequences, using the default settings (18). For homologs that were close but not identical in sequence identity (30-45% identical), we downloaded the complete genome or shotgun sequencing contig containing the coding region for the protein. We wrote a simple computer program to analyze the sequencing files and identify 34 bp sequences with pseudo-palindromic symmetry. The outer 13 bp flanks were compared for inverted repeats. The 8 bp spacer regions were ignored. A user-specified number of mismatches between the putative arm regions were allowed. We had two requirements for a sequence to qualify as an RT prediction. First, we allowed at most one deviation from inverted symmetry in the arm regions. Second, the putative RT sequence had to be located within 2 kbp of the coding region for the homologous gene.

RESULTS

A high-throughput assay of recombinase DNA specificity

Assessing the DNA specificity of a recombinase involves identifying not just the preferred RT sequence, but also the relative impact of base substitutions throughout the RT site. The DNA binding specificities of transcription factors (TFs) have been determined *via* high-throughput sequencing of the bound subset of randomized DNA binding sites (19). In contrast, tyrosine recombinases exhibit their specificity in an enzymatic reaction in which two RT sites are cleaved and religated. We have established an *in vitro* assay to determine the DNA specificity of Cre recombinase using the successful completion of the DNA splicing reaction as a selection method (Figure 3.1). We generate a library of linear DNA substrates (824 bp) for recombination by PCR. The RT sites are incorporated into the PCR primers, allowing for ready substitution of the half-site sequences and different choices of randomized bases. The library is incubated with purified recombinase to initiate the reaction. Successful recombination results in a small 97 bp linear product containing the functionally competent members of the randomized library, and a 727 bp circular product. The reaction mixture is then digested by XbaI to prevent PCR amplification of unreacted substrate in the following step. The 97 bp product is subsequently amplified by PCR to generate the selected sample, and a background sample for comparison is generated by PCR amplification of the unreacted 824 bp substrate. Both the background and selected library samples are subjected to high-throughtput sequencing.

We have tested various library designs and conditions for Cre variants and homologs (Supplemental Table 3.1). We observed no recombination when Cre was incubated with a library in which the entire 13 bp half-site was randomized (13N library), and low sequence specificity

with a library in which the outermost nine bases were randomized (9N library)(Supplemental Figure 3.1A). The low activity was likely the result of the high complexity of the 13N library, and the low sequence specificity was likely the result of non-specific recombination driven by the cooperativity of the Cre complex. As has been shown previously, Cre monomers assemble cooperatively to form a functional tetrameric complex (14) (16). When the three non-randomized RT half-sites are the preferred *loxP* sequence, we observed promiscuous recombination of the 9N library. We then tried to reduce the cooperativity by mutating the half-site next to the 9N library to the *loxM7* sequence (12)}, hypothesizing that the reduced affinity of Cre for this half-site would reduce the overall cooperativity for the reaction. In line with this expectation, we observed increased specificity from the 9N library that is consistent with the expected *loxP* sequence (Figure 3.2).

For experiments with recombinases other than Cre, we do not have known weak RT sites analogous to *loxM7*. Furthermore, we would like to use native RT sites when possible. We found that we could obtain higher specificity profiles when randomized libraries with lower complexity were used, even without changing the *loxP* sequence at the non-randomized half-sites. Therefore, for subsequent specificity assays, we broke the 13 bp arm region into overlapping windows of 5-6 bp, performing independent experiments randomizing one window at a time.

NaCl concentration has a large effect on observed *in vitro* specificity

One obstacle that prevented us from observing highly specific recombination is the cooperativity of DNA binding in Cre tetramers. At pH 7.8, Cre forms tetramers and are more active but less soluble at low salt conditions (<300 mM NaCl) but remains mostly monomeric formations at 700mM NaCl (20). Most previous *in vitro* recombinations were reported in low

salt conditions (~50 mM NaCl) (12) (13, 21). We proposed that at higher NaCl concentration, Cre's interaction between monomers would be compromised, so that Cre would become less cooperative, and the reactivity of the library members would be less dependent on the sequences of the non-randomized half-sites.

We sequenced the recombination product from the randomized substrate library shown in Figure 3.2 after incubation in 4 different salt conditions: 120 mM, 170 mM, 200 mM and 210 mM (Figure 3.2, Supplemental Table 3.1). As expected, Cre's specificity increased with the salt concentration. However, there was no detectable recombination product for this library when we raised the NaCl concentration above 210 mM (Supplemental Table 3.1). In order to obtain highly specific recombination, we tried to raise the salt concentrations as high as possible while still permitting detectable activity in all following assays. Unfortunately, the optimal salt conditions to use are dependent on both the recombinase and the substrate sequences. Therefore we empirically identified the concentration to use for each assay (Supplemental Table 3.1).

Sequence specificity for Cre variants and homologs

We applied the window approach to assess the DNA specificity of wild-type Cre recombinase throughout the full *loxP* arm region (Figure 3.3). The location of the randomized windows relative to the arm region sequence are indicated, and sequence logos for each of these experiments are shown. We also tested the sequence preference of Cre within a window that included the first base pair of the spacer region (Supplemental Figure 3.1D). Cre makes a single-stranded break after the 1st base on the top strand and before the 8th base on the bottom strand in the spacer region. Homology is required for the intervening 6 bp sequence for strand exchange. We attempted to test the specificity of the spacer by randomizing half of the bases in this region,

but did not observe any recombinant product (Supplemental Table 3.1). In all the reported windows in the arm region, the consensus *loxP* sequence was the most enriched sequence. For the first base pair of the spacer region, Cre showed a strong preference against cytosine (Supplemental Figure 3.1D). Although the overlapping bases between different windows show preferences for the same bases, the relative information contents can be quite different. We attribute the discrepancies to the differences in the substrate and salt conditions between assays. The discrepancies prevent us from combining the overlapping windows to show a single logo for the full half-site. Therefore, we report the results for each window independently in this manuscript.

In recent years, Cre homologs targeting different RT sites have been found. Such additional homologs are particularly useful in RMCE applications, where two recombinases can be used simultaneously for genome cassette exchanges (22). We set out to determine detailed sequence specificities for the VCre recombinase (23). VCre recombines *VloxP* sites (Table 3.1) and the purified protein was active in our *in vitro* assays (Figure 3.4, Supplemental Figure 3.1C). In the absence of other detailed studies, we assumed that the 34 bp *VloxP* site possesses the same structure as *loxP*, containing two 13 bp arm regions and an 8 bp spacer region. Interestingly, the two arm regions of *VloxP* site are not strictly inverted repeats containing a mismatch at position 9. Results detailing the sequence specificity of VCre using three randomized windows over the arm region are shown in Firgue 4. By comparing the logos between Cre and VCre, we found that the two recombinases showed strong sequence preferences at different positions. Cre showed strong preference at position 2, 7, 8, 11, 12 and 13, while VCre showed strong preference at position 1, 3, 6, 7, 12, 13, suggesting the possibility of slightly different DNA binding modes.

VCre showed the least sequence preference at position 9, consistent with this being the position of the mismatch between the arm regions in *VloxP*.

Probing the sequence specificity of an evolved recombinase

The ability of directed evolution to generate recombinases that can operate on novel RT sites has been demonstrated repeatedly. However, the process can be time consuming and laborious, especially if the desired RT site is significantly different from any known wild-type RT sequence. To accelerate the process of retargeting recombinases, we propose to use sequence analysis of homologs of Cre. PSI-BLAST searches of bacterial genomes using Cre recombinase as a starting template identify hundreds of Cre homologs. Using custom software to scan the surrounding genomic context for the open reading frames of these homologs, we are able in many cases to predict their RT sequences. Closely related homologs share similarity in both amino acid sequence and predicted RT DNA sequence. Using the crystal structure of Cre bound to *loxP* (PDB code: 1KBU (24)), we are able to infer determinants of specificity by inspecting coordinated amino acid and base changes that map to interacting regions of the structure. As a demonstration, we picked a group of Cre homologs with predicted RT sites that switched from *loxP* to *loxTA* at position 8 and 9 (Figure 3.6). In the Cre-loxP structure, these bases are contacted by an alpha helix. Inspection of a multiple sequence alignment of the residues in this helix suggested the R259P mutation as a good candidate (Figure 3.6C).

We expressed and purified the Cre R259P mutant to assess the validity of the putative R259P specificity determinant. First, we performed *in vitro* activity assays with Cre and Cre259P on the *loxP* and *loxTA* sequences (Figure 3.7A,B). The assays showed that Cre and Cre259P can recombine both *loxP* and *loxTA*, and there was no observable difference in specificity between

Cre and Cre259P. This indicates that the R259P mutation does not have enough of an effect on specificity to prevent recombination at *loxP* sites. We used Rec-Seq to look for more subtle changes in specificity. We designed a 4 bp randomized library centered on the base substitutions in *loxTA* relative to *loxP* and used high-throughtput sequencing to examine Cre259P's DNA specificity in greater detail (Figure 3.7C-E). The resulting logo shows there is indeed a specificity switch caused by Cre259P. We tried additional mutations suggested by the sequence alignment at positions 258, 263 and 266, but did not observe increased specificity for any (data not shown). Although Cre259P does not result in a strict change in specificity, the ready availability of this mutation makes the search for *loxP*-like sequences more tolerant at these positions, and the Cre259P mutant can be used as an alternate starting point for directed evolution if the desired RT site matches *loxTA* better than *loxP*. We expect that the wealth of sequence data for recombinase homologs can be 'mined' to identify many more specificity determinants that can be helpful in this regard as well.

