



**Discovery and evolution  
of novel Cre-type site-specific recombinases  
for advanced genome engineering**

Milica Jeličić

Aus dem Universitäts KrebsCentrum (UCC), Medizinische Systembiologie  
Leiter: Prof. Dr. Frank Buchholz

---

**Discovery and evolution of novel Cre-type  
tyrosine site-specific recombinases for advanced genome  
engineering**

DISSERTATIONSSCHRIFT

zur Erlangung des akademischen Grades

Doctor of Philosophy (Ph.D.)

vorgelegt

der Medizinischen Fakultät Carl Gustav Carus

der Technischen Universität

Dresden

von

B.Sc. M.Sc. Milica Jeličić

aus Belgrad, Serbien

Dresden 2023

1. Gutachter: Prof. Dr. Frank Buchholz
2. Gutachter: Prof. Dr. Konstantinos Anastassiadis

Tag der mündlichen Prüfung: 11.10.2023

gez.: -----  
Vorsitzender der Promotionskommission

Anmerkung:

Die Eintragung der Gutachter und Tag der mündlichen Prüfung (Verteidigung) erfolgt nach Festlegung von Seiten der Medizinischen Fakultät Carl Gustav Carus der TU Dresden. Sie wird durch die Promovenden nach der Verteidigung zwecks Übergabe der fünf Pflichtexemplare an die Zweigbibliothek Medizin in gedruckter Form oder handschriftlich vorgenommen.

# TABLE OF CONTENTS

LIST OF FIGURES .....	I
LIST OF TABLES .....	III
LIST OF ABBREVIATIONS.....	IV
PART I INTRODUCTION	1
CHAPTER 1 ENGINEERING GENOMES.....	2
CHAPTER 2 SITE-SPECIFIC RECOMBINASES .....	4
<b>2.1 Tyrosine site-specific recombinases (Y-SSRs) .....</b>	<b>4</b>
2.1.1 Mechanism of tyrosine-based site-specific recombination.....	5
2.1.2 The Cre/ <i>loxP</i> system .....	6
2.1.3 <i>loxP</i> variants for broadening the Cre/ <i>loxP</i> utility .....	8
2.1.4 Other tyrosine recombinases .....	10
<b>2.2 Application of SSRs.....</b>	<b>11</b>
CHAPTER 3 PROTEIN ENGINEERING BY DIRECTED EVOLUTION .....	14
<b>3.1 Principles .....</b>	<b>14</b>
<b>3.2 Directed evolution of recombinases .....</b>	<b>16</b>
<b>3.3 Challenges and optimization of directed evolution protocols for development of designer recombinases .....</b>	<b>20</b>
CHAPTER 4 GENOMIC INTEGRATION – SSRS VS NUCLEASES .....	22
CHAPTER 5 AIMS.....	24
PART II RESULTS	26
CHAPTER 6 DISCOVERY AND CHARACTERIZATION OF NOVEL CRE-TYPE SSRS .....	27
<b>6.1 In silico prediction of putative recombinases and their target sites.....</b>	<b>27</b>
<b>6.2 Novel Y-SSRs recombine their predicted target sites in bacteria .....</b>	<b>29</b>
<b>6.3 Profiling target-site selectivity of Cre-type recombinases.....</b>	<b>33</b>
<b>6.4 Activity of novel recombinases in human cells .....</b>	<b>36</b>

<b>6.5 Influence of recombinase expression on cell proliferation.....</b>	<b>37</b>
<b>6.6 Directed evolution of YR9 recombinase.....</b>	<b>38</b>
<b>6.7 Prediction of possible recombination target sites in human and mouse genomes .....</b>	<b>41</b>
<b>CHAPTER 7 EVOLUTION OF VIKa RECOMBINASE FOR RECOMBINASE-MEDIATED INTEGRATION IN POTENTIAL SAFE HARBOR LOCUS.....</b>	<b>44</b>
<b>7.1 Preface .....</b>	<b>44</b>
<b>7.2 Vika wt mediates integration into endogenous voxH9 locus .....</b>	<b>45</b>
<b>7.3 Establishing Vika libraries .....</b>	<b>47</b>
<b>7.4 Integration-based SLiDE .....</b>	<b>49</b>
<b>7.5 Screening of best performing clones in bacteria.....</b>	<b>55</b>
7.5.1 IntDEQSeq screen.....	57
<b>7.6 Screening of best performing clones in human cells.....</b>	<b>60</b>
<b>7.7 Competition assay .....</b>	<b>64</b>
<b>7.8 Mutational analysis of the clones .....</b>	<b>66</b>
<b>PART III DISSCUSSION .....</b>	<b>68</b>
<b>CHAPTER 8 NOVEL CRE-TYPE RECOMBINASES FOR MANIPULATION OF THE GENOMES.....</b>	<b>69</b>
<b>8.1 Mining for novel Y-SSRs and other genome editing enzymes .....</b>	<b>69</b>
<b>8.2 Molecular characterization of novel recombinases .....</b>	<b>71</b>
<b>8.3 Limitations.....</b>	<b>73</b>
<b>8.4 What comes next .....</b>	<b>74</b>
<b>CHAPTER 9 NOVEL Y-SSRS AS PLATFORMS FOR EXTENSIVE DEVELOPMENT OF DESIGNER RECOMBINASES .....</b>	<b>77</b>
<b>CHAPTER 10 Y-SSRS AS TOOLS FOR EFFICIENT TARGETED DNA INTEGRATION .....</b>	<b>79</b>
<b>10.1 Caveats of using Y-SSRs for genomic integration .....</b>	<b>81</b>
<b>10.2 Future prospects of DNA integration via voxH9 target sites .....</b>	<b>83</b>
<b>CHAPTER 11 CONCLUSION .....</b>	<b>85</b>
<b>PART IV MATERIAL AND METHODS .....</b>	<b>86</b>

CHAPTER 12 MATERIALS.....	87
12.1 Organisms .....	87
12.2 Synthetic oligonucleotides .....	87
12.3 Molecular biological products.....	88
CHAPTER 13 METHODS .....	90
13.1 Bacterial growth conditions .....	90
13.2 Recombinant DNA techniques .....	90
13.2.1 DNA purification.....	90
13.2.2 High-fidelity polymerase chain reaction (PCR) .....	90
13.2.3 Restriction enzyme digestion .....	92
13.2.4 Gel electrophoresis.....	92
13.2.5 Ligation .....	92
13.2.6 Transformation .....	93
13.2.7 Sequencing.....	93
13.2.8 Test digest.....	94
13.3 Bioinformatics.....	94
13.3.1 Identification of putative recombinases and native target sites .....	94
13.3.2 Identification of pseudo-ox sites in human and mouse genomes .....	95
13.4 Plasmids .....	95
13.5 Plasmid construction .....	96
13.5.1 Expression of recombinases .....	96
13.5.2 Recombination reporters .....	97
13.6 Cell culture .....	98
13.6.1 Cell culture maintenance.....	98
13.6.2 Fluorescent activated cell analysis.....	99
13.6.3 Plasmid transfection .....	99
13.6.4 Lentiviral particle production.....	100
13.6.5 Generation of landing pad cell lines .....	101
13.7 Biochemistry .....	101
13.7.1 Preparation of the protein extract.....	101
13.7.2 SDS-PAGE .....	101
13.7.3 Western blotting.....	102
13.8 Recombination assays.....	102
13.8.1 Recombination assay based on pEVO vector- excision .....	102

