DNA Recombinases as Genome Editing Tools

DNA Recombinases as Genome Editing Tools

A dissertation presented

by

Jeffrey Lawrence Bessen

to

The Department of Chemistry and Chemical Biology

in partial fulfilment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Chemistry and Chemical Biology

Harvard University

Cambridge, Massachusetts

April 2019

DNA Recombinases as Genome Editing Tools

Abstract

Site-specific recombinases (SSRs) have the potential to serve as ideal genome editing agents because they catalyze precise and efficient DNA strand exchange, but their innate specificity limits their applicability to a narrow range of DNA sequences. I have investigated several paths toward developing SSRs as viable genome editing tools. First, I describe the laboratory evolution of ROSACre, a variant of Cre recombinase that recognizes a human genomic target, using phage-assisted continuous evolution (PACE). We developed a PACE selection for recombinases and used it to evolve Cre to target a sequence in a genomic safe harbor. We demonstrated that ROSACre variants possess activity in mammalian cells on a target identical to a sequence within the human ROSA26 locus. Subsequently, I describe several alternative strategies, including adaptations of PACE as well as independent selections, in efforts to improve the activity and specificity of the resulting enzyme variants.

Next I describe the development recCas9, an RNA-programmed small serine recombinase that functions in mammalian cells. We fused a catalytically inactive Cas9 to the catalytic domain of Gin recombinase using an optimized fusion architecture. The resulting recCas9 system recombines DNA sites containing a minimal recombinase core site flanked by guide RNA (gRNA) specified sequences. We show that recCas9 can operate on DNA sites in mammalian cells identical to genomic loci naturally found in the human genome in a manner that is dependent on the gRNA sequences. DNA sequencing reveals that recCas9 catalyzes gRNA-dependent recombination in human cells with efficiency as high as 32% on plasmid substrates. Finally, we demonstrated that recCas9 expressed in human cells can catalyze *in situ* deletion between two genomic sites. Additionally, I describe efforts to improve the first-generation recCas9 construct by fusion of alternative recombinase domains.

The engineering or evolution of SSRs into more versatile genome editing agents is limited in part by an incomplete understanding of SSR protein:DNA specificity determinants. To

address this challenge, I describe the development of Rec-seq, a method for revealing the DNA specificity determinants and potential off-target substrates of SSRs in a comprehensive and unbiased manner. We applied Rec-seq to characterize the DNA specificity determinants of several natural and evolved SSRs including Cre, evolved variants of Cre, and other SSR family members. Rec-seq profiling of these enzymes and mutants thereof revealed previously uncharacterized SSR interactions, including specificity determinants not evident from SSR:DNA structures. Finally, we used Rec-seq specificity profiles to predict off-target substrates of evolved Cre variants Tre and Brec1, including endogenous human genomic sequences, and confirmed their ability to recombine these off-target sequences in human cells.

This dissertation is dedicated to the memory of my uncle Dr. Richard Bessen and my grandfather Howard Bessen, who inspired my scientific curiosity since childhood.

Table of Contents

List of Figures and Tables

Chapter 1:

The Potential of DNA Recombinases as Genome Editing Tools

1.1 Prologue

With rapid-fire publications in prestigious science journals, blockbuster Hollywood movies, and scandals that made front-page headlines across the globe, few scientific topics have garnered as much attention in the last 5 years as genome editing. This cultural phenomenon has formed an exciting backdrop for my graduate research of genome editing proteins. In the fall of 2012, at the same time I was applying to graduate school, the landmark papers describing the CRISPR/Cas9 genome editing technology were published[1-3]. By the time I arrived on campus the next year, the journal *Science* referred to the flurry of follow-up studies as "the CRISPR craze"[4]. With those initial publications, the immense potential of the CRISPR/Cas system thrust genome editing to the scientific and cultural forefront. Dreams of fantastical cures, fears of biotechnology run amok, and serious discussions about ethical quandaries are all heard in the current conversation about gene editing. Cas9 has been hailed as both humanity's savior and its downfall.

The current excitement about genome editing has its roots in decades of research into manipulating the genome. Since the 1940's, scientists have known that the genome encodes proteins[5], and subsequent research has elucidated the many complex mechanisms of gene regulation. Testing hypotheses about genome function, studying the impact of genomic perturbations, and manipulating genomic sequences for therapeutic purposes all require the ability to precisely alter the sequence of DNA bases within the genome. Given that one human genome contains approximately 6 billion base pairs, this alone is no trivial task. But researchers attempting to devise a general tool for genome manipulation also face numerous other complicating factors, including variable cell states and cellular environments, and the challenge of delivering macromolecular genome editing agents into living cells.

The potential payoff for genome editing success is difficult to overstate. Researchers using CRISPR/Cas9 or related technologies have taken steps toward cures for Duchenne muscular dystrophy[6], various forms of cancer[7], HIV[8], metabolic disease[9], Alzheimer's disease[10],

heart disease[11], genetic deafness[12], Huntington's disease[13], sickle cell disease[14], cystic fibrosis[15], and many other unmet medical needs. Genome editing in insects and plants has raised hopes for gene drives that could eliminate the mosquito species responsible for malaria[16] or crops with improved properties[17]. In the lab, the applications of genome editing are nearly limitless, from inquiries into the developmental fate of multipotent cells[18] and the genetic roots of cancer[19] to applications such as cellular computers[20] and metabolic engineering[21]. Some of these applications are already underway; others, meanwhile, await improvements in genome editing technology before they reach consumers or patients.

Efficient, programmable genomic modification, and specifically gene integration, remains a longstanding goal of genetics and genome editing[22]. While many researchers have spent the past 6 years trying to realize the potential of precise and efficient genome modification using CRISPR/Cas9 or similar technology, I have been drawn to a different class of proteins: site-specific recombinases (SSRs). Recombinases possess a tantalizing capability – catalysis of highly precise and efficient genome modification – as well as a critical limitation, an innate DNA preference and thus a barrier to retargetability, which has prevented widespread embrace by the genome editing community. With my studies, and with this dissertation, I seek to answer the following questions: Can laboratory engineering and evolution yield clinically useful recombinase variants? How should the development of SSRs be carried out? And, ultimately, what role will DNA recombinases play alongside the genome editing technologies of the future?

1.2 A brief history of genome editing research

Before large-scale efforts at genomic modification could be attempted, researchers first had to master the technique on a smaller scale, both *in vitro* and in simple model organisms such as *E. coli.* This early molecular biology research was enabled by the development of recombinant DNA in the 1970's[23]. Armed with recombinant DNA technology, genome editing researchers achieved gene integration through random uptake of foreign DNA, viral mediated

gene transfer[24], or transposon mutagenesis[25]. Researchers accomplished precision integration, or “gene targeting”, using a DNA donor capable of homologous recombination with the genome[26,27]. However, gene targeting relies on low-frequency integration events, and thus screening or selection of many cells is required to isolate the desired product[28].

Subsequently, the discovery that double-stranded DNA breaks (DSBs) increase the rate of homologous recombination and local mutagenesis was a critical breakthrough for genome editing research[29]. Relatively efficient genome modification at a specific locus was thus reduced to the challenge of promoting the desired DSB. Following a genomic cleavage event, cellular repair responses are activated at the site of the double-stranded break. In mammalian cells, the major outcomes of DSB repair include error-prone processes such as non-homologous end joining (NHEJ)[30,31], which can introduce insertions or deletions at the DSB site, and homology-directed repair (HDR)[32,33]. Error-prone repair of a DSB located within a gene often results in inactivation of that gene, a desirable outcome for studying genomic knockouts or disabling a disease-causing gene. Alternatively, HDR using a researcher-defined repair template can result in genomic integration, albeit at low efficiency; NHEJ and other error-prone processes occur at much higher rates than HDR, especially in non-mitotic cells[34,35].

The efficient introduction of DSBs and subsequent promotion of the desired repair outcome have become twin goals of modern genome editing research. The first of these goals – introducing DSBs at a desired location in the genome - has been largely achieved. While early studies used homing endonucleases to predictably generate DSBs[36], the arrival of programmable DNA-binding proteins ushered in the current golden age of genome editing. Covalent linkage of a nuclease domain[37] to an array of modular DNA-binding domains, such as Zinc fingers[38,39] or TALEs[40,41], enabled the facile introduction of DSBs at user-defined DNA sequences. These programmable nuclease systems have largely been eclipsed by the widespread adoption of CRISPR/Cas9 beginning in 2013. This system offers an advantage over its predecessors because DNA cleavage by Cas9 is defined by a guide RNA (gRNA) and not

the nuclease itself. Thus, targeting a new DNA sequence requires making the complementary gRNA, and not protein reengineering. The Cas9 protein does have innate DNA preference for a protospacer-adjacent motif (PAM), limiting its applicability. However, researchers continue to discover or engineer new Cas9 variants with different PAM requirements that can target unaddressed regions of the genome[42-44].

Comparatively little progress has been made toward enhancing HDR efficiency under clinically-relevant conditions. HDR rates vary based on the method used, the cell type, the cell state, and the genomic location[33,45]. The most advanced methods for enhancing HDR using Cas9 have maximum efficiencies in the single- or low double-digit percentages, with a concomitant excess of indels at the editing site[46-48]. Further, the use of programmable nucleases in living cells has been associated with unwanted editing at off-target loci, translocations or other DNA arrangements, and p53 activation[49-53]. Finally, recent findings suggest that cellular therapies involving CRISPR components may trigger an immune response in patients[54,55]. While efforts to address these shortcomings are underway, there remains strong demand for a general technology for efficient and predictable homologous recombination at a user-defined locus.

1.3 Introduction to site-specific recombinases

SSRs represent an alternative approach to precise genomic modification. SSRs are a broad class of enzymes that directly catalyze strand exchange between DNA molecules[56]. As implied by their name, SSRs have innate specificity for their cognate target sequence. While some SSRs form multi-protein complexes or have expansive binding and topological requirements[57,58], simpler family members require no accessory proteins and recognize targets that range from approximately 20-50 base pairs in length. My research has focused on these simpler SSR family members.

SSRs are classified as either tyrosine or serine SSRs based on the identity of the catalytic residue. While members of the two families perform recombination through distinct

mechanisms, there are many similarities between the respective recombination processes[56]. For both enzyme classes, recognition targets can be divided into half-sites flanking a core sequence (Figure 1.1). The half-sites often consist of inverted repeats, and during recombination, each half-site is bound by a recombinase monomer. These dimers assemble into a homotetrameric complex and catalyze strand exchange between the core sequences of two recombinase targets. Typically, productive recombination requires that the core sequences of two recombinase targets are complementary[59]. The asymmetric core sequence imparts an overall directionality to the recombinase target, and recombination outcomes are dictated by the orientation and location of the two target sites (Figure 1.1). For example, recombination between two targets in the same orientation on the same DNA molecules results in deletion of the intervening sequence. The reverse of this reaction yields the integration of two DNA molecules. When two targets appear on the same DNA molecule in opposite orientations, the intervening DNA sequence is inverted. Finally, recombination between two orthogonal targets on separate molecules results in cassette exchange.

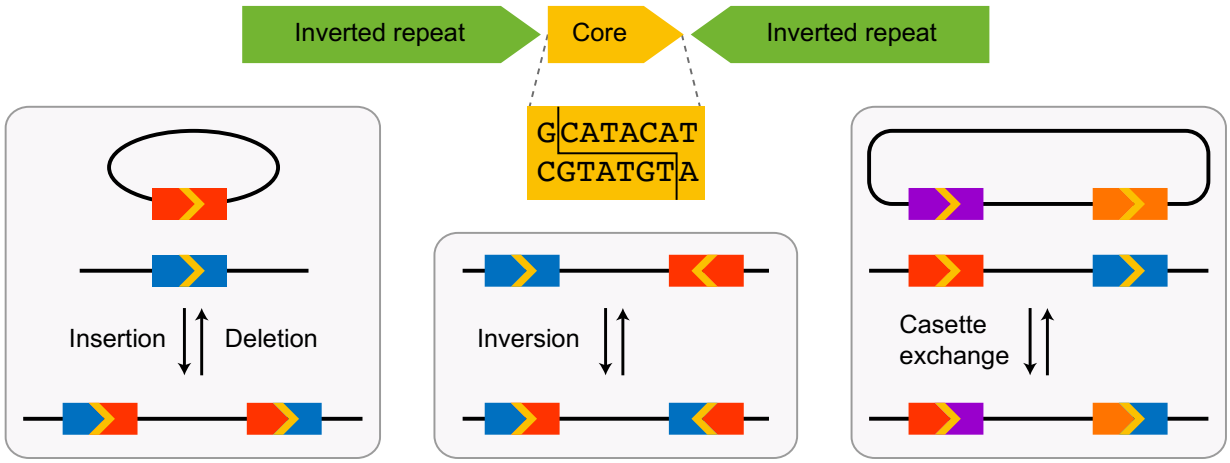


Figure 1.1. Recombination outcomes based on the core sequence orientation. For simple SSR family members, the recognition target is composed of two symmetric half-sites flanking an asymmetric core sequence. The *loxP* core sequence is shown, with the cleavage product indicated (black line). The non-palindromic core sequence imparts a directionality to the recombinase target (yellow arrow), and the relative orientation and location of two targets dictate the result of recombination: deletion, or the reverse reaction, integration; inversion; or cassette exchange between two orthogonal recombinase targets.

The prototypical tyrosine SSR is Cre, which recombines the 34-bp *loxP* target[60] (Table 1.1). For tyrosine recombinases, the catalytic mechanism proceeds via two cycles of 3'-phosphotyrosine linkages that are resolved by attack of the 5' hydroxyl of the adjacent DNA strand, with a Holliday Junction intermediate (Figure 1.2). Tyrosine SSRs perform strand exchange between two identical target sites, and therefore recombination reactions are freely reversible. For deletion/integration reactions, the deletion product is favored for entropic reasons[60].

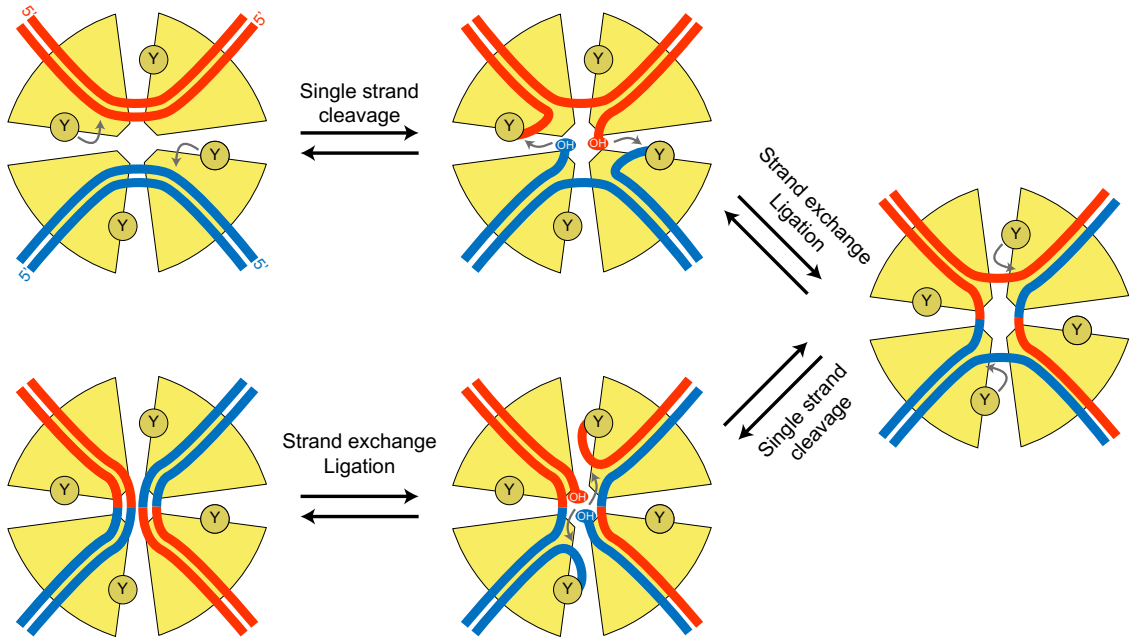


Figure 1.2. Mechanism of recombination of tyrosine SSRs. Recombinase monomers bind to each half-site of a target sequence, with one monomer in the active and one in the inactive conformation. Two dimeric protein:DNA assemblies join with C2 symmetry to form the synaptic complex. The catalytic tyrosines (Y) of the active monomers attack one strand of the DNA duplex, forming 3'-phosphotyrosine linkages which are resolved by attack of the 5' hydroxyl of the adjacent strand. In the Holliday Junction intermediate, the recombinase monomers isomerize, such that the neighboring monomer is now in the active conformation. The steps of single-strand cleavage, exchange, and ligation are then repeated, yielding the recombined product.

Serine SSRs can be further divided into two major groups: small serine resolvases and large serine integrases[61]. Both families share a common mechanism, in which both recombinase targets undergo simultaneous double-stranded cleavage, and strand exchange is

accomplished by a 180° rotation of one half of the tetrameric complex (Figure 1.3). The resolvases, such as the Gin, recombine between two identical *gix* targets, much like Cre:*loxP* (Table 1.1). The serine integrases, however, recognize two distinct substrates, which are often asymmetric in sequence and target length. For example, the integrase Bxb1 recombines between the sequences *attP* and *attB* (Table 1.1), generating the product substrates *attL* and *attR*[62]. Excisive recombination between *attL* and *attR* requires a separate directionality factor protein[63], and the serine integrases are therefore considered unidirectional.

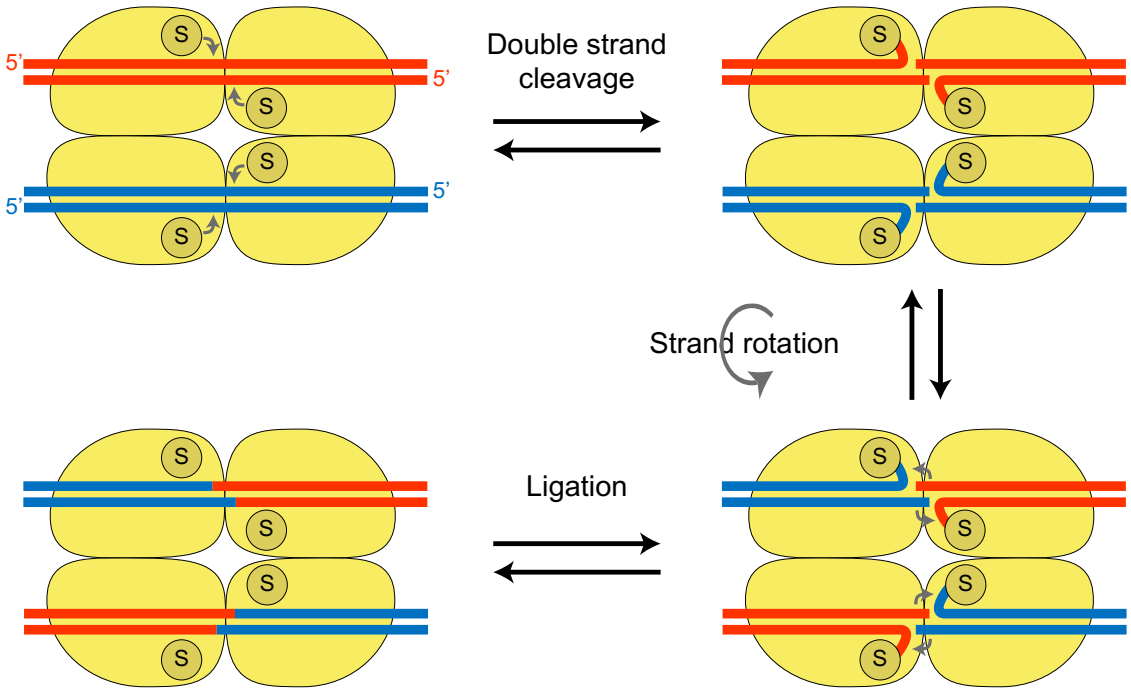


Figure 1.3. Mechanism of recombination of serine SSRs. The recombination complex consists of two DNA molecules bearing recognition targets, which are occupied by a recombinase dimer. The serine nucleophile (S) of each recombinase monomer cleaves the adjacent DNA strand, resulting in double-stranded cleavage of both recombinase targets. Strand exchange is accomplished by a 180° rotation of one half of the tetrameric complex. The free 3' hydroxyl groups at the cleavage site then attack the 5'-phosphoserine linkage, ligating the recombined strands.

Recombinase	Target	DNA sequence
Cre (Tyr)	*loxP*	ATAACTTCGTATAGCATACATTATACGAAGTTAT
Flp (Tyr)	*FRT*	GAAGTTCCTATTCTCTAGAAAGTATAGGAACTTC
Dre (Tyr)	*rox*	TAACTTTAAATAATGCCAATTATTTAAAGTTA
VCre (Tyr)	*loxV*	TCAATTTCTGAGAACTGTCATTCTCGGAAATTGA
Gin (Ser)	*gix*	TTCCTGTAAACCGAGGTTTTGGATAA
Bxb1 (Ser)	*attP*	GGTTTGTCTGGTCAACCACCGCGGTCTCAGTGGTGTACGGTACAAACC
	attB	GGCTTGTCGACGACGGCGGTCTCCGTCGTCAGGATCAT
phiC31 (Ser)	*attP*	GTGCCCCAACTGGGGTAACCTTTGAGTTCTCTCAGTTGGGGG
	attB	TGCGGGTGCCAGGGCGTGCCCTTGGGCTCCCCGGGCGCGTACTCC

Table 1.1. Recombination targets of representative tyrosine and serine SSRs. Crossover sequences (red) are highlighted.

Site-specific recombinases have many appealing properties as genome editing tools. The reactions catalyzed by SSRs can result in the direct replacement, insertion, or deletion of target DNA fragments with efficiencies exceeding those of HDR[56,64]. SSRs are active in a wide variety of cell types and cell states including non-dividing cells[56], and many efficiently operate on mammalian genomes[60,65]. For instance, multiple serine integrases have been shown to efficiently integrate into the human genome[66]. Likewise, the Cre:*loxP* system has been used in transgenic animals for applications including conditional gene regulation[67,68] and lineage tracing[69,70], and evolved variants of Cre have been used to remove HIV provirus from human blood cells engrafted in mice[71,72]. Finally, the catalytic mechanisms of SSRs induce less DNA damage and toxicity than comparable exposure to programmable nucleases[73].

Although SSRs offer many advantages, they are not widely used because they have a strong innate preference for their cognate target sequence. The recognition sequences of SSRs are typically ≥ 20 base pairs and thus unlikely to occur in the genomes of humans or model organisms. Further, the native substrate preferences of SSRs are not easily altered, even with

extensive laboratory engineering or evolution[74]. For example, Buchholz and coworkers required 126 and 145 rounds of laboratory evolution to evolve two Cre variants, Tre[75] and Brec1[72], that recombine sites differing from *loxP* at 50% and 68% of DNA base pairs, respectively; their retargeting efforts likely required decades of total researcher time. Thus, despite continued efforts to develop SSRs, the challenge of altering their DNA specificity to manipulate arbitrary sequences of interest remains a major barrier to their widespread use.

1.4 Prospects for engineering and evolving site-specific recombinases

Programmable nucleases and SSRs have advantages and drawbacks as tools for gene integration, and their characteristics largely mirror one another (Figure 1.4). For example, SSRs catalyze precise and efficient DNA strand exchange, but their innate specificity limits their applicability to a narrow range of DNA sequences. Programmable nucleases can easily be retargeted to a new DNA sequence, but cannot perform efficient gene integration. An ideal gene integration tool would carry out efficient, predictable gene insertion at an arbitrary genetic locus. While many current genome editing researchers are focused on improving the HDR efficiency when using programmable nucleases, I and others who study SSRs have approached this challenge by seeking to develop recombinases with broadened applicability.

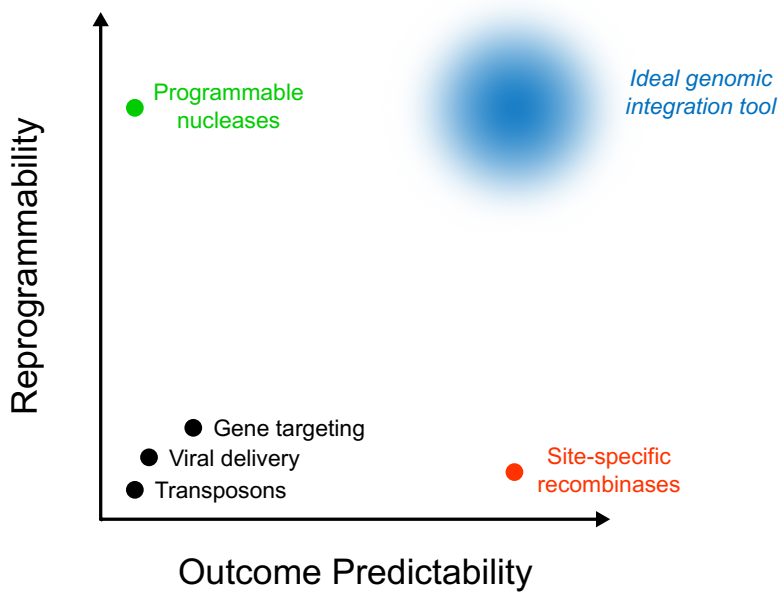


Figure 1.4. Model of the relative strengths and weaknesses of gene integration techniques. Candidate gene editing methods are assessed based on their ability to effect highly efficient and precise gene integration (*i.e.*, outcome predictability) at an arbitrary genomic locus. While early genome editing techniques score poorly in both regards, programmable nucleases and SSRs have opposite strengths and weaknesses. The development of ideal gene integration tools may require overcoming the weaknesses of either programmable nucleases or SSRs.

In principle, targeted gene integration with SSRs could be achieved in one of two ways. One pathway involves laboratory evolution or engineering of a recombinase to specifically target a new sequence of clinical or academic interest, as Buchholz and colleagues demonstrated[72,75]. In **Chapter 2** of this dissertation, I describe the laboratory evolution of a retargeted recombinase using the phage-assisted continuous evolution (PACE) system[76]. Reasoning that PACE could rapidly generate custom recombinases, we developed a PACE selection for SSRs and used it to evolve Cre to target a sequence in a human "safe harbor" genomic locus. Continuous selection generated variants of Cre that possess activity in mammalian cells on a sequence present within the ROSA26 locus[77]. Subsequently, I attempted several methods, including adaptations of PACE as well as independent selections, to improve the activity and specificity of the resulting enzyme variants.

Another strategy to achieve SSR-mediated gene integration is the development of programmable recombinases, by combining the capabilities of SSRs with the versatility of DNA-binding proteins. Previous work established that chimeric fusions of serine resolvases and Zinc finger or TALE DNA-binding domains are active in mammalian cells[78-80]. In **Chapter 3**, I describe the development of recCas9, an RNA-programmed small serine recombinase that functions in mammalian cells[81]. We optimized the chimeric fusion between catalytically inactive Cas9 and an engineered Gin recombinase domain. We then showed that recCas9 can operate on DNA sites in mammalian cells identical to genomic loci naturally found in the human genome in a manner that is dependent on the gRNA sequences. We also showed that recCas9 can operate directly on the genome of unaltered human cells, catalyzing *in situ* deletion between two genomic substrates. I also describe subsequent attempts to improve recCas9 by fusion of alternative recombinase domains.

1.5 Building tools for better understanding determinants of recombinase specificity

In the course of evolving and engineering recombinases as genome editing tools, I encountered several recurring obstacles, including low activity and a lack of specificity among recombinase variants. From my experience and literature reports, it is evident that the an incomplete understanding of SSR protein:DNA specificity determinants has limited the development of recombinases as genome editing tools[60,74,82]. For example, crystal structures of tyrosine-family SSRs demonstrate that recombinases interact with DNA through relatively few direct protein:DNA contacts, and that shape- and charge-complementarity, as well as water-mediated interactions, contribute to SSR specificity[60,83]. Further, mutagenesis studies showed that mutations in Cre can alter its tolerance for mismatches in regions of *loxP* with no direct protein:DNA interactions[84]. These and other findings establish that the relationship between SSR residues and DNA specificity is not straightforward; some residues impact specificity more than others, and some contribute to specificity at distant DNA positions.

While high resolution methods for assaying the DNA binding preferences of programmable nucleases exist[85-88], no analogous method for recombinases has been developed. In **Chapter 4**, I describe Rec-seq[89], a method for profiling the DNA specificity of SSRs in a rapid and unbiased manner using *in vitro* selection and high-throughput DNA sequencing (HTS). We applied Rec-seq to characterize wild-type Cre and Cre mutants, resulting in the identification of known and novel DNA specificity determinants, including long-range interactions not evident from structural studies. We also profiled the sequence preferences of the laboratory-evolved Cre variants Tre and Brec1, as well as three additional orthogonal SSRs, including the directional integrase Bxb1. Finally, the application of Rec-seq to Tre and Brec1 recombinases resulted in specificity profiles that accurately predicted activity at off-target sites, including several pseudo-sites within the human genome, an important consideration when evaluating SSRs as potential research tools or therapeutics.

1.6 Conclusion

I was fortunate to investigate DNA recombinases during a period of skyrocketing interest in genome editing. Compared to when I first arrived in Boston, there has been a major increase in awareness of genome editing technology throughout the scientific community. The rapid pace of genome editing research, catalyzed by the widespread adoption of programmable nuclease-based techniques, has attracted researchers to problems that are adjacent to the question of gene integration; for example, how to deliver genome editing macromolecules *in vivo*, how to detect off-target modifications, etc. Solutions to these problems are unlikely to apply solely to programmable nucleases. Therefore, the development of SSRs is likely to benefit from this increased interest and infrastructure surrounding genome editing.

The future of genome editing technology is far from predetermined. As we search for creative solutions to the translational and research hurdles that remain, the appealing properties of DNA recombinases makes them ideal candidates for gene integration tools. In my graduate

studies, I have investigated several paths toward developing SSRs as viable genome editing agents. I have also established a rapid and general method for profiling the specificity of recombinases, the findings of which may enable the generation of custom recombinase tools. In **Chapter 5** of this dissertation, I summarize the insights from my studies and describe experiments that incorporate these insights. I also highlight overlooked classes of enzymes that may be suitable for development as genome editing tools.

Chapter 2:

Continuous *In Vivo* Directed Evolution of Site-Specific Recombinases

David Thompson designed and performed the experiments described in sections 2.2.1-2.2.2 and figures 2.2-2.4. David Thompson and I designed and performed the experiments described in sections 2.2.3-2.2.4 and figures 2.5-2.6. I designed and performed all remaining experiments.

2.1 Introduction

The ability to retarget the specificity of DNA recombinases would represent a powerful contribution to biomedical and translational research. For example, a site-specific recombinase (SSR) that stably integrates transgenes within a genomic "safe harbor" – i.e. a chromosomal region where foreign DNA is robustly expressed without perturbing endogenous genomic function[90] – could serve as a general tool for creating transgenic cells lines or delivering a gene-based therapy. Alternatively, an SSR engineered to recognize a high-value intragenic target could be used to modify or replace an endogenous sequence while retaining the native genomic regulation. Despite their potential to serve as ideal genome editing agents, the utility of SSRs has been limited by the innate recognition of DNA targets that are typically ≥ 20 base pairs and thus unlikely to occur in high-value regions of human or model genomes. And unfortunately, the substrate preferences of SSRs are not easily altered even after extensive laboratory evolution or engineering[74]. Thus, the challenge of retargeting SSRs to manipulate arbitrary sequences of interest presents a major barrier to realizing the goal of facile gene integration.

Attempts to alter the specificity of SSRs have spanned more than 30 years (reviewed in ref. 74). Of the SSRs, Cre has been subject of the most evolution or engineering attempts. However, few retargeting efforts resulted in changes in specificity at more than a handful of base pairs within *loxP* (*e.g.*, refs. 84, 91). The most extensive retargeting was accomplished by Buchholz and colleagues[72,75,92]. Using the substrate-linked protein evolution technique[92], Buchholz and colleagues performed 126 and 145 rounds of laboratory evolution, yielding Cre variants that recombine sites differing from *loxP* at 50% and 68% of DNA base pairs[72,75]. These research feats demonstrate the feasibility of retargeting Cre toward sequences that greatly differ from its endogenous target. However, the retargeting campaigns required dozens of iterations of labor-intensive experiments, likely entailing decades of total researcher time.

My colleagues and I investigated whether the development of retargeted recombinases could be accelerated using the phage-assisted continuous evolution (PACE) system[76]. In PACE,

the cycle of laboratory evolution – gene diversification, selection of fit variants, and amplification of the resulting population – is mapped onto the life cycle of the M13 bacteriophage[93], allowing evolution to occur at the same rate as phage replication. PACE selections have been developed to modify the properties of a wide range of proteins, including polymerases[76,94,95], programmable nucleases[43,96], proteases[97,98], insecticidal toxins[99], and aminoacyl-tRNA synthetases[100].

We developed a PACE selection for DNA recombinases and used it to retarget Cre toward a sequence present in the human ROSA26 locus[77]. We completed retargeting using a second-generation PACE selection, generating recombinase variants with activity on the ROSA*loxP* target in a transfected reporter in mammalian cells. We implemented several modifications to PACE in attempts to increase the activity and specificity of the resulting Cre variants. Finally, the insights gathered from the efforts to evolve or engineer Cre have informed the design of novel selections that could be used to develop retargeted recombinases.

2.2 Results and Discussion

2.2.1 Developing a selection for DNA recombinases in PACE

In PACE (Figure 2.1a), a population of phage (selection phage, SP) encoding the evolving protein of interest (POI) is continuously diluted in a fixed volume vessel (the lagoon) by host *E. coli* cells. Development of a PACE selection requires linkage between the activity of the POI and survival of the phage that encodes it. This is accomplished by removing an essential phage gene, gene III (gIII), from the SP genome, and inserting it on an accessory plasmid (AP) in the host cells, with expression regulated by the POI selection circuit. Gene III encodes the minor coat protein III (pIII), which is critical for producing infectious progeny phage. SP encoding functional library members restore pIII production from the AP and generate infectious progeny phage at a rate that scales with pIII levels[101]. Because the media in the lagoon is constantly replenished, SP must propagate faster than the rate of dilution, and SP bearing non-functional library members are diluted out of the population. Diversity is generated through induction of a

mutagenesis plasmid (MP) in the host cells that dramatically increases SP mutation rates[102]. Because one complete cycle of phage replication can occur in as short as 10 minutes[103], a typical PACE experiment can involve dozens of rounds of evolution in a single day.

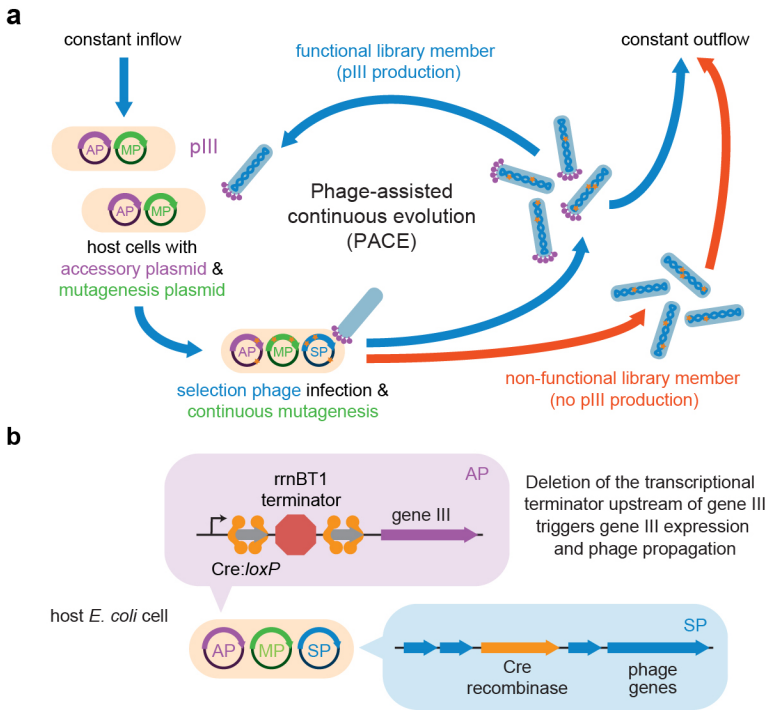


Figure 2.1. Overview of a PACE selection for site-specific recombination. a, General PACE schematic. Host cells contain an accessory plasmid (AP) that expresses gene III (gIII) regulated by the activity of an evolving protein of interest (POI). The selection phage (SP) contains the POI in place of gIII and can therefore only reproduce if it encodes a POI variant that passes the selection established by the AP, thereby triggering pIII production. The gene encoding the POI is mutated by induction of the mutagenesis plasmid (MP). As the SP exist in a continuously diluted fixed-volume vessel, only those SP that propagate faster than the rate of dilution can persist. **b,** Interruption of gIII expression by a transcriptional terminator, flanked by recombinase recognition targets, links DNA recombination to production of pIII. Deletion of the transcriptional terminator restores pIII production and thus the generation of infectious progeny.

We designed a PACE selection for Cre recombinase, which performs strand exchange between two *loxP* sites (see Figures 1.1, 1.2). To link Cre activity to the production of pIII, we constructed a circuit encoded on the AP in which gIII is separated from its upstream promoter by a transcriptional terminator (Figure 2.1b). The terminator is flanked by *loxP* sites (or "floxed") in a deletion orientation, such that recombination between the two *loxP* targets removes the terminator and restores pIII production. Selection pressure for altered specificity is applied by

changing the identity of the sequences that flank the transcriptional terminator, forcing Cre to operate on non-native sequences to pass the selection.

We validated that this PACE circuit is selective for recombinase activity and that selection accompanied by mutagenesis can restore catalytic activity to a population of inactivated enzymes. To demonstrate selective propagation of phage encoding recombinases, but not unrelated enzymes, we initiated a mock PACE experiment in which host cells contained the *loxP* AP and no MP. The lagoon was inoculated with SP encoding wild-type Cre or T7 RNA polymerase (T7RNAP) at a 1:10^6 ratio (Figure 2.2a). The presence of phage encoding Cre or T7RNAP was determined by PCR amplification of the SP genome. While Cre SP was undetectable by PCR in the input phage mixture, Cre SP predominated in the lagoon after 12 hours of PACE, and persisted for an additional 24 hours. This finding suggests that only phage encoding active recombinase can propagate on host cells bearing the recombinase selection AP.

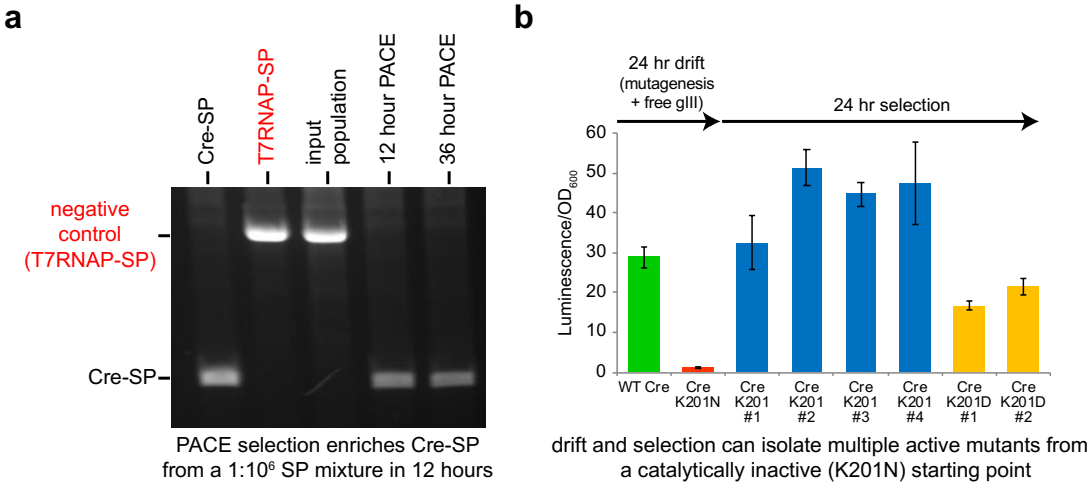


Figure 2.2. Validation of PACE selection for site-specific recombinases. a, SP encoding Cre recombinase are enriched among an excess of SP encoding T7 RNA polymerase (T7RNAP) by propagation on PACE host cells bearing the recombinase selection AP. PCR was used to detect the presence of phage encoding Cre or T7RNAP before selection, and after 12 or 36 hours of PACE. **b,** Catalytically inactive Cre SP was used to inoculate a lagoon that underwent 24 hours of selection-free drift followed by 24 hours of PACE selection. The activity of wild-type Cre, K201N Cre, and the evolved Cre variants was assessed by transcriptional activation assays in *E. coli*. Values and error bars represent the mean and standard deviation of three technical replicates.

Next, we determined whether mutagenesis and PACE selection could restore enzymatic activity to SP encoding catalytically inactive Cre. We inoculated a lagoon with phage bearing catalytically dead Cre with a K201N mutation[104] (Figure 2.2b). Host cells contained the *loxP* AP as well as the drift MP, which produces pIII from a chemically inducible promoter[95,102]. Moderate induction of pIII expression prevents washout of SP encoding inactive variants, while allowing SP with active variants to achieve a fitness benefit by producing additional pIII from the AP. The initial SP population was propagated for 24 hours with intermediate levels of drift, followed by PACE on *loxP* host cells for 24 hours. DNA sequencing of 6 Cre SP that survived the selection revealed that 4 had reverted back to Lys at residue 201, while the remaining 2 encoded an Asp residue at that position.

We measured the activity of the resulting Cre variants using a transcriptional activation assay, in which bacterial luciferase replaces gIII in the AP. We inoculated *E. coli* reporter cells with clonal SP and used the luminescence signal to assess the relative activity of the Cre variants. While the Cre K201N variant was inactive on the *loxP* reporter, wild-type Cre and the K201 PACE variants demonstrated robust activity (Figure 2.2b). Additionally, the Cre K201D variants demonstrated decreased but appreciable *loxP* activity. We attribute the apparent outperformance of wild-type Cre by several of the evolved variants to increased phage fitness, as opposed to superior recombinase activity. Together, these results reveal that PACE selection with the recombinase AP can generate recombinase variants with properties that differ from the input population.

2.2.2 Retargeting Cre recombinase to operate on a sequence present in the human genome using PACE

For the initial SSR retargeting goal, we aimed to generate recombinase variants with broad applications in biomedical and translational research. Therefore, we decided to target the

ROSA26 locus in the human genome[77]. The ROSA26 locus was first discovered in mice, and has become the most popular locus for integration of transgenes in murine models[105]. Foreign DNA integrated at the ROSA26 locus is highly and ubiquitously expressed. In addition, the ROSA26 locus is considered a genomic safe harbor due to its distance from cancer-related genes, microRNAs, and ultra-conserved regions[90], and transgenic mice demonstrate no obvious phenotypic differences[105].

We devised a series of experiments to retarget Cre toward a sequence within the human ROSA26 locus with the greatest similarity to *loxP*. The sequence we chose, termed ROSA*loxP*, contains 15 mismatches (out of 34 total bases) relative to *loxP*; these mismatches are distributed evenly between the left and right half-sites and the core region (Figure 2.3a). Unlike *loxP*, the half-site sequences of ROSA*loxP* are not inverted repeats. Therefore, we devised two series of intermediate substrates, with one series for transitioning preference toward each half-site. Activity on ROSA*loxP* would be achieved by first evolving separate lineages of Cre variants that recognize symmetric left or right half-site intermediates (Figure 2.3a). Upon completion of PACE on the left and right final substrates (LF and RF), we envisioned that the LF- and RF-active Cre could be developed as a heterodimeric pair to specifically recombine the asymmetric ROSA*loxP* target. Alternatively, we could attempt to shuffle the mutations present in LF and RF Cre to generate a singular consensus variant capable of operating on ROSA*loxP*.

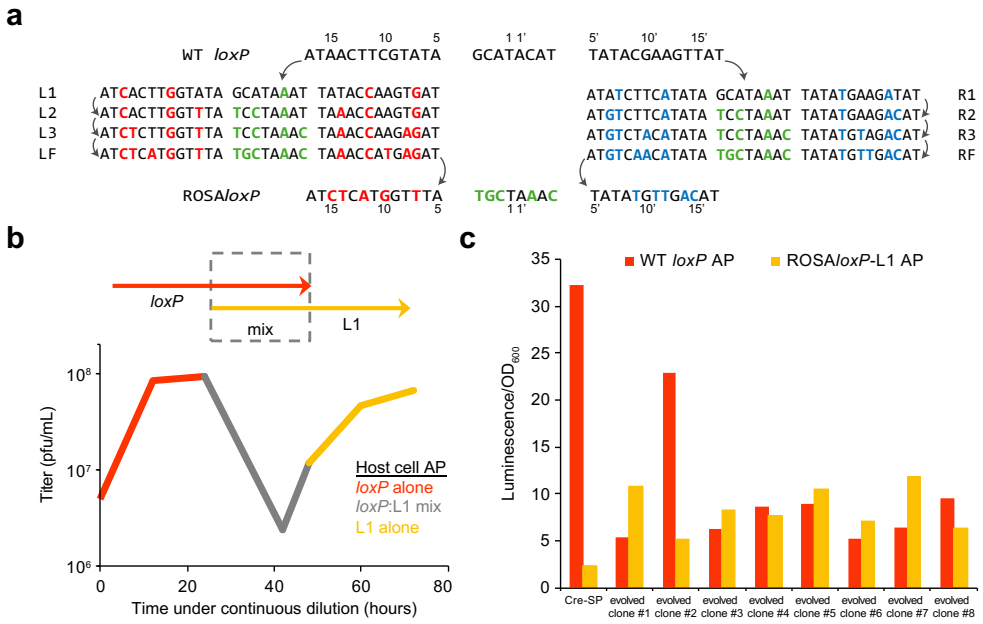


Figure 2.3. Experimental approach for retargeting Cre to the ROSA*loxP* sequence. a, PACE evolutionary trajectory for retargeting Cre recombinase toward the ROSA*loxP* sequence. To evolve activity on an asymmetric target, left and right half-site intermediates were devised, and recombinase variants were selected using APs with symmetric half-sites bearing increasing numbers of mismatches relative to *loxP* (colored bases). **b,** PACE experiments were initiated with wild-type Cre SP, and a mixing strategy was used to transition between selections on different substrate APs. Exemplary data for the transition between wild-type *loxP* and ROSA*loxP*-L1 are shown. The *y* axis shows total phage titer in the lagoon (n=1). **c,** The activity of the L1-evolved Cre variants on *loxP* and ROSA*loxP*-L1 was assessed by transcriptional activation assays in *E. coli*. Values represent the mean of three technical replicates.

Recognition of the ROSA*loxP* intermediate substrates was achieved using a PACE host cell mixing strategy[76]. For example, wild-type Cre SP was propagated on host cells with a *loxP* AP for 24 hours, followed by selection on a 1:1 *loxP*:ROSA*loxP*-L1 mixture of host cells for 24 hours (Figure 2.3b). After the mixing phase, SP were propagated exclusively on host cells bearing the L1 intermediate AP. We isolated phage that survived 72 hours of PACE and assessed their activity on *loxP* and ROSA*loxP*-L1 using the transcriptional activation assay. While the wild-type enzyme exhibited minimal activity on ROSA*loxP*-L1, all 8 clones isolated after 72 hours of PACE were active on the L1 intermediate while retaining activity on *loxP* (Figure 2.3c). This result demonstrates the feasibility of transitioning recombinase specificity in PACE by gradual replacement of the host cell population.

Using the mixing strategy, we carried out selections on intermediate substrates of both half-sites through ROSA*loxP*-L3 and -R2. Successive PACE experiments were inoculated with SP that survived selection on the previous intermediate. In transcriptional activation assays, L3- and R2-evolved Cre variants exhibited activity on all intermediate substrates they had been exposed to (Figure 2.4a,b). We anticipated the possibility of broadened specificity among the Cre variants, as evolving proteins typically acquire substrate promiscuity before gaining specificity for the new target[106]. This broadened specificity may also be a consequence of the mixing strategy, which facilitated the transition between intermediates but may have contributed to expanded substrate preference by simultaneous selection for recognition of two target sequences.

Based on the co-crystal structure of Cre in complex with *loxP*[107], we characterized the potential impact of L3 and R2 mutations on altered substrate specificity. The L3 SP contained 5 converged mutations, including residues proximal to *loxP* positions that were changed during the course of evolution. For example, the M44V mutation occurred at a residue proximal to position 7, the site of an A•T → T•A transversion, and R259C and E262A mutations arose near the C•G → G•C transversion at position 10 (Figure 2.4c). Additionally, consensus mutations A53E and A249V occurred within helices that participate in protein:DNA interactions. The R2 SP contained three converged mutations and several reccurring mutations located near regions of protein:DNA and protein:protein interactions (Figure 2.4d). Of the fixed mutations, only E262A occurred at the protein:DNA interface, proximal to the G•C → A•T transition at position 9. Other mutations, including the I306V and I320M conserved mutations and the high-frequency G342S mutation (not pictured), are located near the C-terminal helix that makes critical inter-monomer contacts during recombination[108]. Collectively, these results suggest that PACE selections on altered substrates can generate recombinase variants with substantial activity on non-native

targets. The varying types of mutations acquired during selection suggests potential roles for altered DNA target recognition and inter-monomer interactions.

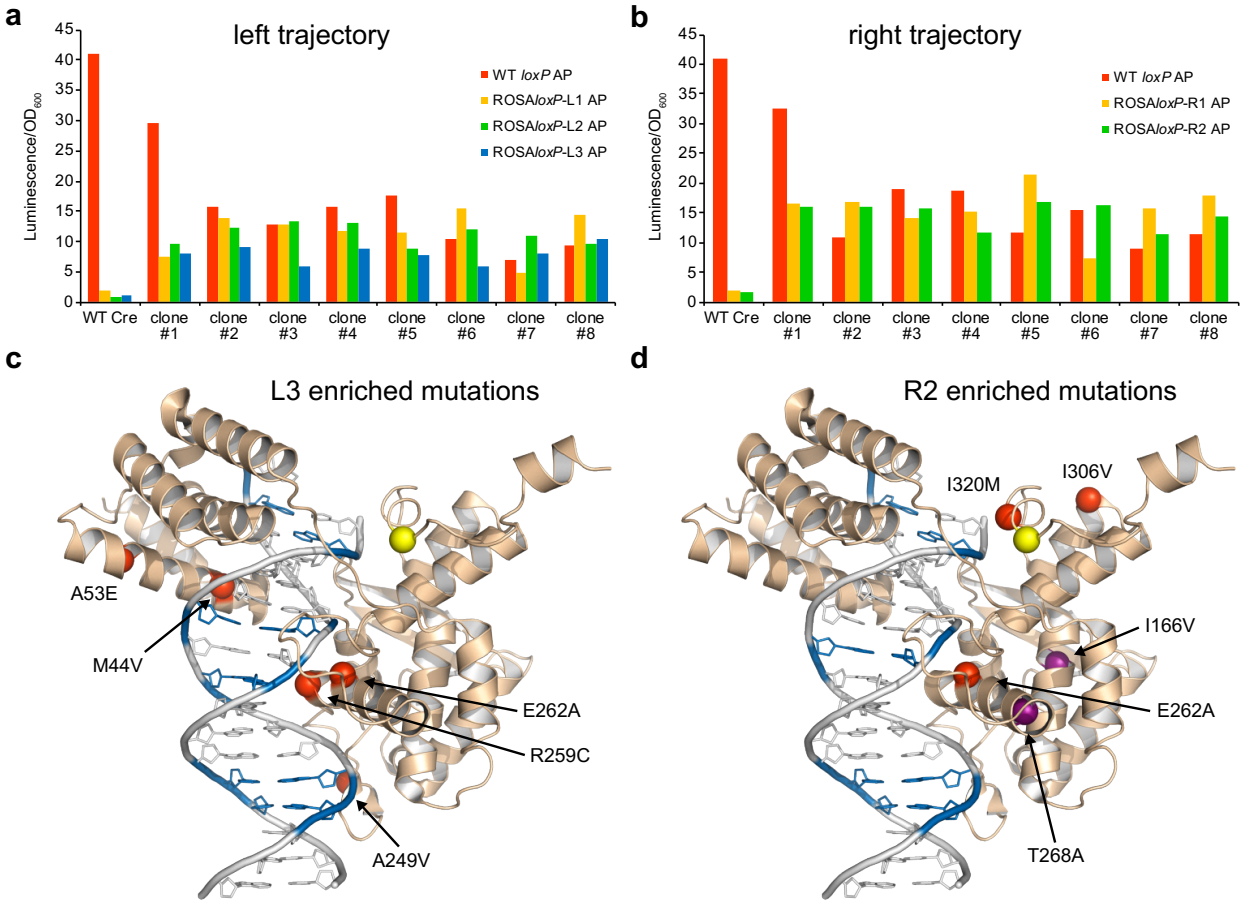


Figure 2.4. Recombinase retargeting PACE through the ROSA*loxP* L3 and R2 intermediate substrates. a, b, The activity of the L3- (**a**) and R2-evolved Cre variants (**b**) on *loxP* and intermediate selection substrates was assessed by transcriptional activation assays in *E. coli*. Values represent the mean of three technical replicates. **c, d,** Mutations accumulated by Cre variants selected on the L3 (**c**) and R2 (**d**) substrates mapped onto the structure of Cre in complex with *loxP*[107]. The catalytic Tyr (yellow), consensus mutations (red), and commonly observed mutations (purple) are depicted as spheres. DNA mismatches relative to *loxP* (blue) are highlighted. One-letter amino acid labels indicate the identity of the Cre residue at that position and the identity of the mutation observed after PACE.