DISCUSSION

Cre recombinase and related homologous recombinases are attractive tools for genome engineering. Here, we have presented a high-throughput method to study the DNA specificity of recombinases. In some instances the native RT sequences of Cre homologs are not difficult to identify. The RT sites are assumed to adopt a *loxP*-like structure, usually located close to the genes that encode the recombinases in the genome (11). However, characterizing the DNA specificity of recombinases involves more than just knowing the native RT sequences. These recombinases may recombine other RT sites with relatively high efficiency, generating off-target activity. Furthermore, if we want to engineer the recombinases to retarget other RT sites, it will

be advantageous to know at which RT positions the recombinases are tolerant of substitutions, and are subsequently easier to retarget.

The Rec-Seq method we have established reports on *in vitro* recombination of randomized DNA substrate libraries. A previous randomized library approach showed Cre can be quite promiscuous *in vivo* (14). In their *in vivo* study, when three out of four arm region were kept constant *loxP* sequence, Cre can recombine randomized arm regions even when 12 out of 13 bases are different from *loxP*. The promiscuity with respect to the randomized region is not entirely surprising, given the strong cooperativity of Cre. However, this feature of Cre makes it a difficult task to accurately assess its DNA specificity. We have addressed this issue *in vitro* in two ways. First, we have intentionally introduced suboptimal sequences(*loxM7* for Cre) in the constant arm regions to reduce binding in the unrandomized half-sites. However, for proteins other than Cre, we will not have sufficient knowledge to design suboptimal, yet still functional, sites. We have explored a second, more general approach in which the salt concentration is tuned to weaken the protein-DNA interactions. This leads to a stronger requirement for recognition between the recombinase and the randomized region of the substrate, yielding results that exhibit stronger specificity. Drawbacks of *in vitro* recombination in the Rec-Seq method include the effort required to obtain purified protein, and the inability to study interesting recombinase variants with known *in vivo* activity, but with insufficient activity *in vitro* to extract meaningful results. We are exploring the possibility of creating an *in vivo* version of Rec-Seq that overcomes the previously observed problem of high cooperativity.

We have adopted a window/tiling approach to characterizing the specificity of recombinases. Assaying the entire arm regions requires high complexity libraries that contain mostly inactive members, and in our case, did not generate detectable recombination product (Supplemental Table 3.1). The windowed approach is particularly suitable for situations where evaluating the specificity of a recombinase for a subset of the arm region is the goal. This is the case for our sequence-based approach to identifying specificity-altering mutations targeting a pair of bases. It is also appropriate for validating recombinase variants coming from directed evolution to target novel RT specificities. When the desired RT has quite a few mismatches to the native RT sequences, specificity alteration is often achieved in a series of intermediate steps, each of which focuses on the specificity switch on only a few base pairs, making it suitable for validation using a window approach (9, 10). The caveat of the windowed approach is that the recombination of each randomized window may be context dependent, and the windows may not be able to be stitched together for presentation of the specificity of a full arm region.

Retargeting recombinases to novel RT sites is a laborious process involving many rounds of directed evolution. In a recent effort, Buchholz and colleagues applied a total of 126 rounds of directed evolution to evolve Cre to target long-terminal repeat of HIV-1 strain (10). Although, the ease of retargeting recombinases will never approach that of CRISPR/Cas systems, the unique capabilities of this family of proteins justify efforts to accelerate the retargeting process. The Rec-Seq method can be very useful in this regard. First, we can use Rec-Seq to identify the bases in the RT sequence for a recombinase that are less difficult to change. Second, with the help of sequence alignment of Cre homologs, we can identify possible candidate mutations for altering specificities. We can then validate these candidate mutations individually using Rec-Seq

to build a database of mutations and their corresponding changes of RT site specificities. By gathering the positions on the RT sequence that are relatively easy to alter and the database of specificity altering mutations, we can screen for DNA elements in a genomic interval to find those targets for which retargeting is most likely to be successful. By prudent selection of RT targets and incorporation of a small number of previously characterized mutations, the burden placed upon and hopefully the experimental effort required for the directed evolution process can be reduced. . Third, Rec-Seq can serve as a criterion for when the cycles of directed evolution selections can be terminated. Finally, in past efforts, recombinases have been evolved for one sequence, but exhibited even stronger activity for another (8). Rec-Seq could identify this problem before efforts are expended in cell culture experiments.

REFERENCES

1. Hoess,R.H. and Abremski,K. (1984) Interaction of the bacteriophage P1 recombinase Cre with the recombining site loxP. *Proc Natl Acad Sci U S A* **81**, 1026-1029.

2. Lee,G. and Saito,I. (1998) Role of nucleotide sequences of loxP spacer region in Cre-mediated recombination. *Gene* **216**, 55-65.

3. Gu,H., Zou,Y.R. and Rajewsky,K. (1993) Independent control of immunoglobulin switch recombination at individual switch regions evidenced through Cre-loxP-mediated gene targeting. *Cell* **73**, 1155-1164.

4. Araki,K., Araki,M. and Yamamura,K. (1997) Targeted integration of DNA using mutant lox sites in embryonic stem cells *Nucleic Acids Res.* **25**, 868.

5. Albert,H., Dale,E.C., Lee,E. and Ow,D.W. (1995) Site-specific integration of DNA into wild-type and mutant lox sites placed in the plant genome *The Plant Journal* **7**, 649-659.

6. Sadowski,P.D. (1995) The Flp recombinase of the 2-microns plasmid of Saccharomyces cerevisiae. *Prog Nucleic Acid Res Mol Biol* **51**, 53-91.

7. Feng,Y.Q., Seibler,J., Alami,R., Eisen,A., Westerman,K.A., Leboulch,P., Fiering,S. and Bouhassira,E.E. (1999) Site-specific chromosomal integration in mammalian cells: highly efficient CRE recombinase-mediated cassette exchange1 *Journal of molecular biology* **292**, 779-785.

8. Bolusani,S., Ma,C.H., Paek,A., Konieczka,J.H., Jayaram,M. and Voziyanov,Y. (2006) Evolution of variants of yeast site-specific recombinase Flp that utilize native genomic sequences as recombination target sites *Nucleic acids research* **34**, 5259.

9. Buchholz,F. and Stewart,A.F. (2001) Alteration of Cre recombinase site specificity by substrate-linked protein evolution *Nature biotechnology* **19**, 1047-1052.

10. Sarkar,I., Hauber,I., Hauber,J. and Buchholz,F. (2007) HIV-1 proviral DNA excision using an evolved recombinase *Science* **316**, 1912.

11. Surendranath,V., Chusainow,J., Hauber,J., Buchholz,F. and Habermann,B.H. (2010) SeLOX--a locus of recombination site search tool for the detection and directed evolution of site-specific recombination systems. *Nucleic Acids Res* **38**, W293-8.

12. Santoro,S.W. and Schultz,P.G. (2002) Directed evolution of the site specificity of Cre recombinase *Proc Natl Acad Sci USA* **99**, 4185-4190.

13. Missirlis,P.I., Smailus,D.E. and Holt,R.A. (2006) A high-throughput screen identifying sequence and promiscuity characteristics of the loxP spacer region in Cre-mediated recombination. *BMC Genomics* **7**, 73.

14. Sheren,J., Langer,S.J. and Leinwand,L.A. (2007) A randomized library approach to identifying functional lox site domains for the Cre recombinase. *Nucleic Acids Res* **35**, 5464-5473.

15. Studier,F.W. (2005) Protein production by auto-induction in high-density shaking cultures *Protein Expr Purif* **41**, 207-234.

16. Zhang,C., Myers,C.A., Qi,Z., Mitra,R.D., Corbo,J.C. and Havranek,J.J. (2015) Redesign of the monomer-monomer interface of Cre recombinase yields an obligate heterotetrameric complex. *Nucleic Acids Res* **43**, 9076-9085.

17. Crooks,G.E., Hon,G., Chandonia,J.M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res* **14**, 1188-1190.

18. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-3402.

19. Liu,J. and Stormo,G.D. (2005) Combining SELEX with quantitative assays to rapidly obtain accurate models of protein-DNA interactions. *Nucleic Acids Res* **33**, e141.

20. Woods,K.C., Martin,S.S., Chu,V.C. and Baldwin,E.P. (2001) Quasi-equivalence in site-specific recombinase structure and function: crystal structure and activity of trimeric Cre recombinase bound to a three-way Lox DNA junction. *J Mol Biol* **313**, 49-69.