13.8.2 Recombination assay based on pEVO vector- integration assay .....	103
13.8.3 Mammalian recombination assays .....	103
<b>13.9 Cross-recombination assay: Nanopore sequencing. ....</b>	<b>105</b>
<b>13.10 Overexpression studies in mammalian cells .....</b>	<b>106</b>
<b>13.11 Substrate-linked directed evolution (SLiDE) .....</b>	<b>106</b>
<b>13.12 Integration-based SLiDE .....</b>	<b>109</b>
<b>13.13 PCR-based genomic integration detection .....</b>	<b>110</b>
<b>13.14 IntDEQSeq screen.....</b>	<b>110</b>
1.1.1 UMI fragment preparation .....	110
13.14.1 Recombinase variant barcoding and integration assay .....	110
13.14.2 Nanopore sequencing and processing of screen libraries .....	111
<b>13.15 Competition assay .....</b>	<b>112</b>
<b>PART V APPENDIX .....</b>	<b>113</b>
<b>SUPPLEMENTARY FIGURES .....</b>	<b>114</b>
<b>SUPPLEMENTARY TABLES.....</b>	<b>126</b>
<b>SUMMARY.....</b>	<b>144</b>
<b>ZUSAMMENFASUNG.....</b>	<b>146</b>
<b>BIBLIOGRAPHY .....</b>	<b>148</b>
<b>ACKNOWLEDGMENTS.....</b>	<b>160</b>

## LIST OF FIGURES

Figure 1. Schematic representation of the Y-SSR recombination mechanism. ....	6
Figure 2. The Cre//loxP system. ....	7
Figure 3. lox-site variants for promoting integration. ....	10
Figure 4. Schematic overview of the process of directed molecular evolution. ....	15
Figure 5. Substrate linked directed evolution (SLiDE). ....	19
Figure 6. Stepwise evolution of recombinases and its optimization. ....	21
Figure 7. Bioinformatical mining of tyrosine recombinases and their target sites. ....	28
Figure 8. Experimental validation of the chosen candidates. ....	30
Figure 9. Novel recombinases' activity as a function of arabinose. ....	31
Figure 10. Expression differences of novel Y-SSRs in <i>E. coli</i> . ....	32
Figure 11. Profiling target-site selectivity of Cre-like recombinases. ....	34
Figure 12. Activity of newly discovered SSRs in mammalian cells. ....	37
Figure 13. Effect on cell growth upon overexpression of Y-SSRs in mammalian cells. ....	38
Figure 14. Directed molecular evolution of YR9 recombinase on <i>lox9</i> . ....	39
Figure 15. Characterization of evolved YR9 variants. ....	40
Figure 16. Prediction of pseudo-recombination target sites in human genome. ....	42
Figure 17. Vika activity on pseudo-vox site (voxH9) and design of the lock-in integration strategy. ....	45
Figure 18. Vika wt can integrate donor DNA into human endogenous <i>voxH9</i> locus. ....	46
Figure 19. Establishment of the Vika libraries. ....	48
Figure 20. Integration based assay. ....	51
Figure 21. Integration-based Substrate-linked directed evolution (IntSLiDE). ....	52
Figure 22. The IntSLiDE evolution progress. ....	53
Figure 23. Quantification of recombinase activity of randomly picked clones. ....	56
Figure 24. Integration-based DNA Editing Quantification Sequencing (IntDEQSeq) screen workflow. ....	58
Figure 25. The results of IntDEQSeq screen. ....	59
Figure 26. Validation of the IntDEQSeq clones. ....	60
Figure 27. Clone screening in human cells. ....	61
Figure 28. Testing the landing pad construct with wild type Vika. ....	62
Figure 29. Screening of the libraries with the landing pad integration assay. ....	63
Figure 30. Validation of the best performing clones in the landing pad integration assay. ....	64
Figure 31. Competition assay. ....	65
Figure 32. Mutational analysis of the clones. ....	67
Supplementary Figure S1. Clustal Omega amino acid sequence alignment of the seventeen recombinases chosen to be experimentally validated. ....	115
Supplementary Figure S2. Plasmid map of pEVO recombination reporter. ....	115
Supplementary Figure S3. Profiling target site selectivity of Y-SSRs. ....	116
Supplementary Figure S4. Gating strategy for recombination assay in HEK293T cell line. ....	117
Supplementary Figure S5. Prediction of pseudo-recombination target sites in the human genome. ....	119
Supplementary Figure S6. Prediction of pseudo-recombination target sites in the mouse genome. ....	120
Supplementary Figure S7. Mutational analysis of Vika libraries. ....	121
Supplementary Figure S8. Plasmid maps of donor and expression vectors used to test integration into endogenous <i>voxH9</i> site in human genome. ....	122
Supplementary Figure S9. Plasmid maps of donor and host vectors for integration assay in <i>E. coli</i> . ....	122
Supplementary Figure S10. Plasmid maps of vectors used for screening of the libraries with the landing pad integration assay. ....	123



Supplementary Figure S11. Sanger sequencing of 5' and 3' junctions..... 124  
Supplementary Figure S12. Gating strategy for confirming integration into landing pad locus.  
..... 125  
Supplementary Figure S13. PCR and cloning scheme of active clone pulls for testing the  
individual clones..... 125

## LIST OF TABLES

Table 1. Mutant variants of <i>loxP</i> site. ....	9
Table 2. Mutations observed in consensus sequencing of the libraries. ....	54
Table 3. Bacterial strains and cell lines used in this work.....	87
Table 4. Molecular biological products and materials used in this work.....	88
Table 5. Reaction mixture for Herculase II Fusion high-fidelity DNA polymerase chain reaction. ....	90
Table 6. Reaction mixture for Q5 high-fidelity DNA polymerase chain reaction. ....	91
Table 7. Different reaction mixtures used for restriction enzyme digestion.....	92
Table 8. Reaction mixture of a typical ligation reaction. ....	93
Table 9. Reaction mixture for pEVO-target PCR. ....	96
Table 10. Pipetting scheme for splitting cells for different formats. ....	99
Table 11. Pipetting scheme for plasmid transfections for different formats and reagents. ...	100
Table 12. Reaction mixture to prepare DNA solution for virus production in a 10 cm dish..	100
Table 13. Reaction mixture for error-prone PCR to generate the starting library.....	106
Table 14. Reaction mixture for evolution PCR. ....	108
Supplementary Table S1. Plasmids used in this work. ....	126
Supplementary Table S2. Primers used in this work. ....	132
Supplementary Table S3. Y-SSR candidates chosen for validation.....	137