2.2.3 Development of a second-generation recombinase selection in PACE

We were unable to carry out PACE selections on the LF and R3 intermediate substrates, prompting us to redesign the recombinase selection circuit. Using plaque assays, we determined that host cells bearing LF and R3 APs were uninfectible by M13 phage, and we

traced the source of this uninfectibility to expression of pIII in the host cells prior to phage infection or exposure to Cre. M13 bacteriophage enter *E. coli* via pIII-mediated interactions with the F-pilus[109], causing the F-pilus to retract[110]. We reasoned that leaky expression of pIII from the LF and R3 APs prior to deletion of the floxed terminator was likely causing retraction of the F-pilus and prevention of SP infection. This leaky expression could be due to cryptic promoters introduced by the LF or R3 targets immediately upstream of gIII. We were unable to identify the promoter sequences in the LF or R3 APs using a predictive algorithm[111], preventing the redesign of the evolutionary intermediates. We therefore opted to redesign the selection circuit to negate the impact of leaky pIII expression prior to phage infection.

We reasoned that relocation of the floxed terminator on the AP could prevent expression of full-length pIII in the absence of AP recombination. For example, insertion of the deletion cassette within the coding sequence of gIII would result in expression of N-terminally truncated pIII from the upstream promoter prior to recombination (Figure 2.5a). Leaky expression from cryptic promoters in the downstream *loxP* site would also generate a truncated protein which would likely be out of frame. However, post-recombination, the AP would produce full-length pIII containing an internal peptide corresponding to the in-frame *loxP* DNA sequence (pIII'). To be compatible with PACE, this selection design requires that the floxed terminator is inserted in a region of gIII such that the resulting pIII' enables the production of infectious progeny phage.

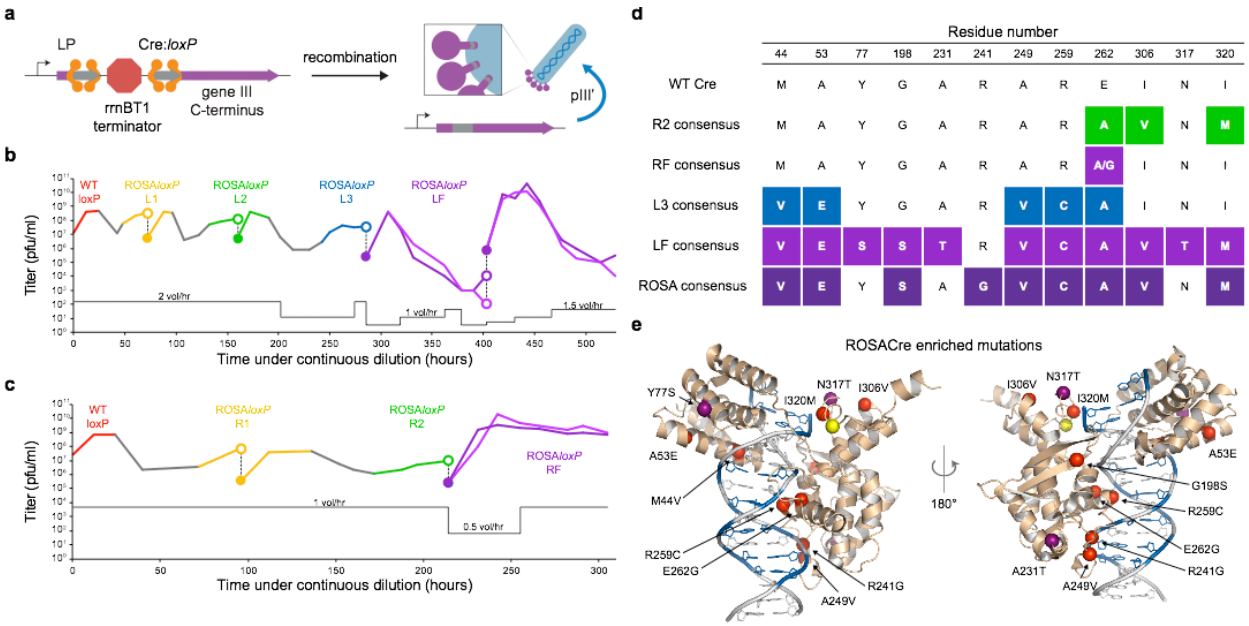


Figure 2.5. Recombinase retargeting PACE on LF, RF, and ROSA*loxP* sequences using a second-generation selection. a, Schematic of second-generation recombinase selection in PACE. The deletion cassette lies within the coding sequence of gIII, in between the leader peptide (LP) and the C-terminal domains. Deletion of the transcriptional terminator restores production of modified pIII (pIII'), containing a peptide corresponding to the recombinase target DNA sequence, which is functionally incorporated by infectious progeny. **b,** PACE toward the ROSA*loxP*-LF target was executed in five segments. Segments 1-3 implemented the mixing strategy of first-generation AP host cells under intermediate levels of mutagenesis (MP4). The final two segments implemented the LF target on the second-generation AP under high levels of mutagenesis (MP6). **c,** PACE toward the ROSA*loxP*-RF target was executed in three segments. Segments 1-2 implemented the mixing strategy of first-generation AP host cells under intermediate levels of mutagenesis (MP4). The final segment implemented the RF target on the second-generation AP under high levels of mutagenesis (MP6). For **b** and **c,** phage titer (colored line) and lagoon flow rate (black line) are shown at all sampled time points. The dotted lines and open circles indicate transfer of evolving phage to a new lagoon fed by the host cell culture containing the indicated AP. **d,** Mutations (colored boxes) accumulated by key Cre variants during PACE. **e,** Mutations present in the ROSACre population mapped onto the structure of Cre in complex with *loxP*[107]. The catalytic Tyr (yellow), consensus mutations (red), and commonly observed mutations (purple) are depicted as spheres. DNA mismatches relative to *loxP* (blue) are highlighted.

To determine the ideal placement of the *loxP* peptide within pIII', we identified candidate regions within gIII where an insertion would disrupt pIII expression prior to phage infection but support SP propagation after recombination. PIII is composed of three domains connected by flexible linker regions[110], as well as an N-terminal leader peptide (LP) that directs the secretion of pIII to the periplasm[112]. We generated plasmids that encoded gIII regulated by the phage

shock promoter[113] (P_{psp}) with *loxP* inserted in the linker regions between the domains of pIII, as well as between the LP and the N1 domain, and tested the ability of these plasmids to support SP propagation in overnight enrichment assays. We found that placement of *loxP* between the leader peptide and the N1 domain resulted in robust overnight phage propagation.

Next, we constructed an AP encoding gIII with a floxed terminator inserted between the LP and the N1 domain and assessed the ability of host cells bearing this second-generation selection circuit to support activity-dependent phage propagation. In overnight enrichment assays, SP bearing wild-type Cre, but not T7RNAP, enriched up to 10^5-fold. We then tested the performance of a series of APs, modifying parameters such as the AP origin of replication and gIII promoter and RBS strength, to observe the impact of varying these parameters on selection stringency. For example, increasing the copy number of the AP increases the number of recombination events required to produce the maximal amount of pIII, representing a more stringent selection. While the first-generation AP was restricted to a low-copy origin to retain infectibility of the host cells, *E. coli* bearing the second-generation AP on a high-copy pUC origin remained infectible and promoted robust overnight phage enrichment. These findings suggest that the second-generation recombinase selection avoids the infectibility issues of the first-generation selection and offers an expanded repertoire of parameters for continued retargeting of Cre.

Following the development of the second-generation AP, we performed selections on the LF and RF intermediate substrates. In overnight enrichment assays, we found that L3-active SP propagated on host cells bearing the LF AP. We initiated PACE on LF host cells inoculated with SP from the overnight enrichment assay. After phage titer dropped in response to an increase in the lagoon flow rate, we used the surviving SP to inoculate a second PACE with the same host cells at a lower initial flow rate (Figure 2.5b). At the end of LF selection, surviving SP had undergone a cumulative 530 hours of PACE and acquired a total of 11 converged mutations (Figure 2.5d). Additionally, we found that R2-evolved SP were capable of propagating

on RF host cells in overnight phage enrichment assays. We performed PACE on RF host cells inoculated with SP from the overnight enrichment at an initial flow rate of 0.5 volumes per hour (Figure 2.5c). Surviving SP encoded fewer fixed mutations than the input R2 SP; variants contained either E262A or G, while we observed common T268A and I320M mutations (Figure 2.5d). Phage that survived the RF selection had undergone a cumulative 305 hours of PACE. These experiments demonstrate the utility of the second-generation selection by enabling selections with substrates that were previously impossible using the first-generation AP.

Finally, we performed selections for activity on the ROSA*loxP* target. In overnight propagation assays, both LF- and RF-evolved SP enriched on host cells bearing the ROSA*loxP* AP. We initiated PACE experiments on ROSA*loxP* host cells, and although the lagoon seeded with RF SP washed out, the LF SP persisted for 40 hours at a flow rate of 0.5 volumes per hour. Isolation and characterization of the surviving Cre variants revealed a consensus mutant with 8 of the 11 LF mutations as well as R241G (Figure 2.5d,e). In crystal structures of Cre in complex with *loxP*[108], R241 is located proximal to position 15, which is mutated in both half-sites of ROSA*loxP*, suggesting a role in altered DNA recognition (Figure 2.5e).

The generation of ROSA-active Cre variants represented the achievement of our initial goal of generating an SSR with activity on a substantially altered substrate using PACE.

*2.2.4 Activity of ROSACre variants on ROSA*loxP *in mammalian cells*

Having demonstrated recombination of the ROSA*loxP* target in bacterial assays, we next assessed the performance of Cre variants in human cells. To monitor recombination, we used a two-plasmid reporter system, in which one plasmid encodes exon 1 of *EYFP* followed by a splice donor sequence, and the second plasmid encodes a splice acceptor followed by *EYFP* exon 2 (Figure 2.6a). In this reporter, recombinase-mediated integration between recognition sequences located in the intronic regions of *EYFP* restores fluorescence expression[114]. We co-transfected HEK293T cells with *loxP* or ROSA*loxP* reporter plasmids and a plasmid expressing

wild-type Cre or a ROSA-evolved variant, then used the fraction of cells exhibiting EYFP fluorescence to assess the relative activity of each variant (Figure 2.6b). The Cre variant with the highest activity on the ROSA*loxP* reporter, termed “ROSACre 1”, contained the 9 consensus ROSACre mutations as well as F142L. ROSACre 1 showed ~10% recombination of the ROSA*loxP* target as well as substantial, albeit lower than wild-type, activity on *loxP*. These data show that the performance of ROSACre variants in bacterial assays is consistent with their ability to recombine these target sequences in a experiments conducted in mammalian cells.

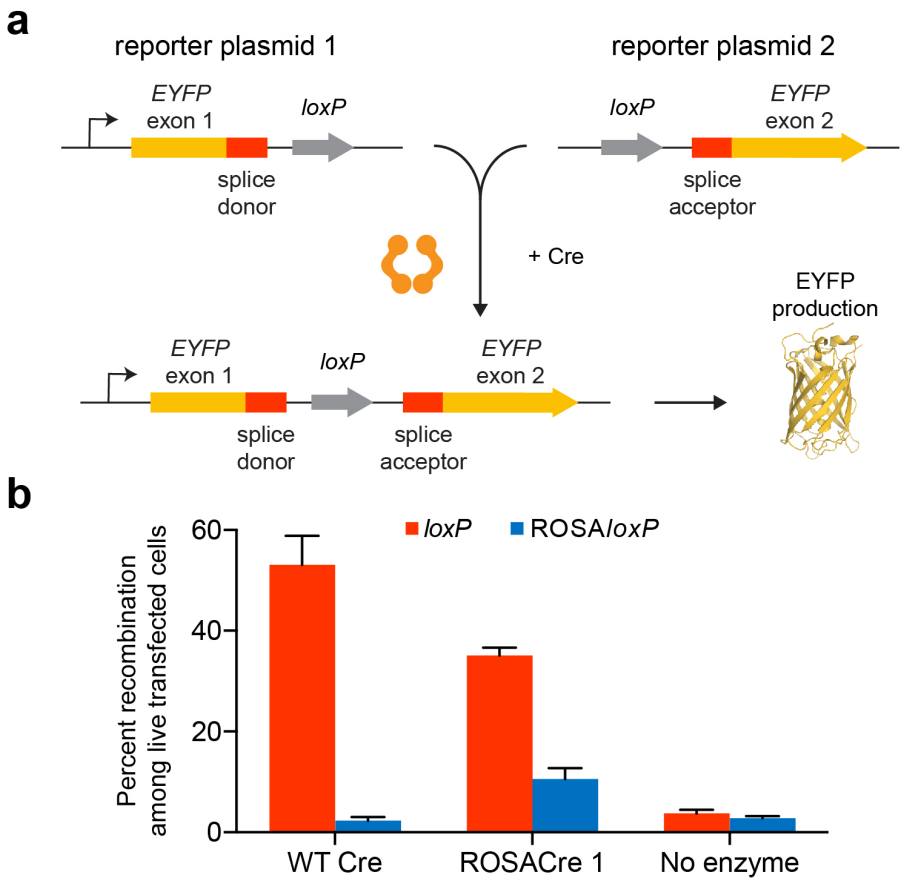


Figure 2.6. ROSACre recombination of the ROSA*loxP* sequence in mammalian cells. a, Cells were transfected with recombinase expression plasmid and two reporter plasmids bearing *EYFP* exons 1 and 2 adjacent to splice donor or splice acceptor sequences, respectively. Recombinase-mediated integration between two target sequences located in the intronic regions of the reporter plasmids results in *EYFP* expression[114]. **b,** Cre and ROSACre 1 activity on *loxP* and ROS*AloxP* was measured as the fraction of cells exhibiting EYFP fluorescence. The percentage of EYFP-positive cells shown is of transfected cells (determined by gating for the presence of co-transfected plasmid constitutively expressing *iRFP*) and 10,000 live events were recorded for each experiment. Values and error bars represent the mean and standard deviation of three independent biological replicates.

To investigate one possible application of ROSACre, we next attempted to integrate foreign DNA into the genome of unmodified human cells using ROSACre 1. We co-transfected HEK293 cells with a plasmid expressing ROSACre 1 and an integration donor plasmid encoding a single ROSA*loxP* target and a neomycin resistance gene. Recombinase-mediated integration of the plasmid into the genome confers geneticin (G418) resistance to the cell and its daughter cells. After transfection, we grew the HEK293 cells in selective media for two weeks, during which period control cells lacking a recombinase expression plasmid were susceptible to G418. Following selection, we harvested the genomic DNA of surviving cells and performed nested PCR with primers internal to the integration cassette paired with primers that bind genomic sequences upstream or downstream or the predicted integration site. We did not detect PCR amplicons of the expected size by gel electrophoresis, and high-throughput sequencing of the amplicons failed to produce evidence of targeted genomic integration at the ROSA26 locus.

Together, these results demonstrate that variants of Cre generated in PACE are active in mammalian cells on a sequence identical to one present in the human genome, but may not integrate efficiently enough to detectably modify the genome.

2.2.5 Addressing low activity and promiscuity of ROSACre variants

I next conducted experiments aimed at improving the activity and specificity of ROSACre variants. Attempts at higher-stringency positive selection of ROSACre SP on the ROSA*loxP* target were frustrated by the emergence of recombinant SP in the PACE lagoons. So-called “cheater phage” were able to propagate on PACE host cells in an activity-independent manner, outcompeting SP that encoded functional library members. Sanger sequencing of SP from different lagoons with a cheating phenotype revealed independent instances of recombination between the AP and SP, with gIII reinserted in the SP genome. I suspected that promiscuous recombination by ROSACre variants was responsible for producing the cheater phage. Upon closer inspection of the mutations accumulated by ROSACre in PACE, the source of this

promiscuity became evident. For example, ROSACre contains multiple substitutions of functionalized amino acids at the protein:DNA interface, such as Arg or Glu, with small hydrophobic amino acids like Ala, Gly, or Val (Figure 2.5d), indicative of broadened rather than retargeted specificity. Indeed, a previous study found that E262A or G mutations, both observed in PACE-evolved variants, were sufficient to increase the mismatch tolerance of Cre[84]. These results suggest that ROSACre weakly recognizes ROSA*loxP* as one of many possible substrates, and I therefore sought to modify our PACE experiments to promote specific recognition of the ROSA*loxP* target.

I first attempted negative selection against residual *loxP* activity among the ROSACre variants (Figure 2.7a). Continuous counterselection in PACE is achieved by linking unwanted activity of the POI to production of pIII-neg, a dominant-negative mutant of pIII that inhibits propagation of progeny phage[95]. PIII-neg production is regulated by the inducible TetA promoter (P_{tet}), allowing for negative selection stringency to be modulated by the small molecule anhydrotetracycline (aTc). I constructed a recombinase negative selection circuit by inserting the floxed terminator between P_{tet} and gene III-neg on a separate AP-neg (Figure 2.7a). In overnight enrichment assays, I observed aTc concentration-dependent defects in ROSACre SP propagation on host cells bearing the ROSA*loxP* AP and *loxP* AP-neg. However, when I attempted PACE selections with the same host cells, I observed washout of SP bearing active Cre variants upon moderate induction of the negative selection circuit. I suspected that counterselection against *loxP* activity was too stringent, given that Cre has multiple indirect mechanisms for recognizing its native target[60]. But without knowledge of alternative DNA targets that would better serve as ROSACre counterselection substrates, I was unable to design additional negative selection experiments, and instead focused on different strategies to retarget ROSACre.

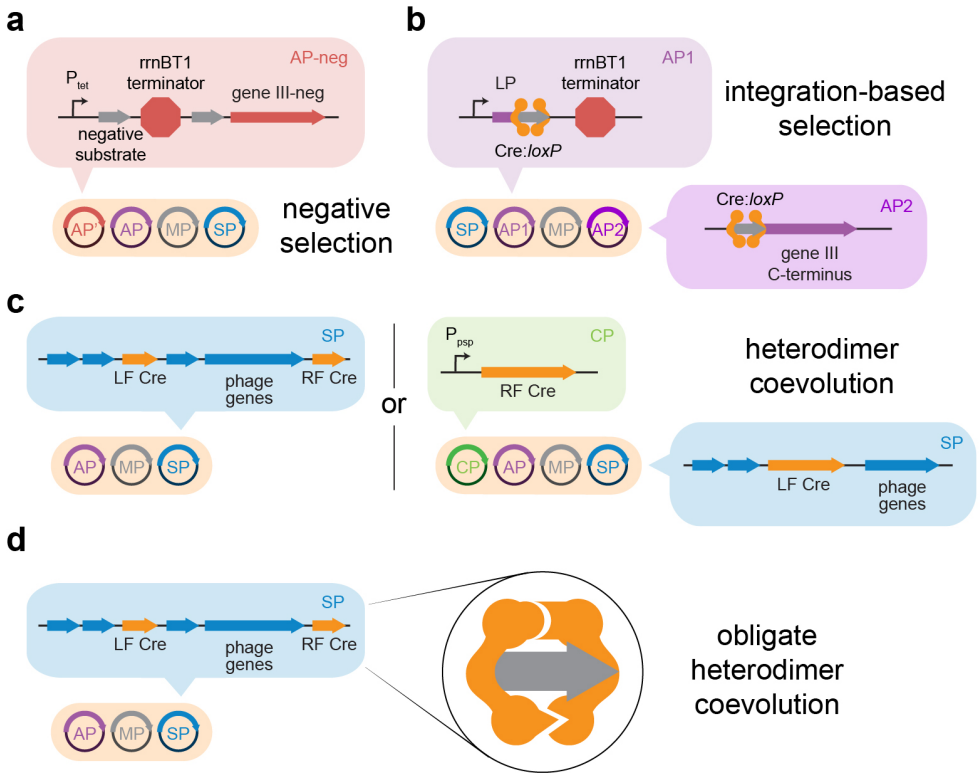


Figure 2.7. Modifications to PACE for promotion of enhanced activity and specificity of ROSACre variants. a, Negative selection in PACE is achieved by linking unwanted recombinase activity to the production of the dominant-negative pIII-neg, regulated by a small molecule-inducible promoter (P_{tet}). **b,** A selection for integrative recombination splits the second-generation recombinase AP between the transcriptional terminator and the downstream recombinase target. Recombinase-mediated integration between two target sequences located on AP1 and AP2 results in production of pIII'. **c,** Coevolution of a heterodimeric ROSACre pair was attempted by expressing LF Cre and RF Cre from a dual SP, and by evolving SP-encoded LF Cre in the presence of RF Cre expressed from a complementary plasmid (CP) regulated by the phage-shock promoter (P_{psp}). **d,** Coevolution of an obligate heterodimeric ROSACre pair was attempted by installation of mutations at the monomer interface of LF and RF Cre encoded on a dual SP.

I next attempted to select for integrative recombination among ROSACre SP in PACE. I designed a integration-based selection circuit as a two-plasmid system (Figure 2.7b), in which AP1 contains a promoter, a sequence encoding the gIII LP, a recombinase target, and a transcriptional terminator, and AP2 contains a recombinase target and the C-terminal domains of gIII with no upstream promoter. Recombinase-mediated integration between the AP1 and AP2 targets results in expression of pIII'. In addition to promoting unidirectional recombination, a selection circuit which splits gIII across two plasmids should theoretically reduce the possibility

of recombinant cheater phage. I confirmed the ability of host cells bearing the ROSA*loxP* integration circuit to support ROSACre SP propagation in overnight enrichment assays, and then attempted PACE on these same host cells. Although I observed the emergence of recombinant SP within 66 hours, one of the ROSACre variants isolated at an earlier time point showed ROSA*loxP* activity comparable to that of ROSACre 1 in the mammalian *EYFP* assay. This variant, termed ROSACre 2, contains the 11 LF consensus mutations (Figure 2.5d) as well as E69A, A112V, V182I, and R241Q. While three of the newly observed mutations are substitutions of small hydrophobic amino acids, and thus unlikely to affect substrate recognition, the R241Q substitution is proximal to the protein:DNA interface and could potentially contribute to altered specificity. These findings suggest that an integration-based selection can generate novel ROSACre variants, but remains susceptible to the emergence of recombinant cheater phage.

The final strategy I explored for promoting retargeted specificity of ROSACre was coevolution of LF and RF Cre variants. I reasoned that a heterodimeric pair consisting of monomers specific for each half-site would be less susceptible to off-target recombination. In overnight enrichment assays, dual SP expressing both LF and RF Cre variants (Figure 2.7c) did not propagate on ROSA*loxP* host cells. To generate compatible pairs of LF and RF Cre, I initiated PACE on host cells with the ROSA*loxP* AP and a complementary plasmid (CP) expressing RF Cre under P_{psp} regulation (Figure 2.7c). In order to survive this selection, LF Cre SP must recombine ROSA*loxP* in the presence of RF Cre, and I expected that cooperative binding with RF Cre might provide a selective advantage. SP propagated at high titers at flow rates up to 1.5 volumes per hour for a total of seven days. I cloned the resulting LF Cre variants into the dual SP with RF Cre, and the resulting dual SP library was active on ROSA*loxP* in overnight enrichment assays.

I conducted PACE coevolution selections with the dual SP and ROSA*loxP* host cells, but these experiments resulted in the emergence of recombinant cheater phage. I reasoned that,

since the LF variants had activity on ROSA*loxP*, there was weak selection pressure for LF Cre to operate as a heterodimeric partner of RF Cre, undermining the rationale for the coevolution. I therefore sought to more deliberately promote cooperation between LF and RF Cre by incorporating insights from studies in which SSRs were engineered to operate as obligate heterodimeric pairs (Figure 2.7d). Several groups have developed heterodimeric Cre variants by modifying the protein:protein interface of neighboring monomers[115,116]. For example, Havranek and colleagues used the Rosetta molecular modeling program[117] and rational mutagenesis to generate a series of Cre variants with increasing heterodimeric behavior[116]. The "HetA1" and "HetB1" Cre variants show reduced but detectable activity on *loxP* in the absence of the partner monomer, but the "HetA2" and "HetA3" variants (incorporating mutations A1+A2 and A1+A2+A3, respectively) demonstrate increasing reliance on the "HetB2" variant (bearing B1+B2 mutations) for recombination (Figure 2.8a).

I inserted the A1-A3 mutations in ROSACre 1 and ROSACre 2, and the B1 and B2 mutations in RF Cre, and assessed the activity of pairwise combinations of the heterodimer variants on *loxP* or ROSA*loxP* reporters in mammalian cells (Figure 2.8b). All pairs of ROSACre and RF Cre showed higher activity on *loxP* than ROSA*loxP*. Activity of the heterodimer pairs on the ROSA*loxP* target was similar to unmodified ROSACre 1 (Figures 2.8b, 2.6b), and decreased with the introduction of additional heterodimer mutations (*i.e.*, HetA1 > HetA2 > HetA3; HetB1 > HetB2). However, the activity of the ROSACre heterodimer variants in the absence of RF Cre did not greatly exceed the level of background signal for the assay, suggesting that the obligate heterodimeric mutations conferred dependence of ROSACre on the presence of RF Cre.

I then attempted coevolution of the obligate heterodimeric ROSACre and RF pairs in PACE (Figure 2.7d). I generated versions of the dual SP with HetA1 or HetA2 ROSACre paired with HetB1 or HetB2 RF Cre. I observed no propagation on ROSA*loxP* host cells in overnight enrichment assays, suggesting low activity of the engineered Cre pair. To enable coevolution, I used drift PACE to select for increased activity of the heterodimeric pair on ROSA*loxP*.

Induction of the drift MP was gradually decreased over 4 days, and dual SP were propagated for an additional 48 hours without drift. Sequencing analysis of surviving phage revealed that the dual SP had lost functional RF Cre through the introduction of premature stop codons, suggesting that ROSACre could operate alone on ROSA*loxP*, even with the heterodimer mutations.

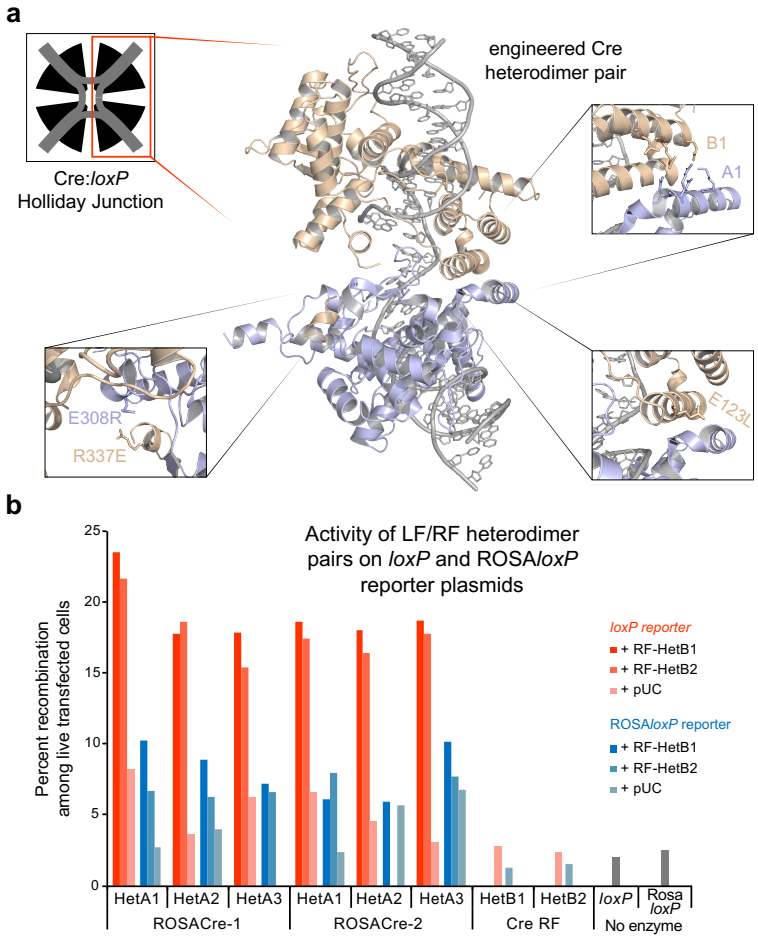


Figure 2.8. Mutations at the interface between Cre monomers promote obligate heterodimeric activity. a, Mutations of engineered obligate heterodimeric Cre variants[116] mapped onto the structure of Cre monomers in complex with *loxP* in the Holliday Junction conformation[118]. A1 and B1 heterodimeric mutations occur at the interface between helices A and C (top right inset). A2 and B2 mutations consist of a reversal of polarity of a salt bridge near the Cre C-terminus (bottom left inset). The A3 mutation occurs at residue 123, where mutation of Glu to Leu is predicted to disfavor homodimeric binding (bottom right inset). **b,** Cells were transfected with reporter plasmids for *loxP* or ROSA*loxP* and expression plasmids for heterodimeric pairs of ROSACre variants and RF Cre. Each variant was also co-transfected with pUC dummy plasmid in place of the partner heterodimer. The percentage of EYFP-positive cells shown is of transfected cells (determined by gating for the presence of co-transfected plasmid constitutively expressing *iRFP*) and 10,000 live events were recorded (n=1 biological replicate).

Together, the attempted modifications to PACE did not result in improvements to the activity or specificity of the ROSACre variants. I therefore considered alternative techniques for continued retargeting of Cre.

2.2.6 Practical challenges of evolving recombinases

Having attempted multiple methods of evolving Cre using PACE, I opted to critically evaluate other directed evolution strategies for retargeting recombinases. The most extensive retargeting of Cre has been accomplished using substrate-linked protein evolution (SLiPE)[92]. In SLiPE, host *E. coli* are transformed with the pEVO plasmid – which encodes a partially mutagenized recombinase variant and a floxed restriction enzyme site[92] – and cultured on agar plates. pEVO is collected from transformants and subjected to restriction enzyme digestion. Recombinase-mediated deletion of the restriction site prevents digestion of the plasmid encoding the functional SSR variant, and PCR is used to amplify and further diversify the variant pool. Performing successive rounds of SLiPE has resulted in Cre variants with activity on targets that differ from *loxP* at greater than 50% of base pairs[72,75].

The PACE recombinase selection appears to offer several advantages over SLiPE. Rounds of selection in PACE occur in as few as 10 minutes and require minimal researcher intervention once the experiment has been initiated. In comparison, dozens of rounds of manual directed evolution were required for retargeting Cre with SLiPE, with each round likely occupying several days of researcher time. PACE also offers the advantage of continuous *in vivo* mutagenesis, resulting in the facile generation of variant libraries that greatly exceed the size of libraries use in typical discrete evolution experiments[102]. Finally, the PACE selection rewards SSR variants which perform multiple recombination events per cell, as the production of progeny phage scales with pIII expression levels[101]. In contrast, a single recombination event permanently modifies the pEVO plasmid and permits survival in SLiPE. As a consequence,

many dozens of surviving variants must be characterized post-selection to determine the most active enzyme, and SLiPE experiments may experience an elevated false-positive rate[119].

The main impediment to carrying out recombinase selections in PACE is the emergence of a promiscuous phenotype, and I was curious as to how SLiPE avoids this problem. I reasoned that one potential reason for the promiscuity observed in ROSACre variants is that broadened specificity is easier to achieve than retargeted specificity, and the current PACE circuit applies limited pressure for the latter. Indeed, my difficulties with subcloning several of the ROSACre variants suggested that PACE yields recombinases which are genotoxic to the *E. coli* host cells. During the course of a PACE selection, Cre variants that recognize sequences in the *E. coli* genome but nevertheless recombine the selection circuit can produce progeny phage even if the host cell is killed due to those genomic modifications. In comparison, a Cre variant with a genotoxic phenotype would not pass a plate-based selection such as SLiPE because the library member relies on the viability of the *E. coli* host cell for its own amplification. I suspect that an underappreciated aspect of SLiPE is implicit negative selection against promiscuous recombination, as SSR variants must not be so broadly active as to recognize sequences within the host genome.

I therefore attempted to design a recombinase selection that maintains the appealing properties of PACE and incorporates the benefits of discrete evolutionary techniques. I envisioned inoculating fluorescent reporter *E. coli* with SP containing Cre, and screening for variants with desirable properties using fluorescence-activated cell sorting (FACS) of the *E. coli*; SP could then be recovered from the sorted cells and subjected to further selection (Figure 2.9a). This system would synergize with the existing PACE selection, as the same SP vectors could be used for PACE or FACS experiments. Additionally, in the proposed scheme, passing the selection is made dependent on the viability of the fluorescent reporter cells, which are grown in the presence of Cre-bearing SP for several hours before FACS and overnight after sorting. Multiple fluorescent reporter plasmids could be devised in order to conduct

simultaneous positive and negative selection. Finally, FACS screens are rapid, high-throughput, and afford fine-tuned control over stringency[120]. Indeed, a FACS-based method has previously been reported for retargeting Cre[91].

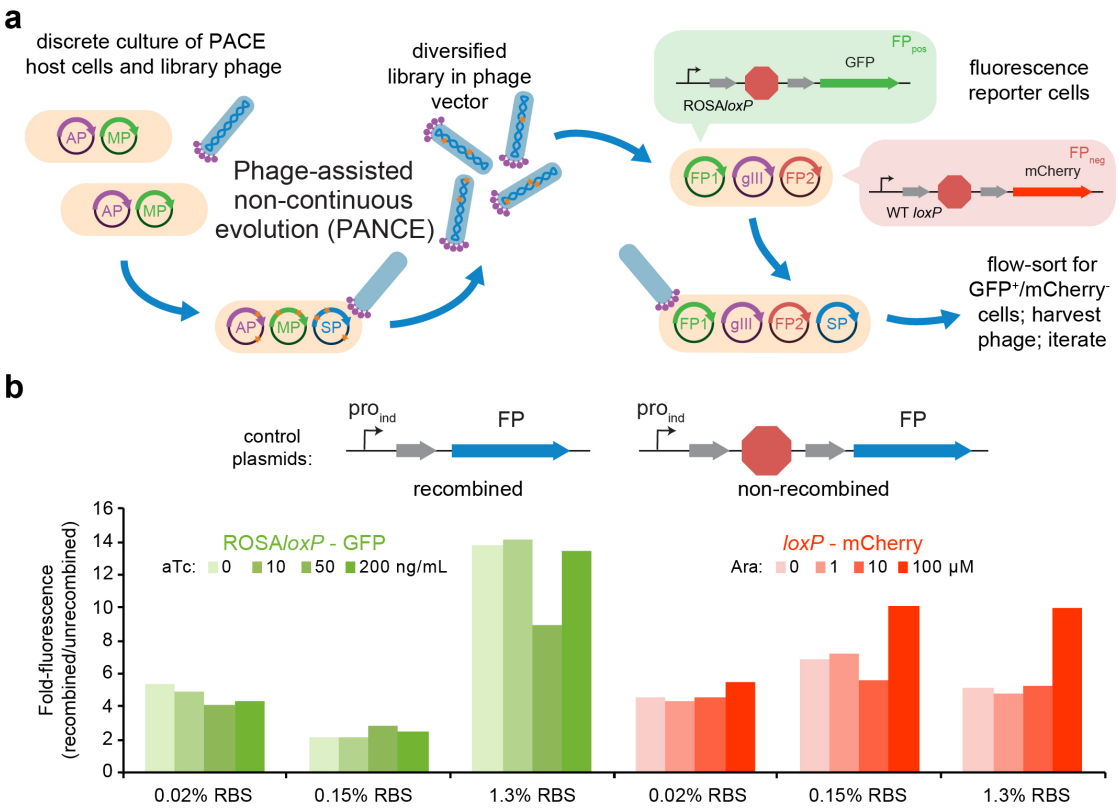


Figure 2.9. Overview of FACS-based method for directed evolution of recombinases. a, Schematic of a FACS-based recombinase selection. In the first phase, PACE host cells containing the recombinase selection AP are infected with SP bearing diverse libraries of recombinase variants and grown in discrete cultures overnight (phage-assisted non-continuous evolution, or PANCE). SP genotypes with higher activity produce more infectious progeny and enrich in the population. In the second phase, SP from PANCE are used to infect fluorescent reporter cells and cultured for 6 hours. Activity on desired and undesired substrates are linked to the expression of different fluorescent proteins using a floxed terminator-based circuit. *E. coli* are sorted on the basis of favorable fluorescent protein expression using FACS and grown overnight. The presence of a gIII-producing plasmid in the fluorescent reporter cells enables the recovery of SP after sorting. The resulting SP populations are then re-screened via the same FACS workflow or used to seed a subsequent round of PACE or PANCE. **b,** *E. coli* were transformed with control plasmids simulating pre- and post-recombination levels of fluorescent protein (FP) expression from two orthogonal series of plasmids, each with inducible promoters (pro_{ind}) and variable ribosome binding site (RBS) strengths upstream of the FP. Mean fluorescence for each condition was calculated among the live (propidium iodide-negative) population of live *E. coli* cells. The fold-fluorescence was calculated for each pre- and post-recombination plasmid pair by dividing the respective mean fluorescence values.

I designed orthogonal plasmids for assessing recombinase activity on two different substrates by monitoring fluorescent protein (FP) expression. The fluorescent reporter plasmids contained a floxed terminator between the FP and an inducible promoter, mirroring the design of the first-generation PACE circuit (Figure 2.1b). To assess the dynamic range of the reporter, I generated control plasmids that lacked the transcriptional terminator, simulating the product of recombination, and measured the fluorescence signal from cells transformed with either the unrecombined or control plasmid (Figure 2.9b). I observed a maximal signal increase of 15-fold from cells bearing the control ROSA*loxP* - GFP plasmid versus cells bearing the unrecombined plasmid, and a 10-fold difference for cells bearing the *loxP* – mCherry plasmids. I anticipated that the theoretical > 10-fold increase in signal would be sufficient to discriminate between active and inactive recombinase variants. However, when fluorescent reporter cells were inoculated with SP, the observed fluorescence signal was far lower than the theoretical maximum, and I was unable to distinguish *loxP* reporter cells inoculated with no phage, Cre SP, or SP bearing an unrelated recombinase. This data suggests that even wild-type Cre activity on the *loxP* site is not sufficient to trigger fluorescence expression approaching maximum levels under the current configuration.

While there are potential improvements to be made to the proposed FACS-based method of evolving recombinases, I opted to pursue other avenues of research that seemed more promising, described in the following chapters of this dissertation.

2.3 Conclusion

Currently, a facile, high-throughput method for retargeting the specificity of recombinases – a promising class of enzymes with potential applications in precision genome editing – does not exist. PACE has been successfully applied to a diverse group of enzymes, and we found it conceptually straightforward to link recombinase activity to SP survival in PACE. We established a PACE selection with the goal of retargeting Cre to recognize a sequence

present in a human genomic safe harbor locus, and used it to generate variants of Cre with substantial activity on the ROSA*loxP* sequence in mammalian cells. Efforts to improve the activity and specificity of the ROSACre variants were unsuccessful, leading to several attempted modifications to PACE and an alternative methodology that incorporated FACS of fluorescent reporter cells. While PACE has many appealing properties, the difficulties I experienced suggest that the recombinase selection circuit is not a viable strategy for retargeting SSRs as currently configured. In the following chapters of this dissertation, I describe alternative approaches to developing recombinase-based genome editing tools, as well as potential future applications of recombinase PACE.

2.4 Methods

General methods

All oligonucleotides and gBlocks were purchased from Integrated DNA Technologies (IDT). All enzymes and buffers were purchased from New England Biolabs (NEB) unless noted. PCR was performed using either Phusion U Green Multiplex PCR Master Mix (ThermoFisher Scientific) or Q5 Hot Start High-Fidelity 2x Master Mix (NEB). All plasmids were generated by USER cloning, blunt-end ligation cloning of 5'-phosphorylated PCR products, or ligase cycling reaction[121] and transformed into One Shot Mach1 T1 *E. coli* (ThermoFisher Scientific). Plasmids for mammalian cell transfection were prepared using an endotoxin-removal plasmid purification system, PureYield Plasmid Miniprep System (Promega).

Preparation and transformation of chemically competent cells

Strain S2060 (ref. 96) was used in all transcriptional activation and overnight phage propagation assays, as well as in all PACE experiments. To prepare competent cells, an overnight culture was diluted 1,000-fold into 50 mL of 2xYT media (United States Biologicals) supplemented with streptomycin and grown at 37 °C until OD_{600} ~ 0.4-0.5. Cells were collected

by centrifugation at 8,000 *g* for 10 minutes at 4 °C. The cell pellet was then resuspended by gentle stirring in 10 mL of ice-cold TSS (LB media supplemented with 5% v / v DMSO, 10% w / v PEG 3350, and 20 mM $MgCl_2$). The cell suspension was aliquoted and frozen dry ice, and stored at -80 °C until use.

To transform cells, 100 µL of competent cells thawed on ice was added to a prechilled mixture of plasmid in 80 µL deionized water and 20 µL KCM solution (500 mM KCl, 150 mM $CaCl_2$, and 250 mM $MgCl_2$ in H_2O). The mixture was incubated on ice for 10 minutes and heat shocked at 42 °C for 45 s before 200 µL of SOC media (NEB) was added. Cells were recovered at 37 °C for 30 minutes and streaked on agar plates containing appropriate antibiotics, and incubated at 37 °C for 16-18h.

Transcriptional activation assay

S2060 cells were transformed with the recombinase circuit of interest as described above. Overnight cultures of single colonies grown in 2xYT media supplemented with maintenance antibiotics were diluted 500-fold into DRM media[76]. Cells were grown at 37 °C until OD_{600} ~ 0.4-0.6, then induced with 100 ng / mL anhydrotetracycline (aTc; Fluka) and 5 µM arabinose (Gold Biotechnology) before incubation for an additional 1 h at 37 °C. 120 µL of cells were transferred to a 96-well black-walled clear-bottomed plate (Costar), and 600 nm absorbance and luminescence were read using a Tecan Infinite M1000 Pro microplate reader. OD_{600}-normalized luminescence values were obtained by dividing the raw luminescence by background-subtracted 600 nm absorbance. The background value was set to the 600 nm absorbance of wells containing DRM only.

Overnight phage propagation assay

S2060 cells were transformed with the AP(s) of interest as described above. Overnight cultures of single colonies grown in 2xYT media supplemented with maintenance antibiotics

were diluted 1,000-fold into DRM media and grown at 37 °C until OD_{600} ~0.4–0.6. Cells were then infected with SP at a starting titer of 1×10^4 pfu/mL. Cells were incubated for another 16–18 h at 37 °C, then centrifuged at 3,000g for 10 minutes. The supernatant containing phage was filtered through a 0.2 μm cellulose acetate syringe filter (Sartorius) and stored at 4 °C until use.

Plaque assays

S2060 cells were transformed with pJC175e[95] as described above. Overnight cultures of single colonies grown in 2xYT media supplemented with maintenance antibiotics were diluted 1,000-fold into fresh 2xYT media and grown at 37 °C until OD_{600} ~0.6–0.8 before use. SP were serially diluted 100-fold (4 dilutions total) in DRM. 40 μL of cells were added to 10 μL of each phage dilution, and to this 200 μL of liquid (55 °C) top agar (2xYT media + 0.6% agar) was added and mixed by pipetting up and down once. This mixture was then immediately pipetted onto one well of a 12-well plate (Costar) already containing 1 mL of solidified bottom agar (2xYT media + 1.5% agar, no antibiotics). After solidification of the top agar, plates were incubated at 37 °C for 16–18 h.

For Sanger sequencing of phage, single plaques were picked into 2xYT and grown at 37 °C for 16–18 h. The cells were pelleted by centrifugation at 3,000g for 10 minutes, and the DNA in the supernatant was amplified using the Illustra TempliPhi 100 Amplification Kit (GE Life Sciences).

Phage-assisted continuous evolution

PACE apparatus, including host cell strains, lagoons, chemostats, and media, were all used as previously described[76,99]. To reduce the likelihood of contamination with gIII-encoding recombined SP, phage stocks were purified as previously described[99].

Chemically competent S2060s were transformed with AP(s) and an MP as described above, and a single colony was grown in 2xYT until OD_{600} ~0.6–0.8. This culture was used to inoculate a chemostat containing 100 mL DRM. The chemostat was grown to OD_{600} ~0.8–1.0,

then continuously diluted with fresh DRM at a rate of 0.5 chemostat volumes/h or higher. The chemostat was maintained at a volume of 80-100 mL.

Prior to SP infection, lagoons were continuously diluted with culture from the chemostat at 0.5 lagoon vol/h or higher and pre-induced with 10 mM arabinose. If a drift MP was used, the lagoons were also pre-induced with aTc. Lagoons were infected with SP at a starting titer of typically > 10^9 pfu and maintained at a volume of 40 mL. Samples (500 µL) of the SP population were taken at indicated times from lagoon waste lines. The mixture of cells and phage was passed through a 0.22 µm cellulose acetate centrifugal filter (Costar) and stored at 4 °C. Lagoon titers were determined by plaque assays using S2060 cells transformed with pJC175e.

FACS-based selection

Chemically competent S2060s were transformed with fluorescent reporter plasmids as described above, and a single colony was grown in DRM until OD_{600} ~0.6–0.8. For control experiments, the cells were centrifuged at 3,000g for 10 minutes, resuspended in PBS supplemented with propidium iodide viability dye (Bio Rad), and analyzed on a BD FACSAria IIIu cell sorter.

For testing the performance of recombinase variants, fluorescent reporter cells containing pJC175e were diluted in DRM to OD_{600} ~0.4, infected with SP (typically > 10^7 pfu), and grown at 37 °C for 6h. Then the cells were prepared for flow analysis as described above.

HEK293T transfection and flow cytometry

HEK293T cells (ATCC CLR-3216) were cultured in Dulbecco's Modified Eagle's Medium (DMEM, Corning) supplemented with 10% fetal bovine serum (FBS; Life Technologies). Cells were seeded into 48-well poly-D-Lysine-coated plates (Corning) in the absence of antibiotic. 12-15h after plating, cells were transfected with 0.5 µL of Lipofectamine 2000 (ThermoFisher Scientific) using 50 ng of recombinase plasmid, 100 ng of each reporter plasmid, and 10 ng of

fluorescent protein expression plasmid as a transfection control. Cells were cultured for 3 d before they were washed with PBS (ThermoFisher Scientific) and detached from plates by the addition of TrypLE Express (ThermoFisher Scientific). Cells were diluted in 250 µL of culture media and run on a BD Accuri C6 analyzer.

Mammalian genomic integration experiments

HEK293 cells (ATCC CLR-1573) were cultured in DMEM supplemented with FBS (full media). Cells were seeded into 6-well poly-D-Lysine-coated plates (Corning) in the absence of antibiotic. 12-15h after plating, cells were transfected with 4 µL of Lipofectamine 2000 using 1 µg of recombinase plasmid and 1 µg of integration donor plasmid. Cells were cultured for 3 d, then passaged in 75 mm^2 flasks (Corning) in full media supplemented with 500 µg / mL geneticin (G418; VWR). Selective media was replaced every 3 d and cells were passaged into new flasks when they reached confluency. After 2 weeks of selection, genomic DNA was harvested using the E.Z.N.A. Tissue DNA Kit (Omega Bio-Tek) and eluted in 100 µL EB.

Nested PCR was carried out using Q5 Hot Start Polymerase 2x Master Mix supplemented with 3% DMSO and diluted with nuclease-free water (GE Life Sciences). DNA was analyzed by electrophoresis on a 1% agarose gel in TAE alongside a 1 Kb Plus DNA ladder (ThermoFisher Scientific).

Chapter 3:

A Programmable Cas9-Serine Recombinase Fusion Protein That Operates on DNA Sequences in Mammalian Cells

Adapted from Brian Chaikind, Jeffrey L. Bessen, David B. Thompson, Johnny H. Hu, and David R. Liu. A programmable Cas9-serine recombinase fusion protein that operates on DNA sequences in mammalian cells. *Nucleic Acids Research*, **44**, 9758-9770 (2016).

Brian Chaikind, David Thompson, and I contributed to the initial design of a programmable recombinase genome editing tool. Brian Chaikind designed and performed the experiments described in sections 3.2.1-3.2.4 and figures 3.1-3.4. Johnny Hu was responsible for writing the software described in section 3.2.2. Brian Chaikind and I designed the experiments described in section 3.2.5. I designed and performed all remaining experiments.

3.1 Introduction

Efficient, programmable, and site-specific homologous recombination remains a longstanding goal of genetics and genome editing[22]. An enzyme that catalyzes recombination at sites specified by the researcher would be a valuable tool for studying the phenotypic effects of genetic alterations, enabling gene integration or gene deletion-based therapeutic strategies. Tyrosine and serine recombinases such as Cre, Flp, and phiC31 integrase have been widely used to catalyze the recombination of exogenous DNA into model organisms[80,122]. However, the use of these enzymes has been limited by their intrinsic, non-programmable DNA sequence specificity. Most small serine recombinases, for example, recognize a partially palindromic DNA sequence of approximately 20 base pairs[61]. Recombination using these enzymes at endogenous DNA sequences only occurs at pseudo-sites that resemble the recombinase's natural DNA recognition sequence, or at genomic sequences for which the recombinase has been experimentally evolved[72,75,80,92,123-126].

To increase the number of sites amenable for targeted recombination in cells, researchers have fused hyperactive variants of small serine recombinases to zinc finger and TALE DNA-binding proteins[78,79,127-129]. Because the catalytic domain and DNA-binding domain are partially modular in some recombinases, replacement of the natural DNA-binding domains with zinc-finger or TALE repeat arrays can partially retarget these enzymes to specified DNA sequences. Although the guide RNA (gRNA)-programmed Cas9 nuclease has quickly grown in popularity due to its relatively unrestricted DNA binding requirements and its ease of use, a gRNA-programmed recombinase has not been reported.

Here we describe the development of recCas9[81], a gRNA-programmed recombinase based on the fusion of an engineered Gin recombinase catalytic domain with a catalytically inactive, or "dead", Cas9 (dCas9). The recCas9 enzyme operates on a minimal core recombinase site (NNNNAAASSWWSSTTTNNNN) flanked by two guide RNA-specified DNA sequences. Recombination mediated by recCas9 is dependent on both gRNAs, resulting in

orthogonality among different gRNA:recCas9 complexes, and recCas9 functions efficiently in human cells on DNA sequences matching those found in the human genome. The recCas9 enzyme can also operate directly on the genome of cultured human cells, catalyzing a deletion between two recCas9 pseudo-sites located approximately 14 kb apart. Finally, I investigate fusions of dCas9 to promiscuous variants of Cre recombinase developed using phage-assisted continuous evolution. This work represents a key step toward engineered enzymes that directly and cleanly catalyze gene insertion, deletion, inversion, or chromosomal translocation with user-defined, single base-pair resolution in unmodified genomes.

3.2 Results

3.2.1 Fusing Ginβ recombinase to dCas9

The Liu group and others demonstrated that the N-terminus of dCas9 could be fused to the FokI nuclease catalytic domain, resulting in a dimeric dCas9-FokI fusion that cleaves DNA sites flanked by two gRNA-specified sequences[130,131]. We used the same linkage orientation to develop the "recCas9" fusion of dCas9 and Ginβ, a highly active catalytic domain of Gin recombinase previously evolved by Barbas and co-workers[132]. Ginβ promiscuously recombines several 20-bp *gix* sequences[132] related to the native *gix* core sequence CTGTAAACCGAGGTTTTGGA[133-135] (Table 1.1). We envisioned that recCas9 dimers would localize to a *gix* site directed by the presence of two flanking gRNA-specified sequences, enabling the Ginβ domains to catalyze DNA recombination in a gRNA-programmed manner (Figure 3.1d).