21. Gelato,K.A., Martin,S.S., Liu,P.H., Saunders,A.A. and Baldwin,E.P. (2008) Spatially directed assembly of a heterotetrameric Cre-Lox synapse restricts recombination specificity *J Mol Biol* **378**, 653-665.

22. Anderson,R.P., Voziyanova,E. and Voziyanov,Y. (2012) Flp and Cre expressed from Flp-2A-Cre and Flp-IRES-Cre transcription units mediate the highest level of dual recombinase-mediated cassette exchange. *Nucleic Acids Res* **40**, e62.

23. Suzuki,E. and Nakayama,M. (2011) VCre/VloxP and SCre/SloxP: new site-specific recombination systems for genome engineering. *Nucleic Acids Res* **39**, e49.

24. Martin,S.S., Pulido,E., Chu,V.C., Lechner,T.S. and Baldwin,E.P. (2002) The order of strand exchanges in Cre-LoxP recombination and its basis suggested by the crystal structure of a Cre-LoxP Holliday junction complex. *J Mol Biol* **319**, 107-127.

LoxP	ATAACTTCGTATA GCATACAT TATACGAAGTTAT
LoxM7	ATAACT**CTA**TATA GCATACAT TATA**TAG**AGTTAT
LoxTA	ATAACTT**TA**TATA GCATACAT TATA**TA**AAGTTAT
VloxP	TCAATTTCCGAGA ATGACAGT TCTCAGAAATTGA

Table 3.1 34 bp full recombinase target sites

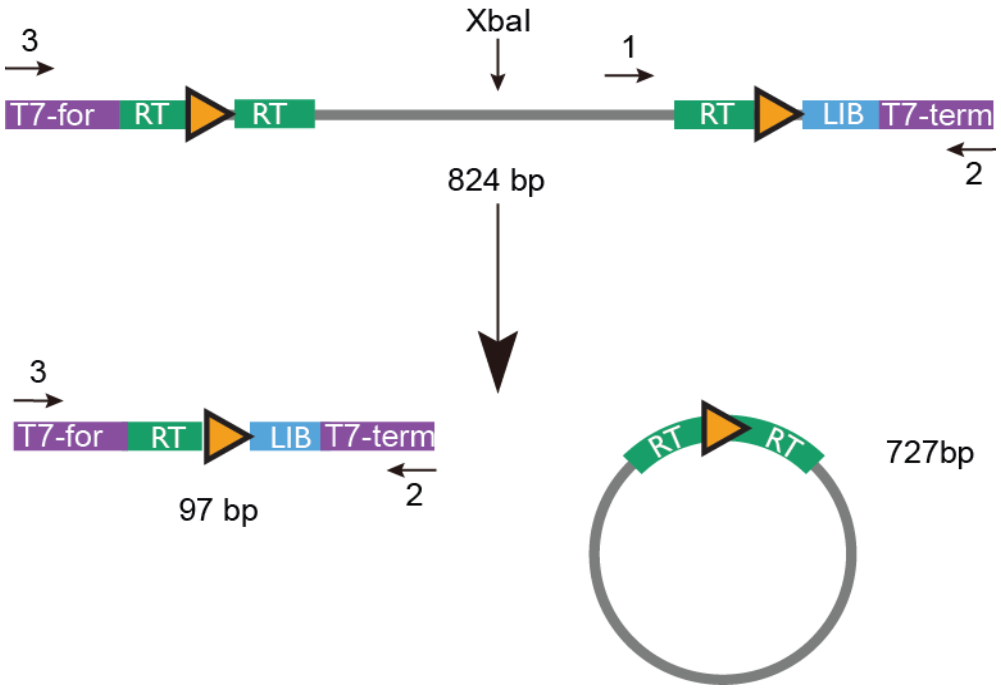


Figure 3.1 Randomized library construction and *in vitro* assay for recombinase specificity.

An 824 bp DNA substrate containing a direct repeat of two target sites is generated by PCR. The asymmetric spacer region is represented as an orange triangle. Three out of the four half-sites are kept constant (green), while one arm region contains randomized libraries (blue). The library is incubated with purified recombinase at 37℃ overnight. After recombination, the substrate is resolved into a 97 bp small linear product containing the primer sites and a 727 bp circular product. Recombined members of the library are amplified by PCR with primers 2 & 3 (black arrows) for subsequent sequencing. For comparison, a background library is generated by PCR using primers 1 & 2, and the starting 824 bp substrate as template.

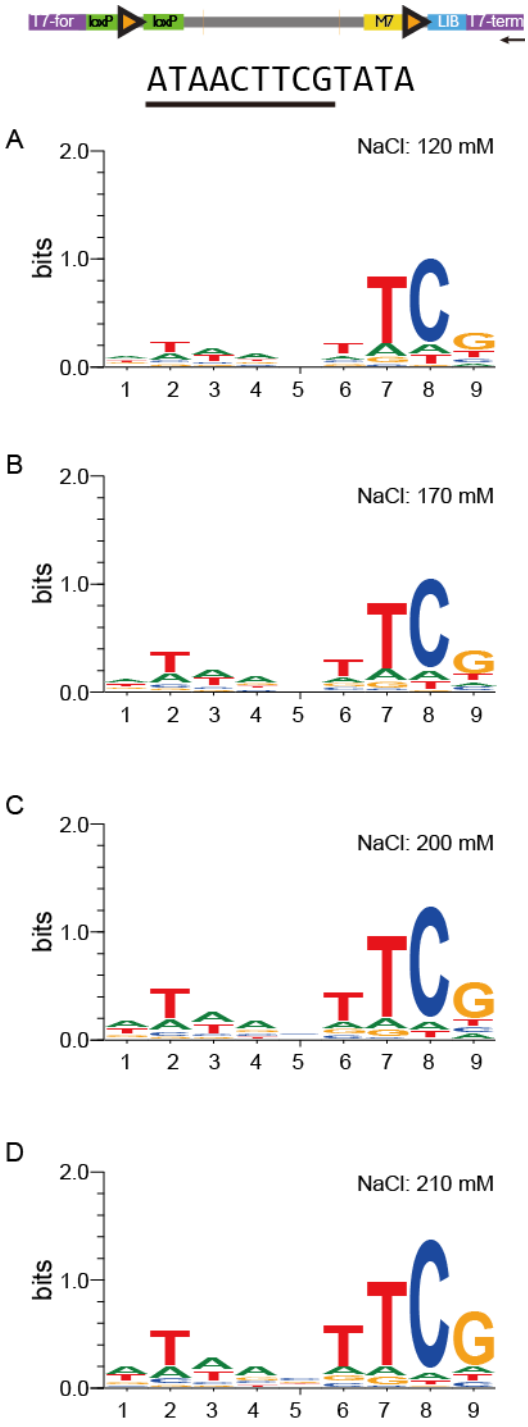


Figure 3.2 Recombinase specificity as a function of salt concentration

Wild-type Cre protein was incubated with a substrate library comprised of a *loxP* RT site (green boxes) and an RT site with a loxM7 half-site (yellow box) and a *loxP* half-site with the

outer nine bases randomized (blue box). The reaction was performed in different NaCl concentrations. The sequence was read from a reverse primer, indicated by the arrow above. The results are reported with the outermost position (1) shown on the left, and the innermost (9) shown on the right. The corresponding region of the *loxP* arm region is shown with the randomized portion underlined. Results are shown for experiments including NaCl concentrations of: (A) 120mM, (B) 170mM, (C) 200mM and (D) 210mM.