## LIST OF ABBREVIATIONS

°C	Degree Celsius
µg	Microgram
µl	Microliter
µM	Micromolar
Amp	Ampicillin
Ara	Arabinose
BFP	Blue fluorescent protein
BLAST	Basic Local Alignment Search Tool
bp	Base pair(s)
Cm	Chloramphenicol
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
Cas9	CRISPR-associated protein 9
DEQSeq	DNA editing quantification sequencing
DMEM	Dulbecco's modified Eagle's medium
DNA	Deoxyribonucleic acid
dNTP	Desoxyribonucleosidtriphosphate
dpt	Days post transfection
dpi	Days post infection
DSB	Double-strand break
DTT	Dithiothreitol
<i>E. coli</i>	<i>Escherichia coli</i>
EDTA	Ethylenediaminetetraacetic acid
FACS	Fluorescence Activated Cell Sorting
fw	Forward (primer)
GFP	Green fluorescent protein
HDR	Homology-directed repair
HEK	Human embryonic kidney
Int	λ integrase
IntDEQSeq	Integration-based DNA editing quantification sequencing
Kan	Kanamycin
kb	Kilo base(s)
kDa	Kilodalton
LB	Lysogeny broth
LE	Left element
mCherry	Monomeric red fluorescent protein
ml	Milliliter
mM	Millimolar
MOI	Multiplicity of infection
ng	Nanogram
NHEJ	Nonhomologous end joining
NLS	Nuclear localization signal
nM	Nanomolar
OD	Optical density
ON	Over Night
ORF	Open Reading Frame
PBS	Phosphate-buffered saline
PCR	Polymerase chain reaction
PEI	Polyethylenimine
Puro	Puromycin
RE	Right element
rev	Reverse (primer)

RMCE	Recombinase-mediated cassette exchange
rpm	Revolutions per minute
RT	Room temperature
SDS	Sodium dodecyl sulfate
SLiDE	Substrate-linked directed evolution
SSR	Site-specific recombinase
S-SSR	Serine site-specific recombinase
TALEN	Transcription activator-like effector nuclease
TBE	TRIS-Borat-EDTA
U	Unit
WT	Wild type
Y-SSR	Tyrosine site-specific recombinase
ZFN	Zinc-finger nuclease

PART I

# INTRODUCTION

### Chapter 1 ENGINEERING GENOMES

In the post-genomic era, the field of genome engineering has witnessed remarkable advancements, playing a pivotal role in the development of modern biology and the biotechnology industry. The introduction of foreign DNA into mammalian cells has enabled numerous applications, including reverse genetics, which has significantly improved our understanding of genotype-to-phenotype relationships through strategies such as insertional mutagenesis (Hardy et al., 2010). Furthermore, the genetic modification of mammalian cells, has facilitated the production of therapeutic or industry relevant proteins, with a global market projected to reach \$389 billion by 2024 (O’Flaherty et al., 2020). Moreover, the first gene therapy treatment was approved in 2017 by the U.S. Food and Drug Administration (FDA), in order to restore vision for patients with an inherited form of blindness (Mullard, 2018). All these applications involve inserting custom-designed foreign DNA segments into mammalian genomes to disrupt (e.g., mutagenesis), expand (e.g., therapeutics production), or restore (e.g., gene therapy) natural cellular functions (Zhang et al., 2021).

The next stage of genome engineering, genome editing, allows for sequence-specific modifications of intracellular DNA, enabling sophisticated biological strategies. These *in vivo* modifications include correcting genetic constructs (e.g. re-excision of selectable markers), application-specific DNA rearrangements, like controllable switching on/off the expression of a gene (Saunders, 2010), targeted mutations and gene tagging. Nuclease-based systems and site-specific recombinases, which function as molecular scissors, primarily carry out these functions, cutting and, in the case of recombinases, re-ligating double-stranded DNA with high specificity.

Nuclease-based approaches are focused primarily on programmable nucleases such as meganucleases, zinc-finger nucleases (ZFNs), transcription activator-like effector nucleases (TALENs), and more recently the CRISPR/Cas9 (Clustered Regularly Interspaced Short Palindromic Repeats/CRISPR-associated protein 9) system (Christian et al., 2010; Urnov et al., 2010; Jinek et al., 2012). All these systems rely on sequence-specific double-strand breaks (DSBs) stimulating nonhomologous end joining (NHEJ) or homology-directed repair (HDR) mechanisms at specific loci via cellular DNA repair pathways (Zhang et al., 2021). Distinct from the other nucleases, the simplicity and speed in design and use of the CRISPR/Cas9 system as well as its comparable low costs revolutionized the genome engineering field. However, some challenges limit the utility of these technologies. While the repair pathway through NHEJ has much lower fidelity and constitutes a more error-prone repair pathway, HDR provides greater accuracy and is therefore preferred. Yet, HDR is

inefficient in mammalian cells and mainly active during DNA replication, thus, usually outnumbered by the error-prone NHEJ events (Zhang et al., 2021). Development of next generation CRISPR -based tools such as Base editors and Prime editors, is facilitating the precise introduction of targeted point mutations without requiring double-strand breaks (DSBs) or relying on homology-directed repair (HDR) (Komor et al., 2016; Gaudelli et al., 2017; Anzalone et al., 2019).

Alternative tools that hold a great promise to address these limitations and expand the genome engineering field are site-specific recombinases (SSRs). Compared to nucleases, SSRs are able to excise, integrate, invert and exchange genomic sequences through autonomous cleavage and re-ligation, thus, circumventing an accumulation of potential toxic DNA double-strand breaks (Meinke et al., 2016). As SSRs operate independently from the cell's own repair machinery, unwanted sequence alterations such as indels are avoided, and the enzymes remain functional in virtually any cell type and stage of the cell cycle. Consequently, SSR-based editing offers a precise, more predictable, and error-free alternative for genome engineering.

## Chapter 2 SITE-SPECIFIC RECOMBINASES

**S**ite-specific recombination encompasses a range of specialized DNA rearrangements involving defined DNA sites, such as excision, integration, deletion, inversion, and translocation. In nature, these processes serve various biological purposes, playing crucial roles in the life cycles of bacteria, bacteriophages, archaea, and yeast. Functions include the excision and integration of phage and viral genomes (Enquist et al., 1979; Groth and Calos, 2004), the transposition of mobile elements, and the equal segregation of phage, plasmid, and bacterial genomic DNA (Austin et al., 1981; Broach et al., 1982; Rice et al., 2010). Additionally, some recombinases are involved in essential regulatory processes in bacteria, such as sporulation in *Bacillus* and heterocyst differentiation in *Anabaena* (Sato et al., 1990; Kolb, 2002). Site-specific recombination always involves two DNA recognition sites and two SSR dimers responsible for strand recognition, DNA breakage, strand exchange, and reunion.

SSRs can be classified into two structurally and mechanistically distinct classes—tyrosine and serine recombinases—based on the amino acid residue of their catalytic domain, which seem to have evolved separately (Meinke et al., 2016; Olorunniji et al., 2016). Tyrosine recombinases cut and rejoin two DNA strands individually, introducing single-strand breaks and generating a Holliday junction intermediate. Conversely, serine recombinases create simultaneous double-strand breaks (DSBs) at each cleaving site before strand exchange and re-ligation. While most tyrosine recombinase reactions are reversible, serine recombinases catalyze their reactions directionally.

### 2.1 Tyrosine site-specific recombinases (Y-SSRs)

Over the last decades, various Y-SSRs from all three biological domains have been identified and applied in genome engineering. What they all have in common is their catalytic domain with a well-conserved protein fold and recognizable sequence motifs, the highly conserved tyrosine nucleophile coupled with a trio of basic amino acids, forming the arginine-histidine-arginine triad (Esposito and Scocca, 1997; Nunes-Düby et al., 1998). These residues are crucial for optimal recombination activity. In most recombinases, the catalytic domain is further preceded by a variable N-terminal domain that contributes to DNA binding (Grindley et al., 2006). Tyrosine-type recombinases comprise proteins such as the  $\lambda$  integrase from bacteriophage lambda (Landy, 1989), bacterial XerC and XerD recombinases (Sherratt et al., 1995), Cre recombinase from bacteriophage P1 (Sternberg and Hamilton, 1981), and FLP recombinase from *Saccharomyces cerevisiae* (Sadowski, 1995) and several additional FLP-related recombinases.