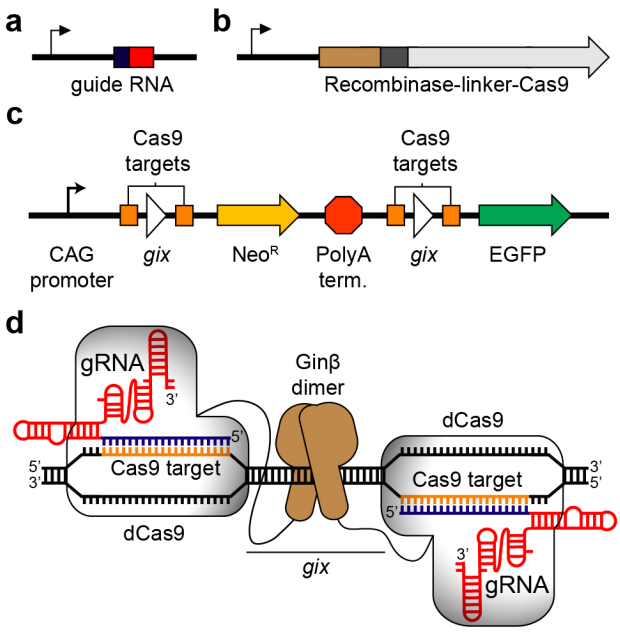


Figure 3.1. **Overview of recCas9 experimental setup**. **a-c,** Cells were transfected with guide RNA (gRNA) expression vectors under the control of the hU6 promoter (**a**), a dCas9-Ginβ expression vector under the control of a CMV promoter (**b**), and a recCas9 reporter plasmid (**c**). **d**, Co-transfection of these components results in reassembly of gRNA-programmed recCas9 at the target sites, mediating deletion of the poly-A terminator and allowing transcription of *EGFP*.

To assay the resulting recCas9 fusions, we constructed a reporter plasmid containing two recCas9 target sites flanking a poly-A terminator that blocks *EGFP* transcription (Figure 3.1c). Each recCas9 target site consists of a *gix* pseudo-site "core" flanked by gRNA binding sites. Recombinase-mediated deletion removes the terminator, restoring transcription of *EGFP*. We co-transfected HEK293T cells with this reporter plasmid, a plasmid transcribing the gRNAs, and a plasmid expressing candidate dCas9-Ginβ fusion proteins (Figure 3.1a-c), and used the fraction of cells exhibiting EGFP fluorescence to assess the relative activity of each fusion construct.

We first sought to optimize the architecture of the recCas9 components for maximal activity and gRNA dependence. We varied parameters such as the spacing between the core *gix* site and the gRNA-binding site (from 0- to 7-bp), as well as the linker length between the dCas9 and Ginβ domains ($(GGS)_2$, $(GGS)_5$, or $(GGS)_8$; Figure 3.2a,b). Most fusion architectures resulted in no observable gRNA-dependent *EGFP* expression (Figure 3.2c,d). However, one

fusion construct containing a linker of eight GGS repeats and 3- to 6-base pair spacers resulted in approximately 1% recombination when a matched, but not mismatched, gRNA was present (Figure 3.2e). Recombination was consistently higher when 5-6 base pairs separated the dCas9 binding sites from the core (Figure 3.2f). These results collectively reveal that specific fusion architectures between dCas9 and Ginβ can result in gRNA-dependent recombination at *gix* core sites in human cells. Unless otherwise noted, use of the term "recCas9" in this dissertation refers to this $(GGS)_8$-linker fusion construct.

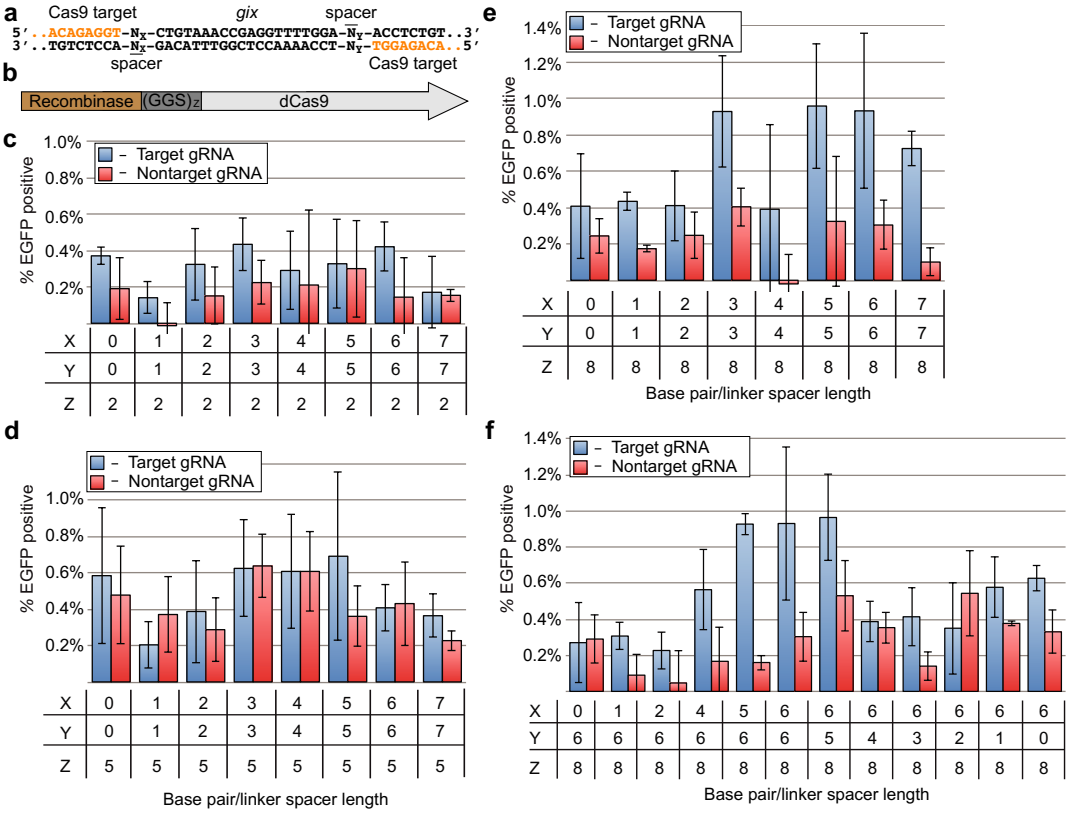


Figure 3.2. **Optimization of recCas9 fusion linker lengths and target site spacer variants**. **a**, A recCas9 target with identical 5' and 3' gRNA target sites (orange) and a *gix* core site (black). Varied parameters included (**a**) the length of the spacers separating the *gix* core site from the 5' and 3' binding sites (X, Y) and (**b**) the number of GGS repeats connecting Ginβ to dCas9 (Z). **c-e,** Cells were transfected with recCas9 reporters bearing targets in which X=Y and expression vectors with recombinase fusions where Z= $(GGS)_2$ (**c**), $(GGS)_5$ (**d**), or $(GGS)_8$ linkers (**e**). **f**, Cells were transfected with recCas9 and reporters bearing target sites composed of uneven base pair spacers (X≠Y). The percentage of EGFP-positive cells shown is of transfected cells (determined by gating for the presence of a co-transfected plasmid constitutively expressing *iRFP*) and at least 6,000 live events were recorded for each experiment. Values and error bars represent the mean and standard deviation of three independent biological replicates.

3.2.2 Targeting DNA sequences found in the human genome with recCas9

We hypothesized that low levels of observed activity may be a result of suboptimal gRNA or core *gix* sequences, consistent with previous reports showing that the efficiency of Cas9:gRNA binding is sequence-dependent[136]. Moreover, although our optimization was conducted with the native *gix* core sequence[133-135], several studies have shown that zinc finger-Gin or TALE-Gin fusions are active, and in some cases more active, on slightly altered core sites[78,79,128,132,137-139]. Therefore, we next sought to target sequences found within the human genome to test whether unmodified human genomic sequences were substrates for recCas9 and whether varying the gRNA and core sequences would increase recCas9 activity.

We identified potential target sites using the previous characterization of evolved Gin variants[132] as well as our above observations. We searched the human genome for sites that contained $CCN_{(30\text{-}31)}$-AAASSWWSSTTT-$N_{(30\text{-}31)}$-GG, where W is A or T, S is G or C, and N is any nucleotide. The $N_{(30\text{-}31)}$ includes the N of the NGG protospacer adjacent motif (PAM) of *S. pyogenes* Cas9 (SpCas9)[1], the 20-bp gRNA binding site, a 5- to 6-bp spacing between the Cas9 and *gix* sites, and the four outermost base pairs of the *gix* core site. The internal 12 base pairs of the *gix* core site (AAASSWWSSTTT) were previously determined to be critical for Ginβ activity[132].

Our search revealed approximately 450 potential recCas9 targets in the human genome (Appendix A). We generated a reporter plasmid bearing a DNA sequence found in *PCDH15* and gRNA expression vectors to direct recCas9 to the *gix* pseudo-site (Figure 3.3a). Co-transfection of the reporter plasmid, gRNA expression vectors, and the recCas9 expression vector resulted in *EGFP* expression in 11%-13% of transfected cells (Figure 3.3b), representing a > 10-fold improvement in activity over the results shown in Figure 3.2. These findings demonstrate that a more judicious choice of recCas9 target sequences can result in substantially improved recombination efficiency at DNA sequences matching those found in the human genome.

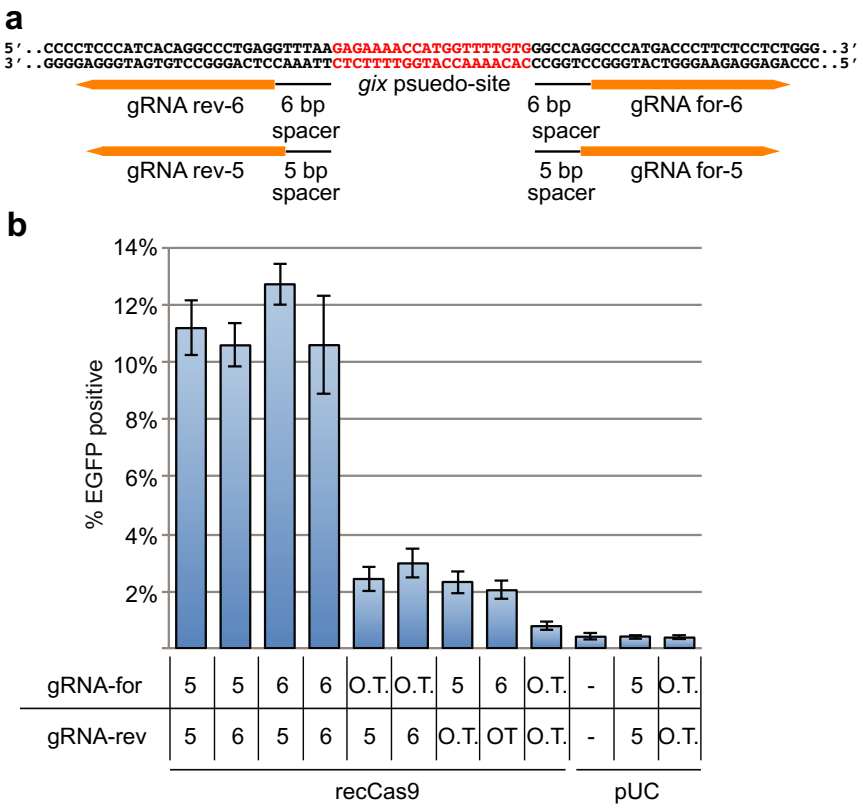


Figure 3.3. **The dependence of recCas9 activity on forward and reverse gRNAs. a**, A recCas9 reporter target bearing a sequence found within *PCDH15*, which contains offset protospacers on both the 5' and 3' side of a pseudo-*gix* core site. **b**, Cells were transfected with a recCas9 expression vector, *PCDH15* reporter plasmid, and all four pairs of gRNA expression vector as well as individual gRNA vectors with off-target (O.T.) gRNA vectors. The percentage of EGFP-positive cells reflects that of transfected (iRFP-positive) cells. At least 6,000 live events were recorded for each experiment. Values and error bars represent the mean and standard deviation of three independent biological replicates.

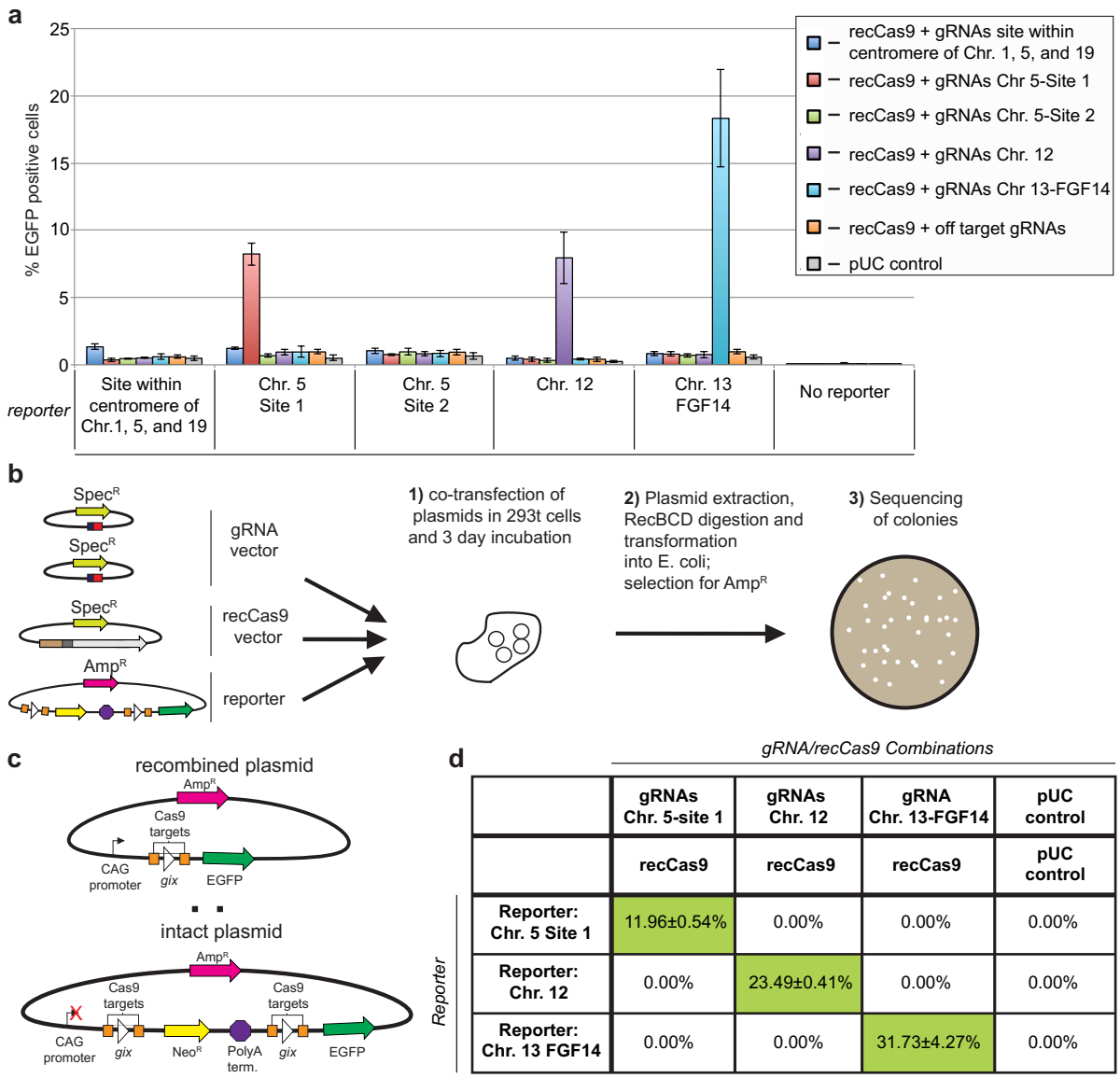
Next we determined whether both gRNA sequences were required for recCas9-mediated deletion. We transfected HEK293T cells with just one of the gRNA vectors targeting either the 5' or 3' target flanking sequences, resulting in 2.5-3% *EGFP* expression (Figure 3.3b). We speculate that the low levels of activity observed upon expression of just one of the targeting gRNAs may be due to the propensity of hyperactivated Gin monomers to spontaneously form dimers[140]; transient dimerization may occasionally allow a single protospacer sequence to localize the dimer to a target site. No recombination was detected when using off-target gRNA vectors or when the recCas9 vector was replaced by a pUC plasmid (Figure 3.3b).

Together, these findings demonstrate that recCas9 has substantial activity on well-matched targets identical to sequences found in the human genome, with maximal recombination dependent on the presence of both gRNAs.

3.2.3 Orthogonality of recCas9

Next, we sought to test the orthogonality of recCas9 for multiple reporter plasmids bearing different recCas9 targets found in the human genome. In choosing these targets, we prioritized sequences with the potential to serve as safe-harbor loci for genomic integration or which bear relevance to human disease. We used an ENSEMBL search[141] to identify which of the approximately 450 predicted recCas9 target sites fall within annotated genes. One target site fell within an intronic region of *FGF14*. Mutations within *FGF14* are believed to cause spinocerebellar ataxia type 27[142-146]. In addition, we identified four genomic targets that matched most of the five criteria for safe harbor loci described by Bushman and coworkers[90]. For these five sequences, we constructed recCas9 reporters with matching gRNA vectors.

To evaluate the orthogonality of recCas9 when programmed with different gRNAs, we tested all combinations of five gRNA pairs with five reporters. Co-transfection of the recCas9 components revealed substantial recCas9 recombination activity on three of the five reporters. Importantly, *EGFP* expression was strictly dependent upon co-transfection with a recCas9 expression vector and gRNA plasmids matching the target sequences on the reporter plasmid (Figure 3.4a). These results demonstrate that recCas9 activity is orthogonal and will only catalyze recombination at a *gix* pseudo-site when programmed with a pair of gRNAs matching the flanking sequences.



	gRNAs Chr. 5-site 1	gRNAs Chr. 12	gRNA Chr. 13-FGF14	pUC control
	recCas9	recCas9	recCas9	pUC control
Reporter: Chr. 5 Site 1	11.96±0.54%	0.00%	0.00%	0.00%
Reporter: Chr. 12	0.00%	23.49±0.41%	0.00%	0.00%
Reporter: Chr. 13 FGF14	0.00%	0.00%	31.73±4.27%	0.00%

Figure 3.4. **RecCas9 can target multiple sequences found in the human genome. a**, Cells were transfected with a recCas9 expression vector, recCas9 reporter plasmids bearing sequences found within the human genome, and pairs of cognate gRNA expression vectors. The percentage of EGFP-positive cells reflects that of transfected (iRFP-positive) cells. At least 6,000 live events were recorded for each experiment. Values and error bars represent the mean and standard deviation of three independent biological replicates. **b**, Transfections were repeated and episomal DNA was extracted and transformed into *E. coli,* and individual colonies were sequenced to determine the number of recombined and fully intact plasmids (**c,d**). Values reflect the mean and standard deviation of two independent biological replicates.

3.2.4 Characterization of recCas9 products

Zinc finger-recombinases have been reported to cause mutations at recombinase core-site junctions[132], and we tested whether such mutagenesis occurs during recCas9

recombination. To determine whether recCas9 activated *EGFP* expression via precise removal of the poly-A terminator sequence or via some other mechanism, we characterized reporter plasmids that had been exposed to recCas9. We co-transfected HEK293Ts with the recCas9 components and reporters for the chromosome 5-site 1, chromosome 12, and chromosome 13 (*FGF14* locus) targets. After 72 hours of incubation, plasmid DNA was extracted and transformed into *E. coli*, and single colonies containing reporter plasmids were sequenced (Figure 3.4b).

Individual colonies were expected to contain either an unmodified or a recombined reporter plasmid (Figure 3.4c). We only observed recombined plasmids for conditions in which reporter plasmids were co-transfected with cognate gRNA plasmids and recCas9 expression vectors (Figure 3.4d). For two biological replicates, the average fraction of recombined plasmid ranged from 12% for chromosome 5-site 1 to 32% for chromosome 13. While the sequencing data from Figure 3.4d showed agreements with the flow cytometry data in Figure 3.4a with respect to the relative activity of recCas9 on each reporter, the absolute levels of recombined plasmid were somewhat higher in the DNA sequencing experiments. We attribute this discrepancy to the lower sensitivity of the flow cytometry assay, in which cells may be transfected with several copies of the reporter plasmid, and one or multiple recombination events within a cell produce the same EGFP-positive phenotype. As a result, the percentage of EGFP-positive cells may correspond to a lower limit on the actual percentage of recombined reporter plasmids. Alternatively, the difference may reflect the negative correlation between plasmid size and transformation efficiency[147]; the recombined plasmid is approximately 5,700 base pairs and may be transformed more efficiently than the intact plasmid, which is approximately 6,900 base pairs.

We found minimal evidence of DNA damage as a result of plasmid exposure to recCas9. Of the 134 recombined sequences examined, all contained the expected recombination products. Further, of a total of 2,317 sequencing reads examined, only two contained potential

indels that could be attributed to recCas9. Theoretically, recCas9 could have caused these indels by catalyzing the excision and then re-integration of the poly-A terminator into the reporter, accumulating errors in the process. However, because excisive recombination is strongly favored over integrative recombination for entropic reasons[60], we suspect that these indels in otherwise non-recombined plasmids are the result of DNA damage that occurred during the transfection, isolation, or subsequent manipulation of the plasmid, and not the activity of recCas9.

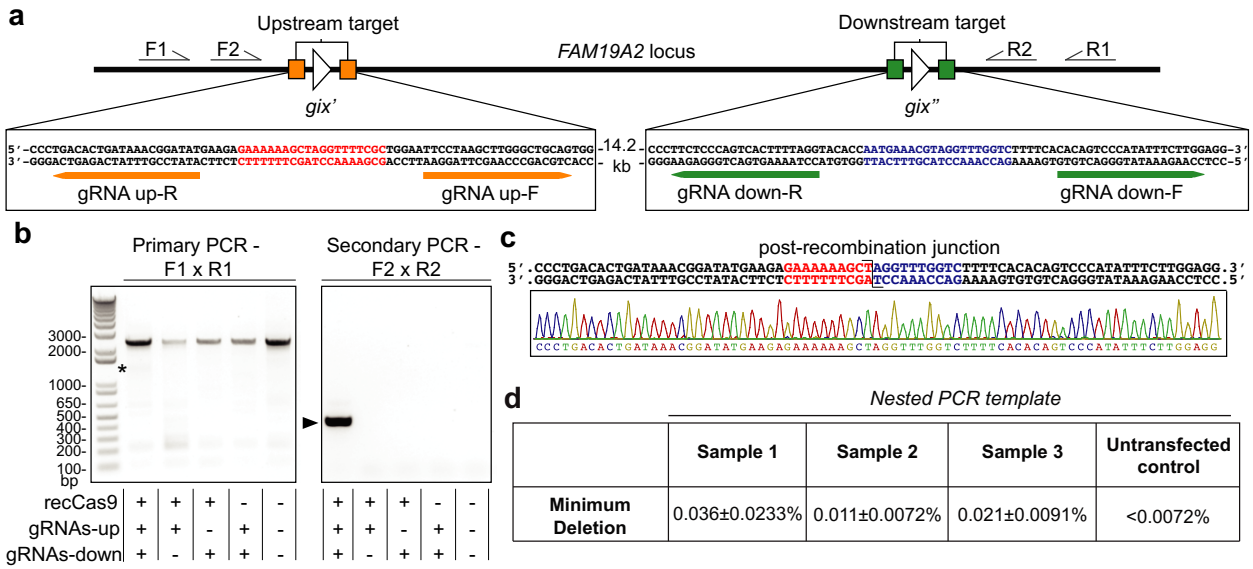
Taken together, these results establish that recCas9 can target multiple sites found within the human genome with minimal cross-reactivity or byproduct formation. Substrates undergo efficient recombination in human cells, but only in the presence of cognate gRNA sequences and recCas9, and generally do not contain mutations that typically result from cellular DNA damage repair.

3.2.5 RecCas9-mediated genomic deletion

We next investigated whether recCas9 is capable of operating directly on the genomic DNA of cultured human cells. First, we attempted to use recCas9 to genomically integrate a plasmid containing a neomycin resistance gene and a recCas9 target – chromosome 13-FGF14, chromosome 12, or chromosome 5-site 1 – with previously demonstrated activity (Figure 3.4a). However, we did not observe an increase in antibiotic resistance indicative of integration into the genome of HEK293 cells. Reasoning that excisive recombination would be higher efficiency than integration, we used the list of potential recCas9 recognition sites in the human genome (Appendix A) to identify pairs of sites that, if targeted by recCas9, would yield chromosomal deletion events detectable by PCR. We designed gRNA expression vectors that would direct recCas9 to targets closest to the chromosome 5-site 1 or chromosome 13 sites. The new target sites ranged from approximately 3 to 23 Mbp upstream and 7 to 10 Mbp downstream of chromosome 5-site 1, and 12 to 44 Mbp upstream of the chromosome 13-

FGF14 site. We cotransfected the recCas9 expression vector with each of these new gRNA pairs and the validated gRNA pairs used for the chromosome 5-site 1 or chromosome 13 targets, but were unable to observe evidence of chromosomal deletions by genomic PCR.

We reasoned that genomic deletion might be more efficient if the recCas9 target sites were closer to each other in the genome. We identified two recCas9 sites separated by 14.2 kb within an intronic region of *FAM19A2*, one of the TAFA-family genes encoding small, secreted proteins that are thought to have a regulatory role in immune and nerve cells[148]. Small nucleotide polymorphisms located in intronic sequences of *FAM19A2* have been associated with elevated risk for systemic lupus erythematosus (SLE) and chronic obstructive pulmonary disease (COPD) in genome-wide association studies[148]; deletion of the intronic regions of this gene might therefore provide insights into the causes of these diseases. I transfected HEK293T cells with plasmids expressing recCas9 and the *FAM19A2*-targeting gRNAs (Figure 3.5a), harvested genomic DNA after incubation for 72 hours, and carried out nested PCR to detect instances of genomic deletion. RecCas9-mediated recombination between the two sites should result in deletion of the 14.2 kb intervening region. Indeed, I detected this deletion event by nested PCR using gene-specific primers that flank the two *FAM19A2* recCas9 targets. I observed the expected PCR product that is consistent with recCas9-mediated deletion only in genomic DNA isolated from cells co-transfected with recCas9 and all four gRNA expression vectors (Figure 3.5b). I did not detect the deletion PCR product in the genomic DNA of cells transfected without either the upstream or downstream pair of gRNA expression vectors, without the recCas9 expression plasmid, or untransfected control cells (Figure 3.5b). Our estimated limit of detection for these nested PCR products is approximately 1 deletion event per 5,500 chromosomal copies. I isolated and sequenced the 415-bp PCR product corresponding to the predicted genomic deletion, and confirmed that it matched the expected product of recCas9-mediated genomic deletion and did not contain any insertions or deletions suggestive of DNA damage repair (Figure 3.5c).



	Sample 1	Sample 2	Sample 3	Untransfected control
Minimum Deletion	0.036±0.0233%	0.011±0.0072%	0.021±0.0091%	<0.0072%

Figure 3.5. RecCas9 mediates gRNA- and recCas9-dependent deletion of genomic DNA in cultured human cells. a, Schematic showing predicted recCas9 target sites located within an intronic region of the *FAM19A2* locus of chromosome 12 and the positions of primers used for nested PCR. **b**, Representative results of nested genomic PCR of template from cells transfected with the indicated expression vectors (n=3 independent biological replicates). The position of the 1.3-kb predicted primary PCR deletion product (asterisk) and the 415-bp deletion product after the secondary PCR (arrow) are shown. **c**, Sanger sequencing of PCR products resulting from nested genomic PCR of cells transfected with all four gRNA expression vectors and the recCas9 expression vector, compared to the predicted post-recombination product. **d**, Estimated lower limit of deletion efficiency of *FAM19A2* locus determined by limiting-dilution nested PCR. The values shown reflect the mean and standard deviation of three technical replicates.

We estimated a lower limit of genomic deletion efficiency by performing nested PCR on serial dilutions of genomic DNA samples[149]. A given amount of genomic DNA that yields the recCas9-specific PCR product must contain at least one edited chromosome. To establish a lower limit of recCas9-mediated genomic deletion, I therefore performed nested PCR on serial dilutions of genomic DNA (isolated from cells transfected with recCas9 and the four *FAM19A2* gRNA expression vectors) to determine the lowest concentration of genomic template DNA that results in a detectable deletion product. These experiments revealed a lower limit of deletion efficiency of 0.023±0.017% (average of three biological replicates; Figure 3.5d), suggesting that recCas9-mediated genomic deletion proceeds with at least this efficiency. Nested PCR of the genomic DNA of untransfected cells resulted in no detectable product, with an estimated limit of

detection of <0.0072% recombination. Together, these results indicate that recCas9 can mediate a targeted, seamless deletion of an endogenous DNA sequence present within the genome of cultured human cells.

3.2.6 Fusing promiscuous ROSACre variants to dCas9

Finally, I investigated whether the recCas9 fusion architecture was compatible with different recombinase domains to enable broader sequence targeting, and whether alternative fusions would display greater activity than the first-generation recCas9. I chose Cre recombinase as the dCas9 chimeric fusion partner because it has undergone extensive structural and biochemical characterization[60], facilitating its further development. Additionally, using phage-assisted continuous evolution (PACE), we generated variants of Cre recombinase with a promiscuous phenotype (see Chapter 2). We developed a PACE selection for recombinases with the goal of retargeting Cre toward a sequence in a human safe harbor locus, and carried out a series of PACE selections to promote recognition the ROSA*loxP* target. The resulting ROSACre 1 variant showed activity on a ROSA*loxP* reporter plasmid in mammalian cells, but it also displayed concomitant recognition of *loxP*, which differs from ROSA*loxP* at 44% of base pairs. While these promiscuous variants were unsuitable for further retargeting using PACE, I reasoned that their broadened substrate tolerance represented ideal behavior for a theoretical recCas9 fusion partner. Therefore, I chose to investigate fusions of Cre and ROSACre with dCas9.

I first optimized the architecture of Cre and ROSACre fusions to dCas9 for maximal activity and gRNA dependence. I constructed plasmids for expressing second-generation recCas9 fusions with various linkers between dCas9 and Cre ($(GGS)_1$, $(GGS)_2$, $(GGS)_5$, $(GGS)_8$, or XTEN[150]). Next, I constructed reporter plasmids that contained the *loxP* or ROSA*loxP* target flanked by the previously validated *FGF14* gRNA sequences in the optimal configuration (6-bp spacing) for Ginβ-based recCas9. Finally, I transfected HEK293T cells with a recombinase

expression plasmid, a reporter plasmid, and expression plasmids encoding on- or off-target gRNA sequences (Figure 3.6a). I observed substantial recombination of the *loxP* reporter by wild-type Cre-dCas9 with longer covalent linkers, but limited dependence on the presence of cognate gRNA sequences. I suspect that, in this case, innate target recognition and cooperative binding between Cre monomers[151] enabled Cre-mediated recombination of the *loxP* reporter independent of dCas9 binding. In contrast, ROSACre 1 fusions to dCas9 showed low overall activity on the ROSA*loxP* reporter, but the $(GGS)_2$ and $(GGS)_5$ linker variants demonstrated moderate gRNA dependence. These data suggest that ROSACre-based recCas9 fusions may have favorable properties as broadly useful programmable recombinases, and that the recCas9 fusion architectures tested for Ginβ are compatible with additional recombinases.

Inspection of the endogenous sequence context of the ROSA*loxP* target revealed the presence of gRNA binding sites flanking the recombinase substrate (Figure 3.6b). To test the suitability of targeting the endogenous ROSA*loxP* locus using the existing ROSACre-dCas9 fusions, I constructed a reporter plasmid that contained the endogenous human ROSA*loxP* sequence context (hROSA*loxP*) as well as plasmids for expressing gRNAs complementary to the upstream and downstream Cas9 binding sites. I transfected HEK293T cells with the hROSA*loxP* reporter, plasmids for expressing on- and off-target gRNAs, and a plasmid for expressing ROSACre 1-dCas9 with varying linker lengths (Figure 3.6c). I also transfected HEK293T cells with an equimolar mixture of all five linker variants, as the spacing between the Cas9 binding sites and ROSA*loxP* differs from the optimal length determined for Ginβ-dCas9 (Figure 3.2), and I hypothesized that ROSACre-dCas9 variants with differing linker lengths might demonstrate synergistic behavior on the non-optimal target. The ROSACre-dCas9 fusion with a $(GGS)_8$ linker showed > 10% activity on the hROSA*loxP* reporter with strong gRNA dependence. This finding demonstrates that ROSACre-dCas9 fusions can operate on the ROSA*loxP* target in its endogenous context within human cells.

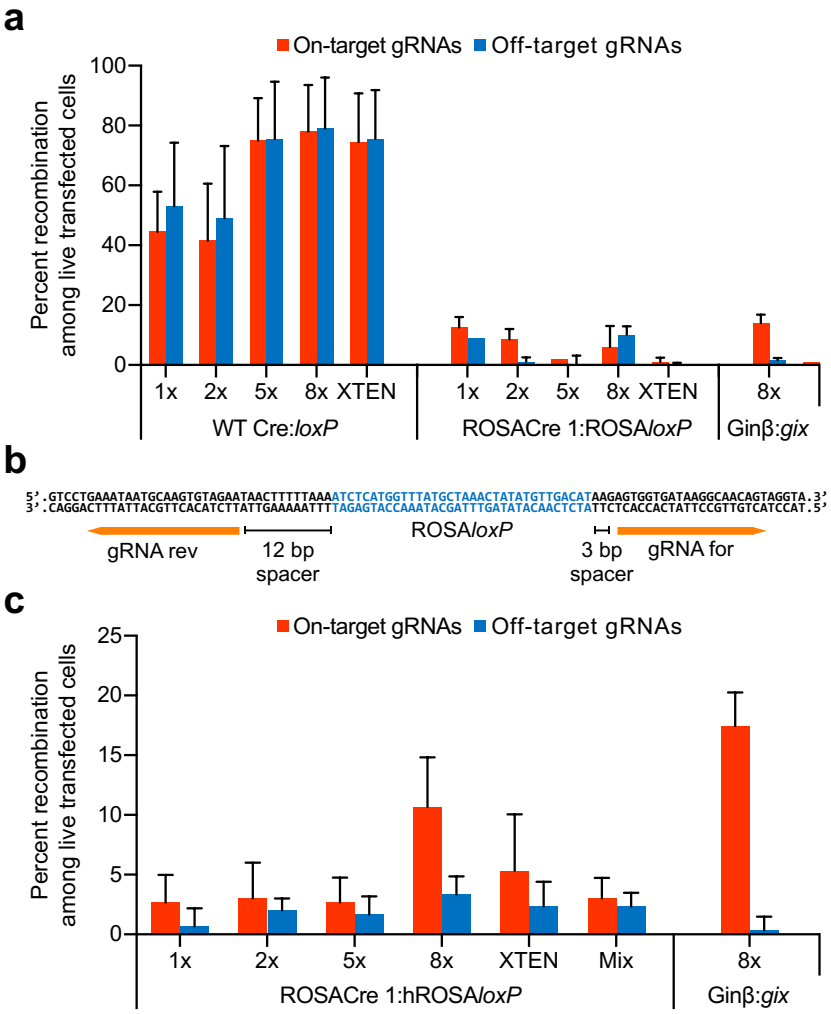


Figure 3.6. Chimeric fusions of dCas9 and promiscuous Cre variants are active on the ROSA*loxP* target. a, Cells were transfected with expression plasmids for recombinase fusions to dCas9 with the indicated covalent linker ($(GGS)_1$, $(GGS)_2$, $(GGS)_5$, $(GGS)_8$, or XTEN), **(Figure 3.6 continued)** expression plasmids for on- or off-target gRNAs, and a reporter plasmid for *loxP* or ROSA*loxP*. **b,** Endogenous human ROSA*loxP* (hROSA*loxP*) reporter target containing native protospacers upstream and downstream of the ROSA*loxP* core site. **c,** Cells were transfected with expression plasmids for ROSACre 1 fusions to dCas9 with the indicated chimeric linker ($(GGS)_1$, $(GGS)_2$, $(GGS)_5$, $(GGS)_8$, XTEN, or an equimolar mixture of all variants), expression plasmids for on- or off-target gRNAs, and a reporter plasmid for the endogenous hROSA*loxP* target from (**b**). Exemplary data for Ginβ-based recCas9 are shown. The percentage of EGFP-positive cells reflects that of transfected (iRFP-positive) cells. Values and error bars represent the mean and standard deviation of two (**a**) or three (**c**) independent biological replicates.

Encouraged by the finding of ROSACre-dCas9 activity on the endogenous hROSA*loxP* target in a mammalian cell assay, I next attempted to integrate foreign DNA directly into the genome of unmodified human cells. I transfected HEK293 cells with a plasmid expressing

ROSACre 1-$(GGS)_8$-dCas9, plasmids expressing the hROSA*loxP* gRNAs, and an integration donor plasmid encoding a single hROSA*loxP* target and a neomycin resistance gene. Recombinase-mediated integration of the plasmid into the genome confers geneticin (G418) resistance to the cell and its daughter cells. After transfection, we grew the HEK293 cells in selective media for two weeks, during which period control cells lacking one of the recCas9 components were susceptible to G418. Following selection, we harvested the genomic DNA of surviving cells and performed a modified version of the GUIDE-seq protocol for unbiased detection of genomic modification[88]. We sheared the genomic DNA, ligated on single-tail adapters, and performed PCR amplification using a primer that binds to the ligated adapter paired with a primer internal to the integration donor plasmid. High-throughput sequencing of the resulting amplicons failed to produce evidence of targeted genomic integration at the ROSA26 locus.

Together, these findings suggest that ROSACre-based recCas9 fusion proteins are promising candidates as programmable recombinase tools, but further improvements to activity and retargetability are needed to efficiently modify the genomes of human cells.

3.3 Discussion

We demonstrated that the optimized fusion of a catalytically inactive Cas9 to the hyperactive catalytic domain of a small serine invertase results in an RNA-programmed recombinase. RecCas9 activity is dependent on the presence of both gRNA sequences complementary to sites that flank a pseudo-*gix* target. Importantly, this fusion can be directed to a variety of endogenous human genomic sequences, resulting in seamless recombination events that rarely contain indels or other mutations at recombinase junctions. Current or future generations of recCas9 could be used to cleanly delete or integrate DNA in studies seeking to develop treatments for genetic diseases.

This work represents the first step toward seamless, RNA-programmed enzymatic recombination of genomic DNA. RecCas9-catalyzed genomic integration has the potential to overcome one of the major limitations imposed by strategies that integrate DNA by homology-directed repair (HDR): in mammalian cells, double-stranded breaks are typically repaired by error-prone processes more frequently than by HDR. Although recombinase-mediated integration is a less favorable process than recombinase-mediated deletion, strategies such as recombinase-mediated cassette exchange (RMCE; Figure 1.1) have been implemented to favor genomic integration[122,152]. Current RMCE strategies require that recombinase substrates be integrated into the target genome prior to integration of exogenous DNA. Our strategy, in principle, overcomes this limitation since the recCas9 system is capable of targeting sequences found endogenously within the human genome.

The findings reported here provide a foundation toward RMCE on native genomic loci, which would require two recCas9 target sites in relative proximity. The estimated 450 human genomic sites identified *in silico* for recCas9 could theoretically be expanded substantially by replacing the Ginβ recombinase catalytic domain with other natural or manmade recombinase domains that recognize different core sequences; many of these related enzymes have also been directed to novel sites via fusion to zinc finger proteins[80,153]. We investigated fusions of promiscuous Cre variants to dCas9, potentially representing an orthogonal enzymatic partner for enabling RMCE. Moreover, recent work altering Cas9 PAM binding specificity and the recent discovery of numerous Cas9 orthologs raise the possibility of further expanding the number of potential recCas9 sites[42-44]. The approach developed here can be expanded upon by other researchers to generate even more tools capable of specific, seamless integration of exogenous DNA into the human genome.

Deletion of the *FAM19A2* intronic sequence in human cells demonstrates that recCas9 is capable of precisely modifying genomic DNA. This is the first demonstration, to our knowledge, of a Cas9-based recombinase tool with direct activity on the human genome. While we carried

out extensive optimization of the chimeric recCas9 to improve its activity, further improvements such as evolution of the chimeric fusion or use of a recombinase domain with a broader sequence tolerance would likely increase the activity and substrate scope of recCas9-mediated genomic modification.

Additionally, further characterization of recCas9 sequence requirements and tolerances may allow a more judicious choice of target sites and ultimately expand the utility of this enzyme. Such characterization may help to explain why recCas9 was inactive on two of the five genomic sequences tested in our plasmid-based assays (Figure 3.4a). The inability of recCas9 to function on these or other sites may be caused by important, but unknown, sequence preferences of Ginβ. Alternatively, poorly active gRNA sequences may also affect recCas9 activity at particular sites. Identifying the Ginβ and gRNA sequence requirements will inform future applications of recCas9.

In principle, programmable recombination-based gene deletion offers advantages over current nuclease-based approaches for generating therapeutic gene knockouts. Unlike mutations induced by programmable nucleases such as ZFNs, TALENs, or Cas9, recCas9 deletion is not dependent on error-prone forms of DNA double-stranded break repair and is theoretically not prone to undesired chromosomal rearrangements or p53 activation[49-51,53]. Indeed, non-programmable recombinase-mediated deletions have already proven effective at removing latent HIV provirus from infected hematopoietic stem cells[71,72], or unwanted vector backbone resulting from *ex vivo* gene therapy[154]. Finally, the requirement of four separate gRNA-programmed binding events as well as a matching dinucleotide core in the recombination substrates may reduce the likelihood of off-target recCas9 modifications, which are commonly observed in nuclease-mediated mutagenesis. Therapeutic applications of recCas9-mediated deletions may be possible once future studies expand the activity and substrate scope of recCas9.

3.4 Methods

General Methods

See Chapter 2 methods section.

Cloning of mammalian recCas9 expression, guide RNA expression, and reporter plasmids

Mammalian expression plasmids for recCas9 were constructed by restriction cloning. Subcloning vectors containing the recCas9 gene were constructed by PCR amplification of a gBlock encoding an evolved, hyperactivated Gin variant (Ginβ)[132], digestion with BamHI and NotI, and ligation into a previously described Cas9 expression vector.[155] PCR was used to generate amplicons containing dCas9-Flag-NLS flanked by BamHI and AgeI and variable-length GGS linkers. The subcloning vectors and Cas9 PCR amplicons were digested with BamHI and AgeI and ligated to create recCas9 (pGinβ-8xGGS-dCas9-FLAG-NLS) and GGS-variants thereof. For plasmid sequencing experiments, the AmpR gene in pGinβ-8xGGS-dCas9-FLAG-NLS was replaced by SpecR using Golden Gate assembly, performed as described previously with Esp3I (ThermoFisher Scientific)[156].

Expression vectors for guide RNAs were generated by blunt-end ligation cloning of 5'-phosphorylated PCR products generated from a previously described plasmid[155]. For plasmid sequencing experiments, the AmpR gene was replaced by SpecR via circular polymerase extension cloning, performed as previously described[157,158].

The pCALNL-GFP subcloning vector, pCALNL-EGFP-Esp3I, was used to clone all recCas9 reporter plasmids and was based on the previously described pCALNL-GFP vector[159]. To create pCALNL-EGFP-Esp3I, pCALNL-GFP vectors were digested with XhoI and MluI and ligated with double-stranded DNA oligonucleotides containing inverted Esp3I sites and compatible overhangs.

pCALNL-EGFP recCas9 reporter plasmids were created by Golden Gate assembly with the pCALNL-EGFP-Esp3I acceptor vector, a PCR product containing neomycin and the poly-A

terminator, and pairs of dsDNA oligonucleotides bearing recCas9 target sites, performed as described previously with Esp3I (ThermoFisher Scientific)[156].

HEK293T transfection, flow cytometry, and plasmid sequencing

HEK293T cells (ATCC CLR-3216) were cultured in Dulbecco's Modified Eagle's Medium plus GlutaMAX-I (Corning) supplemented with 10% fetal bovine serum (FBS; Life Technologies). Cells were seeded into 48-well poly-D-Lysine-coated plates (Corning) in the absence of antibiotic at a density of $3x10^5$ cells per well. 12-15h after plating, cells were transfected with 0.8 µL Lipofectamine 2000 (ThermoFisher Scientific) using 160 ng of recCas9 expression vector, 45 ng of each guide RNA expression vector, 9 ng of reporter plasmid, and 9 ng of fluorescent protein expression plasmid as a transfection control. Cells were cultured for 72 h before they were washed with PBS (ThermoFisher Scientific) and detached from plates by the addition of 0.05% trypsin-EDTA (Life Technologies). Cells were diluted in 250 µL culture media and run on a BD Fortessa analyzer.

For plasmid sequencing experiments, cells were transfected and harvested as described, and episomal DNA was extracted using a modified HIRT extraction involving alkaline lysis and spin column purification as previously described[160,161]. Briefly, after harvesting, HEK293T cells were washed in 500 µL of ice cold PBS, resuspended in 250 µL GTE Buffer (50 mM glucose, 25 mM Tris-HCl, 10 mM EDTA, pH 8.0) and lysed on ice for 5 minutes in lysis buffer (200 mM NaOH, 1% sodium dodecyl sulfate). Lysis was neutralized with neutralization buffer (5 M acetate, 3 M potassium, pH 6.7). Cell debris was pelleted and lysate was applied to EconoSpin columns (Epoch Life Science), washed with ethanol wash buffer, and eluted in TE buffer. Isolated episomal DNA was digested for 2 hours at 37 °C with exonuclease V (10 units) and purified with a Minelute columns (Qiagen) in elution buffer (EB). The DNA was transformed

into One Shot Mach1 T1 E. coli and plated on agar plates containing carbenicillin (50 µg/mL). Individual colonies were Sanger sequenced to determine the rate of recombination.

Analysis of recCas9 catalyzed genomic deletions

HEK293T cells were seeded into 24-well poly-D-Lysine-coated plates (Corning) in the absence of antibiotic at a density of $6x10^5$ cells per well. 12-15h after plating, cells were transfected with 2 µL Lipofectamine 2000 (ThermoFisher Scientific) using 320 ng of recCas9 expression vector, 90 ng of each guide RNA expression vector, and 20 ng of *GFP* expression plasmid as a transfection control. Cells were cultured for 48 h before they were harvested as described above. Cells were diluted in 250 µL culture media and the live, transfected (GFP-positive) cell population was collected using a BD FACSAria III cell sorter. Cells were sorted on purity mode using a 100 µm nozzle and background fluorescence was determined by comparison with untransfected cells. Sorted cells were collected on ice in PBS, pelleted and washed twice with ice cold PBS. Genomic DNA was harvested using the E.Z.N.A. Tissue DNA Kit (Omega Bio-Tek) and eluted in 100 µL EB. Genomic DNA was quantified using the Quant-iT PicoGreen dsDNA kit (ThermoFisher Scientific) measured on a Tecan Infinite M1000 Pro fluorescence plate reader.

Genomic PCR was carried out using Q5 Hot Start Polymerase 2x Master Mix supplemented with 3% DMSO and diluted with nuclease-free water (GE Life Sciences). DNA was analyzed by electrophoresis on a 1% agarose gel in TAE alongside a 1 Kb Plus DNA ladder (ThermoFisher Scientific). Material to be Sanger sequenced was purified on a Qiagen Minelute column according to the manufacturer's instructions. Template DNA from 3 biological replicates was used for three independent genomic nested PCR experiments.

The limit of detection was calculated given that one complete set of human chromosomes weighs approximately 3.6 pg ($3.3 \cdot 10^9 \text{ bp} \times 1 \cdot 10^{-21} \frac{\text{g}}{\text{bp}}$). Therefore, a PCR

reaction seeded with 20 ng of genomic DNA template contains approximately 5500 sets of chromosomes.

For quantification of genomic deletion, nested PCR was carried out using the above conditions in triplicate for each of the 3 biological replicates. A two-fold dilution series of genomic DNA was used as template, beginning with the undiluted sample. The lowest DNA concentration for which a deletion PCR product could be observed was assumed to contain a single deletion product per total genomic DNA.

Identification of genomic target sites

Potential endogenous recCas9 targets within the human genome were identified using custom software written in R and made available online at https://github.com/JohnHHu/recCas9. The GRCh38 human reference genome was scanned for sequences on both DNA strands that match the recCas9 motif $CCN_{(30\text{-}31)}$-AAASSWWSSTTT-$N_{(30\text{-}31)}$-GG. Potential endogenous targets are listed in Appendix A.

Chapter 4:

High-Resolution Specificity Profiling and Off-Target Prediction for Site-Specific DNA Recombinases

Adapted from Jeffrey L. Bessen, Lena K. Afeyan, Vlado Dančík, Luke W. Koblan, David B. Thompson, Chas Leichner, Paul A. Clemons, and David R. Liu. High-resolution specificity profiling and off-target prediction for site-specific DNA recombinases. *Nature Communications,* DOI: 10.1038/s41467-019-09987-0 (2019).

David Thompson and I designed and performed the initial Rec-seq experiments. Lena Afeyan designed and performed experiments described in section 4.2.1 and figure 4.2. Chas Leichner and I wrote the software initially used to analyze Rec-seq data. Vlado Dančík performed the computational and statistical analysis for figures 4.1-4.8. Luke Koblan contributed to the design and execution of experiments described in figure 4.9. Andrew Bohm and Gretchen Meinke provided materials utilized in section 4.2.3 and figure 4.7 I designed and performed all remaining experiments.

4.1 Introduction

Site-specific recombinases (SSRs) have the potential to serve as ideal genome editing agents because they directly catalyze the cleavage, strand exchange, and rejoining of DNA fragments at defined recombination targets[56] without relying on the endogenous repair of double-strand breaks which can induce indels, translocations, other DNA rearrangements, or p53 activation[49-51,53]. The reactions catalyzed by SSRs can result in the direct replacement, insertion, or deletion of target DNA fragments with efficiencies exceeding those of homology-directed repair[56,64]. SSRs are active in a variety of cell states including non-dividing cells[56], and many efficiently operate on mammalian genomes[60,65].

Although SSRs offer many advantages, their native substrate preferences are not easily altered, even with extensive laboratory engineering or evolution[74]. The development of SSRs into more versatile genome editing agents is limited in part by an incomplete understanding of SSR protein:DNA specificity determinants[60,74,82]. Crystal structures of tyrosine-family SSRs demonstrate that Cre and other recombinases interact with DNA through relatively few direct protein:DNA contacts, and that shape- and charge-complementarity and water-mediated interactions contribute to SSR specificity[60,83]. Further, static co-crystal structures do not comprehensively identify key interactions between SSR residues and substrate nucleotides. For example, replacement of Glu262 increases Cre's tolerance for mismatches in regions of *loxP* with no direct protein:DNA contacts[84]. These and other observations establish that the relationship between SSR residues and DNA specificity is not straightforward; some residues impact specificity more than others, and some contribute to specificity at distant DNA positions.

Efforts to engineer or evolve programmable recombinases from existing SSRs would greatly benefit from an enhanced understanding of their DNA specificity. Motivated by this need, we sought to develop a method to rapidly map the determinants of SSR specificity. Such a method could also be used to predict cellular off-target activity of SSRs, an important consideration when evaluating SSRs as potential research tools or therapeutics. Here we

describe Rec-seq, a method for profiling the DNA specificity of SSRs in a rapid and unbiased manner using *in vitro* selection and high-throughput DNA sequencing (HTS). We applied Rec-seq to characterize wild-type Cre and Cre mutants, resulting in the identification of novel DNA specificity determinants, including long-range interactions not evident from structural studies. We also profiled the sequence preferences of the laboratory-evolved Cre variants Tre and Brec1, as well as three orthogonal SSRs, including the directional integrase Bxb1. The application of Rec-seq to Tre and Brec1 recombinases resulted in specificity profiles that accurately predicted activity at off-target sites, including pseudo-sites within the human genome. Our findings suggest that Rec-seq can inform the application of SSRs as well as their further development.

4.2 Results

4.2.1 Development of an in vitro *selection for recombinase substrates*

We sought to develop a system for profiling recombinase specificity through identification of *bona fide* recombinase substrates from a vast *in vitro* library of possible targets. To do so, we designed substrate oligonucleotides such that recombination yields a degradation-resistant DNA product, permitting the selective digestion of non-substrates. We chose Cre as a model recombinase for developing Rec-seq because Cre has been structurally characterized[60], the effects of some Cre mutations on DNA specificity are known[83,84,104,162-166], and researchers have generated Cre variants with altered specificity[74]. Cre's substrate *loxP* consists of two 13-bp half-sites that together form inverted repeats, flanking an asymmetric 8-bp core region where strand exchange occurs (Figure 4.1a).