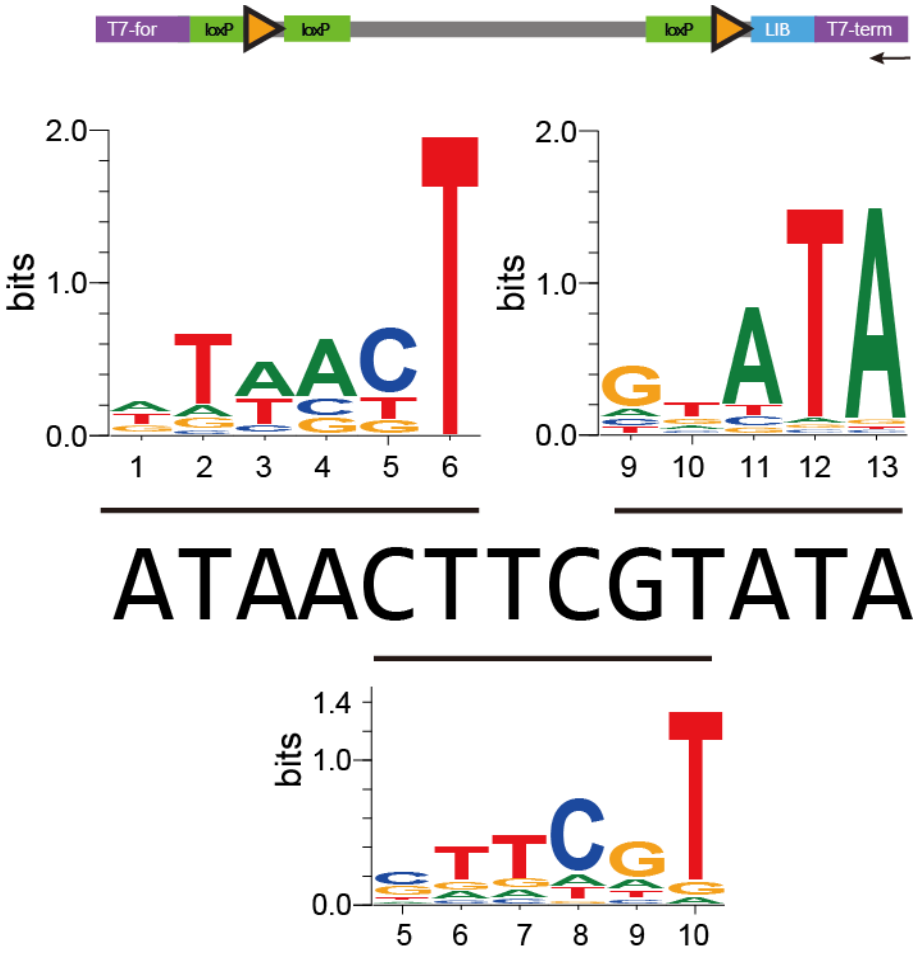


Figure 3.3 Sequence preferences of wild-type Cre recombinase in the arm region

The results of three tiled Rec-Seq experiments that cover the *loxP* arm region are presented. Sequence logos depicting the enrichments at each position are aligned with the corresponding randomized windows of the *loxP* arm region (denoted by black bars). At each position, the most enriched base agrees with the *loxP* sequence, although the relative tolerance for alternate bases varies greatly across the arm region.

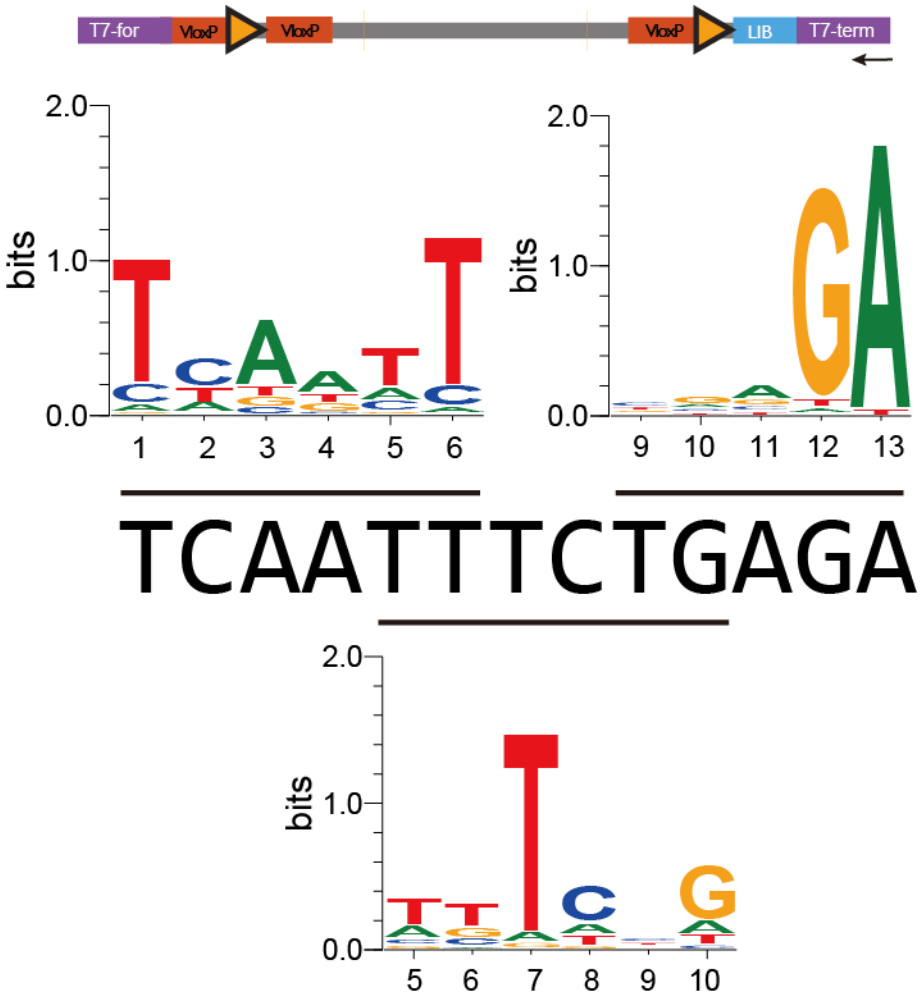


Figure 3.4 Sequence preference of Vcre recombinase in the arm region

The results of three tiled Rec-Seq experiments assaying the specificity of the VCre recombinase over the arm region of *VloxP* are shown. The substrate library (top) includes the *VloxP* sequence for the spacer, 3 unrandomized arm regions, and a randomized arm region, with the location of the sequencing primer annealing region denoted with a black arrow. Sequence logos depicting the enrichments at each position are aligned with the corresponding randomized windows of the *VloxP* arm region (denoted by black bars). The Rec-Seq results for VCre at each position confirm that the most enriched base agrees with the *VloxP* sequence. Position 9 of *VloxP*

is a mismatch between the two arm regions of *VloxP*, with T and C bases present in the left and right arm regions. We find this position to have very little specificity (note the low information content for position 9 in the 5-10 position library).

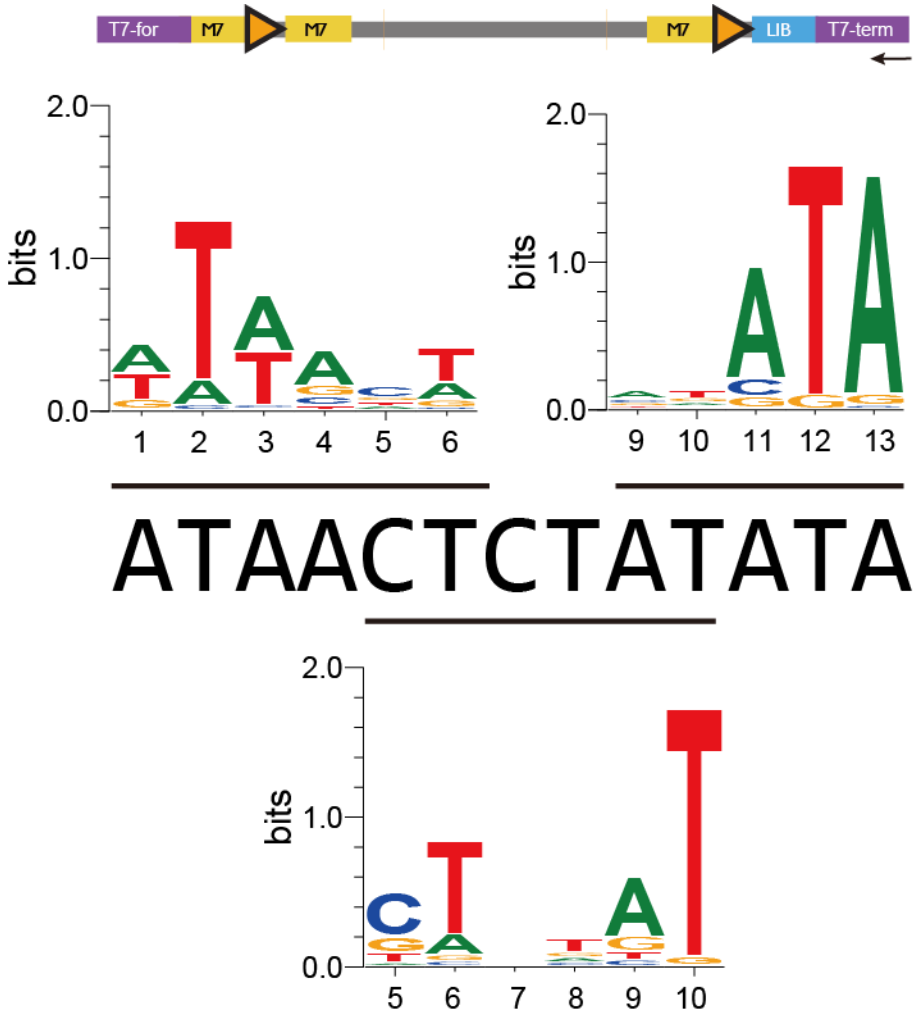


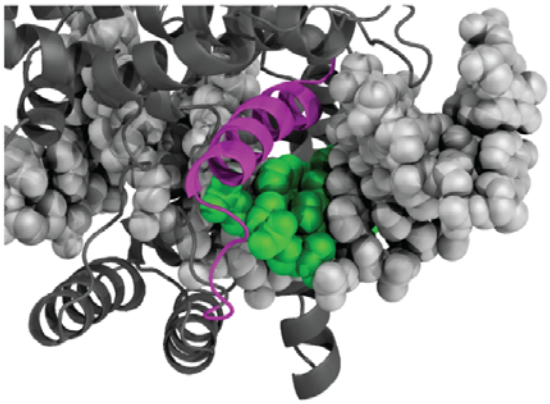
Figure 3.5 Sequence preferences of CreC2#4 recombinase in the arm region

Rec-Seq experiment results are presented as sequence logos for the CreC2#4 altered specificity mutant. The substrate library consisted of three loxM7 arm regions and one arm region that randomized bases in a loxM7 background in three windows. This mutant was selected to recombine the sequence of CTA at positions 7-9 of the arm region. The results show that specificity was indeed switched at positions 8 and 9, but that specificity is lost at position 7.