## SITE-SPECIFIC RECOMBINASES

Many site-specific recombinases (SSRs) necessitate additional host factors for efficient catalysis or directionality, constraining their use in heterologous hosts for *in vivo* applications. However, certain SSRs, such as Cre, FLP, and a few others from the tyrosine family, can recombine their targets effectively without any accessory proteins, making them suitable for DNA rearrangement in living systems (Kolb, 2002).

Among SSRs used for genome engineering, the tyrosine recombinases Cre and FLP are highly popular. The Cre/*loxP* and Flp/FRT systems have proven their utility in animal cells, carrying 34 bp target sites *loxP* and FRT, respectively. Considering these enzymes display distinct enzymatic properties (Buchholz et al., 1996), Cre has demonstrated greater activity in mammalian systems like mice and human cells, whereas Flp was mostly used in yeast models. Consequently, the Cre/*loxP* system remains the most widely employed method for rapid and efficient recombination.

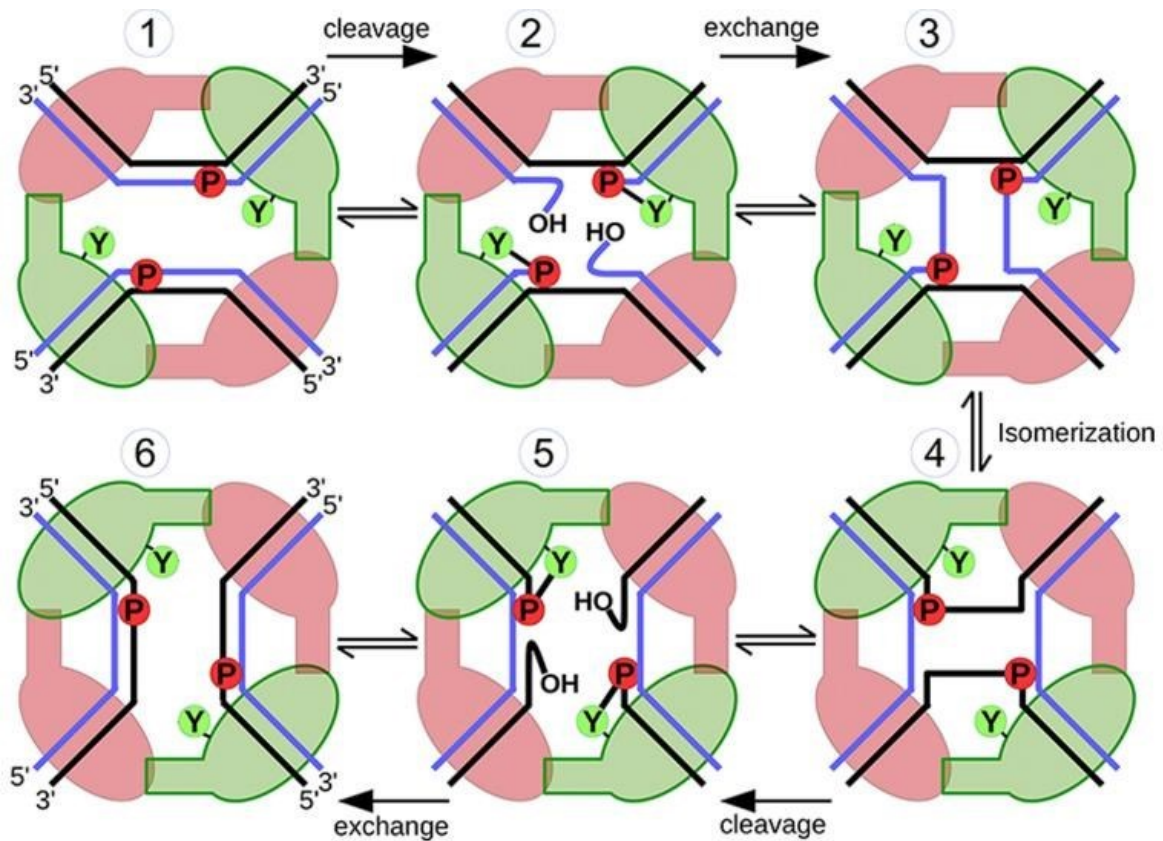
### 2.1.1 Mechanism of tyrosine-based site-specific recombination

The recombination mechanism of Y-SSRs has been elucidated through structural and biochemical analyses, mostly of the Cre/*loxP* recombinase system (Figure 1). Recombinase binds to the *lox* sequence which can be from 30-200 bp long, and consists of two palindromic sequences called the half-sites or binding elements, flanking a spacer region. The process begins with synapsis, where two recombinase dimers bind to one target site, forming a tetrameric complex with two opposing monomers in an active conformation. A conserved tyrosine residue of the active recombinase monomers cleaves one DNA strand in a nucleophilic attack on a phosphate of the spacer region, creating a 3' phosphotyrosine intermediate and a free 5'-hydroxyl group.

Following strand exchange, where the free 5'-hydroxyl attacks the phosphotyrosine linkage of the opposing duplex, a Holliday-junction-like structure is established. The energy freed from the breakage of the phospho-tyrosine bond causes the isomerization of the whole complex, allowing inactive monomers to adopt active conformations. This process repeats for the remaining strands, resulting in recombined DNA products. Notably, the entire process doesn't require external energy-rich cofactors such as ATP, as the energy of the phosphodiester bonds is preserved within the synapse (Duyne, 2001).

The site-specific recombination mechanism has been defined by crystal structures of Cre-*lox* complexes (Guo et al., 1997; Gopaul and Duyne, 1999; Ennifar et al., 2003). These structures reveal that DNA is sharply bent at the spacer region's end, with the bend direction determining which DNA duplex strands are cleaved first (Guo et al., 1999).