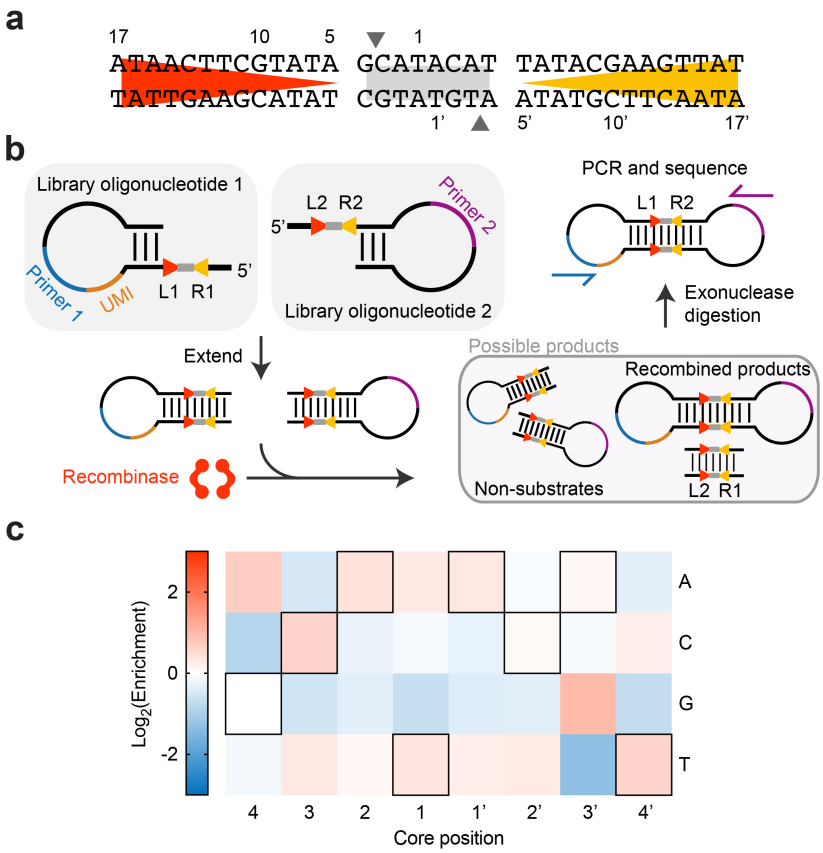


Figure 4.1. Overview of Rec-seq. a, The cognate DNA substrate of Cre, *loxP*. DNA backbone cleavage occurs at the indicated phosphodiester bonds (gray arrows). **b**, In Rec-seq, DNA hairpin oligonucleotides containing partially randomized *loxP* sites and a unique molecular identifier (UMI) are subjected to intramolecular primer extension, exposed to recombinase, and digested with exonucleases to destroy non-recombined DNA. **c,** Heat map of Rec-seq enrichment values for wild-type Cre showing the $\log_2$ of the enrichment value for each nucleotide at each position in the *loxP* core, relative to the canonical base for the forward orientation (black outline). Wild-type Cre was exposed to *loxP* library oligonucleotides in which the half-sites were held constant and the core nucleotides were unbiasedly randomized. Values represent the geometric mean of n=3 independent replicates conducted at 37 °C for 30 minutes at a 1:3 protein:DNA ratio.

To prepare *in vitro* substrate libraries, we extended synthetic DNA containing self-priming 5' overhangs and a partially randomized *loxP* site (Figure 4.1b). The hairpin serves to prime extension across the randomized region of *loxP*, replicating the library member and yielding a double-stranded DNA substrate required by SSRs. We generated two related substrates: left-hairpin substrates (containing left and right half-sites L1 and R1) and right-hairpin substrates (containing half-sites L2 and R2; Figure 4.1b). When Cre protein is exposed to one left-hairpin and one right-hairpin oligonucleotide, successful recombination generates a

double-stranded DNA product with hairpins on both sides. Exonuclease treatment destroys non-recombined library members, and the exonuclease-resistant double-hairpin recombination products are amplified by PCR. High-throughput DNA sequencing of libraries (at a typical depth of 10^5-10^6 reads per experiment) enables quantitation of the frequency of each base at each half-site position before and after selection. Enrichment scores are then determined for each target position (see Chapter 4 Methods), such that higher enrichment scores reflect a stronger preference for a particular base at that half-site position.

In designing the Rec-seq library we considered the optimal degree of *loxP* randomization and the ideal placement of these randomized positions within the Rec-seq oligonucleotides. Since Cre is thought to be highly specific for *loxP*, we hypothesized that a modest number of mutations per half-site would support recombination while allowing the interrogation of many substrate combinations. Randomized positions in *loxP* were varied during DNA synthesis to contain 79% wild-type base and 21% of an equimolar mixture of all three other bases, yielding a library in which each variable half-site contained 2.7 mutations on average. We routinely generated libraries exceeding 10^{11} sequences, sufficient to cover all possible half-sites with up to seven substitutions from the *loxP* sequence. We found no significant differences of enrichment values when performing Rec-seq experiments with a more highly mutagenized *loxP* library (Figure 4.2a). Additionally, the core sequence of *loxP* was held constant because the core regions of two recombining *loxP* substrates must be complementary[59]. Most Cre:*loxP* interactions are thought to involve the half-sites[60,84], and we observed minimal preference among the core nucleotides in experiments in which the half-sites were held constant and the core was mutagenized (Figure 4.1c). Finally, Rec-seq only captures mutations present in L1 and R2, because the product of recombination containing R1 and L2 is degraded (Figure 4.1b). In order to isolate interactions between Cre and a single *loxP* half-site, only L1 or R2 was randomized while R1 and L2 were fixed as the wild-type *loxP* sequence. Enrichment profiles for a full *loxP* target were generated by collecting the enrichment factors from L1 and R2 half-sites.

Next we optimized and validated Rec-seq experimental conditions using wild-type Cre. The Cre specificity profile did not substantially change upon incubation times longer than 30 minutes (Figure 4.2b). A protein:DNA ratio of 1:3 was previously shown to be optimal for recombination[167], and we found that protein:DNA ratios higher than ~1:1 eroded apparent specificity, consistent with excess enzyme enabling the recombination of even non-preferred substrates (Figure 4.2c). Finally, we showed that the Rec-seq enrichment pattern of Cre protein exposed to *loxP* substrate was not dependent on the source of Cre protein (Figure 4.2d). For subsequent experiments, we chose to perform Rec-seq by incubating the *loxP* variant library with Cre *in vitro* at a molar ratio of 1:3 protein:DNA for 30 minutes at 37 °C.

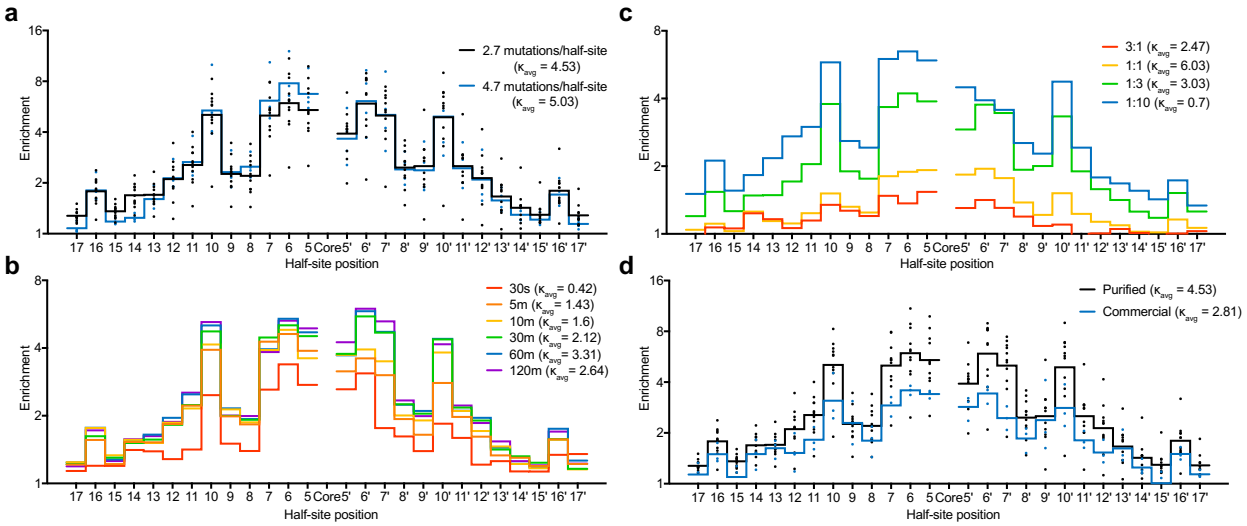


Figure 4.2. Rec-seq parameter optimization. a, Rec-seq profile for wild-type Cre on *loxP* using different levels of *loxP* library randomization. Values represent the geometric mean of n=11 (2.7 mutations/half-site) or n=4 (4.7 mutations/half-site) independent replicates (dots) conducted at 37 °C for 30 minutes at a 1:3 protein:DNA ratio. The differences between Cre enrichment on the two libraries were not significant ($p > 0.05$). **b, c,** Impact of reaction time (**b**) and protein:DNA ratio (**c**) on Rec-seq specificity profile for wild-type Cre reacted with *loxP* substrate. For part (**b**), all reactions were carried out at a 1:3 protein:DNA ratio. For part (**c**), all reactions were carried out for 30 minutes at 37 °C. Values represent the geometric mean of three independent replicates. **d,** Rec-seq profile for purified and commercially available wild-type Cre enzyme (New England Biolabs) on randomized *loxP* substrates. Values represent the geometric mean of n=11 (purified) or n=3 (commercial) independent replicates (dots) conducted at 37 °C for 30 minutes at a 1:3 protein:DNA ratio. The differences between commercial and purified Cre were not significant for any nucleotide position or along the full *loxP* site ($p >> 0.05$).

Before analyzing the resulting enrichment profile, we calculated a quality score for each experiment. Poorly active recombinases or very short exposure to enzyme could result in levels of *bona fide* substrates surviving selection that do not greatly exceed background levels of undigested library material (Figure 4.3a). To identify such instances of poor signal:background ratios, we calculated a quality score, κ, for each experiment. Background amplification for each experiment was measured using quantitative PCR to confirm that SSR-treated samples contained more DNA after selection than a control sample lacking recombinase. To distinguish low activity from poor specificity, we included a unique molecular identifier (UMI) barcode on the left-hairpin library member (Figure 4.1a). The κ value for each experiment was determined by plotting the percent abundance of each DNA sequence variant in the post-recombination library versus the number of UMIs for each sequence variant, with κ being the slope of the best-fit line, divided by 10^4 for ease of comparison (Figure 4.3b). The average κ value among experimental replicates for a given SSR, κ_{avg}, reflects whether its Rec-seq enrichment values are derived from a large number of independent recombination events (a larger κ_{avg} value) or may be subject to undersampling due to low activity (a smaller κ_{avg} value; Appendix B). By comparing Rec-seq outcomes between experimental replicates, we considered experiments to be well-powered if κ_{avg} values exceeded 1.5, modestly influenced by background signal for κ_{avg} values between 1.5 and 0.5, and heavily influenced by background signal for κ_{avg} values below 0.5 (Figure 4.3c).

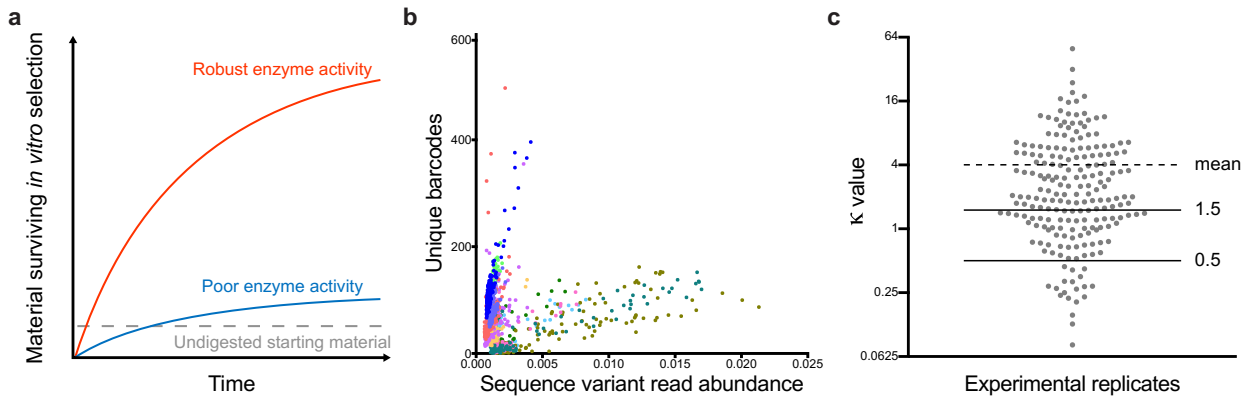


Figure 4.3. Quality score calculation. a, Model for the effect of *in vitro* enzyme activity on apparent SSR specificity. For each experiment, a background level of undigested starting library is present (gray dashed line). This background undigested material is not distinguished from genuine recombination products that survive the *in vitro* selection. Robust enzyme activity produces an excess of genuine recombined products (red line), but poorly-active enzymes (blue line) or shortened reaction times produce lower levels of recombined products that can be similar to the level of background undigested starting material. **b**, To quantify the extent to which apparent specificity of an SSR is affected by its *in vitro* activity, we plotted the fractional abundance of each DNA sequence variant versus the number of unique barcodes for that variant. For DNA sequences with an absolute abundance of 800 or fewer (well below 4,096, the maximum number of unique barcodes), we assumed that each unique barcoded sample represented an independent recombination event. We expect that signal derived from few recombination events or amplification of undigested starting material would have relatively few unique barcodes for a given DNA sequence variant. We plotted the fractional abundance, as opposed to the absolute abundance, of each DNA sequence variant to correct for the effect of sequencing depth. The quality score κ is the slope of the best-fit line for the plot described above, divided by 10^4 for ease of comparison between experiments. The value κ_{avg} was calculated for each SSR variant by averaging the κ values for each experimental replicate. Exemplary data from 11 replicates of wild-type Cre reacted with *loxP* substrate at a 1:3 protein:DNA ratio for 30 minutes at 37 °C (colored dots) are shown. κ_{avg} values for each SSR variant can be found in Appendix B. **c,** Scatter plot showing the distribution of κ values for all Rec-seq experimental replicates on a $\log_2$ axis. We considered experiments to be well-powered if κ_{avg} values exceeded 1.5, moderately influenced by background signal for κ_{avg} values between 1.5 and 0.5, and heavily influenced by background signal for κ_{avg} values below 0.5.

Analysis of the Rec-seq enrichment profile for Cre indicated a preference for the canonical base at every half-site position (Figure 4.4a,b), a surprising finding given the limited direct protein:DNA contacts between Cre and several regions of *loxP*[60]. On average, 22% of post-selection sequences were identical to *loxP*, compared to 6.4% *loxP* abundance pre-selection. Cre's DNA specificity was weakest at the five most distal bases of *loxP* (Figure 4.4a), consistent with previous reports that Cre tolerates mismatches in the distal region of each half-

site[168,169]. In addition, Rec-seq revealed the sequence preference of Cre to be asymmetric, as is evident when the left and right half-site enrichment profiles are superimposed (Figure 4.4c). To ensure that an asymmetric sequence preference is a property of the enzyme and not due to the different DNA sequences flanking the library oligonucleotides (Figure 4.1b), we performed Rec-seq using a substrate library identical to the original except that the non-palindromic *loxP* core was replaced with its reverse complement. The Rec-seq enrichment profile of this "inverted core" *loxP* library mirrored, rather than duplicated, the profile on the original substrate library (Figure 4.4d), indicating that the oligonucleotide sequence context was not responsible for the asymmetry of the Cre specificity profile. These findings establish the utility of Rec-seq for illuminating DNA-recognition properties of Cre that are difficult or impossible to infer solely by structural characterization.

Rec-seq also confirmed previous findings[60] that Cre has a pronounced preference in two regions of *loxP*: half-site positions 5-7 and 10. We observed 5.0-fold enrichment of the canonical base at position 10, consistent with reports that Arg259 participates in hydrogen bonding with the canonical C•G base pair at position 10[108,162] (Figure 4.4a). Rec-seq also identified a 3.9- to 5.4-fold enrichment for the canonical base pair at position 5 in each half-site, consistent with direct interactions between Gln90 and the A•T base pair[108,162]. A final notable interaction at the Cre:DNA interface is between Lys244 and the T•A base pair at positions 16-17, the only major direct contact between Cre and the five most distal bases of *loxP*[108]. Indeed, among positions 13-17, Rec-seq revealed the strongest preference to be at position 16 (Figure 4.4a). Together, these results validate that Rec-seq can identify DNA sequence preferences consistent with known Cre:*loxP* interactions and provide novel context to these preferences, such as the relative specificity of Cre for nucleotides in *loxP*.

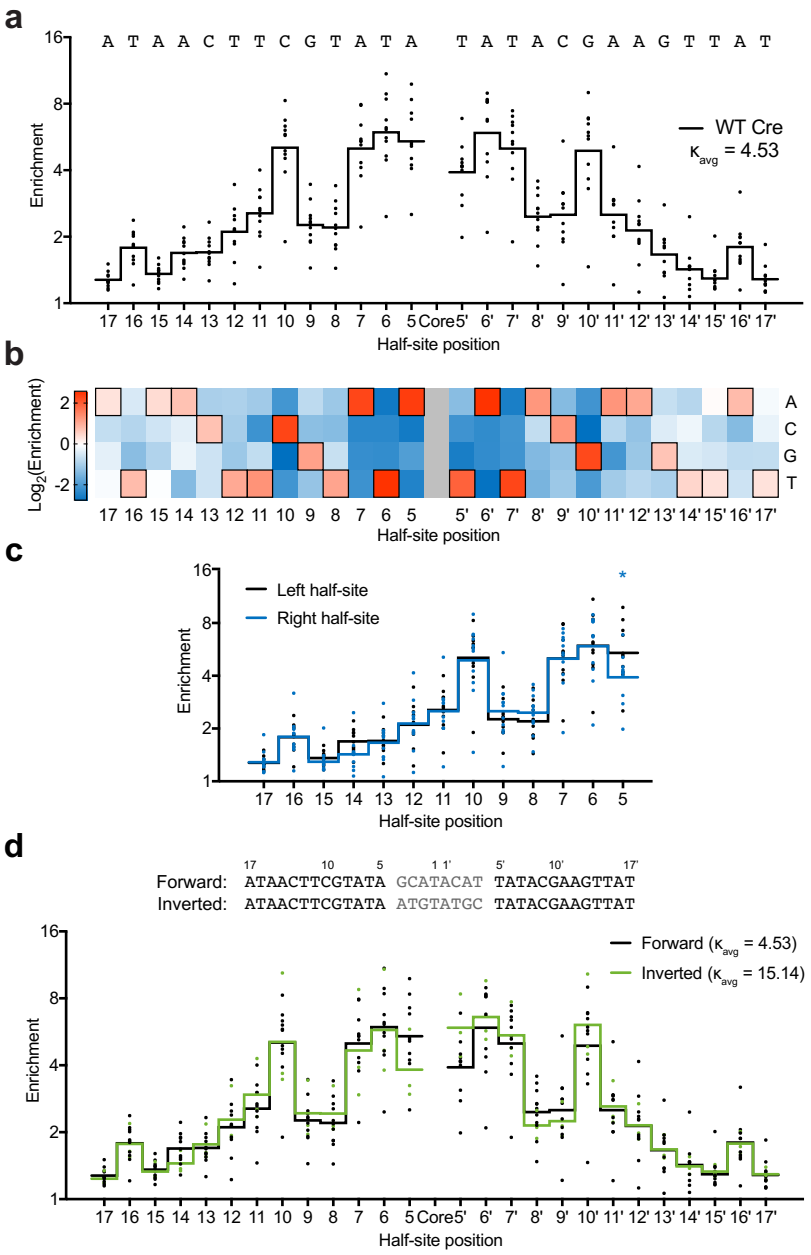


Figure 4.4. Recombinase specificity profiling of wild-type Cre. a, The specificity profile for Cre shows its relative preference for the canonical base at each position in the *loxP* site. The quality score κ_{avg} represents the number of unique recombination events captured by Rec-seq across each experimental replicate, with a value over 1.5 considered a well-powered experiment. **b,** Heat map of Rec-seq enrichment values for wild-type Cre showing the $\log_2$ of the enrichment value for each nucleotide at each position in *loxP* relative to the canonical base (black outline). **c**, Superimposition of the left and right half-site enrichment profiles for purified wild-type Cre on *loxP* library oligonucleotides. Significant differences ($p \leq 0.05$; asterisks) between the log-enrichment values of the left and right half-sites were calculated using a paired t-test. **d**, Rec-seq of wild-type Cre on *loxP* library oligonucleotides with the core sequence in the forward or reverse direction. Values represent the geometric mean of n=11 or n=3 (inverted core) independent replicates (dots) conducted at 37 °C for 30 minutes at a 1:3 protein:DNA ratio.

*4.2.2 Mutational dissection of Cre:*loxP *specificity determinants*

The complexity of Cre:*loxP* interactions has challenged Cre engineering efforts[60,74,82]. To characterize these interactions, we constructed 14 Cre mutants with Ala substitutions at residues known to make contacts with *loxP* (Figure 4.5a), purified each variant, and performed Rec-seq to map the functional relationship between specific residues and the DNA sequence preferences of Cre. Comparison of the Rec-seq profile of Cre mutants and wild-type Cre yielded novel insights into each residue's contribution to DNA specificity across the entire *loxP* site.

Structural and mutagenesis studies[108,162,165] suggested that mutation of Arg259 would affect specificity at half-site position 10. Indeed, the Arg259→Ala variant showed a drop in enrichment at position 10 (from 5.0-fold for wild-type Cre to 1.1-fold for the mutant), with a modest preference for C or T in the left half-site and G or A in the right half-site (Figure 4.5b,c). The Arg259→Ala mutant also showed increased preference at virtually every other position in the *loxP* site, with especially high preferences at positions 5-7 and 16. This observation is consistent with an energetic tradeoff—as we proposed for zinc fingers, TALEs, and Cas9[85,86,170]—in which the loss of binding energy from Ala substitution at Arg259[162] necessitates greater fidelity at other protein:DNA contacts to retain sufficient binding to support recombination, even when these interactions take place far (in this case, >24 Å) from the altered residue. These long-range cannot be inferred from the Cre:*loxP* structure, highlighting the utility of unbiased, high-resolution specificity profiling.

Rec-seq also helped illuminate determinants of specificity at *loxP* positions 5-7, which are less well-understood than the determinants at position 10. Candidate interacting residues are distributed through three regions of Cre: helix B, helix D, and the loop between helices J and K (Figure 4.5a). Rec-seq profiles of Ala mutants at potential interacting residues demonstrate differing impacts of neighboring residues. For example, in helix B, Rec-seq of the Lys43→Ala mutant resulted in a modest drop in specificity relative to wild-type Cre, while Met44→Ala resulted in higher preference at positions 5 and 10 (Figure 4.5d). In helix D, the Lys86→Ala

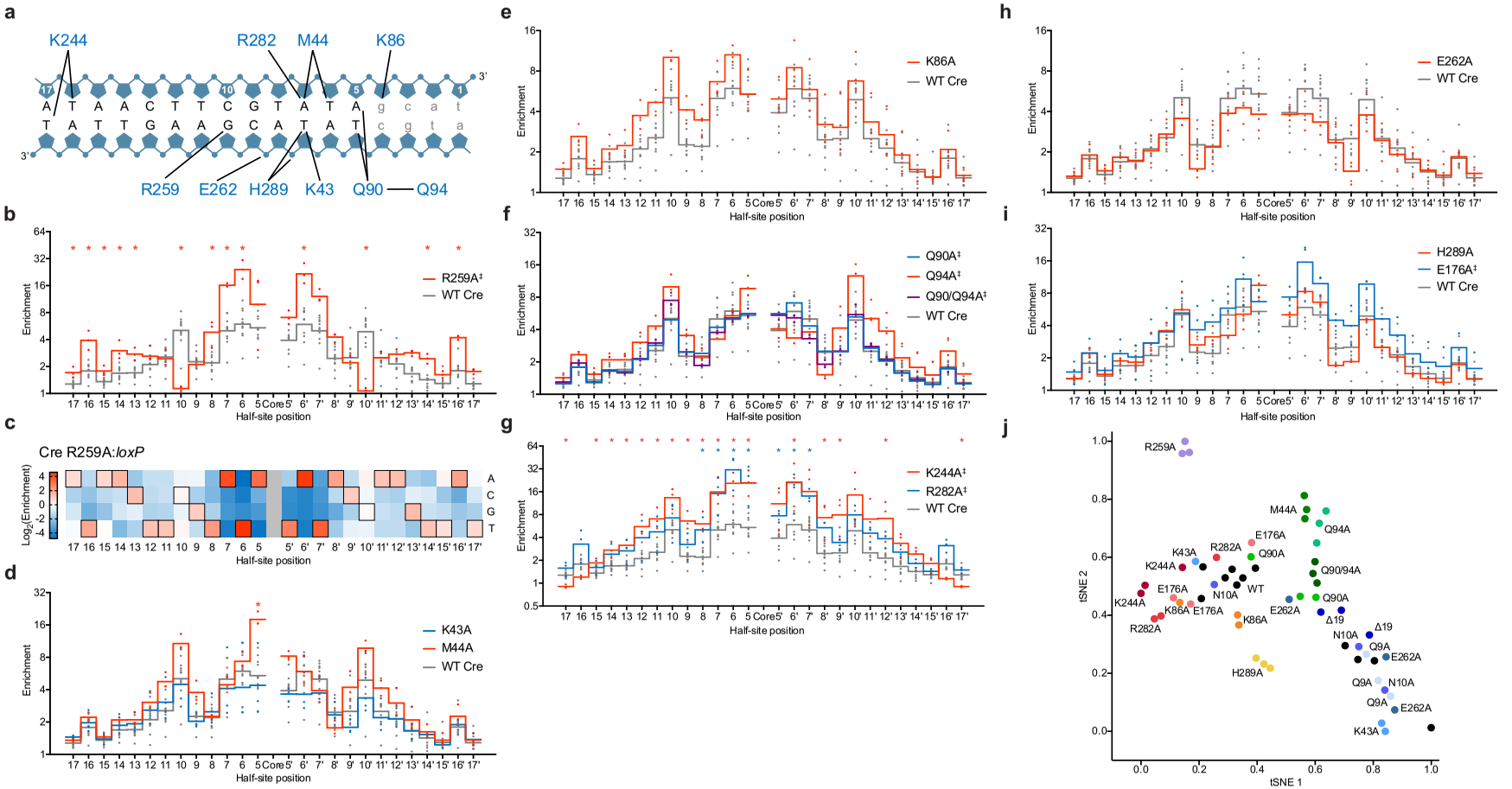


Figure 4.5. Determinants of Cre:*loxP* specificity identified by Rec-seq on wild-type Cre and Ala-substituted Cre variants. **a**, Protein:DNA contacts for Cre:*loxP* inferred from crystal structures[60]. **b**, Rec-seq enrichment profile for the Cre R259A variant (red line) and wild-type Cre (gray line). **c**, Heat map of Rec-seq enrichment values for the Cre R259A variant showing the $\log_2$ of the enrichment value for each nucleotide at each position in *loxP* relative to the canonical base (black outline). **d-i**, Rec-seq enrichment profiles for Ala-substituted Cre mutants (colored lines) and wild-type Cre (gray lines). For **b-i**, values represent the geometric mean of n=3 or n=11 (wild-type Cre) independent replicates (dots) conducted at 37 °C for 30 minutes at a 1:3 protein:DNA ratio. Significant differences ($p \leq 0.05$) relative to wild-type Cre at individual nucleotides (colored asterisks) and across the full log-enrichment profile (‡) are indicated. **j**, The results of Rec-seq experiments on wild-type Cre (black dots) and variants thereof (colored dots) visualized using t-SNE multi-dimensional proximity analysis[173], showing that experimental replicates are clustered by similarity across all specificity features.

variant showed minimal differences from wild-type Cre (Figure 4.5e), while the Gln90→Ala variant showed overall lower enrichment (Figure 4.5f). In the loop between helices J and K, the Arg282→Ala mutant showed higher, rather than lower, DNA specificity across *loxP* (Figure 4.5g). These results demonstrate that Cre's apparent preference at positions 5-7 results from multiple weak or indirect interactions, rather than being strongly determined by residues proximal to these positions.

In addition, Rec-seq identified a contribution from a secondary residue previously unknown to participate in specifying positions 5-7. Ala substitution at Gln94 resulted in lower specificity at positions 6 and 7 but compensatory increases elsewhere (Figure 4.5f), even though Gln94 does not directly contact the DNA, but instead engages in hydrogen bonds with Gln90[171]. Double Ala substitution at both Gln90 and Gln94 performed similarly to the Gln90→Ala single mutant (Figure 4.5f), suggesting that the DNA-contacting residue Gln90 plays the dominant role in defining DNA specificity among the two residues. Together, Rec-seq profiling clarifies the many interactions that together define Cre recognition at positions 5-7, and highlights the important roles of secondary and indirect interactions.

We also applied Rec-seq to examine the role of Glu262, which forms backbone and nucleobase contacts at half-site position 9[108]. Gly or Ala substitutions at Glu262 were previously shown to increase tolerance for mismatches at non-contacted *loxP* positions (*e.g.*, bases 11-12)[84]. The Rec-seq profile of the Glu262→Ala variant showed a drop in specificity at the proximal positions 8-9 (Figure 4.5h), but also decreased specificity at positions 5-7 and 10, consistent with previous findings of Glu262's role in enforcing substrate fidelity[84].

Rec-seq revealed new roles for residues that were not previously known to play a long-range specificity-determining role, such as Lys244 and Glu176. Rec-seq of Lys244→Ala showed a decrease in specificity at the proximal position 17, but otherwise broadly increased specificity for *loxP* (Figure 4.5g). Glu176 is a highly conserved residue among tyrosine recombinases that is proximal to the Cre active site, not the DNA substrate[104], but Rec-seq of

Glu176→Ala showed broadly increased specificity (Figure 4.5i). Another conserved residue, His289, showed a modest decrease in specificity relative to Cre when replaced by Ala (Figure 4.5i). In addition, Rec-seq illuminates contradictory observations about the role of the Cre N-terminus in DNA specificity. While the N-terminus is unresolved in crystal structures and can be truncated with no apparent effect[172], laboratory evolution of Cre yielded mutations at Gln9 and Asn10 that are essential for evolved activity[171]. Rec-seq profiles of Δ19 Cre (lacking the first 19 amino acids), Gln9→Ala, and Asn10→Ala each showed no significant differences compared to wild-type Cre (Figure 4.6). These results suggest that while individual residues in the N-terminus may participate in catalysis, they are unlikely to contribute substantially to *loxP* recognition. Collectively, these findings highlight the ability of Rec-seq to reveal specificity determinants regardless of the proximity between the contributing residue and the DNA base being influenced.

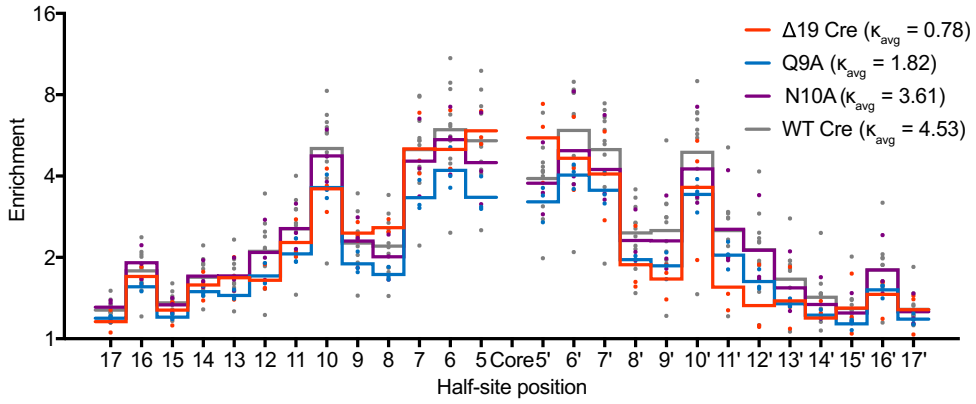


Figure 4.6. Impact of N-terminal mutations on Cre:*loxP* DNA specificity. Rec-seq profiles for the N-terminal truncation (colored lines) relative to wild-type Cre (gray line). Values represent the geometric mean of n=11 (wild-type Cre) or n=3 independent replicates (dots) conducted at 37 °C for 30 minutes at a 1:3 protein:DNA ratio. The differences between N-terminal variants and wild-type Cre were not significant for any nucleotide position or along the full *loxP* site ($p >> 0.05$).

Our understanding of SSR:DNA interactions largely arises from static crystal structures. While structures provide a focused list of possible interactions based on proximity, Rec-seq generates a functional map of residues that contribute to specificity. To visually represent one such map, we used the t-SNE algorithm[173] to correlate the results of individual Rec-seq

experiments using multi-dimensional similarity analysis (Figure 4.5j). The proximity of experiments in the t-SNE visualization relates their similarity across the full Rec-seq profile. For example, the cluster containing Met44 and Gln94 represents the functionally similar residues contributing to specificity at positions 5-7, while other residues proximal to the same bases (Lys43, Lys86, Arg282) appear separately, consistent with their differing roles. Replicates of Rec-seq experiments with wild-type Cre cluster together toward the middle of the graph; Ala-substituted mutants that increase sequence preference appear to the left of the wild-type grouping, while preference-diminishing variants cluster to the right. By revealing and correlating the individual roles of residues in determining DNA recognition across the entire substrate site at single-nucleotide resolution, Rec-seq greatly enhances our understanding of SSR:DNA interactions.

4.2.3 Rec-seq of evolved Cre variants

After confirming that Rec-seq accurately reports known specificity preferences and helps characterize SSR:DNA interactions, we sought to interrogate the basis of specificity for laboratory-evolved Cre variants. The substrate preferences of evolved Cre variants have never been characterized comprehensively, and we reasoned that profiling of these variants at single-nucleotide resolution would illuminate novel specificity determinants and inform their continued development.

We first applied Rec-seq to Tre, which was evolved to recognize *loxLTR*, a sequence that differs from *loxP* at 50% of base pairs[75] (Figure 4.7a). Rec-seq revealed that Tre showed relaxed specificity relative to Cre at multiple positions in *loxLTR*, including positions 9, 10, 12, and 17 in the left half-site and position 14 in the right half-site (Figure 4.7b,c). Tre showed concomitant increased substrate nucleotide preference at positions 5-7, providing further support for the energetic tradeoff model described above. Some of this heightened specificity in Tre occurred at base pairs that were unchanged between *loxP* and *loxLTR* (*i.e.*, 5 of 6 base

pairs among positions 5-7 in both half sites). In addition, Tre maintained enhanced sequence preference at left half-site position 5 and right half-site position 10, which both differ between *loxLTR* and *loxP*. This finding is consistent with the Tre:*loxLTR* co-crystal structure[171], which predicts hydrogen bonding interactions between Gln90 and Arg94 side chains in Tre and the T•A base pair at position 5 (Figure 4.7e). Preferences at these altered positions are consistent with evolved recognition for the *loxLTR* substrate, and are likely necessary to offset the loss of DNA interactions at other positions.

We also applied Rec-seq to Brec1, a Cre variant evolved to recognize the *loxBTR* target, which differs from *loxP* at 68% of base pairs[72] (Figure 4.7a). Similar to Tre, the Rec-seq profile of Brec1 showed evidence of tradeoffs between loss of protein:DNA interactions at some positions within the half-site and enhanced specificity for critical base pairs elsewhere. Brec1 showed diminished preference at position 8 in both half-sites and positions 10 and 12 in the left half-site, and conserved specificity for positions 5 and 6 in both half-sites of *loxBTR* (Figure 4.7b,d). Additionally, Brec1 maintained enhanced specificity for right half-site position 10, which differs between *loxP* and *loxBTR*, suggesting the presence of evolved interactions between Brec1 and this base pair. These regions of high specificity likely represent a mixture of conserved and novel Brec1:*loxBTR* interactions (Figure 4.7f), the presence of which may be required to offset the loss of binding interactions in other regions of the target site.

For both evolved variants, Rec-seq revealed that target recognition arose from a combination of conserved interactions, evolved recognition at important half-site positions, and relaxed specificity. Our results support the findings from structural characterization of Tre:*loxLTR*, and also suggest the presence of novel interactions between Brec1 and *loxBTR*, which have not yet been co-crystallized.

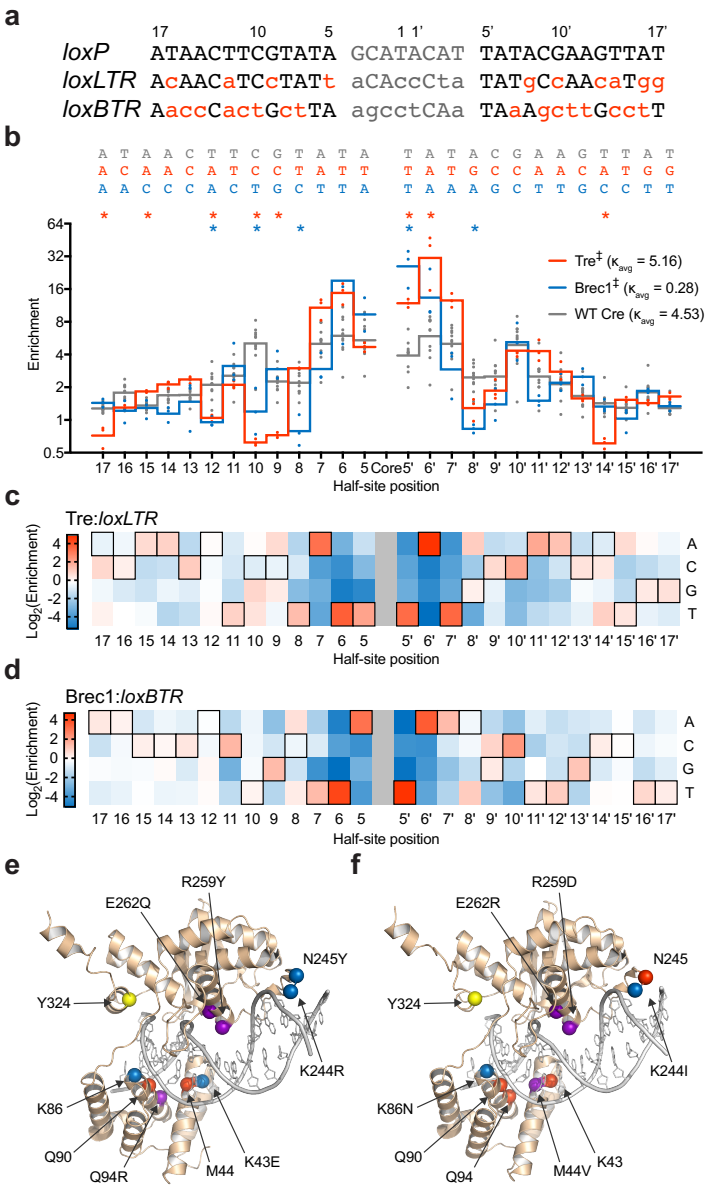


Figure 4.7. DNA specificity of evolved Cre variants revealed by Rec-seq. a, DNA sequences of *loxP*, *loxLTR*, and *loxBTR* showing differences relative to *loxP* (red). **b**, Rec-seq specificity profiles for Tre, Brec1, and wild-type Cre. Values represent the geometric mean of n=3 or n=11 (wild-type Cre) independent replicates (dots) conducted at 37 °C for 30 minutes at a 3:1 protein:DNA ratio. Significant differences ($p \leq 0.05$) relative to wild-type Cre at individual nucleotides (colored asterisks) and across the full log-enrichment profile (‡) are indicated. **c, d,** Heat map of Rec-seq enrichment values for Tre (**c**) and Brec1 (**d**) showing the $\log_2$ of the enrichment value for each nucleotide at each position in *loxLTR* or *loxBTR* relative to the target base (black outline). **e, f**, Specifying interactions mapped onto the structure of Tre in complex with *loxLTR*[171] (**e**) or Brec1 interactions mapped onto the structure of Cre in complex with *loxP*[107] (**f**). The catalytic Tyr (yellow), residues with conserved interactions at unchanged positions relative to *loxP* (red), residues proximal to positions of decreased specificity (blue), and residues that participate in recognition of the new target site (purple) are depicted as spheres. One-letter amino acid labels indicate the Cre residue at that position and the identity of the mutation in Tre or Brec1, if any.

4.2.4 Rec-seq of Dre, VCre, and Bxb1 recombinases

Next, we applied Rec-seq to non-Cre recombinases, most of which remain unexplored as genome editing agents. We performed Rec-seq on Cre relatives Dre[174] and VCre[175] using half-site libraries based on their target substrates *rox* and *loxV*, which differ from *loxP* at 25% and 46% of non-core positions, respectively (Figure 4.8a). Dre and VCre preferred the canonical base at nearly every position in their target sites, similar to wild-type Cre (Figures 4.8b, 4.4a). Though their canonical sequences were enriched in Rec-seq, Dre and VCre profiles revealed several half-site positions with heightened preference relative to neighboring base pairs. Dre showed the strongest preference for half-site positions 6, 7, and 12, while VCre enriched most strongly at positions 5, 6, 10, and 11 (Figure 4.8b). Additionally, VCre showed a unique preference at position 9, which is asymmetric in *loxV* (Figure 4.8a). We observed binary recognition at position 9: T or a C is preferred in the left half-site, with G or A preferred in the right half-site (Figure 4.8c). We hypothesize that these previously unidentified enrichment profile features result from direct interactions between Dre:*rox* and VCre:*loxV,* which may be confirmed by crystallization or in-depth characterization of Dre and VCre.

We also applied Rec-seq to the serine integrase Bxb1[62], which performs strand exchange between two different DNA substrates[56], *attP* and *attB* (Figure 4.8d). Rec-seq with libraries derived from both substrates revealed that Bxb1 maintains two partially overlapping recognition modes to distinguish and selectively recombine two targets that are distinct in sequence and length. We hypothesized that Bxb1 would show the strongest enrichment levels at regions of homology between *attP* and *attB.* Both sites contain a G•C base pair at position 4 and 4', and, in agreement with the literature[176], we observed nearly absolute specificity for these positions in both substrates (Figure 4.8d). Rec-seq profiles also showed enrichment of the ACNAC motif present at positions 6-10 in both the *attP* and *attB* half-sites (Figure 4.8d), consistent with the presence of specifying protein:DNA interactions operating on both targets.

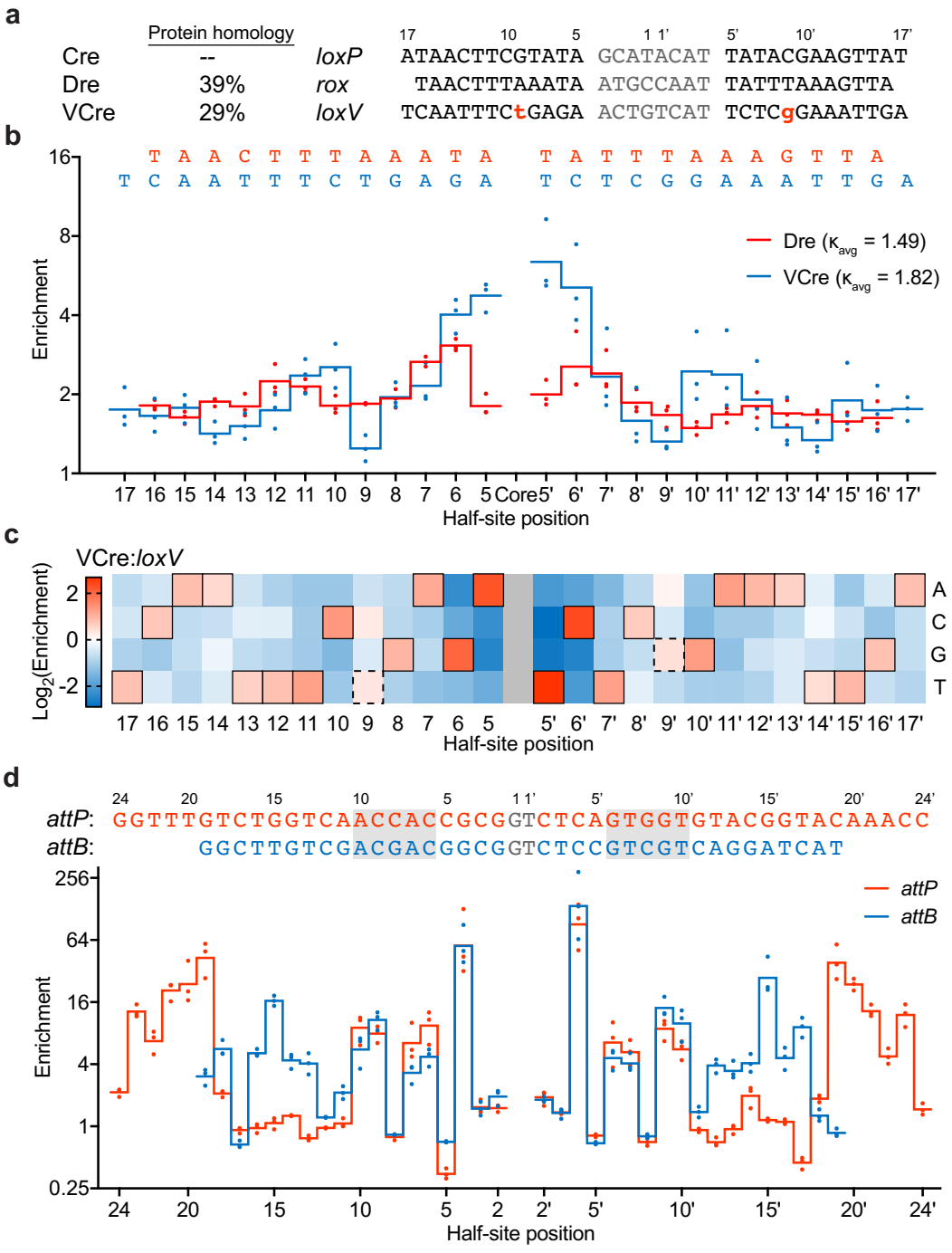


Figure 4.8. Rec-seq profiles of Dre, VCre, and Bxb1 site-specific recombinases. a, Cre, Dre, and VCre differ at the protein sequence level, and bind different recognition targets. **b**, Rec-seq of tyrosine recombinases Dre and VCre. Values represent the geometric mean of n=3 independent replicates (dots) conducted at 37 °C for 30 minutes at a 3:1 protein:DNA ratio. **c**, Heat map of Rec-seq enrichment values for VCre showing the $\log_2$ of the enrichment value for each nucleotide at each position in *loxV* relative to the canonical base (black outline). **d**, Rec-seq of serine integrase Bxb1 on its substrates *attP* and *attB*. Both substrates contain a conserved ACNAC motif (gray box). Values represent the geometric mean of n=3 independent replicates (dots) conducted at 37 °C for 30 minutes at a 3:1 protein:DNA ratio.

Outside of these regions of homology, Bxb1 showed divergent recognition patterns for each substrate. In Rec-seq experiments with *attP* substrates, Bxb1 enriched strongly at half-site positions 19-23 (Figure 4.8d). Enrichment at these positions is consistent with previous reports of a preference for distal bases within *attP* for Bxb1[176] and other integrases[177]. This enrichment largely occurs at positions outside the *attB* minimal site, which consists of two 19-bp half-sites[62]. Bxb1 showed the strongest preference for positions 13-16 in both half-sites of *attB*, but minimal preference for the same region in *attP* (Figure 4.8d). Finally, our observation of pronounced preference at *attB* position 15 is consistent with its reported role in Bxb1's discrimination between *attP* and *attB* substrates[176]. These findings collectively support a model[177] in which Bxb1 enforces fidelity of two asymmetric substrates by adopting overlapping but distinct recognition modes for *attP* and *attB*.

Together, the application of Rec-seq to the characterization of non-Cre recombinases lends support to our model of SSR substrate preferences, uncovers previously unreported specificity determinants, and demonstrates the broad applicability of the Rec-seq method.

4.2.5 Off-target recombinase activity predicted by Rec-seq

Finally, we investigated the ability of Rec-seq to predict off-target activity of SSRs. Before candidate genome editing agents can be used for therapeutic applications, their potential for off-target activity must be assessed[52]. Genomic off-target sequences for these SSRs can be sources of unwanted genomic modification, but also present the opportunity for targeted integration of exogenous DNA, a long-standing goal of recombinase research. Broadened substrate tolerance is anticipated for laboratory-evolved recombinases, as proteins undergoing evolution commonly acquire substrate promiscuity before gaining specificity for the new target[106]. Indeed, we observed relaxed specificity at multiple positions in the Rec-seq profiles of evolved Cre variants Tre and Brec1 (Figure 4.7b). We used Rec-seq data to predict potential off-target substrates for Tre and Brec1, and then assayed the ability of these evolved

recombinases to process predicted substrates, including mismatched “synthetic” substrates enriched from Rec-seq libraries as well as pseudo-sites present in the human genome.

To generate candidate off-target substrates for Tre and Brec1, we first identified non-target half-site sequences that appeared with high abundance in the post-recombinase-treated dataset. For each evolved SSR, we chose four left and right half-site sequences, L1-L4 and R1-R4, that contained 2 or 3 mutations at various half-site positions. The mismatched sequences were observed at 2.7- to 18-fold higher abundance after recombinase treatment versus the input library abundance, compared to the matched *loxLTR* and *loxBTR* sequences, which were enriched 3.0- and 3.4-fold, respectively (Appendix C).

We assessed the activity of Tre and Brec1 on these synthetic substrates in human cells using a reporter plasmid containing pairwise combinations of L1-L4 and R1-R4 half-sites flanking a poly-A terminator that blocks *EGFP* transcription (Figure 4.9a). In this reporter system, recombinase-mediated deletion of the terminator restores *EGFP* expression. We co-transfected HEK293T cells with the reporter plasmid and a plasmid expressing either Tre or Brec1, then used the fraction of cells exhibiting EGFP fluorescence to assess the activity on each target. Both Tre and Brec1 showed comparable or higher activity on the majority of tested synthetic targets relative to their cognate substrate (Figure 4.9b,c), even though these substrates contained up to 5 mismatches. These findings are consistent with relaxed specificities of the evolved variants observed in Rec-seq, and suggest that *in vitro* substrate preferences of SSRs revealed by Rec-seq are predictive of the activity in a reporter plasmid in human cells.

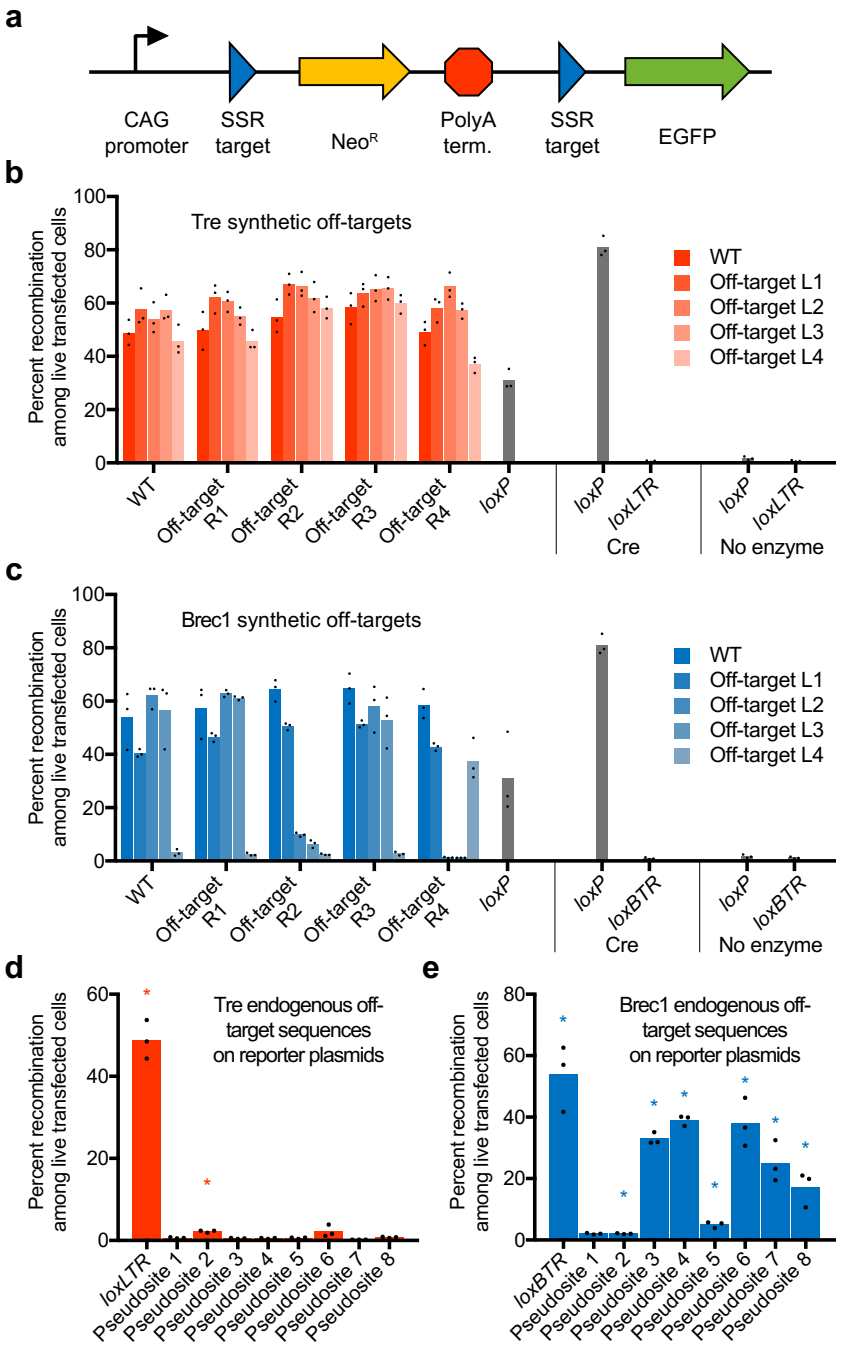


Figure 4.9. Off-target recombinase activity predicted by Rec-seq. a, Cells were transfected with recombinase expression plasmid and an *EGFP* reporter plasmid containing candidate recombinase substrates flanking a poly-A terminator that blocks *EGFP* transcription. Tre and Brec1 activity on synthetic off-target substrates (**b, c**) and predicted endogenous human genomic pseudo-sites (**d, e**) was measured as the fraction of cells exhibiting EGFP fluorescence. The percentage of EGFP-positive cells shown is of transfected cells (determined by gating for the presence of co-transfected plasmid constitutively expressing *mCherry*) and 10,000 live events were recorded for each experiment. Data are represented as the mean (bars) of three independent biological replicates (dots). For **d** and **e**, significant differences ($p \leq 0.05$) relative to no-enzyme control samples are indicated (colored asterisks).