A

Arm Region	Source
ATAACTTCGTATA	loxP
CTAACTTTATATA	γ-proteobacterium WG36
CTGACTTTATATA	Pantoea sp. A4
CTAACTTTACATA	Pantoea sp. At-9b
CTGCCCTTATATA	Pantoea rodasii
ATAACTTTATATA	loxTA

B



C

Residues 256-279 (Cre numbering)	
LSTRALEGIFEATH	Wild-type Cre
LSAPALEAIFARAW	γ-proteobacterium WG36
LTLPALEAIFQNAW	Pantoea sp. A4
LSTPAVEDIFRKAW	Pantoea sp. At-9b
LSAPCLEDIWRRAT	Pantoea rodasii
LSTPALEGIFEATH	Cre R259P

D

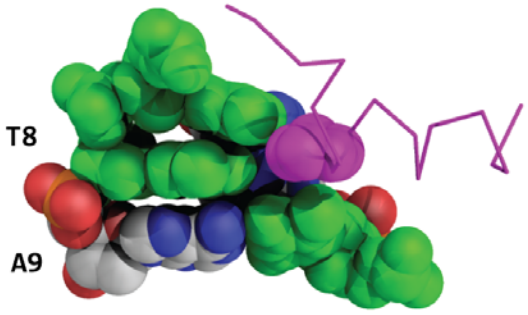


Figure 3.6 Identification of a specificity determinant *via* sequence comparison of recombinase homologs.

(A) An alignment of the known and predicted arm region sequences for Cre recombinase and close homologs are shown, with the source organisms noted. Each homolog recognizes an RT with the sequence TA at positions 8 and 9 (shown in green). The *loxTA* arm region is an altered *loxP* sequence with only these two base substitutions. (B) The model for the Cre-*loxP* complex (PDB code: 1KBU (24)) is shown, with positions 8 and 9 of the arm region rendered in green. Residues 256-279 of the protein (rendered in magenta) are in close proximity to these bases. (C) An alignment of the amino acids for each of the Cre homologs in the range of residues 256-279 (Cre numbering) is shown. The consistent difference between the homologs and Cre itself is an arginine to proline substitution at position 259. We use Cre259P to denote a Cre

variant with this single mutation. (D) A structural model of the region of interaction between CreR259P and *loxTA* was predicted using the Rosetta design program. The substituted proline residue (rendered as spheres and colored magenta) is predicted to make hydrophobic interactions with the methyl groups of the thymine bases at positions 7 and 8, and on the opposite strand from position 9 (shown in yellow).

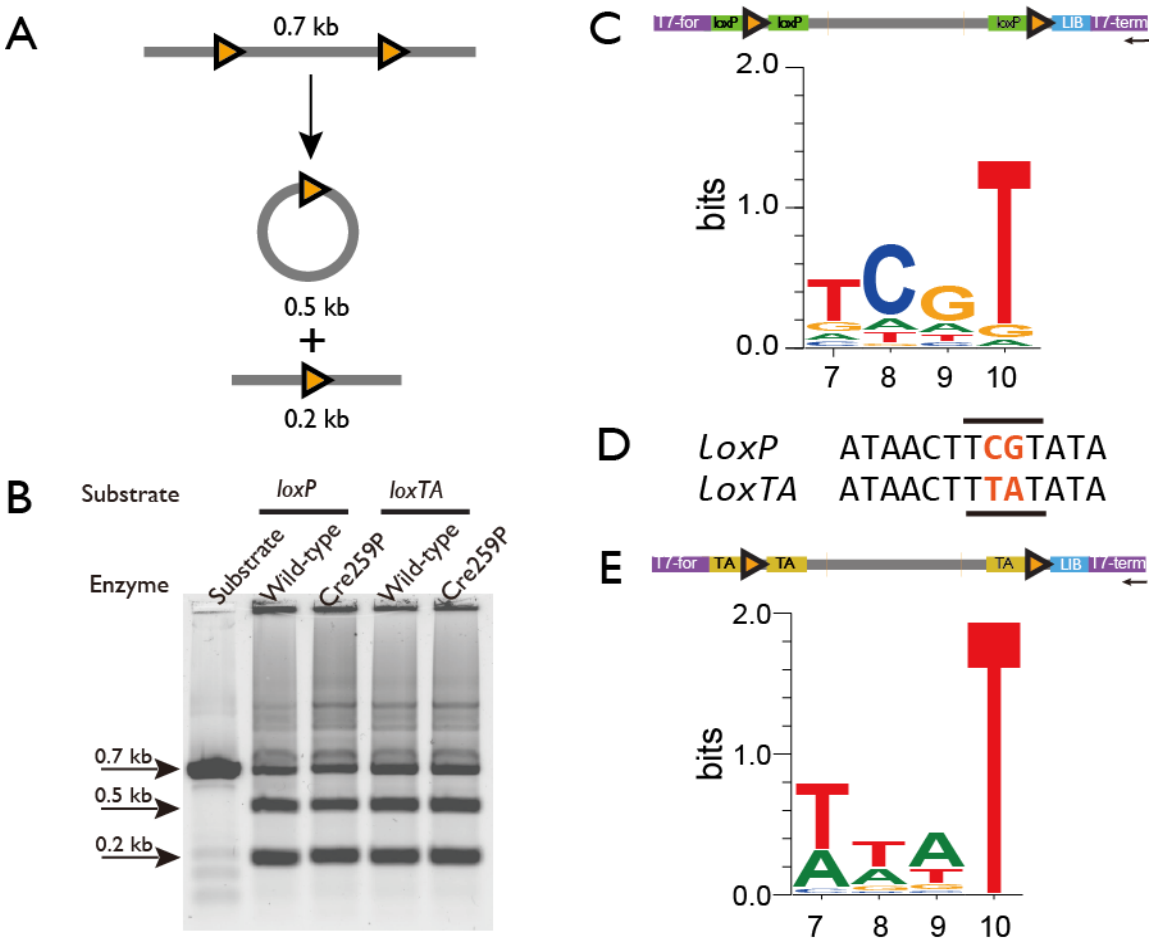


Figure 3.7 Validation of R259P specificity determinant by Rec-seq

(A) *In vitro* assay for recombinase activity. The substrate for the assay is a linear DNA molecule harboring direct repeats of either the *loxP* or *loxTA* RT. Recombination results in generation of smaller circular and linear products. (B) Both wild-type and Cre259P recombinases were incubated with substrates containing both *loxP* and *loxTA* repeats in the presence of 120mM NaCl. We observe activity for each combination, indicating that DNA sequence preferences are too subtle to observe with a bulk assay such as this. (C) The results from a Rec-Seq experiment are shown in logo form for the window covering positions 7-10 of *loxP* (these

columns were extracted from the 6 bp window experiments shown in Fig. 3). (D) An alignment of *loxP* with *loxTA* is shown, with the differing bases colored red, and with black bars indicating the positions shown in the sequence logos. (E) A sequence logo for the results of a Rec-Seq experiment randomizing the four bases at positions 7-10 for *loxTA* is shown. The most enriched sequence corresponds to *loxTA*, indicating that the R259P mutation results in the desired switch in specificity.

Chapter 4

Conclusions and Future Directions

Design of recombinases for asymmetric sites

The tetrameric protein complex of Cre recombinase imposes a symmetric constraint on the RT sites it recombines. As a result, the utility of recombinase mutants with altered DNA specificity is limited if the pool of potential genomic targets must satisfy this constraint. A pair of 'orthogonal' recombinases with independently altered DNA specificities can overcome this challenge, allowing for the recombination of asymmetric sites. In chapter 2, I used both computational and rational design to construct a pair of obligate heterotetrameric Cre monomers: these mutants are active when combined together, but inactive in isolation. I found it relatively easy to reconstruct the protein-protein interface to find a pair of monomers that could work together. However, it is much harder to completely "knockout" the activity of the monomers in isolation. In chapter 2, I found that three successive rounds of design were necessary to achieve a completely orthogonal pair of Cre recombinases.