## SITE-SPECIFIC RECOMBINASES



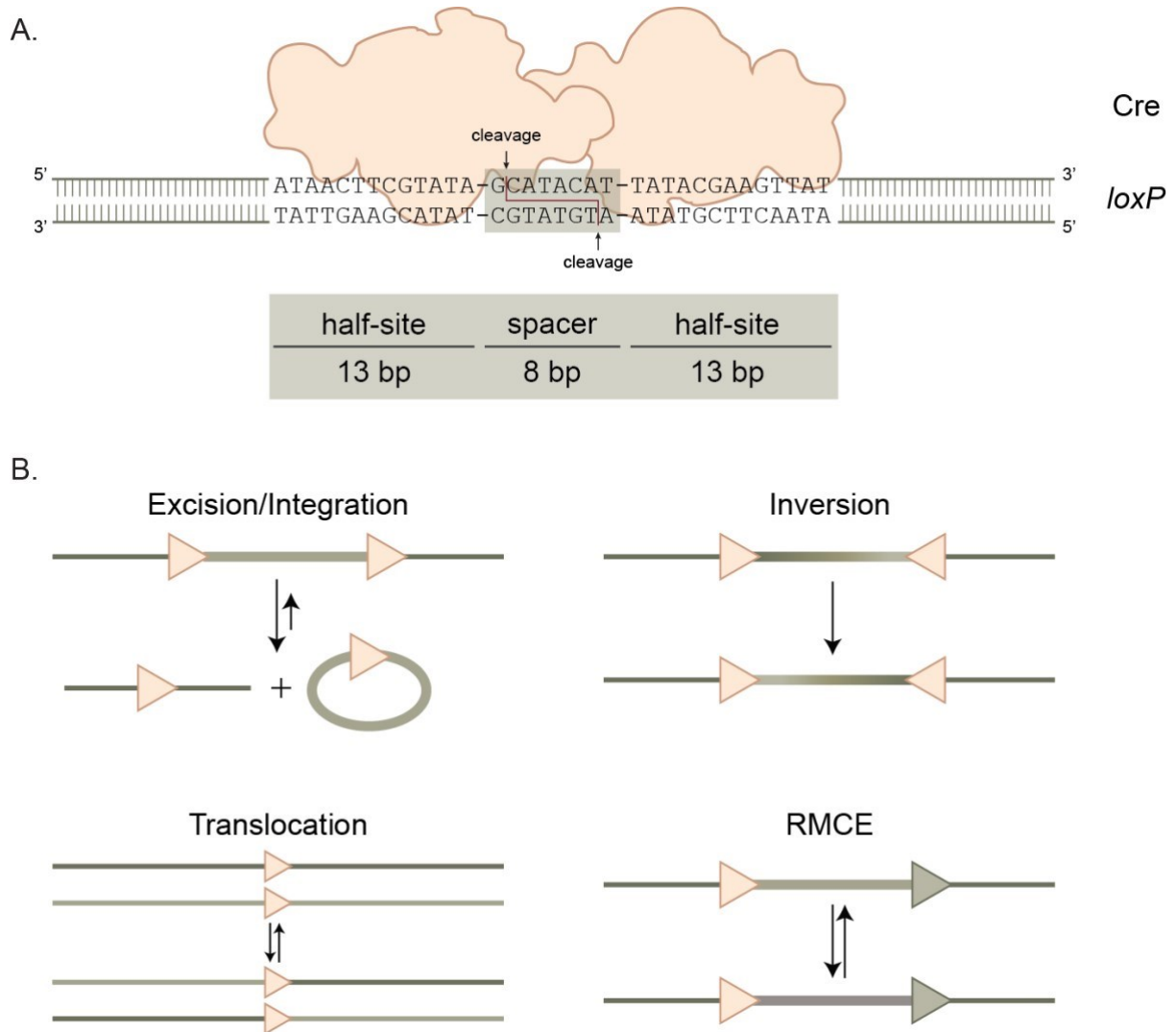
**Figure 1. Schematic representation of the Y-SSR recombination mechanism.**

1) Synapsis: Formation of a tetrameric complex of four recombinases and two target sites. DNA strands are shown as blue and black lines, respectively. Recombinase monomers in active (cleaving) conformation are depicted in green whereas inactive (non-cleaving) monomers are depicted in pink. 2) Cleavage: nucleophilic attack of a tyrosine residue within the cleaving monomers on the spacer, leaving both a 3'-phosphotyrosine and a free 5'-hydroxyl group. 3) Strand exchange: Formation of a Holliday Junction intermediate through interaction between the 5'-hydroxyl group and the respective opposing 3'-phosphotyrosine. 4) Isomerization: Switch from active to inactive conformation and vice versa. 5) Second cleavage: The now active monomers can perform the second round of cleavage. New 3'-phosphotyrosine complexes and free 5'-hydroxyl groups are formed. 6) Second strand exchange: Attack of each free 5'-hydroxyl group on the opposing 3'-phosphotyrosine. Resolution of the final complex. Figure from Meinke et al. 2016 (Meinke et al., 2016).

### 2.1.2 The Cre/*loxP* system

The Cre recombinase, a 38 kDa protein, is encoded by bacteriophage P1. It is thought to play a role in the P1 life cycle by resolving dimeric chromosomes formed after DNA replication into monomeric P1 DNAs and facilitating the circularization of the linear genome (Sternberg and Hamilton, 1981). The DNA sequence where Cre recombinase binds and initiates strand exchange is called *loxP* (locus of crossover in P1). The *loxP* site is a 34 bp palindromic sequence consisting of two inverted repeats (also known as half-sites) that surround a central 8 bp strand exchange region, referred to as the spacer (Figure 2A).

## SITE-SPECIFIC RECOMBINASES



**Figure 2. The Cre/*loxP* system.**

**A.** Schematic representation of Cre recombinase (causes recombination), which mediates site-specific recombination between DNA sequences called *loxP* (locus of crossover (x) in P1) consisting of an 8 bp spacer flanked by two 13 bp palindromic half sites. **B.** Different outcomes of site-specific recombination depending on the position and orientation of the target sites. Arrows indicate the reversibility of the reactions with the larger size of the arrow indicating the kinetically favored excision reaction. A triangle represents one target site.

For the Cre/*loxP* system to execute the recombination reaction, a pair of *loxP* sequences is required. The highly coordinated process involves two recombinase dimers and two target sites, with each dimer binding to a target site and each monomer recognizing a palindromic half-site. These four molecules form a tetrameric complex that catalyzes DNA cleavage, strand exchange, and rejoining at the 5' boundaries of both spacers. The asymmetric spacer provides cleavage sites and directionality to the target site, determining whether integration, deletion, inversion, or translocation occurs based on the relative orientation and position of the target sites. Recombinase-mediated cassette exchange (RMCE) allows for DNA sequence which is flanked by two distinct target sites that do not recombine with each other,

to be exchanged with another DNA sequence that is likewise flanked by the same target sites (Meinke et al., 2016) (Figure 2B). Although reactions are reversible, recombination between closely linked sites is kinetically favored over more distant sites, such as those on different DNA molecules. The three structural outcomes, provided by the simple system of one enzyme and its relatively short DNA target, have found a wide variety of applications.

### 2.1.3 *loxP* variants for broadening the Cre/*loxP* utility

The excision reaction is more kinetically favorable than the insertion reaction, which makes it relatively easy to engineer gene deletion or inactivation experiments by surrounding the target sequence with *loxP* sites. The challenge in achieving DNA insertion is that, after the insertion reaction, two *loxP* sites remain in cis, and can then serve as substrates for Cre and cause the rapid removal of the inserted segment (Figure 3A).

Numerous studies have explored the identity of the *loxP* sequence, demonstrating that by modifying the *wild-type loxP*, efficient Cre recombination can still be achieved (Hoess et al., 1986; Albert et al., 1995; Thomson et al., 2003; Missirlis et al., 2006). Two categories of variant *loxP* sites have been identified that promote stable Cre-*loxP* integrative recombination (Missirlis et al. 2006). Both types rely on sequence mutations in the Cre recognition sequence, either within the 8 bp spacer region or the 13 bp inverted repeats.

Spacer mutants, such as *lox511* (Hoess et al., 1986), *lox5171*, *lox2272* (Lee and Saito, 1998), *m2*, *m3*, *m7*, and *m11* (Langer et al., 2002) can recombine with themselves easily but display a significantly reduced rate of recombination with the *wild-type* site (Table 1). This group of mutants has been utilized for DNA insertion through Recombinase Mediated Cassette Exchange (RMCE) (Seibler & Bode 1997; Schlake & Bode 1994).