We also assessed whether Rec-seq data alone could predict the activity of Tre and Brec1 on endogenous human genomic sequences. To identify potential pseudo-sites, we searched the human genome for sequences that contained the Tre or Brec1 minimal substrate motif, inferred from positions within each half-site with Rec-seq enrichment values greater than 2. Using the RSAT motif scanner[178] and search parameters $A_{14}C_{13}NT_{11}NNT_8A_7T_6T_5NNNNNNNNT_{5'}A_{6'}T_{7'}NNC_{10'}A_{11'}A_{12'}$ for Tre and $C_{11}NG_9NT_7T_6A_5NNNNNNNNT_{5'}A_{6'}A_{7'}NNC_{10'}NT_{12'}G_{13'}$ for Brec1, we identified eight human genomic off-target substrates per SSR, each containing 6-11 non-core mismatches (Appendix C). These candidate pseudo-sites were cloned into the *EGFP* reporter, and Tre and Brec1 activity was assessed in HEK293T cells as described above. Tre showed significant activity on one of eight endogenous pseudo-sites (Figure 4.9d). Brec1, however, showed robust activity (>15%) on five of eight endogenous pseudo-sites, with significant activity on seven (Figure 4.9e). We confirmed previously reported activity of Brec1 on singly mismatched substrates (Appendix C). We also observed Brec1 activity in human cells on human genomic off-target sequences that were previously identified solely on the basis of *loxBTR* sequence similarity, and found to not undergo recombination by Brec1 in bacterial assays[72] (Appendix C). We attribute this discrepancy, as well as our finding of substantial Tre and Brec1 activity on *loxP*, to differences in SSR performance in mammalian cells compared to the *E. coli*-based assays. These findings suggest that Rec-seq can predict the activity of SSRs on off-target loci including endogenous human genomic pseudo-sites using only *in vitro* enrichment data, even when such sequences are absent from Rec-seq substrate libraries.

4.3 Discussion

Rec-seq is a powerful, high-throughput sequencing-based method that reveals the DNA sequence preferences of SSRs, including specificity determinants not evident from structural studies. We validated Rec-seq with Cre:*loxP*, and used it to characterize the specificity

contributions of over a dozen Cre residues. Rec-seq profiling results support a model for recombinase specificity in which productive recombination requires sufficient binding energy, and loss of one protein:DNA interaction necessitates compensatory increases in fidelity at other (often distant) regions of *loxP*. We also used Rec-seq to accurately predict off-target activity of potential therapeutic recombinases Tre and Brec1. Our findings corroborate previous biochemical and structural characterization of recombinases and reveal numerous insights about Cre and other SSRs, including asymmetric substrate preferences of Cre and long-range interactions of unexpected residues.

Rec-seq represents a major improvement over previous approaches to characterizing the specificity of SSRs, which typically require assaying recombinase activity on each substrate of interest in isolation[84,176,179-181]. Such experiments are labor-intensive, making it impractical to test even all doubly mutated substrates, and do not interrogate the relative preference for multiple competing substrates. More sophisticated methods involve generating a pool of randomized substrates with degenerate primers[182-184] or sheared genomic DNA[185], but these methods use bacterial antibiotic selection to isolate recombinase substrates, and the resolution of such profiling methods is therefore limited by the need to DNA sequence many individual colonies.

In contrast, Rec-seq is an unbiased and rapid method for characterizing SSR substrate preferences at high resolution. The experiments are simple and inexpensive, require no specialized training or equipment, and are easily parallelized. Multiple Rec-seq experiments can be conducted by one researcher in a single day beginning with purified protein and synthesized DNA. We demonstrate the generality of Rec-seq by characterizing not only a widely studied recombinase, Cre, but also distantly related tyrosine SSRs with limited biochemical characterization, as well as an unrelated serine integrase.

Rec-seq also enables experimentally driven off-target substrate prediction for recombinases. The predictive ability of computational searches for recombinase pseudo-sites in

a genome of interest[123,186] is limited by the extent of knowledge about recombinase substrate preferences, which have been characterized at modest depth for only a handful of natural enzymes. Empirical methods for detecting SSR pseudo-sites include overexpressing the recombinase in mammalian cells and identifying sites of genomic modification[125,187]. Rec-seq increases the predictive ability of these methods by generating high-resolution, nucleotide-level DNA specificity profiles of recombinases from libraries of DNA sequences that are orders of magnitude larger than the size of typical mammalian genomes, and that contain a much larger fraction of sequences related to cognate DNA substrates. We used these features of Rec-seq to accurately anticipate Tre and Brec1 activity on pseudo-sites present in the human genome. In principle, Rec-seq libraries could be reconfigured to contain a larger fraction of endogenous mammalian sequences. Such libraries could be especially useful when the identification of genomic off-target substrates is more critical than finding DNA specificity determinants.

Despite these significant advantages, Rec-seq has its own limitations. In its current form, Rec-seq is incompatible with recombinases that require supercoiled substrates[56] (*e.g.*, serine resolvases) due to the linear oligonucleotide origins of the substrate variants. Rec-seq also requires that the researcher can generate purified recombinase and can identify conditions that support *in vitro* activity on Rec-seq library substrates. We successfully purified several SSRs not included in this study (including Flp[188], Vika[189], and SCre[175]), but we were unable to detect robust *in vitro* activity under several conditions. Finally, Rec-seq results are derived from experiments in which only one half-site (L1 or R2) contains mutations while the other three half-sites contain the wild-type sequence, preventing Rec-seq from revealing specificity changes that only arise when multiple changes in different half-sites are simultaneously present.

Rec-seq may facilitate the development of therapeutic recombinases with tailor-made specificities. Generating Rec-seq profiles of different SSRs would increase the pool of potential starting points for retargeting SSRs. Thousands of SSRs are predicted to be encoded in sequenced genomes[190,191], and their Rec-seq profiling would require only knowledge of a

cognate substrate sequence and *in vitro* conditions that support SSR activity. Further, we hypothesize that the model for DNA specificity and energetic tradeoffs, developed in part from Rec-seq profiling of Cre and other SSRs, may guide the use of currently uncharacterized recombinases. Broad profiling of diverse SSRs may also uncover family members with desirable traits as genome editing agents, such as the binary specificity of VCre for the asymmetric position 9 in *loxV* and dual substrate recognition by Bxb1 we observed in this study.

In addition to informing the choice of an evolutionary starting point, Rec-seq profiling may also inform *how* to develop custom recombinases. For example, when choosing new recombinase targets for engineering efforts, we showed that mismatches at corresponding positions in different half-sites are not necessarily penalized equally. During the course of laboratory evolution, performing Rec-seq on intermediate mutants would likely inform subsequent retargeting experiments. Finally, when constructing targeted protein libraries for laboratory evolution, our findings suggest it is important to consider distal interactions in order to promote retargeted, as opposed to merely broadened, specificity.

Rec-seq findings show that long-distance compensatory interactions play an underappreciated role in substrate recognition compared to the limited number of direct Cre:*loxP* contacts. Indeed, among all examined residues predicted to make direct protein:DNA contacts, Ala substitution at only one position (at Arg259) resulted in a near-complete loss of specificity for the proximal base. We also observed that extensive laboratory evolution of Tre and Brec1 resulted in few newly evolved interactions. Together, these findings and previous reports suggest that the dominant mode of substrate recognition for SSRs is not direct protein:DNA interactions, but instead a combination of multiple weak interactions and shape- and charge-complementarity.

4.4 Methods

General Methods

See Chapter 2 methods section.

Cloning, expression and purification of Cre and recombinase variants

Ala-substituted Cre variants were generated by blunt-end ligation cloning of 5' phosphorylated PCR products generated from a previously described pET-His-Cre vector[192]. Expression vectors for other proteins were generated by USER cloning using gBlocks (Tre, VCre) or previously described plasmids (Dre[193], Bxb1[194]) as PCR template.

BL21-Star (DE3)-competent *E. coli* cells were transformed with plasmids encoding Cre or other recombinases with a His purification tag. A single colony was grown overnight in 2xYT media containing 50 μg/mL carbenicillin at 37 °C. The cells were diluted 1:250 into 250 mL of the same media and grown at 37 °C until OD_{600} = 0.60. The cultures were incubated on ice for 20 minutes and protein expression was induced with 1 mM isopropyl-β-D-1-thiogalactopyranoside (IPTG, GoldBio). Expression was sustained for 14-16 h with shaking at 16 °C. The subsequent purification steps were carried out at 4 °C. Cells were collected by centrifugation at 8,000 *g* for 20 minutes and resuspended in cell-collection buffer (100 mM tris(hydroxymethyl)-aminomethane (Tris)-HCl, pH 8.0, 1 M NaCl, 20% glycerol, 5 mM tris(2-carboxyethyl)phosphine (TCEP; GoldBio), and 1 cOmplete EDTA-free protease inhibitor pellet (Roche) per 120 mL buffer used). Cells were lysed by sonication (4 minutes total, alternating 1 second on and 1 second off) and the lysate cleared by centrifugation at 12,000 *g* (20 minutes).

The cleared lysate was incubated with His-Pur nickel nitriloacetic acid (nickel-NTA) resin (4 mL resin per liter of culture; ThermoFisher Scientific) with rotation at 4 °C for 60-90 min. The resin was washed with 50 mL of cell-collection buffer before bound protein was eluted with elution buffer (100 mM Tris-HCl, pH 8.0, 0.5 M NaCl, 20% glycerol, 5 mM TCEP, 500 mM imidazole). The resulting protein fraction was injected into a Slide-A-Lyzer dialysis cassette (10-

kDa molecular-weight cutoff; ThermoFisher Scientific) and dialyzed for 14-16 hours at 4 °C in approximately 100-fold excess storage buffer (100 mM Tris-HCl pH, 8.0, 20% glycerol, 5 mM TCEP). The dialyzed protein fraction was then concentrated using a column with a 10-kDa cutoff (Millipore) centrifuged at 3,000 *g*. Proteins were quantified with Reducing Agent Compatible Bicinchoninic acid (BCA) assay (Pierce Biotechnology), snap-frozen in liquid nitrogen and stored in aliquots at -80 °C.

Brec1 protein was provided by Dr. Gretchen Meinke and Professor Andrew Bohm, Tufts University School of Medicine. The protein contained a Leu163Phe stabilizing mutation, and an N-terminal TEV-cleavable His-tag.

In vitro *extension of library oligonucleotides*

DNA oligonucleotides containing the recombinase target sequence and a 3' hairpin were diluted to 1 μM in nuclease-free water (GE Life Sciences) and NEBuffer 2 in a total volume of 25 μL. The oligonucleotides were heated to 95 °C and slow-cooled to 37 °C to anneal the hairpin, before adding 10 nmol dNTP solution mix and 5 units of Klenow Fragment (3'→5' exo-) polymerase and incubating for 60-90 minutes. The extension reaction was stopped by incubation at 75 °C for 20 minutes, and extended DNA was stored at 4 °C for up to one week.

In vitro *recombination assays*

Each recombination reaction contained one left-hairpin and one right-hairpin substrate oligonucleotide with only one randomized half-site per reaction. In a total reaction volume of 50 μL, recombinase (0.66 pmol for a 1:3 ratio of protein:DNA) was mixed with 1 pmol of each oligonucleotide in nuclease-free water and Cre Recombinase Buffer (NEB) for 30 minutes at 37 °C. Addition of PB buffer (200 μL; Qiagen) stopped the reaction, and DNA was purified with Minelute columns (Qiagen). The purified DNA was digested with the addition of NEBuffer 4, 1 mM adenosine 5'-triphosphate (ATP), and exonucleases I (20 units), III (100 units), and V (10

units) and incubated for 45-90 minutes at 37 °C. The reactions were purified with Minelute columns and the remaining DNA was amplified to the middle of linear range by qPCR (1 µL input DNA, 25 µL reaction volume) using iTaq polymerase (Universal SYBR Green Supermix; Bio Rad). PCR conditions were as follows: 98 °C , then repeated cycles of 98 °C, 57 °C, and 72 °C extension for 5 s. Quantitative PCR was used to ensure the library composition was not affected by PCR bias and that the recombinase-treated samples were more abundant than a no-recombinase negative-control sample. Amplified DNA was purified using Minelute columns and barcoded with a second round of qPCR (0.5 µL input DNA) before being prepared for sequencing on an Illumina MiSeq as described below.

The above protocol was modified to reflect the empirical differences in the optimal reaction conditions for assays with evolved Cre variants and unrelated SSR family members. The recombination reactions with Tre, Brec1, Dre, VCre, and Bxb1 were carried out with a 5-fold increase in concentration of both enzyme and substrate DNA. For Tre and Brec1, recombination buffer was supplemented with 100 ng bovine serum albumen (BSA). For Dre and VCre, reactions were supplemented with 100 ng BSA and 1 mM dithiothreitol (DTT). For Bxb1, reactions were carried out in Bxb1 reaction buffer[195] (20 mM Tris-HCl, pH 7.5, 10 mM EDTA, 25 mM NaCl, 10 mM spermidine, and 1 mM DTT) supplemented with 100 ng BSA. All reactions were carried out at 3:1 protein:DNA ratios for 30 minutes at 37 °C.

Sequencing and analysis of DNA amplicons

Sequencing adapters and dual-barcoding sequences are based on the TruSeq Indexing Adapters (Illumina). Barcoded samples were quantified using the Qubit dsDNA HS Kit (ThermoFisher Scientific) according to the manufacturer's instructions. Sequencing of pooled samples was performed using a single-end reads of 225-250 bases on the MiSeq (Illumina) according to the manufacturer's instructions.

Rec-seq data analysis

Sequencing reads were automatically demultiplexed using MiSeq Reporter (Illumina) and Fastq files were analyzed using custom software tools written in Python 3, made available online at https://github.com/broadinstitute/rec-seq. In brief, post-recombination sequencing reads that contained the matched target core sequence were aligned to the native target sequence, with no gaps allowed. After alignment, reads with excessive numbers of mismatches were determined to be the result of sequencing errors, *e.g.*, reads containing indels. Therefore, aligned reads with greater than 6 mismatches relative to the reference sequence were filtered out of subsequent analysis. For the remaining sequences, at each position in the recombinase target, the abundance of the canonical base (A_i) and the sum of the non-canonical bases (B_i) were calculated. The same analysis was performed for the sequencing reads of the input library, but the abundances of the canonical base and the non-canonical bases were expressed as fractions α_i and β_i. The enrichment score for each position was then calculated as the ratio $r_i = (A_i/B_i)/(\alpha_i/\beta_i)$. Analysis was performed separately for the left and right half-sites, using as input the sequencing reads from experiments with either L1- or R2-randomized half-sites (see Figure 4.1b).

Significance of log-enrichment values was calculated by performing the Student's t-test assuming equal variance for each individual position of each SSR variant relative to wild-type Cre, and the effect of multiple comparisons was counteracted using the Bonferroni correction. A paired t-test was used to compare the asymmetry between the left and right half-site log-enrichment values for wild-type Cre (Figure 4.4c). We calculated the significance of differences along the full substrate log-enrichment profile using the two-sided Mann-Whitney U test. To do so, we compared the absolute value of the residuals for wild-type Cre and each enzyme variant, and applied the Bonferroni correction. A list of significance values can be found in Appendix B.

Cloning of mammalian recombinase expression and reporter plasmids

Mammalian expression plasmids were constructed via the ligase cycling reaction method[121] using a pCMV vector and gBlocks encoding Tre and Brec1.

The pCALNL-GFP subcloning vector, pCALNL-EGFP-BsaI, was used to clone all reporter plasmids and was based on the previously described pCALNL-EGFP-Esp3I vector[81]. The BsaI site in the ampicillin gene of the pCALNL-EGFP-Esp3I vector was first removed by Gibson assembly of BsaI-HFv2-digested plasmid and a dsDNA oligonucleotide with Gibson overhangs and a point mutation ablating the BsaI site. The pCALNL-EGFP-BsaI plasmid was created by Golden Gate assembly with the modified pCALNL-EGFP-Esp3I vector and a PCR product bearing a pTET-mRFP cassette flanked by BsaI and Esp3I sites. Golden Gate reactions were set up and performed as described previously with Esp3I (ThermoFisher Scientific)[156]. The donor vector, containing the neomycin-terminator cassette, was constructed by USER cloning using a PCR product of the cassette from pCALNL-EGFP-Esp3I and a pUC-Kan vector.

pCALNL-EGFP *loxP*, *loxLTR*, and *loxBTR* reporter plasmids were created by Golden Gate assembly with the pCALNL-EGFP-BsaI acceptor vector, pBT100-neomycin-terminator donor vector, and pairs of dsDNA oligonucleotides bearing recombinase target sites flanked by BsaI overhangs. Golden Gate reactions contained 0.1-1 pmol of each component, BsaI-HFv2 (20 units; NEB), and T4 DNA Ligase (20 units).

HEK293T transfection and flow cytometry

HEK293T cells (ATCC CLR-3216) were cultured in Dulbecco's Modified Eagle's Medium (DMEM; Corning) supplemented with 10% fetal bovine serum (FBS; Life Technologies). Cells were seeded into 48-well poly-D-Lysine-coated plates (Corning) in the absence of antibiotic. 12-15h after plating, cells were transfected with 1 µL of Lipofectamine 2000 (ThermoFisher Scientific) using 250 ng of recombinase plasmid, 25 ng of reporter, and 10 ng of fluorescent protein expression plasmid as a transfection control. Cells were cultured for 3 d before they

were washed with PBS (ThermoFisher Scientific) and detached from plates by the addition of TrypLE Express (ThermoFisher Scientific). Cells were diluted in 250 µL culture media and run on a BD LSR II analyzer. Significance of recombinase activity measurements relative to no-recombinase control transfections was calculated by performing the Student's two-tailed t-test assuming unequal variance.

Chapter 5:

Insights into the Future Development of Recombinase-Based Genome Editing Tools

Chris Podracky and I designed and performed the experiments described in sections 5.3.3 and 5.4.2 and figures 5.4 and 5.6. I designed and performed all remaining experiments.

5.1 Introduction

Recognizing the potential of site-specific recombinases (SSRs) as genome editing agents, I have undertaken several projects with the goal of developing general recombinase tools for efficient gene integration into human cells. My colleagues and I developed a method for continuous *in vivo* selection of DNA recombinases to retarget Cre toward a sequence present in a human genomic safe harbor locus. We also developed an RNA-programmable recombinase by fusing the Ginβ recombinase catalytic domain to dCas9. In recognition of the difficulties we experienced in developing programmable recombinases, we chose to develop a system for profiling the specificity determinants of SSRs in order to facilitate future retargeting efforts.

While DNA recombinases have so far challenged retargeting efforts, the potential reward for success could be enormous. In principle, a programmable recombinase could accomplish all the same genomic modifications achievable by existing genome editing technologies such as programmable nucleases, base editors[196], and engineered viruses[197] while also catalyzing predictable and efficient gene integration. Solving this problem will likely require a highly interdisciplinary approach, and my own work encompasses three distinct project areas. In the concluding chapter of this dissertation, I describe experiments that explore synergies between the three different approaches I attempted. These experiments include the design of a new PACE retargeting trajectory and rational engineering of recCas9 variants with improved properties. I also describe new selections for improving the activity of the rationally designed recCas9 variants. Finally, I discuss underexplored enzymes that demonstrate promising features as candidate genome editing tools.

5.2 Design of PACE selections informed by recombinase specificity profiling

The Rec-seq method was developed in part to assist with the design of recombinase retargeting experiments. For example, after deciding to retarget Cre toward the ROSA26 locus, we had to choose a specific sequence within ROSA26 to serve as the selection substrate. We

considered questions such as: Which Cre:*loxP* interactions are most important? Should these interactions be conserved when choosing a new target sequence? How should one design evolutionary intermediate sequences, given the initial substrate preferences of Cre? Insights provided by Rec-seq help to answer each of these questions. Rec-seq can also inform the generation of protein libraries for selection experiments by illuminating which regions of Cre are responsible for recognition of a given region of *loxP*. Finally, profiling of Cre variants after selection on a ROSA*loxP* intermediate substrate can be used to evaluate the retargeting progress and design subsequent experiments.

In light of the Rec-seq data, we realized that selection for activity on the ROSA*loxP* target may have encouraged the promiscuous recombinase phenotype we observed. For example, the ROSA*loxP* site contains a mutation at position 10 (Figure 2.3a), the site of critical Cre:l*oxP* interactions (Figure 4.5b). Following selection for activity on the L1 intermediate – which introduces the position 10 transversion – surviving SP contained Cre variants with R259C and E262A mutations (Figure 2.4c). The first of these mutations results in the loss of two energetically-favorable hydrogen bonds between Arg259 and the C•G base pair at position 10[162], while the second was previously shown to increase substrate mismatch tolerance in Cre[84]. In comparison, when Buchholz and colleagues selected for altered recognition at position 10 in *loxLTR* and *loxBTR*, they observed R259Y/E262Q and R259D/E262R mutations in Tre[75] and Brec1[72], respectively. While the nature of the substitutions observed in Tre and Brec1 are suggestive of retargeted recognition, mutations accumulated in PACE likely contribute to increased mismatch tolerance.

In addition, while ROSA*loxP* contained equal numbers of mismatches in the left and right half-sites, the difficulty of evolving recognition of each half-site was not equal. The LF substrate contains transversion mutations at critical positions 7 and 10, while RF mismatches occur at substrate positions with no direct Cre:*loxP* interactions (Figures 2.3a, 4.5a). Accordingly, selection for activity on the LF target resulted in consensus variants with 11 total mutations

(including likely promiscuity-conferring mutations), while RF-active variants converged on a single mutation of E262A or G (Figure 2.5d). Together, the insights from profiling of Cre explain how experimental design choices may have encouraged a promiscuous phenotype resulting from selection for ROSA*loxP* recognition in PACE.

To determine whether retargeting could be successful given a better choice of substrate, I searched for a sequence within the human ROSA26 locus that incorporates the insights from Rec-seq and our initial PACE retargeting efforts. On the basis of Rec-seq profiling of wild-type Cre (Figure 4.4a), I searched for sequences within the ROSA26 locus that did not contain mismatches at critical *loxP* positions 5, 6, and 10 in both half-sites. I also prioritized sequences without transversions at positions 7, 11, and 12. The chosen target, termed ROSA20, contains 20 mismatches relative to *loxP* (Figure 5.1a). I designed two series of intermediate substrates, with one series for transitioning preference toward each half-site. Activity on ROSA20 would be achieved by evolving separate lineages of Cre variants that recognize symmetric left or right half-site intermediates. The mismatched positions with the highest Rec-seq enrichment values, and therefore those likely to have the greatest impact on Cre binding energetics, were targeted in the first selection step, when initial SP binding is at wild-type levels. In addition, I designed the ROSA20 trajectory to mimic the intermediate substrate strategy employed by Buchholz and colleagues, in which each intermediate is subdivided into several sub-sequences, and variants with activity on each sub-sequence are combined and shuffled before selection on the next intermediate (Figure 5.1a).

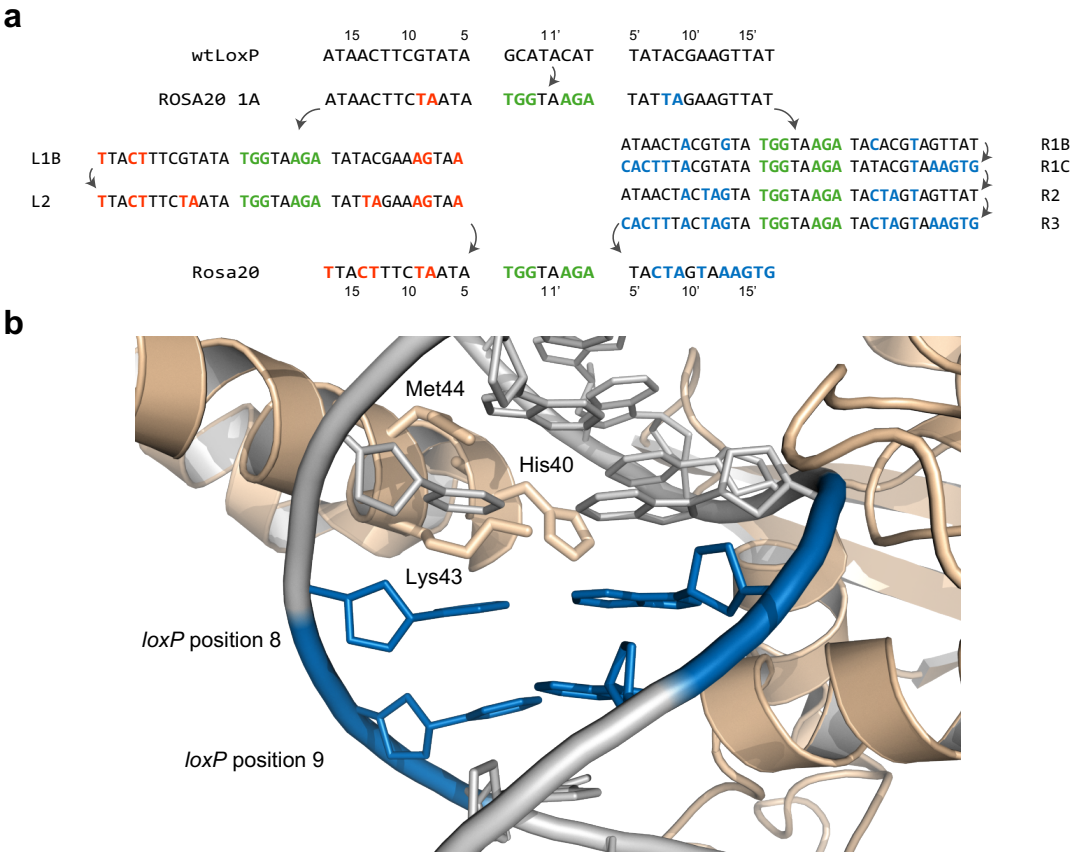


Figure 5.1. ROSA26 retargeting strategy informed by Rec-seq. a, PACE evolutionary trajectory for retargeting Cre recombinase toward the ROSA20 sequence. To evolve activity on an asymmetric target, recombinase variants are first selected for activity on the common intermediate ROSA20-1A. Following selection on ROSA20-1A, variants are selected on left and right half-site intermediates bearing increasing numbers of mismatches relative to *loxP* (colored bases). Each intermediate is also broken down into nested sub-sequences to allow for shuffling of mutations at each step. **b,** PACE lagoons containing ROSA20-1A host cells were seeded with an SP library of Cre with site-saturation mutagenesis at residues 40, 43, and 44. Positions of mismatches within ROSA20 (blue) are highlighted.

I attempted a mixing strategy to evolve Cre recognition of ROSA20 in PACE. Similar to the strategy employed for ROSA*loxP* selections (Figure 2.3b), wild-type Cre SP was propagated on host cells with a *loxP* AP for 24 hours, followed by selection on a 1:1 *loxP*:ROSA20-1A mixture of host cells for 24 hours. After the mixing phase, SP were propagated exclusively on host cells bearing the 1A intermediate AP, and I observed rapid washout of SP. Selections conducted with periods of genetic drift[95] or SP containing Cre variants with site-saturation mutagenesis at residues proximal to positions 8 and 9 (Figure 5.1b) also resulted in washout.

Reasoning that direct selection on the 1A substrate was too stringent, I made APs in which only position 8 or position 9 was mutated, but still observed a defect in Cre SP propagation.

My experience with several PACE selections raises more questions about the choice of retargeting substrates than answers. After ROSA*loxP* retargeting led to promiscuous variants, I concluded that direct Cre:*loxP* interactions should be preserved when choosing a target sequence. However, the ROSA20 experiments demonstrated the difficulty of selecting for recognition of non-contacted positions in *loxP*. In addition, Rec-seq revealed that Tre and Brec1 were successfully evolved to prefer substrates containing mismatches at sites of critical Cre:*loxP* interactions (positions 5, 7, or 10; Figure 4.7). Indeed, selecting for recognition of mismatches at protein:DNA interfaces, as opposed to non-contacted positions, may be preferable, as it remains very difficult to evolve or engineer the indirect interactions required for the latter strategy. Achieving new specificity at non-contacted positions may require a different strategy, such as constructing chimeric fusions of recombinase domains[198,199]. Future research could directly test different strategies for choosing retargeting substrates.

5.3 Further development of recCas9 by protein engineering and evolution

5.3.1 Rational design of recCas9 variants informed by specificity profiling

I also investigated rational design of recCas9 variants based in part on high-resolution profiling of Cre specificity determinants. The ideal programmable recombinase would include two properties: strong reliance on dCas9 for target localization and minimal recombinase target sequence preference. My attempts to improve recCas9 by fusion of dCas9 to ROSACre only partially met these requirements, as the resulting variants demonstrated limited activity (Figure 3.6). I therefore considered whether rational modifications to increase gRNA dependence and recombinase mismatch tolerance would yield recCas9 variants with improved properties.

I attempted to impart Cre dependence on dCas9 binding by generating “hypomorphic” Cre variants that require gRNA-programmed binding events for activity. I reasoned that

engineered heterodimeric Cre pairs could be a source of such hypomorphic mutations, as these Cre monomers were developed to have minimal activity in the absence of the heterodimer partner. For example, Baldwin and colleagues performed a domain swap in the C-terminal helix of Cre[115], yielding the heterodimer pair “HetF” and “HetAA”. Separately, Church and colleagues found that substitution of Arg32, at the interface between helix A and helix C in Cre, with Met or Val resulted in decreased cooperativity between Cre monomers[200]. Therefore, I inserted the candidate hypomorphic mutations in Cre-dCas9 fusion proteins and assessed the variants for gRNA-dependent *loxP* activity in HEK293T cells (Figure 5.2a). Surprisingly, fusions of the candidate hypomorphic variants showed minimal gRNA dependence, similar to wild-type Cre and the PACE-evolved variant ROSACre 1. I attributed this lack of gRNA dependence to the strong innate preference of Cre for *loxP* and the sensitivity of the transfected mammalian cell reporter. I therefore used all four hypomorphic variants for subsequent experiments.

Next, I attempted to design recombinase domains with minimal substrate specificity based on the results of Cre profiling. Previously, we performed Rec-seq with Ala-substituted Cre variants in order to dissect the *loxP* specificity determinants, but in the process we identified residues in Cre where Ala mutations resulted in loss of proximal specificity (Figure 4.5). I reasoned that different combinations of Ala mutations in wild-type Cre might result in the broad reduction of substrate specificity. Insertion of these mutations within the context of a hypomorphic dCas9 fusion could prevent the problems associated with promiscuity we observed in PACE-evolved variants, and gRNA-programmed localization could provide sufficient binding energy for productive recombination. To generate a panel of candidate programmable recombinases, I designed 10 combinations of promiscuity-conferring mutations (Table 5.1) and generated recCas9 variants containing these as well as the four hypomorphic mutations.

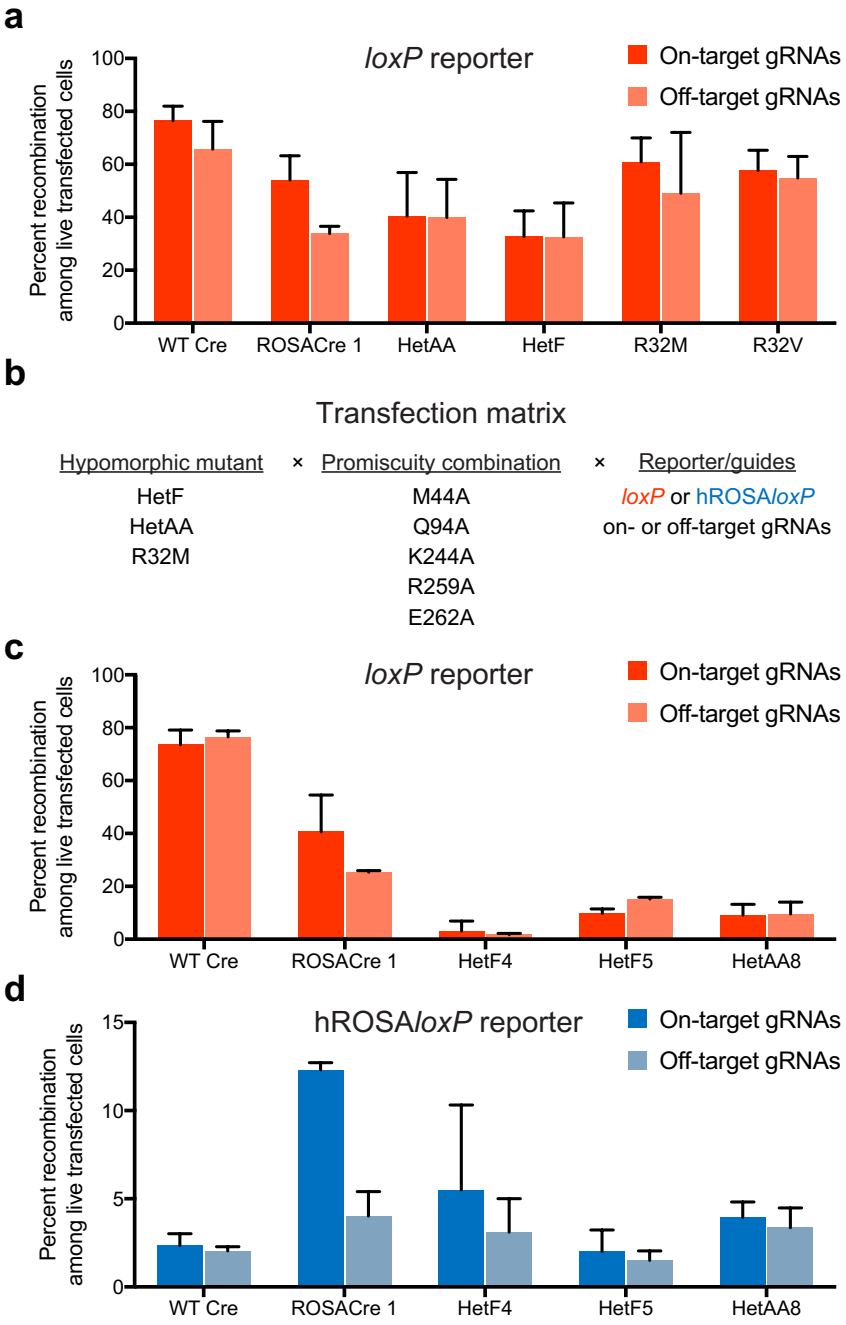


Figure 5.2. Activity of designed recCas9 variants on *loxP* and hROSA*loxP*. a, Cells were transfected with a recCas9 expression vector encoding a hypomorphic Cre variant, *loxP* reporter plasmid, and on- or off-target pairs of gRNA expression vectors. Transfection experiments are described in detail in Figure 3.1. **b,** Summary of transfection experiments for investigating rationally designed recCas9 variants. Cells were transfected with a plasmid expressing dCas9 fusions to Cre variants with hypomorphic and promiscuity-conferring mutations, vectors for expressing on- or off-target gRNAs, and reporter plasmids for *loxP* (**c**) or hROSA*loxP* (**d**). The results of select rational mutants are shown; the list of promiscuity-conferring mutation combinations can be found in Table 5.1. The percentage of EGFP-positive cells reflects that of transfected (iRFP-positive) cells. Values and error bars represent the mean and standard deviation of two independent biological replicates.

Variant	Mutations
1	M44A, Q94A, K244A, R259A, E262A
2	Q94A, K244A, R259A, E262A
3	M44A, K244A, R259A, E262A
4	M44A, R259A
5	M44A, E262A
6	Q94A, R259A
7	Q94A, E262A
8	R259A
9	E262A
10	R259A, E262A

Table 5.1. Combinations of promiscuity-conferring mutations. The chosen residues were implicated in recognition of *loxP* positions 5-7 (M44, Q94), position 10 (R259, E262), and position 16 (K244), based on Rec-seq profiling data (Figure 4.5).

I transfected HEK293T cells with plasmids expressing each rationally designed recCas9 variant, on- or off-target gRNAs, and reporter plasmids with the *loxP* or hROSA*loxP* target (Figure 5.2b). The majority of recCas9 variants were either inactive or showed high gRNA-independent activity on *loxP*. However, several variants displayed minimal *loxP* activity and detectable recombination of the hROSA*loxP* target (Figure 5.2c,d). Encouragingly, the HetF4 variant showed nearly 50% of ROSACre 1 activity on hROSA*loxP*. This variant contains just 3 rational mutations (M44A, R259A, and A334F) and demonstrates activity on a sequence with mismatches at > 40% of *loxP* positions; in comparison, ROSACre 1 accumulated 10 coding mutations over the course of 500+ hours of PACE. While rationally designed recCas9 variants display limited activity and modest gRNA dependence, these results suggest that mismatch tolerance in Cre may be achieved through installation of a limited number of mutations in a fusion context.

5.3.2 Continuous selection of recCas9 variants

Having engineered recCas9 variants with moderate activity on the ROSA*loxP* target, I sought to apply PACE selection to improve levels of recombination. I designed a PACE selection for recCas9 based on the second-generated recombinase selection circuit (Figure

2.5a). Due to the DNA packaging limit of M13 phage[201], I could not encode full-length recCas9 on the SP. I therefore split recCas9 between Cre and dCas9 and fused each half to the *Nostoc punctiforme* DnaE intein[202], with DnaE-dCas9 expressed from a complementary plasmid (CP) within the host cell (Figure 5.3a). I modified the AP by adding a gRNA expression cassette and inserting gRNA binding sequences that flank the ROSA*loxP* target. When translated, these gRNA sequences produce in-frame flexible linker peptides to minimize disruption to pIII' function.

I assessed the activity of recCas9 variants on the PACE selection circuit in overnight enrichment assays. Host cells bearing the ROSA*loxP* AP and DnaE-dCas9 CP were inoculated with SP encoding intein fusions of the top recCas9 variants from the mammalian transfection assays. Compared to the unfused variant, intein-fused ROSACre 1 showed a 1,000-fold decrease in overnight enrichment (Figure 5.3b). This finding suggests that the DnaE intein successfully mediates formation of recCas9 *in vivo*, even if the fusion results in a fitness defect. While the rational recCas9 variants did not substantially enrich overnight, neither did SP bearing wild-type Cre, suggesting that the circuit has low background and may simply require lower selection pressure. However, decreasing stringency by substituting the high-copy pUC origin with origins of intermediate (ColA) or low (sc101) copy number did not improve overnight enrichment (Figure 5.3c). In addition, rational Cre variants showed *improved* enrichment when host cells contained a CP that lacked DnaE-dCas9. This again suggests that recCas9 formation decreases SP fitness. I therefore decided not to attempt PACE, as I suspected that continuous selection would result in premature termination of the dCas9 fusion partner. While I was unable to attempt PACE with recCas9, future engineering of SSR fusions with greater gRNA dependence may provide suitable recCas9 candidates for selection in PACE.

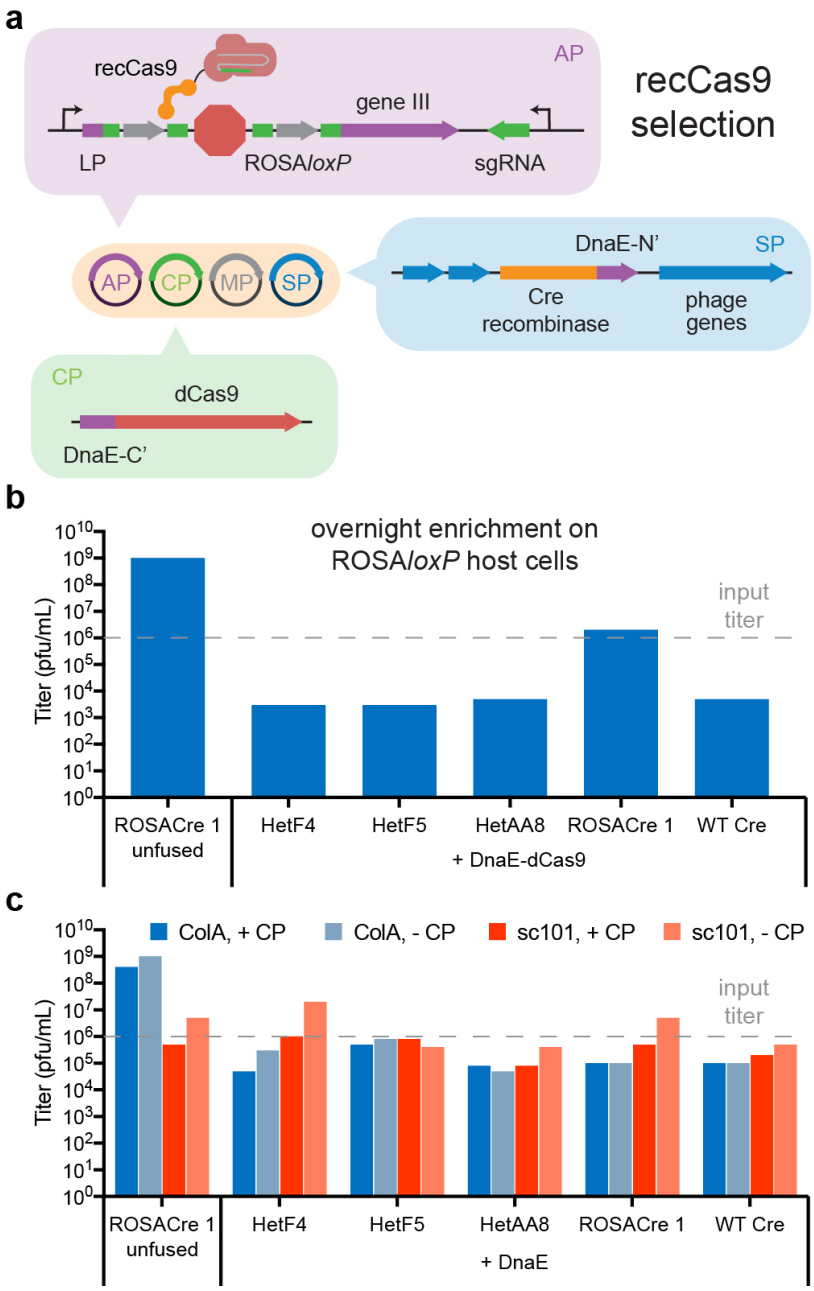


Figure 5.3. PACE selection for recCas9 variants. a, Schematic of recCas9 selection in PACE. The SP encodes the recombinase domain of split-intein recCas9, while the dCas9 half is expressed from a complementary plasmid (CP) in the host cells. The deletion cassette lies within the coding sequence of gIII, in between the leader peptide (LP) and the N1 domain. Deletion of the transcriptional terminator restores production of pIII', containing a peptide corresponding to the gRNA and recombinase target DNA sequences, which is functionally incorporated by infectious progeny. **b,** Host cells bearing the recCas9 ROSA*loxP* AP on a high-copy pUC origin and the DnaE-dCas9 CP were inoculated with 10^6 pfu of SP encoding designed Cre variants. Cells were grown overnight, and SP titer in the supernatant was determined by plaque assay. **c,** Similar overnight enrichment assays were conducted in host cells with the ROSA*loxP* AP on an intermediate-copy (ColA) or low-copy (sc101) origin, and a CP encoding dCas9 or an empty CP.

5.3.3 Eukaryotic selection for improving the activity of programmable recombinases

Besides selection in PACE, I explored alternative approaches for improving the activity of recCas9 variants. The budding yeast *Saccharomyces cerevisiae* is an attractive species for conducting laboratory evolution[203]. Yeast cells contain a highly structured genome within a nucleus, presenting the opportunity to directly select for the desired activity of a programmable recombinase tool: gene integration into a eukaryotic genome. Additionally, conducting plate-based selections or screens in *S. cerevisiae* may incorporate passive negative selection against the emergence of a promiscuous phenotype, as genotoxic variants would get removed from the evolving population.

With my colleague Chris Podracky, I designed a yeast-based fluorescence circuit for detecting recCas9-mediated genomic integration (Figure 5.4). We constructed a host strain with a genomically integrated cassette encoding a promoter upstream of one exon of *GFP*, followed by a splice donor sequence and an intronic recCas9 target. The yeast are transformed with a plasmid expressing a variant library of Cre fused to dCas9 and gRNAs for the intronic target. The library strain is then transformed with a donor cassette containing a matching intronic target and the second exon of *GFP* with a splice acceptor sequence. RNA-programmed integration of the donor cassette into the genomic target results in cellular fluorescence, and active enzyme variants are isolated using fluorescence-assisted cell sorting (FACS). Negative selection against gRNA-independent activity could be accomplished by discarding variants that produce cellular fluorescence in the absence of targeting gRNAs. Enhancing recCas9 activity with the yeast-based genomic circuit is currently the focus of ongoing investigations.

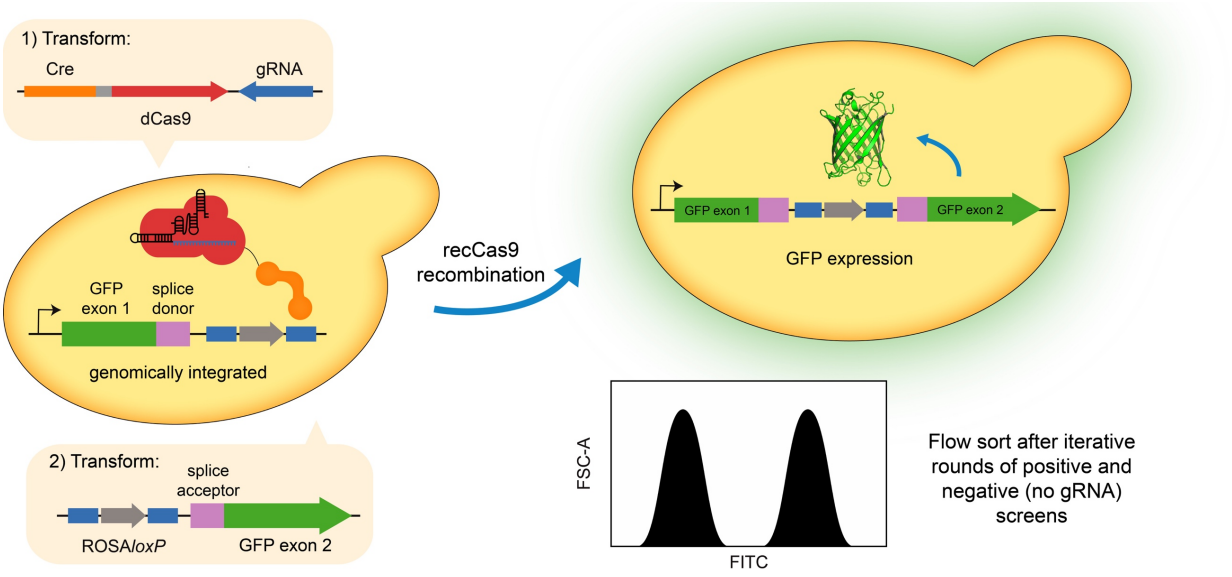


Figure 5.4. Eukaryotic circuit for detecting recCas9-mediated genomic integration. *S. cerevisiae* cells contain a genomically-integrated cassette that expresses one exon of *GFP*, a splice donor sequence, and an intronic target. Cells are transformed with a plasmid expressing gRNAs for the intronic target and a variant library of Cre-dCas9 fusions, followed by transformation with a donor cassette containing a matching intronic target, splice acceptor sequence, and the second exon of *GFP*. RNA-programmed genomic integration of the donor cassette results in cellular *GFP* expression, and active recCas9 variants are identified by flow cytometry. Selection against gRNA-independent recombination could be implemented by discarding GFP-positive cells transformed with off-target gRNAs.

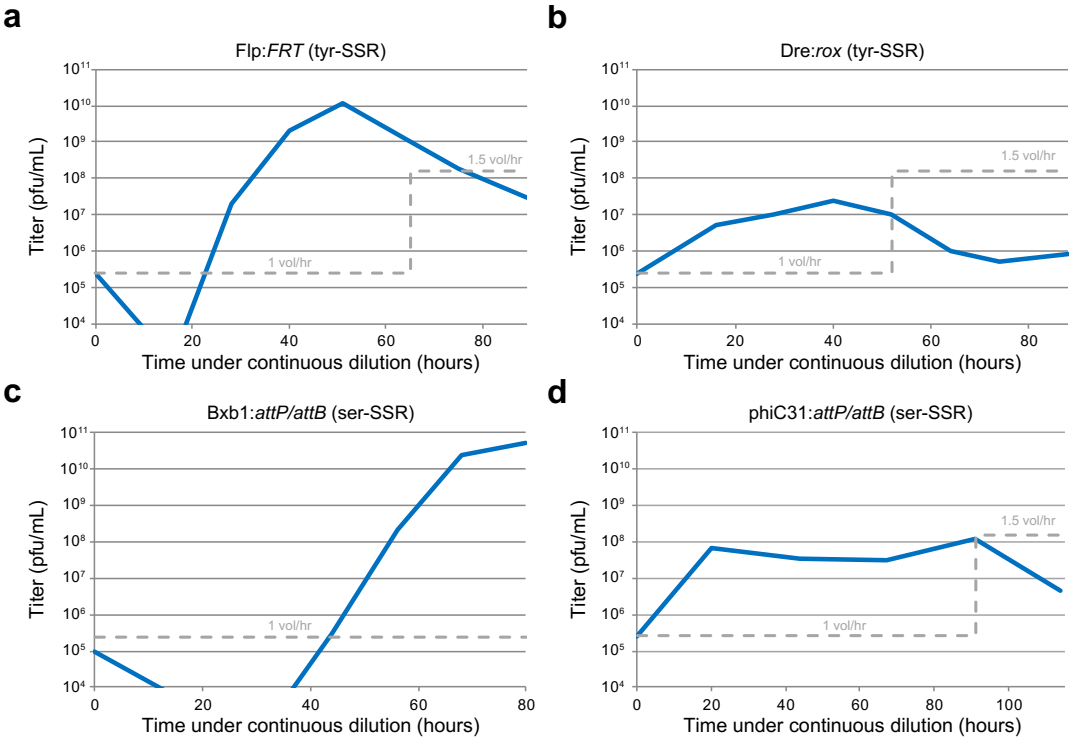
5.4 Promising classes of enzymes for development as genome editing agents

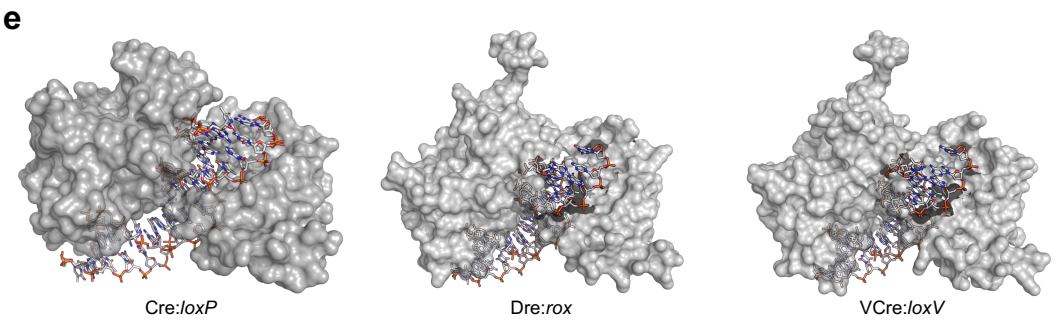
5.4.1 Non-Cre SSRs

In principle, the PACE recombinase selection could be used to evolve non-Cre SSRs. To explore the versatility of the selection circuit, I generated SP encoding a diverse group of SSRs and APs with their cognate recognition sequences. In separate PACE experiments, I observed selective propagation of the tyrosine recombinases Flp and Dre and the serine integrases Bxb1 and phiC31 on their wild-type substrates (Figure 5.5a-d). In each of these experiments, lagoons seeded with SP encoding T7 RNA polymerase resulted in immediate washout. These findings suggest that the recombinase selection circuit is likely general to many more SSRs, expanding the list of possible starting points for retargeting experiments.