Prospects for additional heterotetrameric recombinase pairs

Recombinase-mediated cassette exchange (RMCE) at native loci is a promising strategy to precisely replace defined intervals in the genome with any sequence provided in a DNA template. During the cassette exchange, there are two distinct recombination crossover events, requiring two sets of orthogonal recombinases. For maximal specificity, we require a second pair of obligate heterotetramer recombinase mutants. One possible way to finding such a second pair of recombinases is to redesign another novel interface for Cre monomers and demand that it to be incompatible with both the wild-type interface and the redesigned Cre-A3/Cre-B2 interface. Considering that I found the negative elements of interface design to be more difficult to satisfy than the positive elements, this approach seems difficult to achieve. Another strategy is to engineer homolgous DNA recombinases that will not cross-react with Cre. A group of

researchers have used both Cre and Flp to demonstrate RMCE (1). They found that Flp is not as efficient as Cre *in vivo,* and that a careful balance of the proteins are needed to optimize cassette exchange efficiency.

Several new Cre homologs have been found by screening bacterial genomes. Examples include Vika, SCre and VCre (2, 3). They share around 30% sequence identity to Cre and recombine distinct RT sites. It is likely that they will not cross react with each other, making them ideal candidates to engineer for RMCE. The caveat is that these recombinases are less well studied. In particular, there are no crystal structures available for these homologs. Therefore, interface design for these proteins relies on guessing the contacting residues by mapping their protein sequence onto Cre's crystal structure. Using this strategy, I have tried to engineer the protein interfaces of Vika and VCre recombinase. For Vika recombinase, I used rational design and introduced "charged swaps" into the interface. I found a pair of mutations (R33D and E72R) and an *in vivo* assay showed that they were active when combined together. However, one of the mutants has reduced but clear activity in isolation. Additional mutations are needed to engineer a completely orthogonal pair. For VCre, more aggressive measures were taken. I have tried exchanging pairs of interacting alpha-helical elements of secondary structure between VCre and a close homolog. The resultant VCre variants were non-functional and additional mutations to other residues that intact with the helix are required to rescue the activity.

A future approach to overcome the limited success of sequence-based methods may involve crystallographic efforts to obtain experimentally determined structures for VCre, Vika, or other recombinases. This would allow for the application of the structure-based strategies that I used successfully with Cre recombinase. The availability of many recombinase homologs, with associated predictions for RT sites, will allow us to select recombinases that target diverse RT

sequences for structural characterization, which would simultaneously expand the genomic sequences we can target for cassette exchange.

Rec-Seq is a powerful tool to study the DNA specificity of recombinases

Using recombinases in genome editing not only requires the knowledge of their native RT sites, but also requires a quantitative understanding of base preferences at all of the positions in the RT site. In chapter 3, I described a method call 'Rec-Seq' to fully characterize the DNA specificity in the arm region using high-throughput sequencing. This assay is also crucial for testing and validating recombinase variants obtained from directed evolution.

A possible future direction for Rec-Seq would be to adapt the method for *in vivo* recombination. As described in Chapter 3, the recombination step was performed *in vitro*. Compared to *in vivo* methods, *in vitro* recombination has several advantages. First, the randomized library size is not limited by bacterial transformation. As long as there is detectable product, there is no limit on how many positions can be randomized at the same time. Second, Cre's strong cooperative binding of the RT site (which has hindered previous specificity assays) can be reduced by adjusting the NaCl concentration. It is difficult to change the conditions similarly *in vivo*. On the other hand, recombination *in vitro* requires a time consuming protein purification step, and some recombinases do not express well or are not sufficiently active *in vitro*. As a result, I could not use Rec-Seq to test several recombinases of significant interest. For example, the wild-type Vika recombinase does not express well, and the engineered Fre22 and Tre recombinases have prohibitively weak activity in our *in vitro* assay, although they exhibit activity in bacterial or mammalian cells. Thus, it would be advantageous to develop a Rec-Seq method *in vivo*. The *in vitro* DNA substrate used in chapter 3 can be cloned into a pBAD33

vector where the corresponding recombinases express under the control of arabinose induction. After transformation and arabinose induction in *E. coli*, I can perform a miniprep, use similar procedures to PCR amplify the recombined product and send the amplified DNA to Illumina sequencing. By controlling the recombinase expression with arabinose levels, this proposed *in vivo* method can be used to detect recombinases that have little activity *in vitro*. One drawback of this in vivo method is that the recombinases, especially wild-type Cre, may be too promiscuous to yield highly specific recombination product. It is necessary to tune the arabinose level and modify the substrate sequence to obtain optimal recombination specificity.

Accelerating the retargeting of recombinases

Compared to the ease with which the CRISPR/Cas system can be retargeted, altering the RT site specificity of recombinases is a laborious process, involving many rounds of directed evolution. In a recent effort, Buchholz and colleagues applied a total of 126 rounds of directed evolution to evolve Cre to target long-terminal repeat of HIV-1 strain (4). The number of rounds of directed evolution could be reduced if the starting protein could be made more compatible with the target RT site by rational methods. In chapter 3, I used sequence alignment information from Cre homologs to find candidate mutations that can change the DNA specificity of Cre. I then validated a candidate mutation using Rec-Seq and confirmed the specificity switch. From the sequence alignment, I could build a database of mutations and their corresponding changes of RT site specificities.

Rec-Seq could also improve selection of site for cassette exchange in the human genome. In Chapter 3, I used Rec-Seq to identify the specificity of recombinases at each position of the RT sequence. The combination of knowing the positions on the RT sequence that are difficult to

alter and the database of specificity altering mutations can guide the selection of which DNA elements within a given genomic interval are the best candidates to target with a recombinase. As a demonstration, we searched the sequences upstream and downstream of Huntington's disease genome locus to do RMCE and found candidate target sites. We identified sites in the regions upstream and downstream of the huntingtin locus that we predict could be targeted by the VCre and Cre recombinases, respectively. Early directed evolution results using these recombinases suggest that these sites can indeed be targeted.

Rec-Seq can also serve as a test for when the cycle of directed evolution selections can be terminated. Rec-Seq can distinguish recombinases that are specific from those that are promiscuous within a given window along the arm region. In reported work, a recombinase that was evolved for one sequence exhibited even stronger activity for a different, off-target RT site (5). Rec-Seq could catch this lack of desired specificity before extensive efforts are expended in cell culture experiments.

References

1. Anderson,R.P., Voziyanova,E. and Voziyanov,Y. (2012) Flp and Cre expressed from Flp-2A-Cre and Flp-IRES-Cre transcription units mediate the highest level of dual recombinase-mediated cassette exchange. *Nucleic Acids Res* **40**, e62.

2. Karimova,M., Abi-Ghanem,J., Berger,N., Surendranath,V., Pisabarro,M.T. and Buchholz,F. (2013) Vika/vox, a novel efficient and specific Cre/loxP-like site-specific recombination system. *Nucleic Acids Res* **41**, e37.

3. Suzuki,E. and Nakayama,M. (2011) VCre/VloxP and SCre/SloxP: new site-specific recombination systems for genome engineering. *Nucleic Acids Res* **39**, e49.

4. Sarkar,I., Hauber,I., Hauber,J. and Buchholz,F. (2007) HIV-1 proviral DNA excision using an evolved recombinase *Science* **316**, 1912.

5. Bolusani,S., Ma,C.H., Paek,A., Konieczka,J.H., Jayaram,M. and Voziyanov,Y. (2006) Evolution of variants of yeast site-specific recombinase Flp that utilize native genomic sequences as recombination target sites *Nucleic acids research* **34**, 5259.

Appendix I

Supplemental Materials for Chapter 2

Supplemental Methods

Protein purification of Cre recombinase variants.

Proteins were expressed in BL21(DE3) star cells at 25°C using the autoinduction protocol of Studier (1). The cells were harvested by centrifugation after 48 hours. The cell paste was resuspended in 25 mL buffer A (0.7 M NaCl, 50 mM Tris-HCl pH 7.8, 5 mM Imidazole), lysed by sonication on ice, and separated from cellular debris by centrifugation. The filtered supernatant was applied to a HisTrap^{TM} HP column (Amersham) and washed with 30 mL Buffer A. The column was then washed with 20 mL 15% buffer B (0.7 M NaCl, 50 mM Tris-HCl pH 7.8, 500 mM Imidazole). Cre was eluted with a linear gradient from 15% buffer B to 100% buffer B, with the elution peak starting at roughly 20% buffer B. Approximately 10 mL of the eluted protein was collected and dialyzed overnight at 4°C against 5 L dialysis buffer (0.7 M NaCl, 50 mM Tris-HCl pH 7.8). The protein concentration was then determined by UV absorbance using an extinction coefficient at 280nm of 49 $mM^{-1}cm^{-1}$. The protein retained activity for months when stored at 4° C.