## SITE-SPECIFIC RECOMBINASES

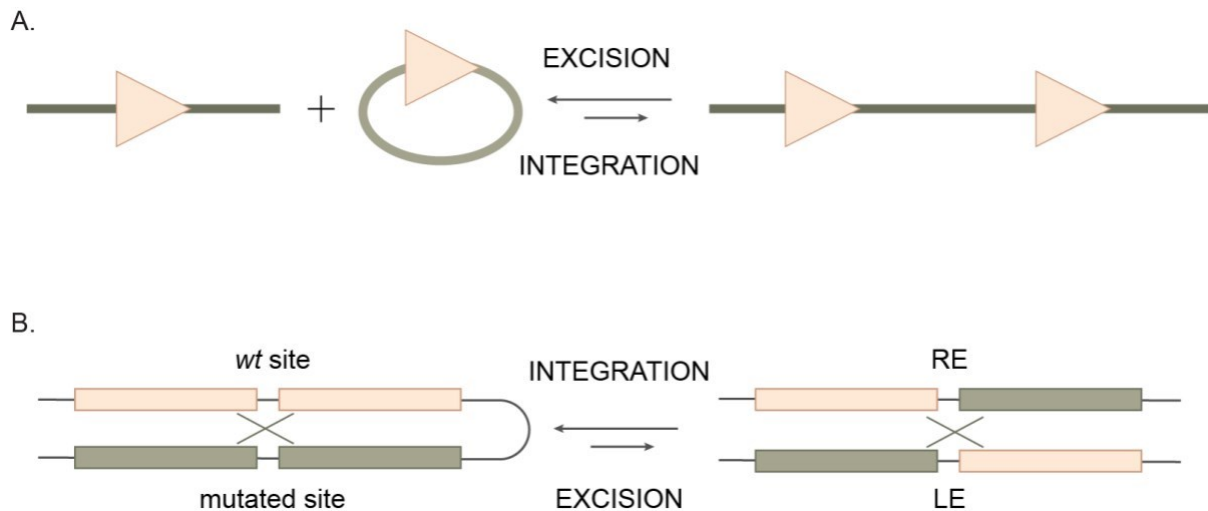
**Table 1. Mutant variants of *loxP* site.**

The mutations in the sites compared to the *wt loxP* are written in lower case

<b><i>loxP</i> site</b>	<b>Left half site</b>	<b>Spacer (5' - 3')</b>	<b>Right half site</b>
<i>Wild-type</i>	ATAACTTCGTATA	ATGTATGC	TATACGAAGTTAT
<i>lox511</i>	ATAACTTCGTATA	ATGTATaC	TATACGAAGTTAT
<i>lox5171</i>	ATAACTTCGTATA	ATGTgTaC	TATACGAAGTTAT
<i>lox2272</i>	ATAACTTCGTATA	AaGTATcC	TATACGAAGTTAT
<i>m2</i>	ATAACTTCGTATA	AgaaAcca	TATACGAAGTTAT
<i>m3</i>	ATAACTTCGTATA	taaTAcca	TATACGAAGTTAT
<i>m7</i>	ATAACTTCGTATA	AgaTAgaa	TATACGAAGTTAT
<i>m11</i>	ATAACTTCGTATA	cgaTAcca	TATACGAAGTTAT
<i>lox71</i>	taccgTTCGTATA	ATGTATGC	TATACGAAGTTAT
<i>lox66</i>	ATAACTTCGTATA	ATGTATGC	TATACGAAcggtA

The second class of mutants, inverted repeat mutants, involves altered bases in the left inverted repeat (LE mutant) or the right inverted repeat (RE mutant). The LE mutant, *lox71*, contains 5 bp at the 5' end of the left inverted repeat, which are changed from the wild-type sequence to TACCG (Albert et al., 1995; Araki et al., 1997). Similarly, the RE mutant, *lox66*, has the five bases at the 3' end altered to CGGTA. Inverted repeat mutants are employed for integrating plasmid inserts into chromosomal DNA, with the LE mutant acting as the "target" chromosomal *loxP* site into which the "donor" RE mutant recombines. After recombination, *loxP* sites are situated in cis, flanking the inserted segment. The recombination mechanism ensures that one *loxP* site post-recombination is a double mutant (with both LE and RE inverted repeat mutations), while the other is wild-type (Van Duyne 2001). The double mutant is sufficiently distinct from the wild-type site that Cre recombinase does not recognize it, and the inserted segment remains unexcised (Figure 3B). Moreover, combining spacer and inverted repeat mutants has been shown to enhance the specificity and stability of integrative recombination (Araki et al., 2002).

## SITE-SPECIFIC RECOMBINASES



**Figure 3. *lox*-site variants for promoting integration.**

**A.** Schematic representation of excision/integration reversible reaction. The excision reaction is kinetically more favorable, as depicted with the larger arrow. Each triangle represents a whole target site. **B.** Schematic depiction of the RE/LE strategy. Light colored squares represent wild type half sites, whereas darker shaded squares are the mutated half sites. The shift in the preferred directionality is depicted with the larger size of the arrow pointing in the direction to the integration outcome. RE – right element, LE – left element.

### 2.1.4 Other tyrosine recombinases

Cre/*loxP* and FLP/FRT systems have become essential tools for DNA and genome engineering since their discovery in 1981 (Sternberg and Hamilton, 1981) and 1995 (Sadowski, 1995), respectively. The Cre/*loxP* system quickly gained widespread use in genetic engineering, followed by the FLP/FRT system. Over the years, researchers have identified additional members of the Cre-like tyrosine family recombinases, such as Dre/*rox* in 2004 (Sauer and McDermott, 2004), which was found by sequencing the homologous region immC of P1 phage in related bacteriophage D6. Characterization of Dre showed that it was a tyrosine recombinase closely related to the P1 Cre recombinase, but that it had a distinct DNA specificity for a 32 bp DNA site (*rox*). Dre was later shown to be highly efficient in mammalian cells and mice (Anastassiadis et al., 2009). The emergence of the collection of sequences and annotated genomes, enabled rational prediction based on the amino acid alignment of the proteins with unknown functions and Cre. At first, this led to the identification of VCre and SCre recombinases, from the *Vibrio* sp. 0908 plasmid p0908 and *Shewanella* sp. ANA-3, respectively. Both enzymes have unique target sites and are offering comparable efficiency to the Cre recombinase without cross-recombination with Cre/*loxP* (Suzuki and Nakayama, 2011). More recently, Karimova et al. added to the toolbox three novel Cre-like Y-SSR systems: Vika/*vox* (Karimova et al., 2012), Nigri/*nox* and Panto/*pox* (Karimova et al., 2016) demonstrating that rational analysis of the available protein and DNA

sequences combined with experimental techniques represent a powerful tactic for straightforward discovery.