The Rec-seq profiling method may further the development of non-Cre recombinases by revealing their specificity determinants with high resolution. Compared to Cre – subject of numerous structural and biochemical characterizations – most SSRs have scarcely been investigated. Thousands of SSRs are predicted to be encoded in sequenced genomes[190,191], each with a unique substrate preference and pattern of protein:DNA interactions. Due to the broad applicability of Rec-seq, profiling an unexplored SSR requires only knowledge of a cognate substrate sequence and *in vitro* conditions that support recombinase activity. Generating a database of Rec-Seq profiles of different SSRs could facilitate the choice of an SSR starting point and evolutionary trajectory for a given retargeting goal.

The findings of Rec-seq analysis of Cre specificity determinants may also translate to other SSRs, allowing for the rapid determination of which residues are responsible for substrate recognition based on protein homology. For example, I used the structural prediction algorithm Phyre2[204] to generate models of the Cre relatives Dre and VCre (Figure 4.8a). While the three enzymes differ substantially at the primary sequence level, the Dre and VCre structures are highly homologous to Cre, with 100% confidence in the backbone alignment covering 94% and 84% of each protein respectively (Figure 5.5e). To predict protein:DNA interactions, I aligned Dre and VCre monomers to a structure of Cre in complex with *loxP*[118], substituted bases in *loxP* to simulate the cognate recombinase target[205], and calculated the hydrogen bonds and van der Waals interactions for the resulting models[206]. One limitation of this approach is that the predicted models do not include solvent interactions, which are important for Cre binding[83]. Nevertheless, the predicted protein:DNA contacts for Dre and VCre include many residues that overlap with important Cre determinants of specificity (Figure 5.5e). Retargeting efforts involving Dre and VCre may benefit from targeted mutagenesis at residues identified as functionally important for Cre substrate recognition.





Cre	Dre	VCre
Met44	Asn45	Arg64
Lys86	Thr86/87	His103
Lys244	Arg245	Lys263/His264
Arg259	Ala260	Glu278
Glu262	Glu262	Glu290

Figure 5.5. Prospects for evolving alternative SSRs using PACE. a-d, PACE experiments were seeded with host cells and SP bearing the recombinase selection circuit for tyrosine SSRs FLP/*FRT* (**a**) and Dre:*rox* (**b**), and serine integrases Bxb1 (**c**) and phiC31 (**d**). The *y* axis shows total phage titer in the lagoon (n=1). **e,** Crystal structure of Cre in complex with *loxP*[118], as well as computational models of Dre and VCre in complex with their cognate DNA sequences, generated using the Phyre2[204] and 3DNA[205] algorithms. Despite substantial differences in primary sequence, high levels of predicted structural homology facilitates the identification of residues in Dre and VCre that are analogous to functionally important Cre residues as identified by Rec-seq.

In particular, the serine integrases are especially promising candidates for further development due to their catalysis of directional recombination. Few retargeting efforts have been attempted with serine integrases[125,199], likely due to limited structural characterization of integrase relatives[181,207] and uncertainty about integrase substrate recognition. To address the latter challenge, we used Rec-seq data to reveal a comprehensive model for integrase specificity, in which Bxb1 enforces fidelity of two asymmetric substrates by adopting overlapping but distinct binding modes for *attP* and *attB* (Figure 4.8b). This in-depth knowledge of the binding preferences of Bxb1 could assist in the identification of endogenous sequences suitable for retargeting. For example, using the RSAT motif scanner[178], I identified over 50 human genomic sequences that contain the highly specified ACNACNGNNNNNNCNGTNGT motif common to both *attP* and *attB* (Appendix D). Informed by Rec-seq, Bxb1 retargeting experiments could be designed to promote recognition of pseudo-*attP* or *attB* sequences with mismatches outside of this conserved motif.

5.4.2 Retroviral integrases

While I have focused my graduate studies on the retargeting of site-specific recombinases, there exist many other classes of enzymes that may be suitable for development as genome editing tools. For example, retroviral integrases (IN) accomplish targeted genomic integration during the life cycle of viruses such as HIV[208] via binding and processing of the long terminal repeat (LTR) ends of proviral DNA[209]. IN have been the subject of extensive biochemical and structural characterization for the purpose of drug discovery, facilitating their potential development as genome editing tools. The dominant mechanism for targeted retroviral integration is association with endogenous genomic features or proteins, with IN often displaying weak DNA sequence preferences (reviewed in ref. 209). For example, HIV IN relies on binding to the nuclear protein LEDGF/p75[210] to direct integration of the HIV provirus toward actively transcribed regions of the genome[211]. The retroviral integration preference can be

influenced by mutating the interface between IN and endogenous binding partners[212]. Several reports have also demonstrated that fusion of IN to DNA binding domains can bias the pattern of genomic integration[213-217]. Due to extensive characterization of IN and the potential for retargeting via protein engineering, I consider retroviral integrases to be promising candidates for further development.

My colleagues and I have conducted preliminary experiments to explore retargeting IN via fusion to dCas9. Several reports demonstrate that HIV IN catalyzes integration into the genome of *S. cerevisiae*[218,219], presenting an opportunity to apply the many molecular biology techniques developed in yeast to the engineering of tools containing IN. We designed a yeast-based selection that could be used in principle to evolve IN-mediated genomic integration programmed by gRNAs (Figure 5.6a). We would first construct a host strain with a genomically integrated *GFP* cassette. The yeast would be transformed with a plasmid expressing a variant library of IN fused to dCas9 and gRNAs targeting sites internal to *GFP*. The library strain would then be transformed with a donor cassette containing *URA3* lacking a promoter and flanked by IN LTR ends. RNA-programmed integration of *URA3* within the *GFP* gene results in loss of cellular fluorescence, and confers survival on selective media lacking uracil. We chose to target a sequence internal to *GFP* because the exact sequence preferences of IN are unclear, and we are unsure exactly where IN would integrate relative to the gRNA sites; targeting the middle of a gene allows for many different integration events to result in the same phenotype. Active enzyme variants could be selected for survival on ura-dropout media, and localized integration could be promoted by FACS enrichment of non-fluorescent cells. Negative selection against gRNA-independent activity could be accomplished by omitting gRNAs and plating cells on media containing the URA3 inhibitor 5-fluorouracil.

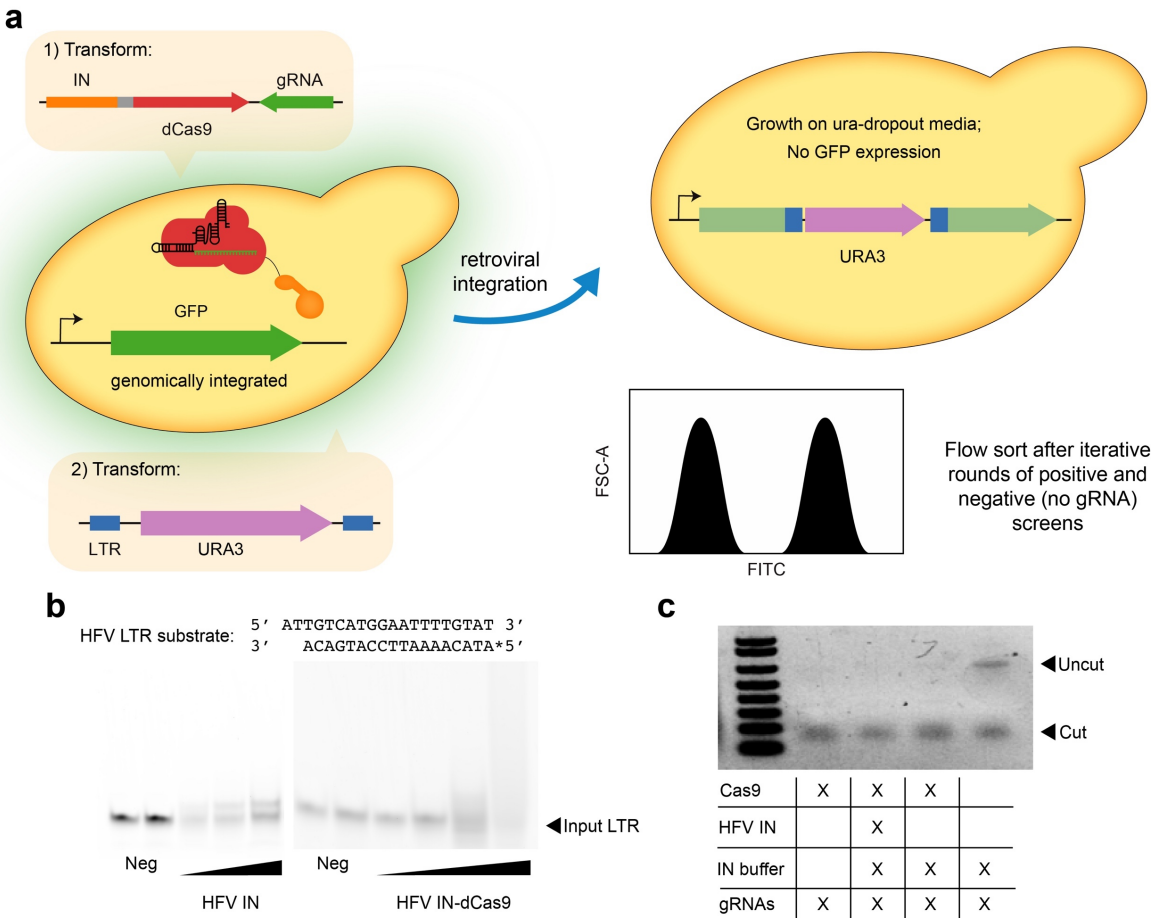


Figure 5.6. Eukaryotic selection for programmable retroviral integrases. a, Schematic for generating active fusions of dCas9 and a retroviral integrase (IN). *S. cerevisiae* host cells contain a genomically-integrated cassette that expresses *GFP*. Cells are transformed with a plasmid expressing *GFP*-targeting gRNAs and a variant library of IN-dCas9 fusions, followed by transformation with a donor cassette containing *URA3* lacking a promoter and flanked by minimal IN long terminal repeats (LTR) ends. RNA-programmed integration of *URA3* into the coding sequence of *GFP* disrupts fluorescence expression and confers survival on ura-dropout media. Counterselection against gRNA-independent integration could be implemented by discarding GFP-positive cells or growing cells transformed without gRNAs on media containing 5-fluorouracil. **b,** *In vitro* LTR integration assay[220] with human foamy virus (HFV) IN alone or as a fusion to dCas9. Increasing amounts of IN or IN-dCas9 were exposed to a fluorescently-labeled (asterisk) LTR substrate, and integration was detected by the appearance of higher MW DNA bands on an agarose gel. **c,** *In vitro* cutting assays were conducted by mixing Cas9 nuclease, gRNA, and plasmid cutting substrate. Nuclease activity in the presence or absence of HFV IN or HFV IN buffer was detected by gel electrophoresis.

We conducted proof-of-principle experiments to determine whether IN and dCas9 could accomplish their respective enzymatic functions in a fusion context, as a precursor to conducting selections in yeast. Specifically, we were unsure of whether IN would retain integrase activity as a chimeric fusion, and whether the presence of a retroviral protein – which

might be expected to bind structured RNA – would disrupt Cas9:gRNA binding. We identified a list of candidate IN proteins for fusion to dCas9, selecting enzymes that have been reconstituted *in vitro*, characterized for integration site preference, and/or fused to DNA-binding domains (Table 5.2). We were able to purify IN from human foamy virus (HFV) alone and as an N-terminal fusion to dCas9, and proceeded to characterize these proteins *in vitro*. We observed that both HFV IN and HFV IN-dCas9 are capable of integration between fluorescently-labelled LTR sequences (Figure 5.6b), indicating that chimeric fusion does not substantially impair IN activity. Next, we observed that Cas9:gRNA complexes are functional in assays conducted in the presence of unfused HFV IN (Figure 5.6c), suggesting that IN does not disrupt Cas9:gRNA binding. Together, these preliminary results are encouraging signs that development of IN-dCas9 fusions as genome editing tools may be feasible.

Integrase	Comments
ASLV	Reports of integration site preferences[221]; successfully tethered to LexA DBD[215]
HFV	Demonstrated activity *in vitro*[220]
HIV	Successfully tethered to λ repressor[213], LexA DBD[214], and Zinc finger proteins[216,217]; operates on genome of *S. cerevisiae*[218,219]
HTLV	Reports of integration site preferences[221]
MLV	Generated mutant integrase deficient in associating with bromodomain extra-terminal proteins[212]
MMTV	Reports of integration site preferences[221]

Table 5.2. Candidate IN proteins for fusion to dCas9. IN were chosen on the basis of reports of reconstitution *in vitro*, characterization of integration site preference, and/or fusion to DNA-binding domains (DBDs).

5.4.3 Additional candidate enzymes

Transposases are a broad class of enzymes which include relatives of both SSRs and retroviral integrases, and they are appealing as tools for gene integration due to catalysis of a similar DNA transformation. Transposons are mobile genetic elements containing cis-regulatory

sequences and a transposase for genomic integration. Similar to retroviral integrases, some transposons demonstrate limited inherent DNA specificity[209,222] and instead their integration pattern is defined by association with host nuclear proteins. Transposases have not been widely studied as genome editing agents, likely due to limited structural and sequence preference information. Nonetheless, transposon integration patterns can be biased by transposase fusions to DNA-binding domains[223,224], including Zinc fingers, TALEs, and dCas9[225]. A theoretical selection or screen for programmable transposition could resemble the yeast-based circuit for dCas9-IN fusions depicted in Figure 5.6a.

Finally, future genome editing tools may today be undiscovered in nature or buried in the literature. For example, a recent investigation of a deep-sea thermophilic archaebacteria included the discovery of the pTN3 mobile genetic element[226]. Study of the TN3 integrase, a tyrosine-type SSR, revealed its ability to not only catalyze site-specific recombination but also homology-mediated recombination between diverse sequences. Much like how early development of CRISPR/Cas9 was accomplished by yogurt manufacturers[227], little-known proteins with tantalizing properties such as TN3 may one day represent the future agents of genome editing.

5.5 Methods

General methods

See Chapter 2 methods section. Plasmids for mammalian cell expression of rationally designed recCas9 variants were generated using the Darwin Assembly method[228].

Phage propagation assay, plaque assays, and phage-assisted continuous evolution

See Chapter 2 methods sections.

HEK293T transfection and flow cytometry

See Chapter 3 methods section.

Acknowledgements

Acknowledgements

I'd first like to thank **David Liu** for his outstanding mentorship and support throughout the course of my Ph.D. I am grateful to have conducted my graduate studies in a lab with tremendous resources, exceptional colleagues, and the freedom to pursue my own research interests. Thank you very much for all the ways you've directly and indirectly supported my growth over the past six years. Thank you as well to **Aleks Markovic, Keenan Holmes,** and **Anahita Hamidi** for your friendship and support of my graduate studies.

I'd also like to thank my committee members, **Keith Joung** and **Dan Kahne**, for their guidance and mentorship. Thank you for challenging me and providing me with a fresh perspective at my committee meetings.

I owe a tremendous debt of gratitude to my research mentors. **David Thompson**, thank you for teaching me just about everything I know about molecular biology. It was apparent as soon as we started working together that you have an infectious enthusiasm for research, and the skills and perspectives you shared with me have stayed with me and shaped me as a scientist. Thank you to **Ahmed Badran** for inspiring me with your incredible work ethic and your razor-sharp experimental design. I'm grateful to you both for answering my many questions and fostering my early growth as a grad student.

I'm very proud to have served as a research mentor to **Lena Afeyan**. You've turned into a bright and promising graduate student, and it was a pleasure to watch your growth, from your first summer doing research through your senior thesis and beyond. Your work on the recombinase profiling project laid the groundwork for my favorite project that I've worked on, and you deserve immense credit for the new discoveries that have resulted.

I was fortunate to collaborate with several excellent scientists on the recombinase profiling project. One of the best decisions I made in graduate school was to partner with **Vlado Dančík**. It was truly amazing to watch my rough analysis of the preliminary data morph into a compelling, mature scientific narrative, thanks to your guidance and expertise. Thank you for your hard work, and for your endless patience in answering my questions during our numerous impromptu meetings in your office. Thank you to **Paul Clemons** for overseeing our partnership and providing helpful feedback. I am tremendously grateful for the research materials provided by **Andrew Bohm** and **Gretchen Meinke**, but I am even more appreciative of the perspective and expertise you shared with me. Finally, without the help of **Chas Leichner**, the recombinase profiling project probably would never have made it off the ground. Thank you for all the time and expertise you contributed, for teaching me everything I know about programming, and for being my lifelong friend.

My favorite memories of graduate school will no doubt be those that I made with my colleagues, both in lab and outside of lab. **Chris Podracky**, I am confident that the recombinase project is in capable hands moving forward. Thank you for your friendship and for help in shaping the next stages of recombinase research. **Luke Koblan**, I've learned a tremendous amount from working with you and being your friend, and I look forward to seeing what you're able to accomplish in your next steps. Thank you to **Ben Thuronyi** for countless thoughtful conversations and useful insights – your future students will be extremely lucky to have you as their professor. I'd also like to thank **David Bryson, Phillip Lichtor, Holly Rees, Tim Roth, Juan Pablo Maianti, Alix Chan,** and **Brian Chaikind** for your support, friendship, and advice. Thank you to all the members of the **PACE subgroup** over the years for serving as a source of ideas, feedback,

thoughtful criticism, and support. Finally, I am grateful to **all of my colleagues in the Liu lab** for contributing to the lab community and making it such a fun and stimulating workplace.

I am extremely grateful for the friendships that I made during grad school. Thank you to **Charlie Margarit** for all the ex-beer-iments and home-cooked meals, and for helping me navigate the ups and downs of grad school and of life. Thank you to **Matt Mitcheltree** for engaging in ridiculous conversations, and for exposing me to art and food and culture. Thank you to **Sam Blau**, **Micah Maetani,** and **Brian Daniels** for helping to relieve the stress of graduate school. Thank you to **Ben Levin** and **Samantha Cassell** for your companionship and support. Thank you to **Eileen Moison** for sharing your insights and helping me to prepare for the real world. Thank you to **Lily Dodge** for helping me improve my writing. Finally, special thanks to my friends **Jordan Wilkerson, Claudia Kleinlein, Nare Janvelyan, Andrew Bendelsmith, David Gygi, Ethan Magno, Matt Volpe** and **Christina Chang.**

I would like to thank **my family** for their love and patience during the last six years. You have been my rock-solid support system, and I'm so appreciative that I can always count on you to be there for me during the good times and the bad. Thank you to **my parents** for providing me with the countless opportunities that have led me to this moment, and thank you to **Erica** and **Sam** putting up with my quirks and always having my back. Thank you as well to the **Malfitano/Palen/Zimmerman's**, who have served as my surrogate East Coast family these past few years. Thank you for welcoming me into your family and smiling politely as I repeatedly told you how there was only "one year left before I graduate!"

Finally, and most importantly, I would like to thank my partner, fiancée, and bride-to-be **Danielle**. My life is so much richer with you in it. You have fully experienced the ups and downs of grad school right alongside me, and I can't overstate how grateful I am that you have always been my most loyal supporter. I love you with all of my heart.

Appendices

Appendix A. RecCas9 genomic targets identified *in silico*

Chromosome	Start	End	Sequence	Pattern ID
chr1	34169027	34169103	CCTTTAGTGAAAAGTAGACAGCTCTGAATATGAAAGGTAG GTTTTCATTTCTGGGAAAGAGACGCCAAGTGATGTGG	2
chr1	51006703	51006780	CCTCCAATAAATATGGGACTATGTGGAAAGACCAAACCTA CGTTTGATTGGTGTACCTGAAAGTGACGGGAAGAATGG	1
chr1	89229373	89229450	CCATTCTGCCCGTCACTTTCAGGTACACCAATCAAACGTA GGTTTAGTCTTTTCACATAGTCCCATATTTCTTGGAGG	1
chr1	115638077	115638154	CCATTCTCCCCGTCACTTTCAGGTACAACAATCAAACGTA GGTTTGGTCTTTTCACATAGTCCCATATTTCTTGGAGG	1
chr1	122552402	122552478	CCTTGTAGTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGACTTGAAACACTCTTGTTGTGG	2
chr1	122609874	122609950	CCTTGTAGTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCATACTTGAAACACTCTTTTTGTGG	2
chr1	122668677	122668753	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGACTTGAAACACTCTTTTTGTGG	2
chr1	123422419	123422495	CCTTGTGTTGTGTTTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGACTTGAAATACTCTTTTTGTGG	2
chr1	123648614	123648690	CCTTGTAGTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCATACTTGAAACACTCTTTTTGTGG	2
chr1	123806335	123806411	CCTTGTATTGTGAGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGACTTGAAACACTCTTTTTGTGG	2
chr1	124078228	124078304	CCTTGTGTTGTGTGTCTTCAACTCACAGAGTTAAACGATG CTTTACACAGAGTAGACTTGAAACACTCTTTTTCTGG	2
chr1	124231074	124231150	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGACTTGTAACACTCTTTTTGTGG	2
chr1	124232435	124232511	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGACGTGAAACACTCTTTTTGTGG	2
chr1	124344781	124344857	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGACTTGAAACACTCTTTTTGTGG	2
chr1	124435716	124435792	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGGAGACTTGTAACACTCTTTTTGTGG	2
chr1	158677186	158677262	CCTGAGGTTTTCCAGGTTTTAAAAGGAAACCTAAAGGTAG GTTTAGCATTAAGTGTCTTGAAGTTTATTTTAAAAGG	2
chr1	167629479	167629554	CCAAAATTCCCACAAAACCGAATGCATCAGTCAAAGCAAG GTTTGAAGAAAAGATTTACCACTTCAGGGAGCTTGG	4
chr1	167783428	167783504	CCTTTTCTGGATATCGTTGATGCTCTGTATGCAAAAGGTA GGTTTTTGGGTTATGTTGTTAAACAGTGATTGAATGG	3
chr1	169409367	169409444	CCTCCAAGAAATATGGAACTATGTGAAAAGACCAAACCTA CGTTTGATTGGTGTACCTGAAAGTGACAGAGAGAATGG	1
chr1	174145346	174145423	CCTCCAAGAAATATGGGACTATGTGAGAAGACCAAACCTA CGTTTGATTGGTGTACCTGAAAGTGATGGGGAGAATGG	1
chr1	183750168	183750245	CCATTCTCCCCATCGCTTTCAGGTACACCAATCAAACGTA GGTTTGGTCTTTTCACATAGTTCCATATTCTTTGGAGG	1
chr1	200801540	200801617	CCATTCTCCCCATCACTTTCAGGTGTACCGATCAAACGTA GGTTTGGTCTTTTCACATAGTCCCATATTTCTTGGAGG	1
chr1	207589936	207590013	CCTCCAAGAAATATGGGACTATGTGAAAAGACCAAACCTA CGTTTGATTGGTGTACCTGAAAGTGACGGGGAGAATGG	1
chr1	209768370	209768445	CCTTCAGGGCAGAAACAGCTCTACTAGCAGAGAAAGCAAG CTTTCAATATTGTGCAATACAAAAACGAGAGCAGGG	4
chr1	218652378	218652455	CCATTCTCCTCATCTCCTTCTGGTACTCCAATCAAACGTA GGTTTGGTCTTTTCTCATAGTCTCATATTTCTTGGAGG	1
chr1	222147250	222147327	CCTCCAAGACATATAGGACTATGTGAAAATACCAAACCTA CGTTTGATTGGTGTACCTGAAAGTGACAGGGAGTATGG	1
chr1	245870710	245870785	CCTGCCAGATACCAGTAGTCACTGTGAATTACAAAGCTAC GTTTCTTCCATAGGGAAAGTTTGGAGTCCAGCCAGG	4
chr2	2376037	2376114	CCATTCTCCCTGTCACTTTCAGGTACACCAATCAAACGTA GGTTTGGTCTTTTCACATAGTCCCATATTTCTTGGAGG	1

chr2	4119629	4119706	CCATTCTCCCCACCACTTTCAGGTACACCAATCAAACGTA GGTTTGGTCTTTTCACATAGTCCCATATTTCTTGTAGG	1
chr2	4909047	4909124	CCTAACCAGAAACTAACTAATAGATATGGGCAGAAAGCAT CCTTTCACTTTTGTTCTGGGAGAGGGAAGAAGCAAAGG	1
chr2	28984877	28984953	CCATTTTGGGGAGGCCTTGATGGGAAGCTGGAAAAGGAAG CTTTCCTCCCAGTCCTGCTGAAGGCCTTGCCAGCTGG	2
chr2	31755833	31755910	CCTCCAAGAAACACAGGACTATGTGAAAAGATCAAACCTA CGTTTGATTGGTGTTCCTGAAAGTGATGGGGAGAATGG	1
chr2	39829583	39829660	CCATTCTCTTCATGACTTTCAGGTACACCATTGAAACGTA GGTTTGGTCTTTTCACATTGTCCCATATTTCTTGGAGG	1
chr2	60205947	60206024	CCATTCTCCCCATCACTTTCAGGTACACCAATCAAACGTA GGTTTGGTCTTTTCACATAGTCCCGTATTTCTTGGTGG	1
chr2	79082362	79082439	CCATTCTCCCTGTCACTTTCAGGTACACCAATCAAACGTA GGTTTGGTCTTTTCACATAGTCCCATATTTCTTGGGGG	1
chr2	79082362	79082438	CCATTCTCCCTGTCACTTTCAGGTACACCAATCAAACGTA GGTTTGGTCTTTTCACATAGTCCCATATTTCTTGGGG	3
chr2	108430915	108430992	CCTCCAAGAAATATGAGATTATATGAAAAGACCAAACCTA CGTTTGATTGGTGTACTTTAAAGTGACGGGGAGAATGG	1
chr2	115893685	115893762	CCATTCTCCCCGTCATTTTCAGGTACACCAATCAAACGTA GGTTTGGTCTTTTCACATAGTCCCAAATTTCTTGGAGG	1
chr2	119620068	119620145	CCCCCAAGAAATGTGGGACTATATGAAAAGACCAAACCTA CGTTTGACTGGTGTACCTAAAAGTGATGGGGAGAATGG	1
chr2	119620069	119620145	CCCCAAGAAATGTGGGACTATATGAAAAGACCAAACCTAC GTTTGACTGGTGTACCTAAAAGTGATGGGGAGAATGG	2
chr2	128495068	128495144	CCCATTGGTGCTGACCAGATGGTGAAGGAGGCAAAGGTTG CTTTGAATGACTGTGCTCTGGGGTGAGCCAGGCCTGG	2
chr2	133133559	133133634	CCCTTTACAGAGGTGAGCTTTGTTATTAGTAAAAAGGTAG GTTTCCCTGTTTTTCTGAAGAAAAGCTGTGAGTGGG	4
chr2	134174983	134175060	CCACTGCCCATTGACAGAGTGGCGAGGTGGGTGAAACCTT GCTTTCCTCCTGGCCCATGGGCAGGGTGGGGCTGTGGG	1
chr2	134174983	134175059	CCACTGCCCATTGACAGAGTGGCGAGGTGGGTGAAACCTT GCTTTCCTCCTGGCCCATGGGCAGGGTGGGGCTGTGG	3
chr2	138069945	138070022	CCATTCTCCCTGTCACTTTTAGATACACCAATCAAACGTA GGTTTGGTCTTTTCACATAGTCCCATGTTTCTTGGAGG	1
chr2	138797420	138797496	CCTCCAAGAAATATCAACTGTGTGAAAAGACGAAACCTAC GTTTGATTAATGTACCTGAAAGTGACAGGGAGAATGG	2
chr2	145212434	145212511	CCATTCTCCCATTAACTTTCAAGTACACCAATCAAAGGTA GGTTTGGTGTTTTCCCATAGTCCCGTATTTCTTGGAGG	1
chr2	147837842	147837919	CCTTTTCATCATGCCCCTTTCACTTTAAGGTGAAAACCTT GCTTTACATGTCAGAGAAAAGAAGAGCCCTCAGCTGGG	1
chr2	147837842	147837918	CCTTTTCATCATGCCCCTTTCACTTTAAGGTGAAAACCTT GCTTTACATGTCAGAGAAAAGAAGAGCCCTCAGCTGG	3
chr2	154152540	154152617	CCATTCACCCCGTCACTTTCAGGTACACCAATCAAACGTA GGTTTGGTCTTTTCACATAGTCCCATATTTCTTGGAGG	1
chr2	157705943	157706019	CCTCCAAGAAATATGGGACTATGTGAAAAGACCAAACCTA CGTTTGATGGTGTACCCGAAAGTGACAGGGAGAATGG	3
chr2	158361152	158361229	CCACCAAGAAATATGGGACTATGTGAAAAGACCAAACCTA CGTTTGATAGGTATACCTGAAAGTGACAGGGAGAATGG	1
chr2	161461006	161461083	CCATTCTCCCCATCACTTTCAGGTGCACCAATCAAACGTA GGTTTGGTCTTTTCACATAGTCCCATATTTCTTGGAGG	1
chr2	179077376	179077453	CCCTCAAGAAATATGAGACTATGTGAAAAGACCAAACCTA CGTTTGACTGGTATACCTGAAAGTGACAGGGAGAATGG	1
chr2	179077377	179077453	CCTCAAGAAATATGAGACTATGTGAAAAGACCAAACCTAC GTTTGACTGGTATACCTGAAAGTGACAGGGAGAATGG	2
chr2	181090699	181090776	CCTCCAACAAATATGGGACTATGTGAAAAGACCAAACCTA CGTTTGATTGGTGTACCTGAAAGTGACGGGGATAATGG	1
chr2	182331957	182332034	CCATTCTCTCCCTCACTTTCAAGTACACCAATCAAACGTA GGTTTGGTCTTTTCACATAGTCTTATATTTCTTGGCGG	1
chr2	183620562	183620638	CCATTCTCCCTGTCACTGTCAGTACACCAATCAAACGTAG GTTTGGTCTCTTCACATAGTCCCATATTTCTTGGAGG	2
chr2	207345927	207346003	CCTCCAAGAAATATGGGACTATGTGAACAGACCAAACCTA	3

			CGTTTGATTGGTGTACCTGAAAGTGATGGCAGAATGG	
chr2	216652047	216652123	CCACCATGCCTGGCCACCACACATTTTTTTCTAAAGCTTG GTTTTGGCCACAGTGAGAGTTTCTTGGGCTGTCAGGG	2
chr2	216652047	216652122	CCACCATGCCTGGCCACCACACATTTTTTTCTAAAGCTTG GTTTTGGCCACAGTGAGAGTTTCTTGGGCTGTCAGG	4
chr2	223780040	223780116	CCCACTAGGTGGCGATATCTGAGGGTCCAATGAAACCATG CTTTTTACTCAGATCTTCCACTAACCACCTCCCCCGG	2
chr2	224486595	224486672	CCTCTAAGAAATATGGGACTATGTGAAAAGACCAAACCTA CGTTTGACTGGTGTACCTGAAAGTGACGGGGAGAATGG	1
chr2	230526902	230526979	CCTCCAAGAAATATGGGACTATGTGAAAAGACCAAACCTA CGTTTGATTAGTGTACCTGAAAGTGACGGGGAGAATGG	1
chr2	232036127	232036204	CCATTCTCCCTGTCACTTTCAGGTACATCAATCAAACGTA GGTTTGGTCTTTTCACATAGTCCCATATTTCTTGGAGG	1
chr3	4072812	4072889	CCTCCAAGAAATATGGGACTATGTGAAAAGACCAAACCTA CGTTTGACTGGTGTACCTGAAAGGGATGGGGAGAATGG	1
chr3	9261677	9261754	CCCCCAAGAAATATGAGACTATGTGAAAAGACCAAACCTA CGTTTGATTGGTGTACCTGAAAGTGACAGGGAGAATGG	1
chr3	9261678	9261754	CCCCAAGAAATATGAGACTATGTGAAAAGACCAAACCTAC GTTTGATTGGTGTACCTGAAAGTGACAGGGAGAATGG	2
chr3	16732146	16732223	CCTCTAAGAAATATGGGACTATGTGAAAAGACCAAACCTA CGTTTGATTGGTGTAACTGAAAGTGACAGGGAGAATGG	1
chr3	17450712	17450789	CCTCCAAGAAATATGCGCCTATGTGAAAAGACCAAACCTA CGTTTGATTGGTATACCTGAAAGTGATGGAGAGAATGG	1
chr3	21559769	21559846	CCATTCTCCCTGTCACTTTGAGGTACACCAATCAAACGTA GGTTTGGTCTTTTCACATATTCGCATATTTCTTGGAGG	1
chr3	23416658	23416735	CCATTCTCCCCGTCACTTTCAGGTACACCAACCAAACGTT GGTTTGGTCTTTTCACATAGTCCCATATTTCTTGGAGG	1
chr3	29984019	29984096	CCATTCTCCCTGTCACTTTCCAGTACACCAGTCAAACGTA GGTTTGGTCTTTTCACATACTCCCATATTTCTTGGAGG	1
chr3	38269551	38269627	CCTGGCCTAATTTTTAATTCTTAGTTTGACTTAAACCTTG CTTTTAGTGTGATGGCGACAAAAGCTGAGCTGAAAGG	2
chr3	40515213	40515288	CCAGTGCTTTTTGGTTTTAAAGGCAAGCCTCCAAACCTTC CTTTCTCCTGGATGCTGTGGTGGTTGCCATGCATGG	4
chr3	49233612	49233687	CCCAACTCCTGCGAGAAGTAGCTCACCATGACAAAGCTAC CTTTGCTTTTATCGTTTTGCAAAACAAAAAAGGGGG	4
chr3	66292894	66292971	CCATTCTCCCCGTCACTTTGAGGTGTGCCAATCAAACGTA GGTTTGGTCTTTTCACATAGTCCTATATTTCTTGGAGG	1
chr3	67541493	67541570	CCTCCAAAAAATATGGGACTACGTAAAAAGACCAAACCTA CGTTTGATTGGTGTACCTGAAACTGACAGGGAGAATGG	1
chr3	82273011	82273088	CCATTCTCCCCGTCACTTTCAGGTACACCAATCAAACGTA GGTTTGGTCTTTTCACATAGTTCCATATTTCTTGGAGG	1
chr3	98683349	98683426	CCTACAAGATATATGGGACTATGTGAAAAGACCAAACCTA CGTTTTACTGGTGTGCCTGAAACTGACGGGGAGAATGG	1
chr3	101923653	101923730	CCATTCTCTCTGTCACTTTCAGGTACACCAATCAAACGTA GGTTTGGTCTTTTCACATAGTCCCATATTTCTTGGAGG	1
chr3	114533467	114533544	CCTCCAAGAAATATGGGACTATGTGAAAAGACCAAACCTA CGTTTCATTGGTGTACCTGAAAGTGATAGGGAGAATGG	1
chr3	132607602	132607679	CCTCCAAAAAATATGGGATGATGTGAAAAGACCAAACCTA GGTTTGACTGGTGTACCTGAAAATGATGGGGAGAATGG	1
chr3	137545176	137545253	CCTCCAAGAAATATGAGACTATGTGAAAAGACCAAACCTA CGTTTGATTGGTGTACCTGAAAGTGACAGGGAGAATGG	1
chr3	137655679	137655756	CCTCCAAGAAATATGGGACTACGTGAAAAGATCAAACCTA CGTTTGATTGTTGTACCTGAAAGTGATGGGGAGAATGG	1
chr3	137662040	137662117	CCTCCAAGAAATATGGGACTATGTGAAAAGACCAAACCTA CGTTTGATTGTTGTACCTGAAAGTGATGGGGAGAATGG	1
chr3	142133796	142133873	CCTCAAAAGTGTTCTGGTTTTGTTTTGTTTTTTAAACCAT GGTTTTACCTCTGGCTTAGTGGGACTAAAAATAGGAGG	1
chr3	146726949	146727026	CCTCCAAGAAATATGGGACTATGTGAAAAGACCAAACCTA CGTTTGACTGGTGTACCTGAAAGTGATGGGGAAAATGG	1
chr3	152421096	152421173	CCTCCAAGAAATATGGGACTGTGTGTAAAGACCAAACCTA CGTTTGATTGGTGTACCTCAAAGTGATGGGGAGAATGG	1

chr3	170620247	170620324	CCATTCTCCCCATCACATTCAGGTACACCAATCAAACGTA GGTTTGGTCTTTTCACATAGTCCCATATTTCTTGGAGG	1
chr3	181166873	181166949	CCCCTGGAAAAGTTGGAGCATCACAGGAAAAGCAAACCAA CCTTTTTTCTCCCCTAGGTAAACTGGGGAGCCAGGGG	3
chr3	181166874	181166949	CCCTGGAAAAGTTGGAGCATCACAGGAAAAGCAAACCAAC CTTTTTTCTCCCCTAGGTAAACTGGGGAGCCAGGGG	4
chr4	6604233	6604309	CCTTCCCCAGTTGCAGCAGACAAGAGTCTCGAAAAGCTTG CTTTGGTTGCTGCAGTGGATGGGTTGGTAGGCACAGG	2
chr4	6626269	6626344	CCCCCACCTCCCAAGCTGCTGGCTTCTCGAATAAAGCTAC CTTTCCTTTTACCAAAACTTGTCTCTCGAATGTCGG	4
chr4	8155396	8155472	CCTTGGCCCTGGACAGCTGCTTTTCCTTCCCTAAACCTTG GTTTCCCCCTTTGTGCAGGTGGGTGGGTTTGGGCTGG	2
chr4	10386803	10386880	CCTCTTCTAGTGAACCCATGGGGTTACCAAGGGAAAGCAA CCTTTTGATAAATATTCCCATCTTTTTATGTTGTCTGG	1
chr4	20701579	20701656	CCACTTGAAAGGGTTACCAAGGATAAGATTTTTAAAGCTT GCTTTCACAAACAACTCATGCTCCAGGCTTGTCAGTGG	1
chr4	29594286	29594363	CCTTTCTCCCCATCACTTTCAGGTACACCAATCAAACGTA GGTTTGATCTTTTCACATAGTCCCATATTTCTTGGAGG	1
chr4	53668422	53668499	CCATTCTCCCCATCAATTTCAGTTACACCAATGAAACGTA GGTTTGGCCTTTTCACATAGTCCCATATTTCTTAGAGG	1
chr4	74914802	74914879	CCATTCTCCCTGTCACTCTCAGGTACACCAATCAAACGTA GGTTTGGTCTTTTCATATAGTCCCATATTTCTTGGAGG	1
chr4	75332783	75332859	CCTCCAAGAAAATTGGGACTATGTGAAAAAACCAAACCTA CGTTTGATTGATGTACCTGAAAGTGACAGGAGAATGG	3
chr4	88123643	88123720	CCTTCAAGAAATATGGGACTATGTGAAAGGACAAAACCTA CGTTTTATTGGTGTACCTGAAAGTGACAGGGAGAATGG	1
chr4	89567192	89567269	CCATTCTCCCCATCACTTTCAGGTACGCTAATCAAACGTA GGTTTGATCTTTTCACATAGTCTTATATTTCTTGGAGG	1
chr4	93556577	93556654	CCTCCAAGAAATATGGGACTATGTGAAAAGACCAAACCTA CGTTTGACTGGTGTACCTCAATGTGACAGGGAGAATGG	1
chr4	100266379	100266456	CCATTCTCCCTGTCACTTTTAGGTACACCAATCAAACGTA CGTTTGGTCTTTTCACATAGACCCATATTTCTTGGAGG	1
chr4	103486234	103486311	CCTTCAAGAAATATGGGACTGTGTGAAAAGACCAAAGCTA GGTTTGATTGGTGTACCTGAAAGTGATGGGGAGAATGG	1
chr4	105923129	105923204	CCTACTATTCACAGAGTAATGCAGTTTGCTGAAAAGGTTG GTTTTTGCTGACCTCTGAGAGCTCACATTACAGTGG	4
chr4	106874711	106874788	CCATTCTCTCTGTCACTTTCTGGTACACCAATCAAACGTA GGTTTGCTCTTTTCACATAATCCCATATTTATTGAAGG	1
chr4	115805791	115805867	CCATAACATGTATTTGCTGGTGCTAGACTCTCCAAAGCTA GGTTTCTTTCTACAACAATGGCTGGAAGTCTTCTTGG	3
chr4	122033277	122033354	CCATTCTCCCCATCACTTTCAGGTACACCAATCAAACGTA GGTTTGGTCTTCTCACACAGTCCCATATTTCTTGGAGG	1
chr4	129125132	129125209	CCATTCTTCCCATTACTTTCAGGTACACCAATCAAACGTA GGTTTGGTCTTTTCACATAGTCCCACATTTCTTGGAGG	1
chr4	135472562	135472639	CCATTCTCCCCCTCACTTTCAGGTACACCAATCAAACGTA GGTTTGGTCTTTTCACATTGTCCCATATTTCTTGGAGG	1
chr4	138507099	138507176	CCATTCTCCCCAGCACTTACAGGTACACCAATCAAACGTA GGTTTGGTCATTTCACATAGTCCCATATTTCTTGGAGG	1
chr4	144249093	144249170	CCATTCTCCCTGTCACTTTCAGGTACAGCAATCAAACGTA GGTTTGGTCTTTTCACATGGTCCCATATTTCTTGGAGG	1
chr4	144436406	144436483	CCTCCAAGAAATATGAGACTATGTGAAAAGACCAAACCTA CGTTTGATTGGTGTACCTGAAAGTGACGGGGAAGATGG	1
chr4	154110259	154110336	CCTCCAAGAAATATGAGACTATGTGAAAAGACCAAACCTA CGTTTGATTGGTGTACCTGAAAGTGACAGGGAGAATGG	1
chr4	154893438	154893515	CCTCCAAGAGATATGAGACTATGTAAATAGACCAAACCTA CCTTTGATTGGTGTACGTGAAAGTGACAGGAAGAATGG	1
chr4	161116854	161116931	CCATTCTCCCCATCACTTTCAGGTACACCAACCAAACGTA GGTTTGGTCTTTTCACATAGTCTCATATTTCTTGGAGG	1
chr4	165140748	165140823	CCTCCATTGACTACTCCTTATCATTGGCTAGAAAACCTAC CTTTCAACCAGTTTCTAAGGCCAAGAAACTTGGAGG	4
chr4	181928508	181928585	CCACCAAGAAATATGGGACTACGTGAAAAGACCAAACCTA	1

			CGTTTGATGGGTGTGCCTGAAAGTGACGGGAAGAATGG	
chr4	187521958	187522035	CCTCCAAGAAATAAGGGACTATGTGAAAAGACCAAACCTA CGTTTGATTGGTGTACCTGAAGGTGACAGGGAGAATGG	1
chr5	12675639	12675715	CCAAAGGGCCTTTGTGATTCTACTTTGTAATATAAAGGAT GGTTTCTTACTACGGTTGGTGTCCTTGCAGGAGTGGG	3
chr5	29271804	29271881	CCTCCAAGAAATATGGGACTATGTGAAAAGACCAAACCTA CGTTTGATTGGTGTACCTGAAAGTGATGGGGAGAATGG	1
chr5	35352660	35352737	CCATTCTCCCCGTTACTTTCAGGTACACCAATAAAACCTA GGTTTGGTCTTTTCACATAGTCCCATATTTCTTGGAGG	1
chr5	38723235	38723310	CCCATATCTCTGGCAAGGGCAGCTCTCTGGCTAAACCAAG CTTTCCTGTAGAGCTTGAGTTCCAAGGCAGCGTTGG	4
chr5	47358339	47358415	CCTTGTAGTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGACTTGAAACACTCTTGTTGTGG	2
chr5	47415811	47415887	CCTTGTAGTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCATACTTGAAACACTCTTTTTGTGG	2
chr5	47474614	47474690	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGACTTGAAACACTCTTTTTGTGG	2
chr5	48228356	48228432	CCTTGTGTTGTGTTTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGACTTGAAATACTCTTTTTGTGG	2
chr5	48454551	48454627	CCTTGTAGTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCATACTTGAAACACTCTTTTTGTGG	2
chr5	48612272	48612348	CCTTGTATTGTGAGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGACTTGAAACACTCTTTTTGTGG	2
chr5	48884165	48884241	CCTTGTGTTGTGTGTCTTCAACTCACAGAGTTAAACGATG CTTTACACAGAGTAGACTTGAAACACTCTTTTTCTGG	2
chr5	49037011	49037087	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGACTTGTAACACTCTTTTTGTGG	2
chr5	49038372	49038448	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGACGTGAAACACTCTTTTTGTGG	2
chr5	49150718	49150794	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGACTTGAAACACTCTTTTTGTGG	2
chr5	49241653	49241729	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGGAGACTTGTAACACTCTTTTTGTGG	2
chr5	88582714	88582790	CCTTTTCATAAGAAGAAAATCGACTCATCATTGAAACCAA GCTTTGGTACAATTTCATTGATGTTTCCAGAAGCAGG	3
chr5	93497156	93497231	CCCATAGACTATGATAGAAACAAAATAACCCAAAAGCTAG CTTTCTGATTGAGTTTCCATAAATGCAATGTGAAGG	4
chr5	94295029	94295105	CCATTCACTTGTCACTTTCTGGTACACCAATCAAACGTAG GTTTGGTCTTTTCACATAGTCTCATATTTCTTGGAGG	2
chr5	94956746	94956823	CCTCCAAGAAATATGGGACTCTGTAAAGAGACCAAACCTA CGTTTGATTGGTGTACCTGAAAGTGAAGGGGAGAATGG	1
chr5	106003488	106003565	CCATTCTCCCCGTCATTTTCAGGTACACCAATCAAACCTA GGTTTGGTCTTTTTACATAGTCCCATATTTCTTGGAGG	1
chr5	118727905	118727982	CCTCCACGAAACATGGGACTATGTGAAAAGACCAAACCTA CGTTTGATTGGTGTACCTGAAAGTGACAGGGAGAATGG	1
chr5	132156032	132156109	CCAATTTCCCCCTCACTTTCAGATACACCAATCAAACGTA GGTTTGGTCTTTTCACATAGTTCCATATTTCCTGGAGG	1
chr5	152037951	152038028	CCATTCTCCCCATCACTTTCAGGTACACCAATCAAACGTA GGTTTGGTCTTTTCACATATTCCCATATGTCTTGGAGG	1
chr5	155183064	155183141	CCCACCGGCTCATGAGAGGTAGAGCTAAGGTCCAAACCTA GGTTTATCTGAGACCGGAACTCATGTGATTAACTGTGG	1
chr5	155183065	155183141	CCACCGGCTCATGAGAGGTAGAGCTAAGGTCCAAACCTAG GTTTATCTGAGACCGGAACTCATGTGATTAACTGTGG	2
chr5	163148211	163148288	CCTTCAAGAAATATGGGACTATGTGAAGAGACCAAACCTA CGTTTGATTGGTGTAGCCAAAAGTGATGGGGAAAATGG	1
chr5	165889537	165889614	CCTCAGATTAGATTTACTTGCAAAGAGACATTTAAAGGAT CGTTTTGATACTATTTTGAAAGTACTATACAAAGATGG	1
chr5	169395198	169395274	CCTTAAGAACATAAATCCCCAGGAATTCACAGAAACCTTG GTTTGAGCTTTGGATTTCCCGCAGGATGTGGGATAGG	2
chr5	171021380	171021457	CCATTCTCTCTGTCACTTTCAGGTACACCAATCAAACGTA GGTTTGGTCTTTTCTCATAGTCCCATATTTCTTGGAGG	1

chr5	173059898	173059973	CCATTTACCATCATTCTCTGTCATGGCAGGTGAAAGCAAG CTTTTATATAGACAATGTTCTACTTAGTTTACAGGG	4
chr5	174102359	174102435	CCCAAAGTTAATTTTACTCTTTTTCTGAATCAAAAGGAAC CTTTCCTCCATGAGAAGAATCCTGCCATATTTCTAGG	2
chr5	180927811	180927888	CCTCCAAGAAATATGGGACTATGTGAAAAGACCAAACCTA CGTTTGATTGCTATACATGAAAGTGACGGGGAGAATGG	1
chr6	1752363	1752440	CCTTCAAGAAATATGGGACTATGTGAAAAGACCAAACCTA CCTTTGATTGGTGTACCTGAAAGTGATGGGAAGAATGG	1
chr6	20595279	20595356	CCATTCTCCCCATCACTTTCAGGTACACCAATCAAACGTA GGTTTGGTCTTTTCACATAGTCCCATAGTTCTTGGAGG	1
chr6	23431370	23431447	CCATTCTCCCCGTCACTTTCAGGGACAACAATCAAACGTA GGTTTGGCCTTTGCACATAGTCTTATATTTCTTGGAGG	1
chr6	29190624	29190701	CCATTCTCCCCATCACTTTCAGGTACACCAATCAAACGTA GGTTTGGTCTTTTCACATAGTCCCATATTTCTTGGAGG	1
chr6	61533266	61533343	CCTCCAAAAAATATGGGACTATGTGAGAAGACCAAACCTA CGTTTTATTAGTGTACCTCAAAGTGACAGGGAGGATGG	1
chr6	101052764	101052841	CCATTCTCCCCATCACTTTCAGGTACACCAATGAAACGTA GGTTTGGCCTTTTCACATAGTTTCATATTTCTTGGAGG	1
chr6	117176355	117176432	CCTCCAAGAAATATGGGACTATGTGAAAAGACCAAACCTA CGTTTGATTGGTGTACCTGAAAGTGATGGGGAGAATGG	1
chr6	117747073	117747149	CCTACAAGAAATATGGAACTTGTAAAAAGACCAAACCTAC GTTTGATTGGTGTACCTGAAAGTGACGGGGAGAATGG	2
chr6	118422508	118422585	CCTCCAAGAAATATGGGACAATGTGAAAAGGCCAAAGCTA CGTTTGATTGGTGTACCTGAAAGTGACAGGGAGAATGG	1
chr6	122035019	122035096	CCTTTCAAACTTAGAGGTAAACAAAAGTCCTGAAAACCTA GGTTTGACCATAAGTTGGGACCATACGAGCATAGAAGG	1
chr6	134445210	134445287	CCAAAAATAAAAAAAAAATTGACTTATAAGTAAGAAAGGTT CGTTTTCTCACATTCAGAAAGAGAACCCACATGTTGGG	1
chr6	134445210	134445286	CCAAAAATAAAAAAAAAATTGACTTATAAGTAAGAAAGGTT CGTTTTCTCACATTCAGAAAGAGAACCCACATGTTGG	3
chr6	135154944	135155021	CCATTCTCCCCATCACTTTCAGGTACACCAATCAAACGTA GGTTTGGTCTTTTCACATAGTCCCATATTTCTTGGAGG	1
chr6	137889995	137890072	CCATTCTCCCCGTCACTTTCAGGTACACCAATCAAACGTT GGTTTAGTCTATTCACATAGTCCCATATTTCTTGGAGG	1
chr6	143993904	143993981	CCGAAAAGAATAAGACTATCAGCTGAAGTCTTAAAACGAT CCTTTGGCCCCCAGTACTCTATATGCAGGATAGAAAGG	1
chr6	152610473	152610549	CCTACAAAAATAGGGGACTATGTGATAAGACCAAACCTAC GTTTGATTGGTGTACCTGAAAGTGATGGGGAGAATGG	2
chr6	160372604	160372681	CCATTCTACCCATCACTTTCAGGTACACCAATCAAACGTA GGTTTGGCCTTTTCATATAGTCTCATATTTCTTGGAGG	1
chr6	169352478	169352555	CCATTCTCCCCATCACTTTCTGGTATACCAATCAAACGTA GGTTTGGTCTTTTCACATAGTCCCATATTTCTTAGAGG	1
chr6_GL000251v2_alt	677196	677273	CCATTCTCCCCATCACTTTCAGGTACACCAATCAAACGTA GGTTTGGTCTTTTCACATAGTCCCATATTTCTTGGAGG	1
chr6_GL000252v2_alt	456242	456319	CCATTCTCCCCATCACTTTCAGGTACACCAATCAAACGTA GGTTTGGTCTTTTCACATAGTCCCATATTTCTTGGAGG	1
chr6_GL000253v2_alt	456202	456279	CCATTCTCCCCATCACTTTCAGGTACACCAATCAAACGTA GGTTTGGTCTTTTCACATAGTCCCATATTTCTTGGAGG	1
chr6_GL000254v2_alt	456371	456448	CCATTCTCCCCATCACTTTCAGGTACACCAATCAAACGTA GGTTTGGTCTTTTCACATAGTCCCATATTTCTTGGAGG	1
chr6_GL000255v2_alt	456225	456302	CCATTCTCCCCATCACTTTCAGGTACACCAATCAAACGTA GGTTTGGTCTTTTCACATAGTCCCATATTTCTTGGAGG	1
chr6_GL000256v2_alt	500011	500088	CCATTCTCCCCATCACTTTCAGGTACACCAATCAAACGTA GGTTTGGTCTTTTCACATAGTCCCATATTTCTTGGAGG	1
chr7	5256551	5256627	CCACCACACCCAGCCTTATGGGATGGTTTTCAAAAGCATC CTTTTTTAGAAGTGGATTCTGATATATAATCGGATGG	2
chr7	7392583	7392660	CCATTCTCAATGTCACTTTCAGGTACACCAATCAAACGTA GGTTTGGTCTTTTCACATAGTCCCATATTTCTTGGAGG	1

chr7	8737741	8737818	CCATTCTCTCTGTCACTTTCAGGTACACCAGTCAAAGGTA GGTTTGTTTTATTCACACGTTCACATATTTCTTGGAGG	1
chr7	11352226	11352303	CCATTCGCCCCATCACTTTCAGGTACACTAGTAAAACGTA GGTTTGGTCTTTTCACATAGTTCCATATTTCTTGGAGG	1
chr7	15519145	15519222	CCTCCAAGAAATATGGGACTATGTGAAGAGATCAAACCTA GGTTTGATTGTTGTACCTGAAAGTGATAAGAAGAATGG	1
chr7	19228341	19228418	CCTCCAATAAATATGGGGCTATGTGAAAAGACCAAACCTA CGTTTGATTGGTGTACCTGAAAGTGACAGGGAGAATGG	1
chr7	23778445	23778522	CCCTTTTCCCTGTCACTTTCAGGTACACCAGTCAAACGTA GGTTTGGTCTTTTCACATAGTCGAATATTTCTTCAAGG	1
chr7	23778446	23778522	CCTTTTCCCTGTCACTTTCAGGTACACCAGTCAAACGTAG GTTTGGTCTTTTCACATAGTCGAATATTTCTTCAAGG	2
chr7	26769065	26769142	CCATTCTCCCTGTCACTTTCAGGTACACTAATCAAACGTA GGTTTGGTGTATTCACACAGTCCCATATTTCTTGGAGG	1
chr7	42864035	42864112	CCATTCTTCCTGTCACTTTCAGGTATACCAATCAAACGTA GGTTTGGTCTTTTCACATAGTCCCATGTTTCTTGGAGG	1
chr7	46498923	46499000	CCTCCAAGAAATATGAGACTATATGAAAATACCAAACCTA CGTTTGATTGGTGTACCTGAAAGAGACAGGGAGAATGG	1
chr7	51535360	51535437	CCATTCTCCCTATCACTTTCAGGTACACCAATCAAACGTA GGTTTGGTCTTTTCATGTAGTCCCATATTTCTTGGAGG	1
chr7	51927106	51927183	CCATTCTGCCCGTCACTTTCAGGTACACCAATCAAACGTA GGTTTGGTCTTTTCACATAGTCCCATATTTCTTGGAGG	1
chr7	56976942	56977018	CCGTCCGATTATATATCAGAATCTACTTCTAAAAAAGGAT GCTTTTGAAAACCATCCCATAAGGCTGGGTGTGGTGG	3
chr7	80021598	80021675	CCTACAAGGAATATAGGACTATGTGAAAATACCAAACCTA CGTTTCACTGCTGTACCTGAAGGTGACAGGGAGAATGG	1
chr7	89673853	89673930	CCATTCTCCCCATCATTTCCAGGTAAACCAATCAAAGGTA GGTTTGGTCATTTCACATAGTCCCATATTTCTTGGAGG	1
chr7	103404790	103404867	CCATTCTCCCCGTCACTTTCAGGTACACCAGTCAAACGTA GGTTTGGTCTTTTCACACAGTCCCATATTTCCTGGAGG	1
chr7	113053651	113053728	CCATTCTCCCCATCACTTTCAGGTACAGCAATCAAACGTA GGTTTGGTCTTTTCACATAGTCCCATATTTCTTGGAGG	1
chr7	125765204	125765279	CCACTACAGATTCTTGGGTCAAGATGTGTGCAAAAGGATG CTTTAGGGTGATGGATATGAGTGGGATGAAATGAGG	4
chr7	128042158	128042234	CCTGAAAAAAAAACCCTGCCAGCCAGCAACTCTGAAAGGAT GCTTTGTGTGAGTGAGCAGTGTCTGAGATGGACAGGG	3
chr7	130637332	130637409	CCATTCTCCCCATCACTTTCAGGTACGCCAATCAAACGTA GGTTTGGTCTTTTGACATAGTCCCATATTTCTTGGAGG	1
chr7	136983050	136983127	CCGTTCTCCCCATCACTTTTAGGTACACCAATCAAACGTA GGTTTGGTCTTTTCACATAGTCTCATATTTCTTGGAGG	1
chr7	143579507	143579584	CCATTCTCCTGGTCACTTTCAGGTATACCAATCAAACGTA GGTTTGGTCTTTTCATGTAGTCCCATATTTCTTGGAGG	1
chr7	143749881	143749958	CCTCCAAGAAATATGGGACTACATGAAAAGACCAAACCTA CGTTTGATTGGTATACCTGAAAGTGACCAGGAGAATGG	1
chr8	2338364	2338441	CCTCCAAGAACTATGGGACTATGTGAAAAGACCAAACCTA CGTTTGATTGGTGTACCTGAAAGTGACGGGGAGAATGG	1
chr8	2383289	2383366	CCATTCTCCCCGTCACTTTCAGGTACACCAATCAAACGTA GGTTTGGTCTTTTCACATAGTCCCATAGTTCTTGGAGG	1
chr8	8414568	8414645	CCATTCTCCCCGTCACTTTCAGGTACACCAATCAAACGTA GGTTTGGTCTTTTCACAGAGTCCCATATTTCTTGGAGG	1
chr8	24163142	24163219	CCATTCTCCCCGTCACTTTCATGTACACCAAGCAAACGTA GGTTTGATCTTTCCACATAGTCCCGTGTTTCTTGGAGG	1
chr8	34299051	34299128	CCTCCAAGAAATATGGGACTATGTGAAAAGACCAAACCTA CGTTTGATTGGTGTACTTGAAAGTGACAGGGAGAATGG	1
chr8	40965485	40965562	CCTCCAAGAAATATGGGACTATGTGAAAAGACAAAACCTA CGTTTCACTGGTGTACCTGAAAGTGACAGGGAGGATGG	1
chr8	48371659	48371735	CCCCCACCTTTTAAAAACATGCATACATACGGAAACGTTG CTTTCTGCACGATTTCATTTTAATGGAACAGAACAGG	2
chr8	82534960	82535037	CCATTTCCCCTGTCACTTTCAGGTACACCAATCAAACGTA GGTTTGGTCTTTTCACATAGTATCATATTTCTTGGAGG	1
chr8	109217624	109217700	CCATTCTCCCCGTCACTTTCAGGTACACCAATCAAACGTA	3