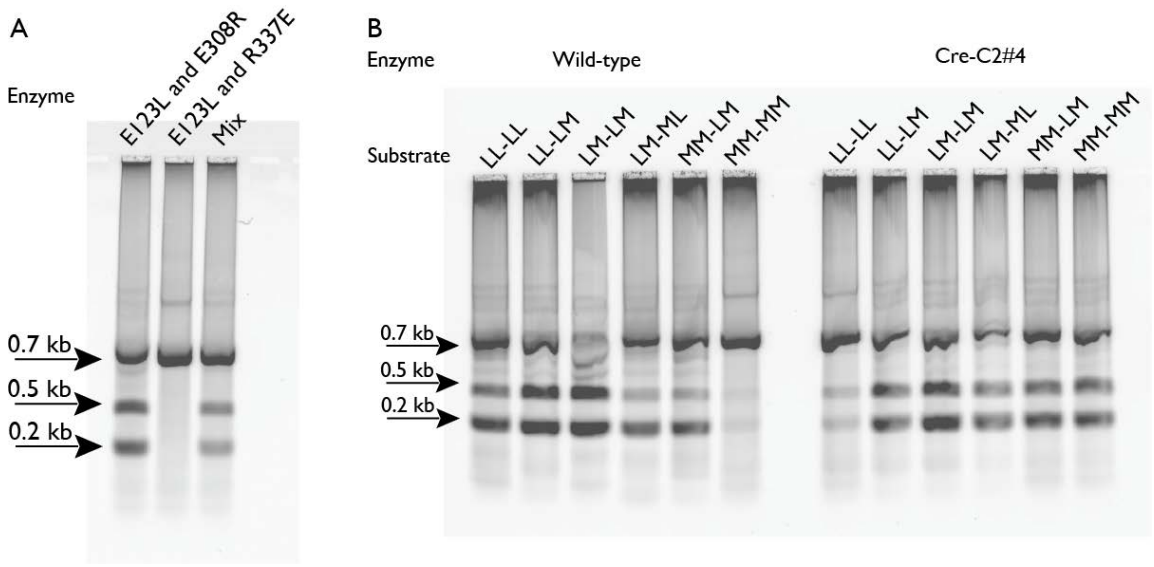
References

1. Studier,F.W. (2005) Protein production by auto-induction in high-density shaking cultures *Protein Expr Purif* **41**, 207-234.

		Replicate		
		1st	2nd	3rd
total # of cells sorted		7000	7000	7000
Cre-A1	hbb	414	378	391
	hbb+cmv	3852	3528	3687
	hbb+sp1	3750	3419	3501
Cre-B1	hbb	97	102	85
	hbb+cmv	1237	1258	1120
	hbb+sp1	1150	1080	1202
A1+B1	hbb	1117	1212	1324
	hbb+cmv	5866	6029	6358
	hbb+sp1	5702	6121	5987
Cre-A2	hbb	47	52	41
	hbb+cmv	1127	1116	1052
	hbb+sp1	1053	1002	1119
Cre-B2	hbb	0	0	1
	hbb+cmv	2	2	4
	hbb+sp1	2	1	3
A2+B2	hbb	573	528	607
	hbb+cmv	3180	3409	3698
	hbb+sp1	3221	3336	3593
Cre-A3	hbb	0	0	1
	hbb+cmv	0	1	1
	hbb+sp1	1	0	0
A3+B2	hbb	256	233	284
	hbb+cmv	1598	1652	1701
	hbb+sp1	1503	1527	1606
WT	hbb	372	391	408
	hbb+cmv	3914	4223	4312
	hbb+sp1	3815	3799	4021

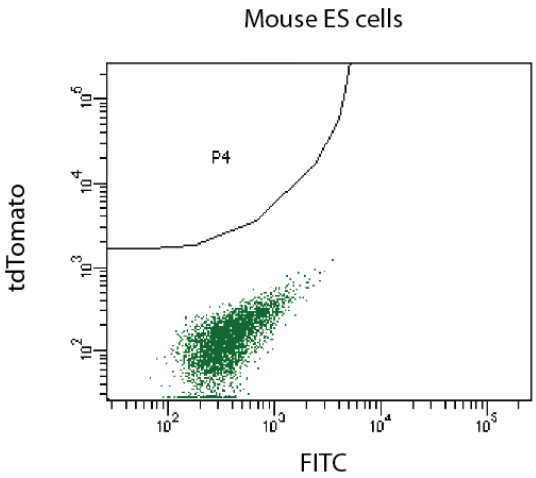
Supplemental Table 2.1. Cell sorting data from mouse ES cells

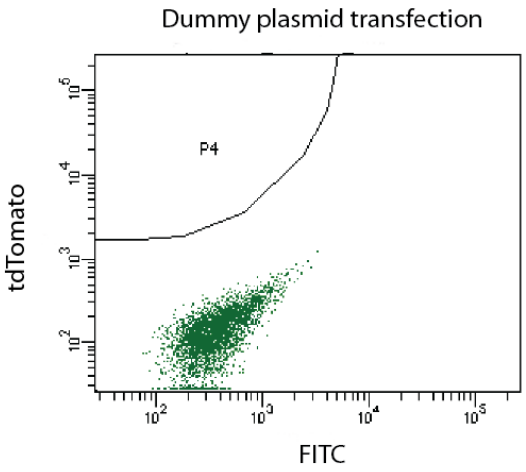
Plasmids with the hbb minimal promoter alone or with either the cmv and sp1 enhancers driving different cre variants were co-transfected into Ai14 mouse embryonic stem (ES) cells containing a reporter cassette with tdTomato preceded by a floxed stop codon. The same total amount of DNA was used for all transfections, and 3 independent transfections were performed for each Cre variant. The number of tdTomato positive cells was measured by flow cytometry.

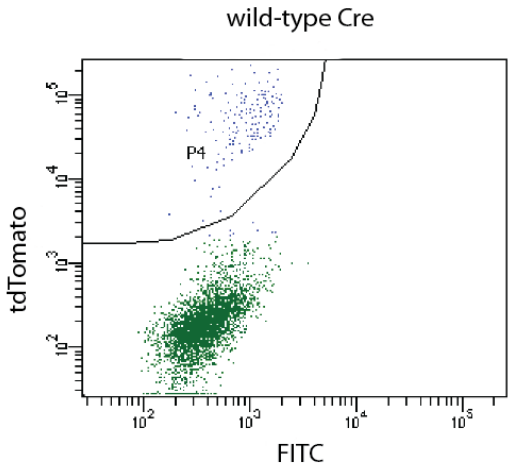


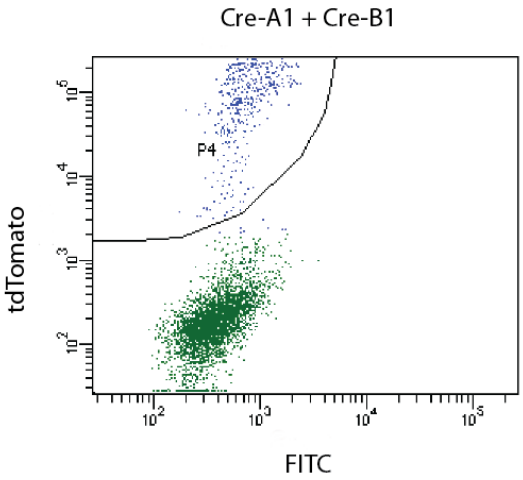
Supplemental Figure 2.1

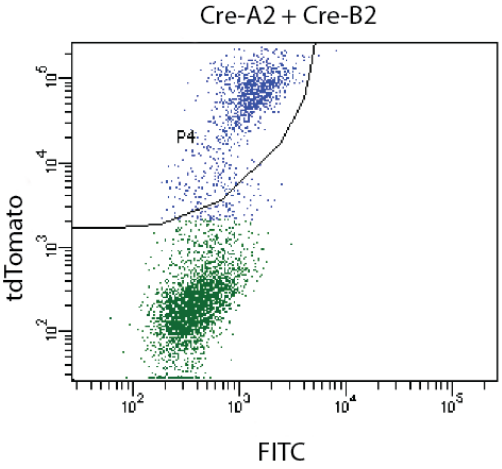
(A) *In vitro* assay results for Cre mutant pairs lacking computationally designed mutations. Linear DNA substrate (0.7 kb) with direct loxP repeats was incubated with Cre mutants. Lane 1: Cre-E123L/E308R; lane 2: Cre-E123L/R337E; lane 3: A 1:1 mixture of above two Cre mutants. The E123L/E308R mutations are insufficient to eliminate activity in this monomer, indicating that additional mutations are necessary to achieve the goal of obligate heterotetramers. (B) *In vitro* assay results for Cre proteins with wild-type monomer-monomer interfaces. Wild-type Cre and Cre-C2#4 were assayed for recombination activity against six loxP/M7 hybrid RT sites. The left panel: wild-type Cre recombined robustly on all six RT sites except for all M7 site. The right panel: Cre-C2#4 recombined all six RT sites, although with diminished activity with increased number of loxP half-sites.


Mouse ES cells
P4
tdTomato
FITC


Dummy plasmid transfection
P4
tdTomato
FITC


wild-type Cre
P4
tdTomato
FITC


Cre-A1 + Cre-B1
P4
tdTomato
FITC


Cre-A2 + Cre-B2
P4
tdTomato
FITC

Supplemental Figure 2.2.

Representative raw data from flow sorting experiments. Each point shows the fluorescence in the red channel (tdTomato) versus green channel (FITC). The cell-only and dummy plasmid experiments exhibit roughly identical autofluorescence. The gating for identifying RFP-positive is the region of each plot labeled 'P4'.