Several FLP-related recombinases encoded by 2-micron circle-like plasmids of yeasts have also been described, including Arg (Yang and Jayaram, 1994), Kw (Ringrose et al., 1997), KD (Bianchi, 1992), B2, B3, and R (Onouchi et al., 1991)). While these systems were discovered some time ago, they have not gained wide utility in engineering genomes of higher organisms, likely due to their relatively low recombination efficiencies. Additionally, the temperature optimum of 30°C, dictated by their yeast origin, may limit the utility of these FLP-like recombinases. Nonetheless, FLP recombinase has undergone several modifications to improve its intrinsic properties, showcasing the potential for these systems in genome engineering (Buchholz et al., 1998; Raymond and Soriano, 2007).

### 2.2 Application of SSRs

Both site-specific recombinase families have garnered significant attention in recent years due to their diverse applications in genome engineering and synthetic biology. These versatile enzymes are characterized by their natural roles, which primarily involve functioning as phage integrases, resolvases or invertases, making them excellent tools for specific DNA recombination. On one side, applicative potential of serine recombinases (S-SSRs) lies in the unidirectional nature of their target site architecture. This feature allows these multifaceted enzymes to execute diverse reactions such as integration, excision, and inversion with exceptional specificity and control. In this context, serine SSRs have been explored for a wide range of applications, driving innovation and expanding our understanding of biological systems.

The use of serine recombinases in genome engineering and synthetic biology has been well-documented in the literature. In a study by Grindley et al. (2006), the authors demonstrated how serine recombinases could be utilized for site-specific genome modifications, enabling precise gene insertion, deletion, or inversion. More recently, researchers have turned to serine recombinases to engineer synthetic genetic circuits, by creating a library of logic gates and memory devices in bacterial cells (Bonnet et al., 2013). Following up to that, Siuti et al. (2013) used serine recombinases to design synthetic circuits that integrated logic and memory functions in living cells. Most recently, Ba et al. (2022) developed SYMBIOSIS, a synthetic biology toolkit that leverages orthogonal serine integrase systems to create manipulable bio bricks. This approach enables researchers to efficiently assemble and reconfigure genetic circuits, thus expanding the range of potential applications in synthetic biology.

## SITE-SPECIFIC RECOMBINASES

In addition to the studies previously mentioned, more recent studies by Durrant et al. and Yarnall et al. highlight the potential of serine recombinases as a tool for gene therapy and complex genome engineering tasks, emphasizing the importance of enzyme specificity in these applications. In the study by Durrant and colleagues, machine learning and hierarchical clustering were employed to identify groups of recombinases with distinct integration site preferences. By comparing the integration preferences of these groups, they were able to determine the minimal set of recombinases needed to target over 85% of the human genome. They further demonstrated that these recombinases could integrate large DNA sequences (up to 200 kilobases) without any loss of efficiency. The study also showed that the identified recombinases could be used in primary human cells, including hematopoietic stem cells and T cells, which are commonly targeted in gene therapy applications (Durrant et al., 2022).

The others developed PASTE by engineering a fusion protein consisting of Cas9, reverse transcriptase, and integrase, allowing efficient integration (5-50%) of diverse cargos at specific target locations in the human genome. This versatile tool can be used for gene tagging, replacement, delivery, and protein production and secretion (Yarnall et al., 2022). PASTE combines CRISPR nuclease engineering with the discovery of various serine integrases, enabling efficient, multiplexed transgene integration in dividing, non-dividing cells, and animal models. This platform expands the scope of genome editing and supports new applications in basic biology and therapeutics (Yarnall et al., 2022).

On the other hand, tyrosine recombinases have found wide application in various fields, enabling researchers to gain a deeper understanding of human diseases, cellular processes, and genetic lineage tracing. A multitude of Cre mouse lines have been developed, expressing enzymes in particular cell types via tissue-specific or inducible promoters. This "Cre zoo" allows for spatial and cell type-specific mutagenesis, as well as inducible knockouts, contributing to studies of embryonic development and lineage tracing in fields such as developmental biology, neuroscience, immunology, and stem cell and regeneration biology.

One of the innovative approaches in this area is the use of orthogonal recombinases, which enable researchers to decipher cell fate with enhanced precision (Weng et al., 2021). In recent years, the development of genetic lineage tracing techniques using multiple DNA recombinases has allowed for more precise cell fate mapping studies (Liu et al., 2020; Jin et al., 2021). The application of dual recombinases, such as Cre and Dre, has further advanced



## SITE-SPECIFIC RECOMBINASES

this field, allowing for synchronized lineage tracing and cell subset ablation *in vivo* (Wang et al., 2022).

Cre mouse lines also serve as valuable tools for modeling and studying human diseases in mice (Justice et al., 2011). For instance, a mouse model using the *Cre/loxP* strategy was developed to investigate the role of the human breast-tumor suppressor gene *Brca1*, which proved essential in understanding its role in mammary gland development and neoplasia (Xu et al., 1999).

Y-SSRs have found applications in synthetic biology too, where they're used to develop recombinase-based gene circuits for complex spatiotemporal regulation of gene expression in bacteria (Siuti et al., 2013), mammalian cells (Weinberg et al., 2017), and even *Arabidopsis* (Lloyd et al., 2022). Spatiotemporal control of recombination was furthermore enabled by developing light-inducible Cre recombinase as a novel optogenetic switch, offering rapid and controlled gene manipulation in response to light stimulation (Duplus-Bottin et al., 2021; Takao et al., 2022).

Finally, recombinases like Cre and Flp have been engineered to recognize a wide range of different target sequences, increasing their versatility. This offers more options for customization and have widen the range of available tools, making them more accessible to researchers. Engineering successes opened the doors for recent efforts to employ recombinases as gene surgery tools, which highlight their potential in future gene therapy applications.

## Chapter 3 PROTEIN ENGINEERING BY DIRECTED EVOLUTION

## 3.1 Principles

Throughout history, humans have been breeding animals and plants with desired properties for thousands of years, even before Charles Darwin's influential theory of evolution emerged in the mid-1800s. This age-old practice has inspired a modern approach called directed molecular evolution, which adapts the principles of natural evolution to rapidly alter the properties of macromolecules in a laboratory setting. In recent years, this method has gained significant recognition, with its pioneers Gregory P. Winter, George P. Smith, and Frances H. Arnold being awarded the Nobel Prize in Chemistry in 2018.

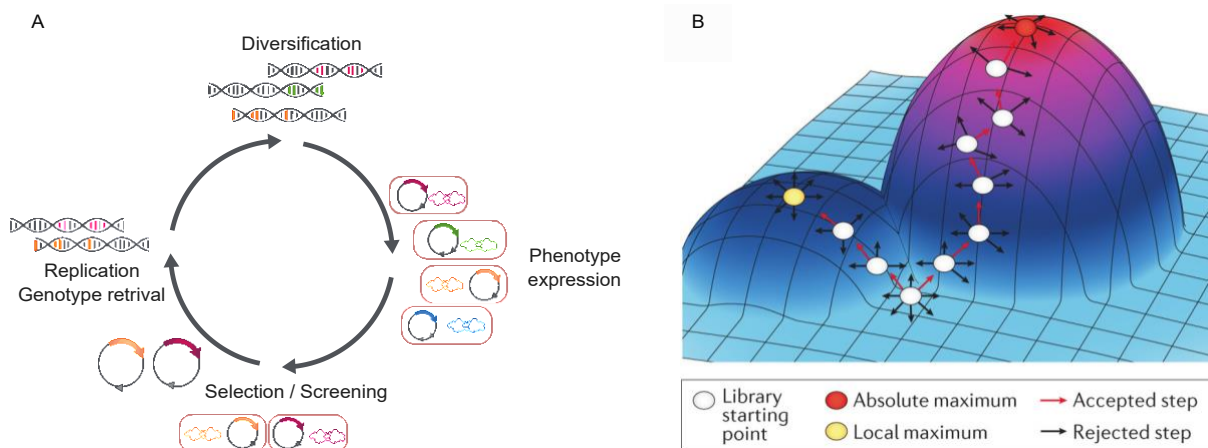
Directed molecular evolution offers a powerful alternative to traditional rational design methods, which require knowledge of structural properties to modify macromolecules. Instead, it allows for the improvement of enzymes or other proteins without detailed knowledge of their structure or function. By employing genetic diversification and natural selection, this approach can increase stability, modify substrate specificity, and adapt proteins to industrial or non-native environments.