			GGTTTGGTCTTTTCACATAGTCCCATATTTCTGGAGG	
chr8	134790285	134790361	CCTTTTGTTAAAGTAATAGAATTCTGCTTCTTAAAGGAAC CTTTCAGGCAAGATGGTGGTTAGAGCACCTAAATGGG	2
chr8	134790285	134790360	CCTTTTGTTAAAGTAATAGAATTCTGCTTCTTAAAGGAAC CTTTCAGGCAAGATGGTGGTTAGAGCACCTAAATGG	4
chr8_KI270821v1_alt	519635	519712	CCTCCAAGAACTATGGGACTATGTGAAAAGACCAAACCTA CGTTTGATTGGTGTACCTGAAAGTGACGGGGAGAATGG	1
chr8_KI270821v1_alt	564557	564634	CCATTCTCCCCGTCACTTTCAGGTACACCAATCAAACGTA GGTTTGGCCTTTTCACATAGTCCCATAGTTCTTGGAGG	1
chr9	14951207	14951283	CCTCCAAGAAATATGGGACTGGTGAAAAGACCAAACCTAC GTTTGACTGGTGTACCTGAAAGTGACGGGGAGACTGG	2
chr9	23249218	23249295	CCTCCAAGAAACATGGGAATGTGTGAAAAGACCAAACCTA CGTTTGATTGGCGTACCTGAAAGTGACGGGGAGTATGG	1
chr9	26278896	26278973	CCTCCAAGAAATATGGGACTGTGTGAAAAGACCAAACCTA CGTTTGATTGGTATACCTGAAAGTGACAGAGAGAATGG	1
chr9	27323237	27323314	CCATTCTCCCCTTCACTATCAGGTACACCAATCAAACGTA GGTTTAGTCTTTTCACATAGTCCCATATTTCTTGGAGG	1
chr9	31517993	31518070	CCATTCTCCCCGTCACTTTCAGATACACCAGTCAAACGTA GGTTTGGTCTTTTCACATAGTCCCATATTTCTTGGAGG	1
chr9	39694860	39694937	CCATCTTACTTTGTACTACACTGTTCTTTAGAGAAAGCTT CCTTTTGGAGACCAACCAGGACTCCTTAGAAGCAGAGG	1
chr9	42451132	42451209	CCATCTTACTTTGTACTACACTGTTCTTTAGAGAAAGCTT CCTTTTGGAGACCAACCAGGACTCCTTAGAAGCAGAGG	1
chr9	60776573	60776650	CCTCTGCTTCTAAGGAGTCCTGGTTGGTCTCCAAAAGGAA GCTTTCTCTAAAGAACAGTGTAGTACAAAGTAAGATGG	1
chr9	62647482	62647559	CCTCTGCTTCTAAGGAGTCCTGGTTGGTCTCCAAAAGGAA GCTTTCTCTAAAGAACAGTGTAGTACAAAGTAAGATGG	1
chr9	66682030	66682107	CCTCTGCTTCTAAGGAGTCCTGGTTGGTCTCCAAAAGGAA GCTTTCTCTAAAGAACAGTGTAGTACAAAGTAAGATGG	1
chr9	82264427	82264503	CCACCACTGTGCCTGGCCATTTTCACTATTCTTAAAGGAA GCTTTGGTTTACAAAGGTTTGCTACTGTACTTCCAGG	3
chr9	84042684	84042761	CCATTCTCCCTGTCACTTTCAGGTACACCATTCAAACGTA GGTTTGGTCTTTTCTCATAGTCCCATATTTCTTGGAGG	1
chr9	95256012	95256089	CCTCCAAGAAATTCGGGACTATGTGAAAAGACAAAACCTA CGTTTAATTGGTGTGTGGTGTACCTGAAAGTGACAAGG	1
chr9	101816988	101817065	CCTCCAAGAAATATGGGACTATGTGAAAAGACCAAACCTA CGTTTGATTGGTGTACCTGAAAGTGACCAGAAGAATGG	1
chr9	135842327	135842403	CCTCCAAGAAATATGGGACTATGTGAAAAGCCCAAACCTA CGTTTGACTGATGTACCTAAAGTGACGGGGAGAATGG	3
chr9	136910865	136910940	CCCGCACTGTGAGCTTGGCCGAGTGCTGTCTGAAAGCATC CTTTCCCTTCACCTGGAGACTGGAGCGCCATAGAGG	4
chr10	13710312	13710389	CCTGTCTCCCCCATTCCATGCAAAATAAAACACAAACCAA GCTTTGCTTTAAGTGCTCCCTGATGCAGTTCAGCGTGG	1
chr10	18938129	18938206	CCATTCTTCCCGTCACATTCAGGTACACCAATCAAACGTA GGTTTGGTCTTTTCCCATAGTCCCATATTTCTTAGAGG	1
chr10	22712838	22712914	CCCCCTGCTCAGCTTGGGGAAGAAAAATACAAAAACGATG CTTTTAGGCATTTTAAACAACTTCACTACATTGAGGG	2
chr10	22712838	22712913	CCCCCTGCTCAGCTTGGGGAAGAAAAATACAAAAACGATG CTTTTAGGCATTTTAAACAACTTCACTACATTGAGG	4
chr10	40160932	40161009	CCTTTGTGTTGTGTGTATTCAACTCACAGAGTGAAACCTT CCTTTATTCAGAGCAGTTTTGAAACACTCTTTTTGTGG	1
chr10	40390136	40390213	CCTTTGTGTTGTGTGTATTCAACTCACAGAGTGAAACCTT CCTTTATTCAGAGCAGTTTTGAAAAACACTTTTTGTGG	1
chr10	40409152	40409229	CCTTTGTGTTGTGTGTATTCAACTCACAGAGTGAAACCTT CCTTTATTCAGAGCAGTTTTGAAAAACTCTTTTTGTGG	1
chr10	40433940	40434017	CCTTTGTGTTGTGTGTATTCAACTCACAGAGTGAAACCTT CCTTTATTCAGAGCAGTTTTGAAACACTCTTTTTGTGG	1
chr10	40588155	40588232	CCTTTGTGTTGTGTGTATTCAACTCACAGAGTGAAACCTT CCTTTATTCAGAGCAGTTTTGAAATACTCTTTTTGTGG	1

chr10	41146207	41146284	CCTTTGTGTTGTGTGTATTCAACTCACAGAGTGAAACCTT CCTTTATTCAGAGCAGTTTTGAAACACTCTTTTTGTGG	1
chr10	43835183	43835260	CCATTCTCCCTGTCACTTTCAAGTACACCAATCAAACCTA GGTTTGGTCTTTTCACATAGTTCCATATTTCTTGGAGG	1
chr10	54913222	54913299	CCCCTCCCATCACAGGCCCTGAGGTTTAAGAGAAAACCAT GGTTTTGTGGGCCAGGCCCATGACCCTTCTCCTCTGGG	1
chr10	54913222	54913298	CCCCTCCCATCACAGGCCCTGAGGTTTAAGAGAAAACCAT GGTTTTGTGGGCCAGGCCCATGACCCTTCTCCTCTGG	3
chr10	54913223	54913299	CCCTCCCATCACAGGCCCTGAGGTTTAAGAGAAAACCATG GTTTTGTGGGCCAGGCCCATGACCCTTCTCCTCTGGG	2
chr10	54913223	54913298	CCCTCCCATCACAGGCCCTGAGGTTTAAGAGAAAACCATG GTTTTGTGGGCCAGGCCCATGACCCTTCTCCTCTGG	4
chr10	58035951	58036028	CCATTCTCCCCATCACTTTCAGGTACACCAATCAAACGTA GGTTTCATCTTTTCACATAGTCCCACGGTTTTTGGAGG	1
chr10	58677525	58677602	CCTCCAAGATATATGGGACTATGTGAAAAGACCAAACCTA CGTTTGATTGGTGTACCTGAAATTGATGGGGAGAATGG	1
chr10	84021390	84021467	CCTCCAAGAAATATGGGACTGTGTGAAAAGAACAAACCTA CGTTTGATTGGTGTACGTGAAAGTGATGGGGAGAATGG	1
chr10	91442692	91442769	CCATTCCTCCCGTCACTTTCAGATACACCAAAAAAACGTA GGTTTGGTCTCTTCACATAGTCCCACATTTCTTGGAGG	1
chr10	91446848	91446925	CCTCCAAGAAATGTGGGACTATGTGAAGAGACCAAACCTA CGTTTTTTTGGTGTATCTGAAAGTGACGGGAGGAATGG	1
chr10	116928784	116928860	CCTCCAAGGGGAATCTGAGTTCTCTGAAGACAAAAAGCAT GGTTTCTTTTCTTCTGTATTTCTTATTGTTTCCTAGG	3
chr10	116937771	116937848	CCATTCTCCCTATCACTTTCCAGTACACCAATCAAACGTA GGTTTGGTCTTTTCACATAGTCCCATATTTCTTGGAGG	1
chr11	31182070	31182147	CCTCCAAGAAATATGGGACTATGTGAAAAGACCAAACCTA CGTTTGATTGGTATACTTGAAATTGACAAGGAGAATGG	1
chr11	34739273	34739350	CCTCCAAGAAATATGGGACTATGTGGAAAGACCAAACCTA CGTTTGACTGGTGTACCTGAAAGTGATGGGGAGAATGG	1
chr11	86646529	86646606	CCTCTAAGAAATATGGGACTATGTGAAGAGATGAAACCTA CGTTTGATTGGTGTACCTGAAAGTGACGAGGAGAATGG	1
chr11	90469791	90469867	CCCTCGTATACTACATGCTATAGTCAAAGCAGTAAACCTT CCTTTCCTTAAGCAGACCACACTCTTTCATGCCTGGG	3
chr11	90469792	90469867	CCTCGTATACTACATGCTATAGTCAAAGCAGTAAACCTTC CTTTCCTTAAGCAGACCACACTCTTTCATGCCTGGG	4
chr11	92429985	92430062	CCATTCTCCCCATCACTTTCAGGTATACTAATCAAAGGTA GGTTTGGTCTTTTCACATAGTCCCATATTTCATGGAGG	1
chr11	102818498	102818574	CCATTCCCCCGTCACTTTCAGGTACACCAATCAAACGTAG GTTTGGTCTTTTCACATAGTCCCATATTTCTTGGAGG	2
chr11	120765065	120765142	CCATTCTCCCCGTCACTTTCAGGTACACCAATCAAACGTA GGTTTTGTCTTTTCTTATAGTCCCATATTTCTTGGAGG	1
chr11	123131901	123131978	CCACTGCACCTGACCAAGATCCTTAATTTTTCTAAACCTA CGTTTATCATCTATAAAATGAGCCATCTTTTCACATGG	1
chr11	129468520	129468597	CCTCCGAGAAATATGGGACTATGTGAAAAGACCAAACCTA CGTTTGATTGTTGTACCTGAAAGTGACAGGGAGAATGG	1
chr11	131272361	131272438	CCATTCTCCCCATCACTTTTAGGTACACCAATCAAACGTA GGTTTGGTCCTTTTGCATAGACCCATATTTCTTGGAGG	1
chr11	132761415	132761492	CCATTTTCCCCGTCAGTTTCATATACACCTATCAAACGTA GGTTTACTGTTTTCACATAGTCCCTTATTTCTTGGAGG	1
chr12	22367416	22367493	CCTCCAAGAAATATGGGACTATGTGAAAAGACCAAACCTA CCTTTGATTGGTGTACCTGAAAGTGACGGGCAGGATGG	1
chr12	33146384	33146461	CCATTCTTCTCGTCATTTTCAAGTACACCAATCAAACGTA GGTTTGGTCTTTTCGCATAGTCCCATATTTCTTGGAGG	1
chr12	33198476	33198553	CCATTCTTCTCGTCACTTTCAAGTACACCAATCAAACGTA GGTTTGGTCTTTTCACATAGTCCCATATTTCTTGGAGG	1
chr12	46038332	46038409	CCTCCAAGAAATATAGGACTATGTGAAAAGACCAAACCTA CGTTTGATTGGTGTACTTGAAAGTGACAGGGAGAATGG	1
chr12	60236126	60236203	CCTCCAAGAAATGTGGAACTATGTGAAAAGACCAAACCTA CGTTTGATTGGTGTACCTGAAAGTGACAGGGAGAATGG	1
chr12	62098359	62098434	CCCTGACACTGATAAACGGATATGAAGAGAAAAAAGCTAG	4

			GTTTTCGCTGGAATTCCTAAGCTTGGGCTGCAGTGG	
chr12	62112591	62112668	CCCTTCTCCCAGTCACTTTTAGGTACACCAATGAAACGTA GGTTTGGTCTTTTCACACAGTCCCATATTTCTTGGAGG	1
chr12	62112592	62112668	CCTTCTCCCAGTCACTTTTAGGTACACCAATGAAACGTAG GTTTGGTCTTTTCACACAGTCCCATATTTCTTGGAGG	2
chr12	62418577	62418652	CCACTCCCTCTCCCCCAAAAAGTAAAGGTAGAAAACCAAG GTTTACAGGCAACAAATAGCACAATGAATGGAATGG	4
chr12	71732311	71732388	CCAAACCCGCATCGCACACCCTGTGAGGGGGACAAAGGAA CCTTTCCGTTCCAACATCAAGGTTGTTTTGACCCAAGG	1
chr12	78047816	78047893	CCATTCTTTCTGTCACTTTCAGGTATACCAGTCAAACCTA GGTTTGGTCTTTTCACATAGTCCCATATTTCTTGGAGG	1
chr12	81480016	81480093	CCATTCTCCCCATCACTTTCAGGTACACCAATCAAACGTA GGTTTGGTCTTTTCACATAGTCCCATATTTCTTGGAGG	1
chr12	96840231	96840307	CCACACGGTAGAGGATAAACTAGGTGGATTCTCAAAGCAA CCTTTGAAATAATCTATGCAGTTTTTCTGGGTACTGG	3
chr12	99187165	99187242	CCACCAAGAAACATGGGACTATGTGAAAAGACCAAACCTA CGTTTGGTTGGTGTACCTGGAAGTGACGGGGAGAGTGG	1
chr12	107860841	107860918	CCTCCAAGAAATATGGGACCATGTGAAAAGACCAAACCTA CGTTTGATTGGTGTACCTGAAAGTGACAGGGAGAATGG	1
chr12	110882809	110882885	CCTGTAAAAAGGTCACATGGTCAGGTGTGCCTAAACGATC CTTTTATTTATTTATTTATTTATTTTTAAGAAACAGG	2
chr12	119063321	119063397	CCAGCCCCAAAATGTCAGGGGCTTAGAACAACAAAGGTTC CTTTTCATGTTTATACTACATGTTTGTCATGGGCTGG	2
chr13	35320704	35320781	CCGTTTTCCCCATCACTTTCAGGTACACCAGTCAAACGTA GGTTTGGTCTTTTCACATGGTCCCACATTTCTTGGAGG	1
chr13	53133477	53133554	CCTGGAATAGCTTTCCTGACTGTCTGACTTCAAAAACCTT GGTTTGACCACTTCGTCTATATCATGAGGAAGGACTGG	1
chr13	53184880	53184956	CCCTACTCTGAACCTACCTTGATAAAGCCTAGAAAACCAA GCTTTGACAAGATTTGACAAGAGATGGAATTTGGAGG	3
chr13	53184881	53184956	CCTACTCTGAACCTACCTTGATAAAGCCTAGAAAACCAAG CTTTGACAAGATTTGACAAGAGATGGAATTTGGAGG	4
chr13	57896962	57897038	CCCTTATAAAACTGAAAACTTTAACCTTTTTTAAAGCATG CTTTTGAATAAATTCTTTTATTACAAAAAAGACCAGG	2
chr13	62610100	62610177	CCATTCTCCCTGTCACTTTCAGGTACACCAATCAAACGTA GGTTTGGTCTTTTCACGTAGTCCCATATTTCTTGGAGG	1
chr13	77004382	77004458	CCCTTTATTATCCAAGTGGTTTCCTGCTCTTCAAACCTTC CTTTCAAAATTTTGTCTCCTACTTAAAACAAGTTAGG	2
chr13	81646075	81646151	CCTTCTGTTGAGACCTACTGCTAAGAAAACAAAAAAGGTT CCTTTCAAATATTATTGTGAATCAATAATGTACCTGG	3
chr13	83755854	83755931	CCTCCAAGAAATATGGGACTATGTGAAAAGACCAAACCTA CGTTTCATTGATGGACCTGAAAGTGATGGGGAGAATGG	1
chr13	89719199	89719275	CCATTCTCCCTTCACTTTCAGTTACACCAATCAAACGTAG GTTTGGTCTTTTCACATAGTCCCATATTTCTTGGAGG	2
chr13	102010574	102010650	CCTAGGGAAGTGATCATAGCTGAGTTTCTGGAAAAACCTA GGTTTTAAAGTTGAGGAGACTTAAGTCCAAAACCTGG	3
chr13_KI270841v1_alt	124240	124316	CCATTCTCCCTTCACTTTCAGTTACACCAATCAAACGTAG GTTTGGTCTTTTCACATAGTCCCATATTTCTTGGAGG	2
chr14	25980646	25980723	CCTCCAAGAAATATGGGACTATGTGAAAAGACTAAACCTA CGTTTGATTGGTGTACCTGAAAGTGACAGGGAGAATGG	1
chr14	35842786	35842863	CCATTCTCCCTGTCACTTTCAGGTATGCCAGTCAAACGTA GGTTTGGTCTTTTCACATAGTCCCATATTCCTTGGAGG	1
chr14	42646400	42646477	CCTCCAAGAAATATGGGACTATGTAAAAAGACGAAACCTA CGTTTGATTGGTGTACTTAAAAGTGACGAGGAGAATGG	1
chr14	49063242	49063319	CCTCCAAGAAATATGGGACTATGTGAAAAGACCAAACCTA CGTTTGATTTGTGTACCTGAAAGTGATGGGGAGAATGG	1
chr14	49130379	49130456	CCATTCTCCCCGTCACTTTCAGGCACACCAATCAAACGTA GGTTTAGTCTTTTCACATAGTCCCATATTTCTTAGAGG	1
chr14	51352342	51352418	CCTTAATGCATTCATATTTCATATTTTAAATAAAACCATG GTTTCCCACAGAGTGACTTCTACTCTAAGAAATGGGG	2

chr14	51352342	51352417	CCTTAATGCATTCATATTTCATATTTTAAATAAAACCATG GTTTCCCACAGAGTGACTTCTACTCTAAGAAATGGG	4
chr14	60835842	60835919	CCGTTCTTTCCGTCACTTTCAGGTACACCAGTCAAACGTA GGTTTGGTCTTTTCACATAGTCCCATATTTCTTGGAGG	1
chr14	66529072	66529148	CCATTCTCCCCATCACTTTCATGTACACCAATCAAACGTA GGTTTGGTCTTTGTTAACATAGTCCCATATTTCTTGG	3
chr14	79210873	79210949	CCCTATAAAGCTTAGAGAAACACAGGGCTCTTTAAACGAT CCTTTTTCTCTTTTCTGTTTTAAATTTCATCACTTGG	3
chr14	79210874	79210949	CCTATAAAGCTTAGAGAAACACAGGGCTCTTTAAACGATC CTTTTTCTCTTTTCTGTTTTAAATTTCATCACTTGG	4
chr14	85371541	85371618	CCATTCTCCCCATCACTTTCAGGTACACTAATCAAAGGTA GGTTTGGTCTTTTCACATGGTCCTATATTTCTTGGAGG	1
chr14	92918713	92918790	CCCCATAGCACGATCACATGGGACATTCAGGGGAAAGCAA CCTTTTCCAGGAAGGAAAACCCAATGCTGGGACCCAGG	1
chr14	92918714	92918790	CCCATAGCACGATCACATGGGACATTCAGGGGAAAGCAAC CTTTTCCAGGAAGGAAAACCCAATGCTGGGACCCAGG	2
chr14	103386821	103386897	CCCTTTCAGCGCTCACAGGCTATGGTTTTATAAAAGGAAC CTTTGATTTTGTTCATGTGAAACTACAAAATGCCAGG	2
chr14_KI270847v1_alt	33275	33352	CCCCATAGCACGATCACATGGGACATTCAGGGGAAAGCAA CCTTTTCCAGGAAGGAAAACCCAATGCTGGGACCCAGG	1
chr14_KI270847v1_alt	33276	33352	CCCATAGCACGATCACATGGGACATTCAGGGGAAAGCAAC CTTTTCCAGGAAGGAAAACCCAATGCTGGGACCCAGG	2
chr15	20630566	20630643	CCTCCAAGAAATATTGGAGTATGTGATAAGACCAAACCTT CGTTTGACTGGTGTACCTGAAAGTGATGGGGAGAATGG	1
chr15	21675103	21675180	CCATTCTCCCCGTCACTTTCAGGTACACCAATCAAACGTA GGTTTGGTCTTTTCACATAGTCCCATATTTCTTGGAGG	1
chr15	22117571	22117648	CCATTCTCCCCGTCACTTTCAGGTACACCAATCAAACGTA GGTTTGGTCTTTTCACATAGTCCCATATTTCTTGGAGG	1
chr15	22369744	22369821	CCATTCTCCCCATCACTTTCAGGTACACCAGTCAAACGAA GGTTTGGTCTTATCACATACTCCAATATTTCTTGGAGG	1
chr15	42302832	42302909	CCTCCAAGATATATGGGACTATGTGAAAAGGCCAAACCTA CCTTTGATTGATACACCTGAAAATGACAGGGAGAATGG	1
chr15	49967601	49967678	CCTCCAAGAAATATGCGACTATGTGAAAAGACCAAACCTA CGTTTCATTGGTGTACCTGAAAGTGATGGGGAGAATGG	1
chr15	83964501	83964577	CCTCCAAGAAATATGGGACTATGTGGAAAGACCAAACCTA CGTTTGTTTGGTGTACCTGAAAGTGAGGGGAGAATGG	3
chr15	87261388	87261465	CCATTCTCCTCATCACTTTCAAGTACACCAATCAAACGTA GGTTTGGTCTTTTCACATAGTCTTATATTTCTTGGAGG	1
chr15_KI270727v1_random	409348	409425	CCATTCTCCCCGTCACTTTCAGGTACACCAATCAAACGTA GGTTTGGTCTTTTCACATAGTCCCATATTTCTTGGAGG	1
chr15_KI270851v1_alt	14235	14312	CCATTCTCCCCATCACTTTCAGGTACACCAGTCAAACGAA GGTTTGGTCTTATCACATACTCCAATATTTCTTGGAGG	1
chr15_KI270852v1_alt	440099	440176	CCATTCTCCCCGTCACTTTCAGGTACACCAATCAAACGTA GGTTTGGTCTTTTCACATAGTCCCATATTTCTTGGAGG	1
chr16	22123671	22123748	CCAGCAGAAGAATCTGGGGCACAGTCTGTGAAAAAAGGTA CCTTTCTTAAGCAGGGTTCTTATCCTTCATGGGTCTGG	1
chr16	25557623	25557700	CCTCCAAGAAATATGGGACTATGTGAAAAGACCAAACCTA CGTTTGATTGTTGTACCTGAAAGTGAGGGGGAGAATGG	1
chr16	36427179	36427255	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	36476450	36476526	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2

chr16	36512469	36512545	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	36520964	36521040	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACACAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	36524704	36524780	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	36566812	36566888	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	36573603	36573679	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	36667694	36667770	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	36677320	36677396	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	36683096	36683172	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	36691251	36691327	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	36710951	36711027	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	36750364	36750440	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	36791455	36791531	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACACAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	36856683	36856759	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	36926655	36926731	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	36931752	36931828	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	36948058	36948134	CCTTGTGTTGTGTGTATTCAACTCACCGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	36974541	36974617	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	36981331	36981407	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	36990839	36990915	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	37021075	37021151	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	37042812	37042888	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	37085971	37086047	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	37129462	37129538	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	37146110	37146186	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACACAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	37157309	37157385	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	37183118	37183194	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	37190924	37191000	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	37221808	37221884	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	37259501	37259577	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	37272409	37272485	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	37281923	37281999	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC	2

			CTTTACACAGAGCAGATTTGTAACACTGTTTTTCTGG	
chr16	37346472	37346548	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	37357000	37357076	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	37373301	37373377	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	37419498	37419574	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	37430714	37430790	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	37455845	37455921	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	37458558	37458634	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	37486127	37486203	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	37525183	37525259	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTGTGG	2
chr16	37536735	37536811	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	37554730	37554806	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	37575784	37575860	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	37577483	37577559	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	37583598	37583674	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	37696368	37696444	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTCCACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	37704524	37704600	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	37706223	37706299	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	37708941	37709017	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	37763622	37763698	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	37772115	37772191	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	37791815	37791891	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	37796229	37796305	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	37797928	37798004	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	37843453	37843529	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	37848548	37848624	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	37864846	37864922	CCTTGTGTTGTGTGTATTCAACTCACCGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	37902550	37902626	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	37907307	37907383	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	37928033	37928109	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	37959262	37959338	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2

chr16	37964355	37964431	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	37974881	37974957	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAAAACTGTTTTTCTGG	2
chr16	37987789	37987865	CCTTGTGTTGTGTGTATTCAACTCACAGAGTAAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	37994586	37994662	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTGTGG	2
chr16	38006479	38006555	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	38011567	38011643	CCTTGTGTTGTGTGTATTTAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	38040096	38040172	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	38041456	38041532	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	38062179	38062255	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	38102937	38103013	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	38128412	38128488	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	38131809	38131885	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	38144723	38144799	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	38168845	38168921	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	38209287	38209363	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	38210986	38211062	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	38229667	38229743	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGATTTGAAACACTGTTTTTCTGG	2
chr16	47424037	47424114	CCATTCTCCCTATCACTTTCAGGTACACCAATCAAACGTA GGTTTGGTCTTTTCACATAGTCCCATATTTCTTGGAGG	1
chr16	60730549	60730625	CCTCGTCACTGCCAGATTTTGTGGCTACCAGCAAAGGATC GTTTTAAGCTGCAACTCAGGAAATTGAGAAAATATGG	2
chr16	72545014	72545091	CCTCCAAGAAATATGGGACTATGTGAAAAAACCAAACCTA CGTTTGATTGGTGTACCTGAAAGTGACAGGGAGAATGG	1
chr16	81945503	81945579	CCCTGTGTTCTTTTATACTAAAACAAGCCAGCAAACCAAC CTTTGAGATGTGTTGCCTTAAACATTACTGAATGGGG	2
chr16	81945503	81945578	CCCTGTGTTCTTTTATACTAAAACAAGCCAGCAAACCAAC CTTTGAGATGTGTTGCCTTAAACATTACTGAATGGG	4
chr17	16474024	16474100	CCGAGAAACGGCTTTAGCAACAAATAAATATCAAAAGGAT GCTTTCTCTTCAGAATAATCTAAAGTAAGTTGGGAGG	3
chr17	34438512	34438589	CCATGTTACTCCGGATAAGGACAGCAAAGGAGGAAAGGAA CCTTTTCTGGGCCACCAGAAGGATGAGCTTGGGCTTGG	1
chr17	43690782	43690859	CCCAGGGATATGCTGGCCACGGGGAGGAGCCGGAAACCAA CCTTTGTGTCACTGTGTAGTGACAAGTGCCTTTGGAGG	1
chr17	43690783	43690859	CCAGGGATATGCTGGCCACGGGGAGGAGCCGGAAACCAAC CTTTGTGTCACTGTGTAGTGACAAGTGCCTTTGGAGG	2
chr17	69156298	69156375	CCTTAGGGACCCATAATGGCCACAACCAGGAGAAAAGCAA GCTTTGATGCTTAAACACTACTTACAGACATGTACAGG	1
chr17	74595228	74595305	CCTGCCTCTGTTCCTCCTTCCTGATGGTGGCGGAAAGGAT GCTTTTGCCAGATCAACAGTCACACACAACACACCAGG	1
chr17	83191644	83191721	CCTGACTCCAGCCCTCCTTGACAAGGTCTCCGTAAAGCAT GCTTTCTCTTAGGGACCCTCAGAGGGAGGCTTGGTGGG	1
chr17	83191644	83191720	CCTGACTCCAGCCCTCCTTGACAAGGTCTCCGTAAAGCAT GCTTTCTCTTAGGGACCCTCAGAGGGAGGCTTGGTGG	3
chr18	35135224	35135300	CCTTATTTGGAATGTGACAAGACCCATTTGTTTAAACCTT	3

			GGTTTTTATGCAGAAAGAAAAGGAAGGCTGCAGTGGG	
chr18	38918861	38918938	CCATTCTCCCTGTCACTTTCAGGTACACTAATCAAACGTA GGTTTGCTGTTTTTACATAGGCTCATATTTCTTGGAGG	1
chr18	45476589	45476666	CCATTCTCCCCATCACTTTCAGGTACACCAGTCAAACGTA GGTTTGGTCTTTTCACATAGTCCCATATTTCTTGGAGG	1
chr18	48640821	48640896	CCTGTTTGTTATTTTAGCTAATGTCAAAAAGAAAACCTTG CTTTTTCTGAACCCTTTCAGAGGCAGAAAGTGGGGG	4
chr18	71096732	71096808	CCATTTTCCCCACCACTTTCACGTACAGCAATCAAACGTA GGTTTGGTCTTTTCACTAGTCCCATATTTCTTGGAGG	3
chr19	24957844	24957920	CCTTGTAGTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGACTTGAAACACTCTTGTTGTGG	2
chr19	25015316	25015392	CCTTGTAGTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCATACTTGAAACACTCTTTTTGTGG	2
chr19	25074119	25074195	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGACTTGAAACACTCTTTTTGTGG	2
chr19	25827861	25827937	CCTTGTGTTGTGTTTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGACTTGAAATACTCTTTTTGTGG	2
chr19	26054056	26054132	CCTTGTAGTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCATACTTGAAACACTCTTTTTGTGG	2
chr19	26211777	26211853	CCTTGTATTGTGAGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGACTTGAAACACTCTTTTTGTGG	2
chr19	26483670	26483746	CCTTGTGTTGTGTGTCTTCAACTCACAGAGTTAAACGATG CTTTACACAGAGTAGACTTGAAACACTCTTTTTCTGG	2
chr19	26636516	26636592	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGACTTGTAACACTCTTTTTGTGG	2
chr19	26637877	26637953	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGACGTGAAACACTCTTTTTGTGG	2
chr19	26750223	26750299	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGCAGACTTGAAACACTCTTTTTGTGG	2
chr19	26841158	26841234	CCTTGTGTTGTGTGTATTCAACTCACAGAGTTAAACGATC CTTTACACAGAGGAGACTTGTAACACTCTTTTTGTGG	2
chr19	28517220	28517297	CCAGGAAAAAATTTAAACTTTCTTAACTTGATAAAAGGTA GCTTTCAAAACCTACAATAAATAACATACTTAGAGTGG	1
chr19	34566821	34566898	CCATTCTCCTCGTCACTTTCAGGTACACCAAACAAACGTA GGTTTGGTCTTTTTACGTAGTCCCATATTTCTTGGAGG	1
chr19	52261770	52261847	CCCTCTTGAAGTTAGGGAAGTAGCATTTAAGGGAAACGTA GCTTTACTATTAAGAATTTCAAACAGCACTTGTCAGGG	1
chr19	52261770	52261846	CCCTCTTGAAGTTAGGGAAGTAGCATTTAAGGGAAACGTA GCTTTACTATTAAGAATTTCAAACAGCACTTGTCAGG	3
chr19	52261771	52261847	CCTCTTGAAGTTAGGGAAGTAGCATTTAAGGGAAACGTAG CTTTACTATTAAGAATTTCAAACAGCACTTGTCAGGG	2
chr19	52261771	52261846	CCTCTTGAAGTTAGGGAAGTAGCATTTAAGGGAAACGTAG CTTTACTATTAAGAATTTCAAACAGCACTTGTCAGG	4
chr20	11151392	11151469	CCATTCTCCCCGTCACTTTCAGGTACACCAATCAAACGTA GGTTTGGTCTTTTCACATATTCCCATATTTCTTGGAGG	1
chr20	14027067	14027143	CCATTCTCCCTTCACTTTCAGGTACACCAATCAAACGTAG GTTTGGTCTTTTCACATAGTCCCATATTTTTTGGAGG	2
chr20	50615399	50615476	CCTATAGTCTCAGTTACTTGGGAGGCTGAGGTAAAAGGAT CGTTTGAGCCCAGGAGGTGGAGGTTGCAGTGAGCCGGG	1
chr20	50615399	50615475	CCTATAGTCTCAGTTACTTGGGAGGCTGAGGTAAAAGGAT CGTTTGAGCCCAGGAGGTGGAGGTTGCAGTGAGCCGG	3
chr20	60909414	60909490	CCTTTCCCAACTCTGCTATTGCCCCCACATCCTAAAGGAA CCTTTCTTTTTTTATATATTTTATTTTAAGTTCCAGG	3
chr21	16226086	16226163	CCTCCAAGAAATATGGAACTATGTGAAAAGACCAAACCTA CGTTTGATTGACGTACCTGAAAGTGACAGGGAGAATGG	1
chr21	17835234	17835309	CCTCTTCTGAAAGCATTGATAATCAACATTTTAAACGTAG CTTTTCCCCATATTGCTAGGAAGGCTCATTCCCGGG	4
chr21	19425636	19425713	CCTCCAAGAAATATGGGACTATGTGAAAAGGCCAAACCTA CGTTTGATTGCTGTACCCGAGAGTGACGGGGAGAATGG	1
chr21	32220958	32221035	CCTCCAAGAAATATGGGACTATGTGAAAAGACCAAACCTA CGTTTGATTGGTGTACCTGAAAGTGATGGGGAGAATGG	1

chr21	34335877	34335953	CCCGGGGCCTGGGTGCCCAGTGCCAGTGGTCAGAAAGGTT GCTTTGGTGTTTTTCATTGTTAGTGAGACAGAGATGG	3
chr21	34335878	34335953	CCGGGGCCTGGGTGCCCAGTGCCAGTGGTCAGAAAGGTTG CTTTGGTGTTTTTCATTGTTAGTGAGACAGAGATGG	4
chr21	36315276	36315353	CCATTCTCCCCATCATTTTCAGGTACACCAATCAAACGTA GGTTTGATCTTTTCACATAGCCCCATATTTCTTGGAGG	1
chr21	41547952	41548028	CCACCAGCACTTCTGTTAGAAGTTGCAGCAGAGAAAGGAT CCTTTAGGCACATCTCCCAGATCCTTGCGAAGAGGGG	3
chr22	18973194	18973271	CCTGTGCCAGGGTCCTTCCACTGGGACTGGCAGAAACGTA GGTTTGCATGGAGTGAGAAGCAGGGGAGAGGTTGAGGG	1
chr22	18973194	18973270	CCTGTGCCAGGGTCCTTCCACTGGGACTGGCAGAAACGTA GGTTTGCATGGAGTGAGAAGCAGGGGAGAGGTTGAGG	3
chr22	20265462	20265539	CCCTCAGCCTCTCCCCTGCTTCTCACTCCATGCAAACCTA CGTTTCTGCCAGTCCCAGCAGAAGGACCCTGGCACGGG	1
chr22	20265462	20265538	CCCTCAGCCTCTCCCCTGCTTCTCACTCCATGCAAACCTA CGTTTCTGCCAGTCCCAGCAGAAGGACCCTGGCACGG	3
chr22	20265463	20265539	CCTCAGCCTCTCCCCTGCTTCTCACTCCATGCAAACCTAC GTTTCTGCCAGTCCCAGCAGAAGGACCCTGGCACGGG	2
chr22	20265463	20265538	CCTCAGCCTCTCCCCTGCTTCTCACTCCATGCAAACCTAC GTTTCTGCCAGTCCCAGCAGAAGGACCCTGGCACGG	4
chrX	27300998	27301075	CCTCCAAGAAATATGGGGCTATGTGAAAAGACCAAACCTA CCTTTGATTGGTGTATCTGAAAGTGACGGGGAGAATGG	1
chrX	28456666	28456743	CCTCCAAGAAATATGGGACTATGTGAAAAGACCAAACCTA CGTTTGATTTGTGTACCTGAAAGTGATGGGGAGAATGG	1
chrX	35634985	35635062	CCATTCTCCCCGTCACTTTCAGGTACACCAATCAAACGTA GGTTTGGTCTTTTCTCATTGTCCCATATTTCTTGGAGG	1
chrX	39460148	39460223	CCCATCAAGAGCGGTTGTGCATGGCAACAGTAAAAGGATG GTTTGTTACACTAGTACAAAAAGAGGTGGCCAGAGG	4
chrX	43926403	43926480	CCATTCTCTCTGTCACTTTCAGGTACACCAATCAAACGTA GGTTTGGTCTTTTCACATAGTCCCATATTTCTTGGAGG	1
chrX	44254600	44254677	CCTCCAAGAAATACGGGACTATGTGAAAAGACCAAACGTA CGTTTGATTGGTGTACCTGAAAGTGATAGGGAGAATGG	1
chrX	46088602	46088679	CCTCCAAGAAATATGGGACTATGTGAAAAGACCAAACCTA CGTTTGATTGGTGTACCTGAAAGTGACTGGGAGAATGG	1
chrX	50222874	50222951	CCATTCTCCCTGTCACTTTCAGGTACACGAATCAAACGTA GGTTTCATCTTTTCACATAGTCCCATATTTCTTAGAGG	1
chrX	57416835	57416911	CCATTCTCTCTGTCACTTTCTGGTACACCAATCAAACGTA GGTTTGGTCTTTTCACATAGTTTCACATATTTCTTGG	3
chrX	57856466	57856543	CCTCCAAGAAATATGGGACTATGTGAAAAGACCAAACCTA CGTTTGATTGGTGTACCTGAAAGTGACAAGGAAAATGG	1
chrX	62702479	62702556	CCTGAAAAACATTGTTTCCAACCTGGTAAATCAAAAGGAA GGTTTAACTTTGTTAGATAAGTCCACATATCACCAAGG	1
chrX	63067129	63067206	CCTCCAAGAAATGTGGGACTATGGGAAAAGACCAAACCTA CCTTTGTTTGGTGTACCTGAAAGTGACGGGGAGAAAGG	1
chrX	64936250	64936327	CCTCCAAGAAATATGGGACTATGTGAAAAGACCAAACCTA CGTTTCATTGGTGTACCTGAAAGTGATGGGTAGAATGG	1
chrX	66720099	66720176	CCTACAAGAAATATGGGACTATGGGAAAAGACCAAACCTA CGTTTGATTGGTACACTGGAAAGTGACAGGGATAATGG	1
chrX	68529086	68529163	CCATTCTCCCTGTCACTTTCTGGTACACCAATCAAAGGTA GGTTTGGTCTTTTCACATAGTCCCATATTTCTTGGAGG	1
chrX	73893994	73894071	CCTCCAAGAAATATGGGACTATGTGAAAAGACCAAACCTA CGTTTGATTGGTGTACCTGAAAGTGATGGGGAGAATGG	1
chrX	75723201	75723278	CCATTCTCTTTGTCACTTTCAGGTATACCAATCAAACGTT GGTTTGGTCTTTTTGCATAGTCCCATATTTTGTGGAGG	1
chrX	75815659	75815736	CCTCCAAGAAATATGAGACTATGTGAAAAGACCAAACCTA CGTTTGATTAGTGTACCTGAAAATGATGGGGAGAATGG	1
chrX	80967103	80967180	CCATTCTTTCTGTCACTTTCAGGTACACCAATCAAACGTA GGTTTGGTCTTTTCACATAGTCCCATATTTCTTGGAGG	1
chrX	89936425	89936502	CCATTCTCCCTGTCACTTTCAGGTACACCAATCAAACGTA GGTTTGTTCTTTTCACATAGTCCCATATTTCTTGGAGG	1
chrX	91038768	91038845	CCATTATCCCCATCACTTTCAGGTACACCAATCAAACGTA	1

			GGTTTGGTTTTTTCACATAGTTCAATATTTCTTTGAGG	
chrX	91471271	91471348	CCTCCAAGAAATATGGGACTATCTGAAAAGATCAAACCTA CGTTTGATTGGTGTACCTGAAAGTGACAGGGAGAATGG	1
chrX	96428180	96428257	CCTTTCTCCCCATCACTTTCAGGTACACCAATCAAACGTA GGTTTGGTCTTTTCATATAGTCCCATATTTCTTGGAGG	1
chrX	100268291	100268368	CCTCCAAGAAATATGGGACTATGTGCAAAGATCAAACCTA CGTTTGATTGCTGTACCTGAAAGTGATGGGGAGAATGG	1
chrX	105811046	105811123	CCATTCTCCCCATCACTTTCAGGTACACCAGTCAAACGTA GGTTTGGTCTTTTCACATAATCCCATATTTCTTGGAGG	1
chrX	115673065	115673141	CCTCCAAGAAGTATGGGACCATGGAAAAGATCAAACCTAC GTTTGACTGGTGTACCTGAAAGTGACTGGGAGAATGG	2
chrX	117269846	117269923	CCTCCAAGAAATATGGGACTATGTGAAAAGACCAAACCTA CGTTTGATTGGAGTACTTGAAAATGACAGGGATAATGG	1
chrX	139191369	139191445	CCTTTAAAGACATGCTCTTTGTGCCAGAAATTCAAAGGTT GCTTTTATGTCCAGTGGGGTGGAGGGAGGAAGCTCGG	3
chrX	147988614	147988691	CCATTCTCCCCGTCACTTTCAGGGACCTCAATCAAACGTA GGTTTTGTCTTTTCACATAGTCCCATATTTCTTGGAGG	1
chrX	155321041	155321118	CCTCCAAGAAATATAGGACTATGTGAAAAGACCAAACCTA CGTTTGACTGGTGTACCTGAAAGTGACAGGGAGAATGG	1
chrY	15109391	15109468	CCATTCTCCCCATCACTTTCAGGTACACCAATCAAAGGTA GGTTTGGTCTTTTCACATAGTCCGATATTTCCTGCAGG	1

Chromosomal sites were identified by searching for $CCN_{(30\text{-}31)}$-AAASSWWSSTTT-$N_{(30\text{-}31)}$-GG where W is T or A and S is G or C. Pattern 1 is $CCN_{(31)}$-AAASSWWSSTTT-$N_{(31)}$-GG, 2 is $CCN_{(30)}$-AAASSWWSSTTT-$N_{(31)}$-GG, 3 is $CCN_{(31)}$-AAASSWWSSTTT-$N_{(30)}$-GG and 4 is $CCN_{(30)}$-AAASSWWSSTTT-$N_{(30)}$-GG. Only the + strand is shown and the start and end corresponds to the first and last base pair in the chromosome (GRCh38) or alternate assembly when applicable. Source code is described in Chapter 3 Methods.

Appendix B. Rec-seq quality scores and significance values

κ_{avg} values for Rec-seq experiments

Enzyme Variant	κ_{avg}
Brec1	0.28
Bxb1 attB	1.01
Bxb1 attP*	2.36
Δ19 Cre	0.78
Dre	1.49
E176A	1.41
E262A	1.34
H289A	1.09
K244A	3.48
K43A	1.61
K86A	2.32
M44A	4.90
N10A	3.61
Q90/94A	15.21
Q90A	7.40
Q94A	6.36
Q9A	1.82
R259A	5.61
R282A	11.20
Tre	5.17
VCre	1.82
WT Cre	4.53
WT Cre (4.7 mut./half-site)	5.03
WT Cre (inv. core)	15.14
WT Cre (commercial)	2.81

*κ_{avg} values for experiments with Bxb1 attP – L1 randomized oligonucleotides could not be calculated, as the unique molecular identifier was omitted due to DNA synthesis size limits.

Student's t-test significance values

Enzyme variant	Half-site position	Bonferroni-corrected p value
Brec1	12	0.01256975
Brec1	10	0.00368103

Brec1	8	0.00178504
Brec1	5'	5.32E-05
Brec1	8'	0.00040415
K244A	17	0.00041764
K244A	15	0.01015747
K244A	14	0.03528123
K244A	13	0.00159779
K244A	12	0.00517888
K244A	11	0.00332736
K244A	10	0.04740163
K244A	9	0.0004642
K244A	8	0.00203837
K244A	7	0.0116214
K244A	6	0.01010555
K244A	5	0.00385061
K244A	6'	0.04993283
K244A	8'	0.00872169
K244A	9'	0.04073499
K244A	12'	0.01620585
K244A	17'	0.04073309
M44A	5	0.00532987
R259A	17	0.0101429
R259A	16	0.00133818
R259A	15	0.03089689
R259A	14	0.00731429
R259A	13	0.0157074
R259A	10	0.00075264
R259A	8	0.02012748
R259A	7	0.00397227
R259A	6	0.00257851
R259A	6'	0.0145799
R259A	10'	0.00639154
R259A	14'	0.04347332
R259A	16'	0.00305847
R282A	8	0.01524759
R282A	7	0.03535704
R282A	6	0.00173287
R282A	5	0.01078905
R282A	6'	0.01441939

R282A	7'	0.02595199
R282A	8'	0.01561694
Tre	17	0.00026323
Tre	15	0.00439622
Tre	12	0.03084838
Tre	10	2.46E-05
Tre	9	7.83E-05
Tre	5'	0.00270499
Tre	6'	0.00512937
Tre	14'	0.00092103

Significance of log-enrichment values was calculated by performing the Student's t-test assuming equal variance for each individual position of each SSR variant relative to wild-type Cre, and the effect of multiple comparisons was counteracted using the Bonferroni correction.

Paired t-test significance values

Enzyme variant	Half-site position	Bonferroni-corrected p-value
WT Cre	5/5'	0.02522168

Significance of log-enrichment values between the left and right half-sites of wild-type Cre was calculated by performing a paired t-test, and the effect of multiple comparisons was counteracted using the Bonferroni correction.

Mann-Whitney U test significance values

Enzyme variant	Bonferroni-corrected p-value
Brec1	0.01583792
E176A	1.03E-11
K244A	5.39E-17
Q90/94A	1.31E-05
Q90A	0.00019859
Q94A	0.01969353
R259A	7.17E-10
R282A	2.99E-11
Tre	3.49E-6

Significance of full substrate log-enrichment profiles was calculated using the two-sided Mann-Whitney U test. We compared the absolute value of the residuals for wild-type Cre and each enzyme variant, and applied the Bonferroni correction.