Appendix II

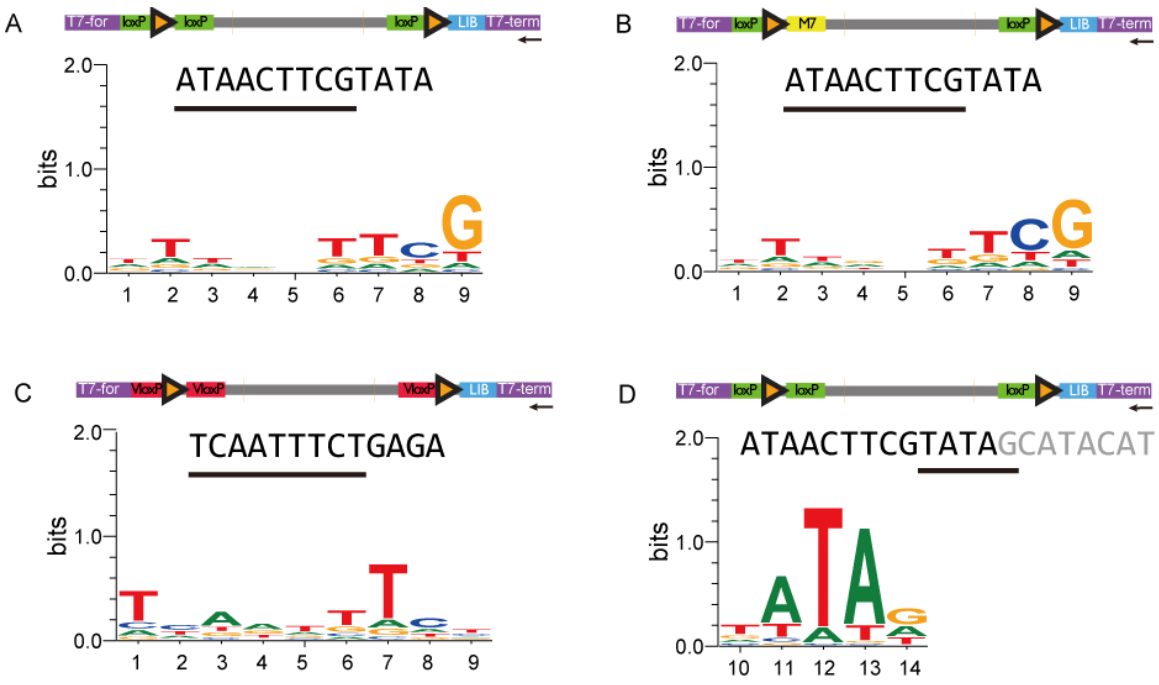
Supplemental Materials for Chapter 3

	Protein	Left Target	Right Target(reverse complement)	NaCl(mM)	Result
1	CreWT	LoxP-LoxP	**NNNNNNNNNNNNN** GCATACAT TATACGAAGTTAT	120	No recombination
2	CreWT	LoxP-LoxP	**NNNNNNNNN**TATA GCATACAT TATA**TAG**AGTTAT	120	Figure 3.2A
3	CreWT	LoxP-LoxP	**NNNNNNNNN**TATA GCATACAT TATA**TAG**AGTTAT	170	Figure 3.2B
4	CreWT	LoxP-LoxP	**NNNNNNNNN**TATA GCATACAT TATA**TAG**AGTTAT	200	Figure 3.2C
5	CreWT	LoxP-LoxP	**NNNNNNNNN**TATA GCATACAT TATA**TAG**AGTTAT	210	Figure 3.2D
6	CreWT	LoxP-LoxP	**NNNNNNNNN**TATA GCATACAT TATA**TAG**AGTTAT	220	No recombination
7	CreWT	LoxP-LoxP	**NNNNNNNNN**TATA GCATACAT TATACGAAGTTAT	460	Supplemental figure 3.1A
8	CreWT	LoxP-loxM7	**NNNNNNNNN**TATA GCATACAT TATACGAAGTTAT	420	Supplemental figure 3.1B
9	CreWT	LoxP-LoxP	**NNNNNN**TCGTATA GCATACAT TATACGAAGTTAT	490	Figure 3.3
10	CreWT	LoxP-LoxP	ATAA**NNNNNN**ATA GCATACAT TATACGAAGTTAT	490	Figure 3.3
11	CreWT	LoxP-LoxP	ATAACTT**NNNNNN** GCATACAT TATACGAAGTTAT	120	No recombination
12	CreWT	LoxP-LoxP	ATAACTTC**NNNNN** GCATACAT TATACGAAGTTAT	470	Figure 3.3
13	CreWT	LoxP-LoxP	ATAACTTCG**NNNN N**CATACAT TATACGAAGTTAT	500	Supplemental figure 3.1D
14	CreWT	LoxP-LoxP	ATAACTTCGTAT**N NNNN**ACAT TATACGAAGTTAT	120	No recombination
15	CreC2#4	LoxM7-LoxM7	**NNNNNNNNN**TATA GCATACAT TATA**TAG**AGTTAT	370	Non-specific
16	CreC2#4	LoxM7-LoxM7	**NNNNNNNNN**TATA GCATACAT TATACGAAGTTAT	340	Non-specific
17	CreC2#4	LoxM7-LoxM7	**NNNNNNCTA**TATA GCATACAT TATA**TAG**AGTTAT	370	Figure 3.5
18	CreC2#4	LoxM7-LoxM7	ATAA**NNNNNN**ATA GCATACAT TATA**TAG**AGTTAT	420	Figure 3.5
19	CreC2#4	LoxM7-LoxM7	ATAACT**CTNNNNN** GCATACAT TATA**TAG**AGTTAT	280	Figure 3.5
20	Cre259P	LoxTA-LoxTA	ATAAC**NNNNNN**TA GCATACAT TATA**TA**AAGTTAT	120	No recombination
21	Cre259P	LoxTA-LoxTA	ATAACT**NNNN**ATA GCATACAT TATA**TA**AAGTTAT	120	Figure 3.7D
22	VCre	Vlox-Vlox	**NNNNNNNNN**GAGA ACTGTCAT TCTCGGAAATTGA	300	Supplemental figure 3.1C
23	VCre	Vlox-Vlox	**NNNNNN**TCTGAGA ACTGTCAT TCTCGGAAATTGA	300	Figure 3.4
24	VCre	Vlox-Vlox	TCAA**NNNNNN**AGA ACTGTCAT TCTCGGAAATTGA	300	Figure 3.4
25	VCre	Vlox-Vlox	TCAATTTC**NNNNN** ACTGTCAT TCTCGGAAATTGA	300	Figure 3.4
26	VCre	Vlox-Vlox	TCAATTTC**NNN**GA **N**CTGTCAT TCTCGGAAATTGA	70	No recombination

Supplemental Table 3.1 Summary of recombination experiments.

Each line corresponds to a single recombination experiment, as described in Figure 1 of the manuscript. The first column indicates the recombinase or recombinase mutant used in the

experiment. The second column gives the sequence of the (constant) recombinase target (RT) site on the ‘left-hand’ side of the DNA substrate, where each half-site of the RT sequence is indicated, as in some cases an asymmetric RT sequence was used. The third column gives the sequence of the ‘right-hand’ RT site on the substrate, with randomized positions denoted by ‘N’. The fourth column indicates the salt concentration used in the experiment. The fifth column indicates the result of the experiment. ‘No recombination’ denotes experiments where no recombinant product was observed. ‘Non-specific’ denotes experiments without a clear product of the expected size. Otherwise, the figure that presents the results of successful experiments is reported.



Supplemental figure 3.1 Extended library assays for the Cre and VCre recombinases.

(A)-(C) Results from recombination specificity assays with high complexity libraries that randomize the outer nine bases of the RT sequence are presented. (A) Results for wild-type Cre recombinase with the indicated DNA substrate, in which all non-random half-sites were the loxP sequence. (B) Results for wild-type Cre recombinase where one of the half-sites consisted of the loxM7 sequence. (C) Results for the VCre recombinase where all non-random half-sites are the cognate VloxP sequence.

(D) Results from wild-type Cre specificity assays that randomizes four bases in the arm region and one base in the spacer region of the RT sequence are presented. All non-random half-sites are the loxP sequence.

Primers for library generation (See Supplemental Table 1 for RT site sequence)

Library-forward

cccgcgaaattaatacgactcactatagggg**-RT-site-**ccaattgtccatattgcatcagac

Library-reverse

gggttatgctagttattgctcagcggtggcag**-RT-site-with-randomized-library–**caacagataaaacgaaaggcccag

Primers for Illumina sequencing preparation

Selected-sequencing-forward

AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTcccgcg aaattaatacgactcactatagggg

Sequencing-reverse

CAAGCAGAAGACGGCATACGAGAT-index-GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTgggttatgctagttattgctcag

Unselected-sequencing-forward

AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTgccgta gcgccgatggtagtgtggggtctccc