Appendix C. Rec-seq predicted synthetic and endogenous off-target sequences

Synthetic Tre substrates and fold-enrichment relative to input-library abundance

Name	Left half-site	Fold-enrichment
LTR	ACAACATCCTATTACAC	2.32
L1	ACAACATAATATTACAC	9.39
L2	ACAACTTGCTATTACAC	10.37
L3	CCAACATTCTATTACAC	10.32
L4	ACAACATTCTATAACAC	3.58

Name	Right half-site	Fold-enrichment
LTR	CCTATATGCCAACATGG	3.77
R1	CCTATATGCCAAGTTGG	17.57
R2	CCTATATACCAACTTGG	13.62
R3	CCTATATGGCAACTTGG	8.58
R4	CCTATATGCCAACAATA	> 39.0

Mismatches relative to *loxLTR* (red) and core sequences (gray) are highlighted. Off-target R4 was not detected in sequencing of the pre-selection library, so the fold enrichment was calculated on the basis of the theoretical abundance of a triply-mutated sequence in the synthesized library.

Synthetic Brec1 substrates and fold-enrichment relative to input-library abundance

Name	Left half-site	Fold-enrichment
BTR	AACCCACTGCTTAAGCC	3.10
L1	AACCCTCCGCTTAAGCC	14.74
L2	AACGCACTGTTTAAGCC	6.04
L3	AACCCACAGATTAAGCC	6.36
L4	AACCCCCTGATTAAGCC	7.19

Name	Right half-site	Fold-enrichment
BTR	TCAATAAAGCTTGCCTT	3.78
R1	TCAATAAACCTTGGCTT	6.03
R2	TCAATAATGCATGCCTT	17.01
R3	TCAATAAAGCTTGTATT	2.73
R4	TCAATAATGGGTGCCTT	> 159.8

Mismatches relative to *loxBTR* (red) and core sequences (gray) are highlighted. Off-target R4 was not detected in sequencing of the pre-selection library, so the fold enrichment was calculated on the basis of the theoretical abundance of a triply-mutated sequence in the synthesized library.

Human genomic off-targets for Tre

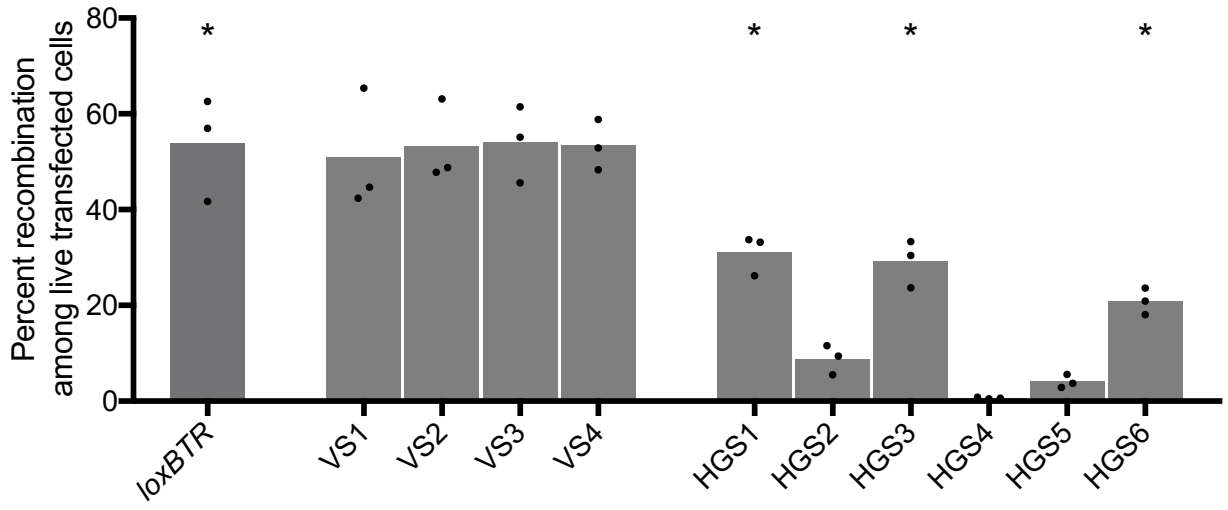
Name	Sequence	Non-core mismatches	Genomic location
LTR	ACAACATCCTATTACACCCTATATGCCAACATGG	--	--
LTR-off 1	TGAACTTAATATTTTTAATAGTATTGCAAATTGA	10	chr14 - 20878251, chr3 + 5904926
LTR-off 2	GCAACATGGTATTAGCTACTTTATCTCAATATGT	7	chr14 - 46653232, chr8 + 106953135
LTR-off 3	AAAACTTTATATTGAAGGAAATATGCCAAATGCA	9	chr3 + 53100634
LTR-off 4	TCAACCTTCTATTGATTTCTCTATTTCAATGGCT	10	chr7 + 43208243, chr4 + 135884591
LTR-off 5	AAAACATTATATTGAGTATAATATTCCAAAATAT	7	chr18 - 36924190, chr7 + 82176261
LTR-off 6	TGAACTTTATATTAATGGAATTATACCAAATGCA	11	11 instances
LTR-off 7	AGAACATGATATTACTCTCAATATCGCAAAAAGT	8	101 instances
LTR-off 8	GTAACATTATATTAATTTTAATATGACAAATCTA	10	6 instances

Mismatches relative to *loxLTR* (red) and core sequences (gray) are highlighted.

Human genomic off-targets for Brec1

Name	Sequence	Non-core mismatches	Genomic location
BTR	AACCCACTGCTTAAGCCTCAATAAAGCTTGCCTT	--	--
BTR-off 1	TATACACTGCTTACTAAGCTGTAAGACTTGGTGT	8	chr12 + 90808809
BTR-off 2	ATGCCTCAGTTTATCCATCTGTAAAACATGGATT	11	23 instances
BTR-off 3	CTCCCGCTGCTTACGTGTCTTTAAACCATGTTCC	9	chr1 - 159864674
BTR-off 4	TCCATACAGGTTAGCATGTAATAAATCATGGCTT	9	chr3 - 167733225
BTR-off 5	CCGGCGCTGCTTATTTCGGCCTAACTCTTGGTTT	9	chr4 + 13484892
BTR-off 6	AACTGTCTGCTTAAGGAAATATAACTCTTGCTTT	6	chr7 - 125265273
BTR-off 7	ATCAAACTGTTTAGTTTAGAATAAAACATGCTAT	8	8 instances
BTR-off 8	AAAGGACTGGTTAACACCCCCTAATTCCTGCCCA	9	chr12 + 103496569

Mismatches relative to *loxBTR* (red) and core sequences (gray) are highlighted.



Brec1 activity on previously-reported off-target sequences. Cells were transfected with Brec1 expression plasmid and a reporter plasmid bearing recombinase targets flanking a poly-A terminator that blocks *EGFP* transcription. Brec1 activity on *loxBTR*, singly-mismatched substrates (VS1-4), and potential genomic pseudo-sites (HGS1-6) was measured as the fraction of cells exhibiting EGFP fluorescence. The percentage of EGFP-positive cells shown is of transfected cells (determined by gating for the presence of co-transfected plasmid constitutively expressing *mCherry*) and 10,000 live events were recorded for each experiment. Data are represented as the mean (bars) of three independent biological replicates (dots). For HGS1-6, significant differences ($p \leq 0.05$) relative to no-enzyme control samples are indicated (asterisks).

Previously reported Brec1 off-targets[72]

Name	Sequence	Non-core mismatches	Genomic location
BTR	AACCCACTGCTTAAGCCTCAATAAAGCTTGCCTT	--	--
VS1	AACCCACTGCTTAAGCTTCAATAAAGCTTGCCTT	0	--
VS2	AACCCACCGCTTAAGCCTCAATAAAGCTTGCCTT	1	--
VS3	GACCCACTGCTTAAGCCTCAATAAAGCTTGCCTT	1	--
VS4	AGCCCACTGCTTAAGCCTCAATAAAGCTTGCCTT	1	--
HGS1	AAGCCCTTGCTTAAAGGATTTAAAGAATGTTTA	8	4 instances
HGS2	AAATTATTGCTTATGAAGAAATAAAGCCAGCATT	7	chr4 – 138478069
HGS3	ATCCGATAGCTTATTTAATAATAAAGTTTGTATA	8	3 instances
HGS4	ATCCCACTGCTGAATATCCTCTAAAGCTTTCTGT	5	chr6 – 60734964, chr6 - 57983493
HGS5	GACGCATTCCTTATTCTTGAAAAAAGCTTGCATA	7	chr2 – 87894680, chrX + 144143849
HGS6	CACAATCTTCTTACACTGTAGTAAAGCTTGCTTG	7	4 instances

Mismatches relative to *loxBTR* (red) and core sequences (gray) are highlighted.

Appendix D. Human genomic Bxb1 minimal substrate sequences identified *in silico*

Sequence ID	Sequence	Genomic location
--	ACNACNGNNNNNNCNGTNGT	*Minimal Bxb1 recognition motif*
Bx1	ACTACAGGTTTTTCTGTGGT	chr2 - 170079896
Bx2	ACCACTGCAGAAACTGTTGT	chr2 - 44171015
Bx3	ACAACAGGCTGGGCGGTGGT	chr7 - 6826752, chr7 + 5898235
Bx4	ACCACAGTGGTAGCCGTGGT	chr22 - 38845180
Bx5	ACCACTGTTATTTCTGTGGT	chr9 + 36189957
Bx6	ACAACGGGAGAACCAGTGGT	chr2 - 112838613
Bx7	ACCACTGCAGAGGCAGTGGT	chr9 - 21181598, chr9 - 21234254
Bx8	ACAACAGAGACCACTGTTGT	chr21 - 5063399, chr21 + 44198732
Bx9	ACCACAGAAAAATCAGTGGT	chr22 - 38507823
Bx10	ACCACTGGAGACCCCGTAGT	chr8 - 94553684
Bx11	ACAACTGGCAGCACAGTAGT	chr1 - 151996034
Bx12	ACCACAGTTTTTCCTGTGGT	11 instances
Bx13	ACGACAGGACTTCCTGTCGT	chr11 - 4213921
Bx14	ACCACTGCACCTACAGTAGT	chr2 - 219179205
Bx15	ACCACCGTCCCCACAGTGGT	chr14 - 70419208
Bx16	ACCACAGAAGTAACTGTGGT	4 instances
Bx17	ACCACTGGTTCTCCCGTTGT	chr2 + 112838613
Bx18	ACAACTGTTTCTTCAGTAGT	chr4 - 124560488
Bx19	ACAACTGAACAAACAGTTGT	chr6 + 26250630
Bx20	ACCACTGTGCACACCGTGGT	chr4 - 765784
Bx21	ACCACGGATGTGTCTGTGGT	chr11 - 66959236
Bx22	ACTACAGATAAAACTGTAGT	chr16 - 50395770
Bx23	ACAACTGCTTGAACTGTGGT	chr11 - 86194192
Bx24	ACTACGGAATAAGCGGTAGT	chr4 + 127880517
Bx25	ACCACAGGTCGACCTGTGGT	chr9 - 23688065, chr19 + 23688065
Bx26	ACCACAGTTCAAGCAGTTGT	chr11 + 86194192
Bx27	ACTACAGAGTCATCTGTTGT	chr20 + 59121057
Bx28	ACTACTGCATGCACAGTGGT	chr19 + 57611438
Bx29	ACAACTGTAATCCCAGTAGT	chr3 + 112331004
Bx30	ACAACAGGTTGGGCGGTGGT	chr7 - 97972805
Bx31	ACAACTGTTTGTTCAGTTGT	chr6 - 26250630
Bx32	ACTACCGTGGGACCTGTTGT	chr2 + 120239885
Bx33	ACTACAGTTTATCCTGTTGT	chr4 + 133707554
Bx34	ACTACAGTGGATGCTGTTGT	chr17 + 51831358
Bx35	ACCACAGAGAGAGCTGTGGT	chrX - 150489623

Bx36	ACAACAGTGACAACAGTAGT	chr14 - 59467983
Bx37	ACTACGGGGTCTCCAGTGGT	chr8 + 94553684
Bx38	ACAACGGCATCTTCAGTGGT	chr8 - 59894946
Bx39	ACAACAGAACATTCTGTTGT	chr18 + 59685779
Bx40	ACTACAGTGTCTGCCGTGGT	chr2 + 64643342
Bx41	ACTACCGCCACTACTGTTGT	chr9 - 21080635
Bx42	ACCACTGGCTATACAGTTGT	chr1 + 91947315
Bx43	ACCACAGAAACATCAGTTGT	chr9 - 110728587
Bx44	ACGACAGTGTGCACTGTTGT	chr21 - 26573641
Bx45	ACCACTGTTAGGACAGTAGT	chr16 - 12560068
Bx46	ACTACTGGGCCTGCGGTTGT	chr8 + 144435850
Bx47	ACAACCGCAGGCCCAGTAGT	chr8 - 144435850
Bx48	ACAACAGATTATTCAGTAGT	chr13 + 19862056
Bx49	ACCACAGATTTTACGGTTGT	chr12 - 46389049
Bx50	ACCACAGTCCCTACCGTGGT	chr3 - 10628734
Bx51	ACTACCGTCACAGCTGTAGT	chr15 - 78622044
Bx52	ACCACAGTAATATCAGTAGT	chr6 - 4604784
Bx53	ACTACTGTGAGGACAGTAGT	chr7 - 124035577

Bibliography

1. Jinek, M. et al. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816-821 (2012).

2. Cong, L. et al. Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819-823 (2013).

3. Mali, P. et al. RNA-guided human genome engineering via Cas9. *Science* **339**, 823-826 (2013).

4. Pennisi, E. The CRISPR craze. *Science* **341**, 833-836 (2013).

5. Avery, O.T., Macleod, C.M. & McCarty, M. Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types : Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated from Pneumococcus Type Iii. *J Exp Med* **79**, 137-158 (1944).

6. Crone, M. & Mah, J.K. Current and Emerging Therapies for Duchenne Muscular Dystrophy. *Curr Treat Options Neurol* **20**, 31 (2018).

7. Huang, C.H., Lee, K.C. & Doudna, J.A. Applications of CRISPR-Cas Enzymes in Cancer Therapeutics and Detection. *Trends Cancer* **4**, 499-512 (2018).

8. Cornu, T.I., Mussolino, C., Bloom, K. & Cathomen, T. Editing CCR5: a novel approach to HIV gene therapy. *Adv Exp Med Biol* **848**, 117-130 (2015).

9. Villiger, L. et al. Treatment of a metabolic liver disease by in vivo genome base editing in adult mice. *Nat Med* **24**, 1519-1525 (2018).

10. Rohn, T.T., Kim, N., Isho, N.F. & Mack, J.M. The Potential of CRISPR/Cas9 Gene Editing as a Treatment Strategy for Alzheimer's Disease. *J Alzheimers Dis Parkinsonism* **8** (2018).

11. Jiang, C. et al. A non-viral CRISPR/Cas9 delivery system for therapeutically targeting HBV DNA and pcsk9 in vivo. *Cell Res* **27**, 440-443 (2017).

12. Gao, X. et al. Treatment of autosomal dominant hearing loss by in vivo delivery of genome editing agents. *Nature* **553**, 217-221 (2018).

13. Yang, S. et al. CRISPR/Cas9-mediated gene editing ameliorates neurotoxicity in mouse model of Huntington's disease. *J Clin Invest* **127**, 2719-2724 (2017).

14. Ye, L. et al. Genome editing using CRISPR-Cas9 to create the HPFH genotype in HSPCs: An approach for treating sickle cell disease and beta-thalassemia. *Proc Natl Acad Sci U S A* **113**, 10661-10665 (2016).

15. Marangi, M. & Pistritto, G. Innovative Therapeutic Strategies for Cystic Fibrosis: Moving Forward to CRISPR Technique. *Front Pharmacol* **9**, 396 (2018).

16. Hammond, A.M. & Galizi, R. Gene drives to fight malaria: current state and future directions. *Pathog Glob Health* **111**, 412-423 (2017).

17. Gao, C. The future of CRISPR technologies in agriculture. *Nat Rev Mol Cell Biol* **19**, 275-276 (2018).

18. McKenna, A. et al. Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science* **353**, aaf7907 (2016).

19. Zhan, T., Rindtorff, N., Betge, J., Ebert, M.P. & Boutros, M. CRISPR/Cas9 for cancer research and therapy. *Semin Cancer Biol* (2018).

20. Tang, W. & Liu, D.R. Rewritable multi-event analog recording in bacterial and mammalian cells. *Science* **360** (2018).

21. Cho, S., Shin, J. & Cho, B.K. Applications of CRISPR/Cas System to Bacterial Metabolic Engineering. *Int J Mol Sci* **19** (2018).

22. Doudna, J.A. & Charpentier, E. Genome editing. The new frontier of genome engineering with CRISPR-Cas9. *Science* **346**, 1258096 (2014).

23. Cohen, S.N., Chang, A.C., Boyer, H.W. & Helling, R.B. Construction of biologically functional bacterial plasmids in vitro. *Proc Natl Acad Sci U S A* **70**, 3240-3244 (1973).

24. Rogers, S., Lowenthal, A., Terheggen, H.G. & Columbo, J.P. Induction of arginase activity with the Shope papilloma virus in tissue culture cells from an argininemic patient. *J Exp Med* **137**, 1091-1096 (1973).

25. Finnegan, D.J. Transposable elements. *Curr Opin Genet Dev* **2**, 861-867 (1992).

26. Thomas, K.R., Folger, K.R. & Capecchi, M.R. High frequency targeting of genes to specific sites in the mammalian genome. *Cell* **44**, 419-428 (1986).

27. Capecchi, M.R. Altering the genome by homologous recombination. *Science* **244**, 1288-1292 (1989).

28. Mansour, S.L., Thomas, K.R. & Capecchi, M.R. Disruption of the proto-oncogene int-2 in mouse embryo-derived stem cells: a general strategy for targeting mutations to non-selectable genes. *Nature* **336**, 348-352 (1988).

29. Carroll, D. Genome Editing: Past, Present, and Future. *Yale J Biol Med* **90**, 653-659 (2017).

30. Moore, J.K. & Haber, J.E. Cell cycle and genetic requirements of two pathways of nonhomologous end-joining repair of double-strand breaks in Saccharomyces cerevisiae. *Mol Cell Biol* **16**, 2164-2173 (1996).

31. Chapman, J.R., Taylor, M.R. & Boulton, S.J. Playing the end game: DNA double-strand break repair pathway choice. *Mol Cell* **47**, 497-510 (2012).

32. Sancar, A., Lindsey-Boltz, L.A., Unsal-Kacmaz, K. & Linn, S. Molecular mechanisms of mammalian DNA repair and the DNA damage checkpoints. *Annu Rev Biochem* **73**, 39-85 (2004).

33. Mao, Z., Bozzella, M., Seluanov, A. & Gorbunova, V. DNA repair by nonhomologous end joining and homologous recombination during cell cycle in human cells. *Cell Cycle* **7**, 2902-2906 (2008).

34. Branzei, D. & Foiani, M. Regulation of DNA repair throughout the cell cycle. *Nat Rev Mol Cell Bio* **9**, 297-308 (2008).

35. Heyer, W.D., Ehmsen, K.T. & Liu, J. Regulation of homologous recombination in eukaryotes. *Annu Rev Genet* **44**, 113-139 (2010).

36. Choulika, A., Perrin, A., Dujon, B. & Nicolas, J.F. Induction of homologous recombination in mammalian chromosomes by using the I-SceI system of Saccharomyces cerevisiae. *Mol Cell Biol* **15**, 1968-1973 (1995).

37. Wah, D.A., Hirsch, J.A., Dorner, L.F., Schildkraut, I. & Aggarwal, A.K. Structure of the multimodular endonuclease FokI bound to DNA. *Nature* **388**, 97-100 (1997).

38. Carroll, D. Progress and prospects: zinc-finger nucleases as gene therapy agents. *Gene Ther* **15**, 1463-1468 (2008).

39. Maeder, M.L., Thibodeau-Beganny, S., Sander, J.D., Voytas, D.F. & Joung, J.K. Oligomerized pool engineering (OPEN): an 'open-source' protocol for making customized zinc-finger arrays. *Nat Protoc* **4**, 1471-1501 (2009).

40. Miller, J.C. et al. A TALE nuclease architecture for efficient genome editing. *Nat Biotechnol* **29**, 143-148 (2011).

41. Joung, J.K. & Sander, J.D. TALENs: a widely applicable technology for targeted genome editing. *Nat Rev Mol Cell Biol* **14**, 49-55 (2013).

42. Cebrian-Serrano, A. & Davies, B. CRISPR-Cas orthologues and variants: optimizing the repertoire, specificity and delivery of genome engineering tools. *Mamm Genome* **28**, 247-261 (2017).

43. Hu, J.H. et al. Evolved Cas9 variants with broad PAM compatibility and high DNA specificity. *Nature* **556**, 57-63 (2018).

44. Nishimasu, H. et al. Engineered CRISPR-Cas9 nuclease with expanded targeting space. *Science* **361**, 1259-1262 (2018).

45. Hartlerode, A.J. & Scully, R. Mechanisms of double-strand break repair in somatic mammalian cells. *Biochem J* **423**, 157-168 (2009).

46. Paquet, D. et al. Efficient introduction of specific homozygous and heterozygous mutations using CRISPR/Cas9. *Nature* **533**, 125 (2016).

47. Bak, R.O. et al. Multiplexed genetic engineering of human hematopoietic stem and progenitor cells using CRISPR/Cas9 and AAV6. *Elife* **6** (2017).

48. Nishiyama, J., Mikuni, T. & Yasuda, R. Virus-Mediated Genome Editing via Homology-Directed Repair in Mitotic and Postmitotic Cells in Mammalian Brain. *Neuron* **96**, 755-768 e755 (2017).

49. Lukacsovich, T., Yang, D. & Waldman, A.S. Repair of a specific double-strand break generated within a mammalian chromosome by yeast endonuclease I-SceI. *Nucleic Acids Res* **22**, 5649-5657 (1994).

50. Rouet, P., Smih, F. & Jasin, M. Introduction of double-strand breaks into the genome of mouse cells by expression of a rare-cutting endonuclease. *Mol Cell Biol* **14**, 8096-8106 (1994).

51. Jeggo, P.A. DNA breakage and repair. *Adv Genet* **38**, 185-218 (1998).

52. Koo, T., Lee, J. & Kim, J.S. Measuring and Reducing Off-Target Activities of Programmable Nucleases Including CRISPR-Cas9. *Mol Cells* **38**, 475-481 (2015).

53. Haapaniemi, E., Botla, S., Persson, J., Schmierer, B. & Taipale, J. CRISPR-Cas9 genome editing induces a p53-mediated DNA damage response. *Nat Med* **24**, 927-930 (2018).

54. Kim, S. et al. CRISPR RNAs trigger innate immune responses in human cells. *Genome Res* (2018).

55. Wagner, D.L. et al. High prevalence of Streptococcus pyogenes Cas9-reactive T cells within the adult human population. *Nat Med* **25**, 242-248 (2019).

56. Grindley, N.D., Whiteson, K.L. & Rice, P.A. Mechanisms of site-specific recombination. *Annu Rev Biochem* **75**, 567-605 (2006).

57. Smith, M.C. & Thorpe, H.M. Diversity in the serine recombinases. *Mol Microbiol* **44**, 299-307 (2002).

58. Jayaram, M. et al. An Overview of Tyrosine Site-specific Recombination: From an Flp Perspective. *Microbiol Spectr* **3** (2015).

59. Rajeev, L., Malanowska, K. & Gardner, J.F. Challenging a paradigm: the role of DNA homology in tyrosine recombinase reactions. *Microbiol Mol Biol Rev* **73**, 300-309 (2009).

60. Meinke, G., Bohm, A., Hauber, J., Pisabarro, M.T. & Buchholz, F. Cre Recombinase and Other Tyrosine Recombinases. *Chem Rev* **116**, 12785-12820 (2016).

61. Grindley, N.D.F., Whiteson, K.L. & Rice, P.A. Mechanisms of site-specific recombination. *Annual Review of Biochemistry* **75**, 567-605 (2006).

62. Kim, A.I. et al. Mycobacteriophage Bxb1 integrates into the Mycobacterium smegmatis groEL1 gene. *Mol Microbiol* **50**, 463-473 (2003).

63. Rutherford, K. & Van Duyne, G.D. The ins and outs of serine integrase site-specific recombination. *Curr Opin Struct Biol* **24**, 125-131 (2014).

64. Wang, B. et al. Highly efficient CRISPR/HDR-mediated knock-in for mouse embryonic stem cells and zygotes. *Biotechniques* **59**, 201-202, 204, 206-208 (2015).

65. Brown, W.R., Lee, N.C., Xu, Z. & Smith, M.C. Serine recombinases as tools for genome engineering. *Methods* **53**, 372-379 (2011).

66. Xu, Z. et al. Accuracy and efficiency define Bxb1 integrase as the best of fifteen candidate serine recombinases for the integration of DNA into the human genome. *BMC Biotechnol* **13**, 87 (2013).

67. Nagy, A. Cre recombinase: the universal reagent for genome tailoring. *Genesis* **26**, 99-109 (2000).

68. Hayashi, S. & McMahon, A.P. Efficient recombination in diverse tissues by a tamoxifen-inducible form of Cre: a tool for temporally regulated gene activation/inactivation in the mouse. *Dev Biol* **244**, 305-318 (2002).

69. Kretzschmar, K. & Watt, F.M. Lineage tracing. *Cell* **148**, 33-45 (2012).

70. Jensen, P. & Dymecki, S.M. Essentials of recombinase-based genetic fate mapping in mice. *Methods Mol Biol* **1092**, 437-454 (2014).

71. Hauber, I. et al. Highly significant antiviral activity of HIV-1 LTR-specific tre-recombinase in humanized mice. *PLoS Pathog* **9**, e1003587 (2013).

72. Karpinski, J. et al. Directed evolution of a recombinase that excises the provirus of most HIV-1 primary isolates with high specificity. *Nat Biotechnol* **34**, 401-409 (2016).

73. Olorunniji, F.J., Rosser, S.J. & Stark, W.M. Site-specific recombinases: molecular machines for the Genetic Revolution. *Biochem J* **473**, 673-684 (2016).

74. Bogdanove, A.J., Bohm, A., Miller, J.C., Morgan, R.D. & Stoddard, B.L. Engineering altered protein-DNA recognition specificity. *Nucleic Acids Res* **46**, 4845-4871 (2018).

75. Sarkar, I., Hauber, I., Hauber, J. & Buchholz, F. HIV-1 proviral DNA excision using an evolved recombinase. *Science* **316**, 1912-1915 (2007).

76. Esvelt, K.M., Carlson, J.C. & Liu, D.R. A system for the continuous directed evolution of biomolecules. *Nature* **472**, 499-503 (2011).

77. Irion, S. et al. Identification and targeting of the ROSA26 locus in human embryonic stem cells. *Nat Biotechnol* **25**, 1477-1482 (2007).

78. Gordley, R.M., Gersbach, C.A. & Barbas, C.F., 3rd Synthesis of programmable integrases. *Proc Natl Acad Sci U S A* **106**, 5053-5058 (2009).

79. Mercer, A.C., Gaj, T., Fuller, R.P. & Barbas, C.F., 3rd Chimeric TALE recombinases with programmable DNA sequence specificity. *Nucleic Acids Res* **40**, 11163-11172 (2012).

80. Gaj, T., Sirk, S.J. & Barbas, C.F., 3rd Expanding the scope of site-specific recombinases for genetic and metabolic engineering. *Biotechnol Bioeng* **111**, 1-15 (2014).

81. Chaikind, B., Bessen, J.L., Thompson, D.B., Hu, J.H. & Liu, D.R. A programmable Cas9-serine recombinase fusion protein that operates on DNA sequences in mammalian cells. *Nucleic Acids Res* **44**, 9758-9770 (2016).

82. Van Duyne, G.D. Cre Recombinase. *Microbiol Spectr* **3**, MDNA3-0014-2014 (2015).

83. Baldwin, E.P. et al. A specificity switch in selected cre recombinase variants is mediated by macromolecular plasticity and water. *Chem Biol* **10**, 1085-1094 (2003).

84. Rufer, A.W. & Sauer, B. Non-contact positions impose site selectivity on Cre recombinase. *Nucleic Acids Res* **30**, 2764-2771 (2002).

85. Pattanayak, V., Ramirez, C.L., Joung, J.K. & Liu, D.R. Revealing off-target cleavage specificities of zinc-finger nucleases by in vitro selection. *Nat Methods* **8**, 765-770 (2011).

86. Pattanayak, V. et al. High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. *Nat Biotechnol* **31**, 839-843 (2013).

87. Pattanayak, V., Guilinger, J.P. & Liu, D.R. Determining the specificities of TALENs, Cas9, and other genome-editing enzymes. *Methods Enzymol* **546**, 47-78 (2014).

88. Tsai, S.Q. et al. GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nat Biotechnol* **33**, 187-197 (2015).

89. Bessen, J.L. et al. High-resolution specificity profiling and off-target prediction for site-specific DNA recombinases. *Nat Commun* DOI:10.1038/s41467-019-09987-0 (2019).

90. Sadelain, M., Papapetrou, E.P. & Bushman, F.D. Safe harbours for the integration of new DNA in the human genome. *Nat Rev Cancer* **12**, 51-58 (2012).

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

92. Buchholz, F. & Stewart, A.F. Alteration of Cre recombinase site specificity by substrate-linked protein evolution. *Nat Biotechnol* **19**, 1047-1052 (2001).

93. Husimi, Y. Selection and evolution of bacteriophages in cellstat. *Adv Biophys* **25**, 1-43 (1989).

94. Dickinson, B.C., Leconte, A.M., Allen, B., Esvelt, K.M. & Liu, D.R. Experimental interrogation of the path dependence and stochasticity of protein evolution using phage-assisted continuous evolution. *Proc Natl Acad Sci U S A* **110**, 9007-9012 (2013).

95. Carlson, J.C., Badran, A.H., Guggiana-Nilo, D.A. & Liu, D.R. Negative selection and stringency modulation in phage-assisted continuous evolution. *Nat Chem Biol* **10**, 216-222 (2014).

96. Hubbard, B.P. et al. Continuous directed evolution of DNA-binding proteins to improve TALEN specificity. *Nat Methods* **12**, 939-942 (2015).

97. Dickinson, B.C., Packer, M.S., Badran, A.H. & Liu, D.R. A system for the continuous directed evolution of proteases rapidly reveals drug-resistance mutations. *Nat Commun* **5**, 5352 (2014).

98. Packer, M.S., Rees, H.A. & Liu, D.R. Phage-assisted continuous evolution of proteases with altered substrate specificity. *Nat Commun* **8**, 956 (2017).

99. Badran, A.H. et al. Continuous evolution of Bacillus thuringiensis toxins overcomes insect resistance. *Nature* **533**, 58-63 (2016).

100. Bryson, D.I. et al. Continuous directed evolution of aminoacyl-tRNA synthetases. *Nat Chem Biol* **13**, 1253-1260 (2017).

101. Rakonjac, J. & Model, P. Roles of pIII in filamentous phage assembly. *J Mol Biol* **282**, 25-41 (1998).

102. Badran, A.H. & Liu, D.R. Development of potent in vivo mutagenesis plasmids with broad mutational spectra. *Nat Commun* **6**, 8425 (2015).

103. Calendar, R. The Bacteriophages, Edn. 2nd. (Oxford University Press, Oxford ; New York; 2006).

104. Gibb, B. et al. Requirements for catalysis in the Cre recombinase active site. *Nucleic Acids Res* **38**, 5817-5832 (2010).

105. Casola, S. Mouse models for miRNA expression: the ROSA26 locus. *Methods Mol Biol* **667**, 145-163 (2010).

106. Aharoni, A. et al. The 'evolvability' of promiscuous protein functions. *Nat Genet* **37**, 73-76 (2005).

107. Ennifar, E., Meyer, J.E., Buchholz, F., Stewart, A.F. & Suck, D. Crystal structure of a wild-type Cre recombinase-loxP synapse reveals a novel spacer conformation suggesting an alternative mechanism for DNA cleavage activation. *Nucleic Acids Res* **31**, 5449-5460 (2003).

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

109. Deng, L.W. & Perham, R.N. Delineating the site of interaction on the pIII protein of filamentous bacteriophage fd with the F-pilus of Escherichia coli. *J Mol Biol* **319**, 603-614 (2002).

110. Karlsson, F., Borrebaeck, C.A., Nilsson, N. & Malmborg-Hager, A.C. The mechanism of bacterial infection by filamentous phages involves molecular interactions between TolA and phage protein 3 domains. *J Bacteriol* **185**, 2628-2634 (2003).

111. de Avila, E.S.S., Echeverrigaray, S. & Gerhardt, G.J. BacPP: bacterial promoter prediction--a tool for accurate sigma-factor specific assignment in enterobacteria. *J Theor Biol* **287**, 92-99 (2011).

112. Rohrer, J. & Kuhn, A. The function of a leader peptide in translocating charged amino acyl residues across a membrane. *Science* **250**, 1418-1421 (1990).

113. Brissette, J.L., Weiner, L., Ripmaster, T.L. & Model, P. Characterization and sequence of the Escherichia coli stress-induced psp operon. *J Mol Biol* **220**, 35-48 (1991).

114. Fenno, L.E. et al. Targeting cells with single vectors using multiple-feature Boolean logic. *Nat Methods* **11**, 763-772 (2014).

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

116. Zhang, C. et al. Redesign of the monomer-monomer interface of Cre recombinase yields an obligate heterotetrameric complex. *Nucleic Acids Res* **43**, 9076-9085 (2015).

117. Leaver-Fay, A. et al. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol* **487**, 545-574 (2011).

118. Martin, S.S., Pulido, E., Chu, V.C., Lechner, T.S. & Baldwin, E.P. 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 (2002).

119. Abi-Ghanem, J. et al. Engineering of a target site-specific recombinase by a combined evolution- and structure-guided approach. *Nucleic Acids Res* **41**, 2394-2403 (2013).

120. Yang, G. & Withers, S.G. Ultrahigh-throughput FACS-based screening for directed enzyme evolution. *Chembiochem* **10**, 2704-2715 (2009).

121. de Kok, S. et al. Rapid and reliable DNA assembly via ligase cycling reaction. *ACS Synth Biol* **3**, 97-106 (2014).

122. Turan, S., Zehe, C., Kuehle, J., Qiao, J.H. & Bode, J. Recombinase-mediated cassette exchange (RMCE) - A rapidly-expanding toolbox for targeted genomic modifications. *Gene* **515**, 1-27 (2013).

123. Thyagarajan, B., Guimaraes, M.J., Groth, A.C. & Calos, M.P. Mammalian genomes contain active recombinase recognition sites. *Gene* **244**, 47-54 (2000).

124. Sclimenti, C.R., Thyagarajan, B. & Calos, M.P. Directed evolution of a recombinase for improved genomic integration at a native human sequence. *Nucleic Acids Res* **29**, 5044-5051 (2001).

125. Thyagarajan, B., Olivares, E.C., Hollis, R.P., Ginsburg, D.S. & Calos, M.P. Site-specific genomic integration in mammalian cells mediated by phage phiC31 integrase. *Mol Cell Biol* **21**, 3926-3934 (2001).

126. Shah, R., Li, F., Voziyanova, E. & Voziyanov, Y. Target-specific variants of Flp recombinase mediate genome engineering reactions in mammalian cells. *FEBS J* **282**, 3323-3333 (2015).

127. Akopian, A., He, J., Boocock, M.R. & Stark, W.M. Chimeric recombinases with designed DNA sequence recognition. *Proc Natl Acad Sci U S A* **100**, 8688-8691 (2003).

128. Gersbach, C.A., Gaj, T., Gordley, R.M., Mercer, A.C. & Barbas, C.F. Targeted plasmid integration into the human genome by an engineered zinc-finger recombinase. *Nucleic Acids Research* **39**, 7868-7878 (2011).

129. Prorocic, M.M. et al. Zinc-finger recombinase activities in vitro. *Nucleic Acids Research* **39**, 9316-9328 (2011).

130. Guilinger, J.P., Thompson, D.B. & Liu, D.R. Fusion of catalytically inactive Cas9 to FokI nuclease improves the specificity of genome modification. *Nat Biotechnol* (2014).

131. Tsai, S.Q. et al. Dimeric CRISPR RNA-guided FokI nucleases for highly specific genome editing. *Nat Biotechnol* (2014).

132. Gaj, T., Mercer, A.C., Sirk, S.J., Smith, H.L. & Barbas, C.F., 3rd A comprehensive approach to zinc-finger recombinase customization enables genomic targeting in human cells. *Nucleic Acids Res* **41**, 3937-3946 (2013).

133. Plasterk, R.H., Brinkman, A. & van de Putte, P. DNA inversions in the chromosome of Escherichia coli and in bacteriophage Mu: relationship to other site-specific recombination systems. *Proc Natl Acad Sci U S A* **80**, 5355-5358 (1983).

134. Klippel, A., Mertens, G., Patschinsky, T. & Kahmann, R. The DNA Invertase Gin of Phage Mu - Formation of a Covalent Complex with DNA Via a Phosphoserine at Amino-Acid Position-9. *Embo Journal* **7**, 1229-1237 (1988).

135. Mertens, G. et al. Site-specific recombination in bacteriophage Mu: characterization of binding sites for the DNA invertase Gin. *EMBO J* **7**, 1219-1227 (1988).

136. Xu, H. et al. Sequence determinants of improved CRISPR sgRNA design. *Genome Res* **25**, 1147-1157 (2015).

137. Gordley, R.M., Smith, J.D., Graslund, T. & Barbas, C.F., 3rd Evolution of programmable zinc finger-recombinases with activity in human cells. *J Mol Biol* **367**, 802-813 (2007).

138. Gersbach, C.A., Gaj, T., Gordley, R.M. & Barbas, C.F., 3rd Directed evolution of recombinase specificity by split gene reassembly. *Nucleic Acids Res* **38**, 4198-4206 (2010).

139. Gaj, T., Mercer, A.C., Gersbach, C.A., Gordley, R.M. & Barbas, C.F. Structure-guided reprogramming of serine recombinase DNA sequence specificity. *P Natl Acad Sci USA* **108**, 498-503 (2011).

140. Gaj, T. et al. Enhancing the Specificity of Recombinase-Mediated Genome Engineering through Dimer Interface Redesign. *Journal of the American Chemical Society* **136**, 5047-5056 (2014).

141. Cunningham, F. et al. Ensembl 2015. *Nucleic Acids Res* **43**, D662-669 (2015).

142. van Swieten, J.C. et al. A mutation in the fibroblast growth factor 14 gene is associated with autosomal dominant cerebellar ataxia [corrected]. *Am J Hum Genet* **72**, 191-199 (2003).

143. Brusse, E. et al. Spinocerebellar ataxia associated with a mutation in the fibroblast growth factor 14 gene (SCA27): A new phenotype. *Mov Disord* **21**, 396-401 (2006).

144. Shimojima, K. et al. Spinocerebellar ataxias type 27 derived from a disruption of the fibroblast growth factor 14 gene with mimicking phenotype of paroxysmal non-kinesigenic dyskinesia. *Brain Dev* **34**, 230-233 (2012).

145. Coebergh, J.A. et al. A new variable phenotype in spinocerebellar ataxia 27 (SCA 27) caused by a deletion in the FGF14 gene. *Eur J Paediatr Neurol* **18**, 413-415 (2014).

146. Choquet, K., La Piana, R. & Brais, B. A novel frameshift mutation in FGF14 causes an autosomal dominant episodic ataxia. *Neurogenetics* **16**, 233-236 (2015).

147. Hanahan, D. Studies on transformation of Escherichia coli with plasmids. *J Mol Biol* **166**, 557-580 (1983).

148. Parker, M.M. et al. Admixture mapping identifies a quantitative trait locus associated with FEV1/FVC in the COPDGene Study. *Genet Epidemiol* **38**, 652-659 (2014).

149. Sykes, P.J. et al. Quantitation of targets for PCR by use of limiting dilution. *Biotechniques* **13**, 444-449 (1992).

150. Schellenberger, V. et al. A recombinant polypeptide extends the in vivo half-life of peptides and proteins in a tunable manner. *Nat Biotechnol* **27**, 1186-1190 (2009).

151. Ringrose, L. et al. Comparative kinetic analysis of FLP and cre recombinases: mathematical models for DNA binding and recombination. *J Mol Biol* **284**, 363-384 (1998).

152. Turan, S. et al. Recombinase-mediated cassette exchange (RMCE): traditional concepts and current challenges. *J Mol Biol* **407**, 193-221 (2011).

153. Sirk, S.J., Gaj, T., Jonsson, A., Mercer, A.C. & Barbas, C.F. Expanding the zinc-finger recombinase repertoire: directed evolution and mutational analysis of serine recombinase specificity determinants. *Nucleic Acids Research* **42**, 4755-4766 (2014).

154. Dormiani, K. et al. Long-term and efficient expression of human beta-globin gene in a hematopoietic cell line using a new site-specific integrating non-viral system. *Gene Ther* **22**, 663-674 (2015).

155. Fu, Y. et al. High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nat Biotechnol* **31**, 822-826 (2013).

156. Sanjana, N.E. et al. A transcription activator-like effector toolbox for genome engineering. *Nat Protoc* **7**, 171-192 (2012).

157. Quan, J. & Tian, J. Circular polymerase extension cloning of complex gene libraries and pathways. *PLoS One* **4**, e6441 (2009).

158. Quan, J. & Tian, J. Circular polymerase extension cloning for high-throughput cloning of complex and combinatorial DNA libraries. *Nat Protoc* **6**, 242-251 (2011).

159. Matsuda, T. & Cepko, C.L. Controlled expression of transgenes introduced by in vivo electroporation. *Proc Natl Acad Sci U S A* **104**, 1027-1032 (2007).

160. Hirt, B. Selective extraction of polyoma DNA from infected mouse cell cultures. *J Mol Biol* **26**, 365-369 (1967).

161. Motmans, K., Thirion, S., Raus, J. & Vandevyver, C. Isolation and quantification of episomal expression vectors in human T cells. *Biotechniques* **23**, 1044-1046 (1997).

162. Kim, S.T., Kim, G.W., Lee, Y.S. & Park, J.S. Characterization of Cre-loxP interaction in the major groove: hint for structural distortion of mutant Cre and possible strategy for HIV-1 therapy. *J Cell Biochem* **80**, 321-327 (2001).

163. Hartung, M. & Kisters-Woike, B. Cre mutants with altered DNA binding properties. *J Biol Chem* **273**, 22884-22891 (1998).

164. Lee, L., Chu, L.C. & Sadowski, P.D. Cre induces an asymmetric DNA bend in its target loxP site. *J Biol Chem* **278**, 23118-23129 (2003).

165. Martin, S.S., Chu, V.C. & Baldwin, E. Modulation of the active complex assembly and turnover rate by protein-DNA interactions in Cre-LoxP recombination. *Biochemistry* **42**, 6814-6826 (2003).

166. Gelato, K.A., Martin, S.S., Wong, S. & Baldwin, E.P. Multiple levels of affinity-dependent DNA discrimination in Cre-LoxP recombination. *Biochemistry* **45**, 12216-12226 (2006).

167. Cantor, E.J. & Chong, S. Intein-mediated rapid purification of Cre recombinase. *Protein Expr Purif* **22**, 135-140 (2001).

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

169. Araki, K., Araki, M. & Yamamura, K. Targeted integration of DNA using mutant lox sites in embryonic stem cells. *Nucleic Acids Res* **25**, 868-872 (1997).

170. Guilinger, J.P. et al. Broad specificity profiling of TALENs results in engineered nucleases with improved DNA-cleavage specificity. *Nat Methods* **11**, 429-435 (2014).

171. Meinke, G., Karpinski, J., Buchholz, F. & Bohm, A. Crystal structure of an engineered, HIV-specific recombinase for removal of integrated proviral DNA. *Nucleic Acids Res* **45**, 9726-9740 (2017).

172. Rongrong, L., Lixia, W. & Zhongping, L. Effect of deletion mutation on the recombination activity of Cre recombinase. *Acta Biochim Pol* **52**, 541-544 (2005).

173. Van der Maaten, L. & Hinton, G. Visualizing Data using t-SNE. *Journal of Machine Learning Research* **9**, 2579-2605 (2008).

174. Sauer, B. & McDermott, J. 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 (2004).

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

176. Singh, S., Ghosh, P. & Hatfull, G.F. Attachment site selection and identity in Bxb1 serine integrase-mediated site-specific recombination. *PLoS Genet* **9**, e1003490 (2013).

177. Smith, M.C.M. Phage-encoded Serine Integrases and Other Large Serine Recombinases. *Microbiol Spectr* **3** (2015).

178. Nguyen, N.T.T. et al. RSAT 2018: regulatory sequence analysis tools 20th anniversary. *Nucleic Acids Res* **46**, W209-W214 (2018).

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

180. Warth, L. & Altenbuchner, J. The tyrosine recombinase MrpA and its target sequence: a mutational analysis of the recombination site mrpS resulting in a new left element/right element (LE/RE) deletion system. *Arch Microbiol* **195**, 617-636 (2013).

181. Li, H., Sharp, R., Rutherford, K., Gupta, K. & Van Duyne, G.D. Serine Integrase attP Binding and Specificity. *J Mol Biol* **430**, 4401-4418 (2018).

182. Langer, S.J., Ghafoori, A.P., Byrd, M. & Leinwand, L. A genetic screen identifies novel non-compatible loxP sites. *Nucleic Acids Res* **30**, 3067-3077 (2002).

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

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

185. Lei, X., Wang, L., Zhao, G. & Ding, X. Site-specificity of serine integrase demonstrated by the attB sequence preference of BT1 integrase. *FEBS Lett* **592**, 1389-1399 (2018).

186. Semprini, S. et al. Cryptic loxP sites in mammalian genomes: genome-wide distribution and relevance for the efficiency of BAC/PAC recombineering techniques. *Nucleic Acids Res* **35**, 1402-1410 (2007).

187. Keravala, A. et al. A diversity of serine phage integrases mediate site-specific recombination in mammalian cells. *Mol Genet Genomics* **276**, 135-146 (2006).

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

189. Karimova, M. et al. Vika/vox, a novel efficient and specific Cre/loxP-like site-specific recombination system. *Nucleic Acids Res* **41**, e37 (2013).

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

191. Van Houdt, R., Leplae, R., Lima-Mendez, G., Mergeay, M. & Toussaint, A. Towards a more accurate annotation of tyrosine-based site-specific recombinases in bacterial genomes. *Mob DNA* **3**, 6 (2012).

192. Zuris, J.A. et al. Cationic lipid-mediated delivery of proteins enables efficient protein-based genome editing in vitro and in vivo. *Nat Biotechnol* **33**, 73-80 (2015).

193. Hermann, M. et al. Binary recombinase systems for high-resolution conditional mutagenesis. *Nucleic Acids Res* **42**, 3894-3907 (2014).

194. Zhu, F. et al. DICE, an efficient system for iterative genomic editing in human pluripotent stem cells. *Nucleic Acids Res* **42**, e34 (2014).

195. Ghosh, P., Kim, A. & Hatfull, G.F. The Orientation of Mycobacteriophage Bxb1 Integration Is Solely Dependent on the Central Dinucleotide of attP and attB. *Mol Cell* **12**, 1101-1111 (2003).

196. Rees, H.A. & Liu, D.R. Base editing: precision chemistry on the genome and transcriptome of living cells. *Nat Rev Genet* **19**, 770-788 (2018).

197. Chen, X. & Goncalves, M.A. Engineered Viruses as Genome Editing Devices. *Mol Ther* **24**, 447-457 (2016).

198. Shaikh, A.C. & Sadowski, P.D. Chimeras of the Flp and Cre recombinases: tests of the mode of cleavage by Flp and Cre. *J Mol Biol* **302**, 27-48 (2000).

199. Farruggio, A.P. & Calos, M.P. Serine integrase chimeras with activity in E. coli and HeLa cells. *Biol Open* **3**, 895-903 (2014).

200. Eroshenko, N. & Church, G.M. Mutants of Cre recombinase with improved accuracy. *Nat Commun* **4**, 2509 (2013).

201. Zinder, N.D. & Boeke, J.D. The filamentous phage (Ff) as vectors for recombinant DNA— a review. *Gene* **19**, 1-10 (1982).

202. Zettler, J., Schutz, V. & Mootz, H.D. The naturally split Npu DnaE intein exhibits an extraordinarily high rate in the protein trans-splicing reaction. *FEBS Lett* **583**, 909-914 (2009).

203. Gonzalez-Perez, D., Garcia-Ruiz, E. & Alcalde, M. Saccharomyces cerevisiae in directed evolution: An efficient tool to improve enzymes. *Bioeng Bugs* **3**, 172-177 (2012).

204. Kelley, L.A., Mezulis, S., Yates, C.M., Wass, M.N. & Sternberg, M.J. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc* **10**, 845-858 (2015).

205. Lu, X.J. & Olson, W.K. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res* **31**, 5108-5121 (2003).

206. Luscombe, N.M., Laskowski, R.A. & Thornton, J.M. NUCPLOT: a program to generate schematic diagrams of protein-nucleic acid interactions. *Nucleic Acids Res* **25**, 4940-4945 (1997).

207. Rutherford, K., Yuan, P., Perry, K., Sharp, R. & Van Duyne, G.D. Attachment site recognition and regulation of directionality by the serine integrases. *Nucleic Acids Res* **41**, 8341-8356 (2013).

208. Craigie, R. HIV Integrase, a Brief Overview from Chemistry to Therapeutics. *Journal of Biological Chemistry* **276**, 23213-23216 (2001).

209. Sultana, T., Zamborlini, A., Cristofari, G. & Lesage, P. Integration site selection by retroviruses and transposable elements in eukaryotes. *Nat Rev Genet* **18**, 292-308 (2017).

210. Llano, M. et al. An essential role for LEDGF/p75 in HIV integration. *Science* **314**, 461-464 (2006).

211. Desfarges, S. & Ciuffi, A. Retroviral integration site selection. *Viruses* **2**, 111-130 (2010).

212. El Ashkar, S. et al. BET-independent MLV-based Vectors Target Away From Promoters and Regulatory Elements. *Mol Ther Nucleic Acids* **3**, e179 (2014).

213. Bushman, F.D. Tethering human immunodeficiency virus 1 integrase to a DNA site directs integration to nearby sequences. *Proc Natl Acad Sci U S A* **91**, 9233-9237 (1994).

214. Goulaouic, H. & Chow, S.A. Directed integration of viral DNA mediated by fusion proteins consisting of human immunodeficiency virus type 1 integrase and Escherichia coli LexA protein. *J Virol* **70**, 37-46 (1996).

215. Katz, R.A., Merkel, G. & Skalka, A.M. Targeting of retroviral integrase by fusion to a heterologous DNA binding domain: in vitro activities and incorporation of a fusion protein into viral particles. *Virology* **217**, 178-190 (1996).

216. Bushman, F.D. & Miller, M.D. Tethering human immunodeficiency virus type 1 preintegration complexes to target DNA promotes integration at nearby sites. *J Virol* **71**, 458-464 (1997).

217. Tan, W., Dong, Z., Wilkinson, T.A., Barbas, C.F., 3rd & Chow, S.A. Human immunodeficiency virus type 1 incorporated with fusion proteins consisting of integrase and the designed polydactyl zinc finger protein E2C can bias integration of viral DNA into a predetermined chromosomal region in human cells. *J Virol* **80**, 1939-1948 (2006).

218. Caumont, A.B. et al. Expression of functional HIV-1 integrase in the yeast Saccharomyces cerevisiae leads to the emergence of a lethal phenotype: potential use for inhibitor screening. *Curr Genet* **29**, 503-510 (1996).

219. Desfarges, S. et al. Chromosomal integration of LTR-flanked DNA in yeast expressing HIV-1 integrase: down regulation by RAD51. *Nucleic Acids Res* **34**, 6215-6224 (2006).

220. Valkov, E. et al. Functional and structural characterization of the integrase from the prototype foamy virus. *Nucleic Acids Res* **37**, 243-255 (2009).

221. Faschinger, A. et al. Mouse mammary tumor virus integration site selection in human and mouse genomes. *J Virol* **82**, 1360-1367 (2008).

222. Craig, N.L. Target site selection in transposition. *Annu Rev Biochem* **66**, 437-474 (1997).

223. Maragathavally, K.J., Kaminski, J.M. & Coates, C.J. Chimeric Mos1 and piggyBac transposases result in site-directed integration. *FASEB J* **20**, 1880-1882 (2006).

224. Owens, J.B. et al. Chimeric piggyBac transposases for genomic targeting in human cells. *Nucleic Acids Res* **40**, 6978-6991 (2012).

225. Luo, W. et al. Comparative analysis of chimeric ZFP-, TALE- and Cas9-piggyBac transposases for integration into a single locus in human cells. *Nucleic Acids Res* **45**, 8411-8422 (2017).

226. Cossu, M. et al. Flipping chromosomes in deep-sea archaea. *PLoS Genet* **13**, e1006847 (2017).

227. Barrangou, R. et al. CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**, 1709-1712 (2007).

228. Cozens, C. & Pinheiro, V.B. Darwin Assembly: fast, efficient, multi-site bespoke mutagenesis. *Nucleic Acids Res* **46**, e51 (2018).