This method of directed evolution is highly versatile and can be tailored to a wide range of applications, from optimizing enzymes for industrial-scale production to developing enzymes with completely new features and properties. The experimenter can control various parameters such as selection pressure, mutation rate, and library size to guide the evolution process in a specific direction, typically towards increased activity (Figure 4A) (Packer and Liu, 2015). This "evolutionary walk" can be visualized as navigating a landscape with rare peaks representing the desired enzyme properties. Directed evolution usually starts somewhere in the flat areas of the evolutionary landscape with a library of an enzyme that shows some weak initial activity. During many cycles of evolution, the enzyme libraries 'walk' through the sequence space of the evolutionary landscape and eventually end up on one of the peaks, selecting for the 'fittest' variants (Figure 4B).

Directed evolution can be performed using both *in vivo* and *in vitro* methods, with the first experimental step involving the generation of a library of protein sequence variants. Since the occurrence rate of spontaneous mutations is insufficient for laboratory evolution, genetic diversification techniques are required to generate a gene library harboring different variants (Packer and Liu 2015). Typical methods include random mutagenesis (e.g. through error-prone PCR (Cadwell and Joyce, 1992)), focused mutagenesis to change specific amino acids, and genetic recombination methods such as DNA shuffling (Stemmer, 1994; Cramer

## PROTEIN ENGINEERING BY DIRECTED EVOLUTION

et al., 1998). The subsequent step involves identifying the desired enzyme variants by coupling the genotype to the phenotype using screening-based or selection-based approaches.



**Figure 4. Schematic overview of the process of directed molecular evolution.** **A.** First, a diverse library of DNA fragments is generated (depicted by different colors). The DNA library is translated into a corresponding protein library that is subjected to screening or selection for a desired property. Replication of the respective genes creates a new pool of DNA fragments following the repetition of several rounds until a desired property is achieved. **B.** Schematic depiction of the process of directed evolution through a given sequence space. Each circle presents another generation cycle of the evolution process. The arrows around the circles show the possible trajectories of the evolution. The red arrows show the positive selection guided trajectory of the evolution towards a given maximum. The 'hills' show different maxima of a given protein feature (e.g., activity). Figure from Packer & Liu 2015

Screening-based methods often rely on identifying catalytic activity through fluorescent or colorimetric-based reporters using manual inspection, Fluorescence Activated Cell Sorting (FACS) or microplate readers (Packer and Liu, 2015), and cell surface or phage display methods, while selection-based methods link enzyme activity to survival (e.g. antibiotic resistance or metabolic genes). Screening-based directed evolution can be limited by throughput, as each variant needs to be individually selected and collected. In contrast, selection-based directed evolution depletes non-active variants from the library, allowing the surviving variants to proceed to the next round of evolution.

The selected enzymes are then diversified, enabling the exploration of the next part of the evolutionary sequence landscape. Directed evolution continues until the experimenter is satisfied with the evolved enzyme's activity. Finally, a last step of screening is performed to identify a single newly evolved enzyme that exhibits all the desired features.

In conclusion, directed molecular evolution offers a powerful and versatile approach to engineering enzymes and other proteins with enhanced properties, without requiring detailed knowledge of their structure or function. By leveraging the principles of natural evolution, this

technique enables a wide range of applications across various fields, including industrial-scale production, basic biology, and therapeutic development.

### 3.2 Directed evolution of recombinases

Site-specific recombinases possess the ideal characteristics for genome editing in therapeutic applications. Their small size makes them easy to deliver, and they can precisely modify the genome without requiring cofactors or producing indels. Additionally, they are effective in both mitotic and post-mitotic cells and have been utilized in animal models for over 25 years (Meinke et al., 2016). However, their use in a therapeutic context is significantly limited, as their target site must first be integrated at the desired location within the recipient cell's genome. While feasible in cell lines and animal models, this is not possible in humans. To overcome this, there were several attempts to fuse recombinases with other, more modular, DNA binding domains such as Zinc fingers (Gaj et al., 2013), TALEs (Mercer et al., 2012; Voziyanova et al., 2020) or recently even CRISPR-Cas systems (Chaikind et al., 2016; Standage-Beier et al., 2019) as these domains are easier to program to alter their target site preference. Nevertheless, these efforts mostly had modest results in the terms of efficiency. As a result, the other viable option for employing recombinases therapeutically is to modify their target site specificity to match a naturally occurring sequence within the human genome.

Although the Cre/*loxP* system is the most widely used site-specific recombinase system, and its molecular details are well understood, altering its specificity remains a challenge. This is primarily because the Cre protein lacks a specific DNA recognition domain, which is present in other genome editing enzymes like ZFN or TALENS (Gaj et al., 2011). Studies of the crystal structure of Cre bound to *loxP* have revealed that only few amino acids are in direct contact with DNA. Those known residues contribute to target site specificity, but many more are also involved and their role is often less clear (Abi-Ghanem et al., 2012; Meinke et al., 2016). Non direct contacts and epistatic relationships with different residues modulate the specificity and activity to the high extent. Protein-protein interactions between the recombinase monomers add another layer of complexity. Our current understanding of these mechanisms contributing to the substrate specificity of Cre is insufficient for rational design approaches or targeted mutagenesis to generate novel Cre-like recombinases with altered DNA specificity.

Therefore, in the early 2000s, two research groups pursued the creation of novel Cre-like enzymes using directed evolution approaches (Buchholz and Stewart, 2001; Santoro and Schultz, 2002). In 2007, the first Cre-like enzyme with altered DNA specificity towards a

therapeutically relevant target site was introduced (Sarkar et al., 2007). Building on this achievement, a recombinase called Brec1, capable of excising the HIV provirus from human cells *in vivo*, was developed (Karpinski et al., 2016). As a testament to its success, Brec1 is set to become the first recombinase used in a clinical trial aiming to cure HIV.

In yet another recent work, the heterodimeric recombinase system RecF8 was generated, which is able to correct an inversion of the F8 gene often found in patients suffering from Hemophilia A (Lansing et al., 2022). This work of Lansing and colleagues demonstrated that one can break the constraint of the complex *lox* site architecture when uncoupling the binding of the two half sites. Particularly, it reports for the first time, that there is a possibility to develop heterodimeric recombinases that are then able to recombine asymmetrical target sites if evolved to work together. Furthermore, by linking the two heterospecific monomers (Lansing et al., 2022), or introducing specific mutations to make the two hetero-monomers obligate (Hoersten et al., 2021), they demonstrated high specificity for the asymmetric target site.

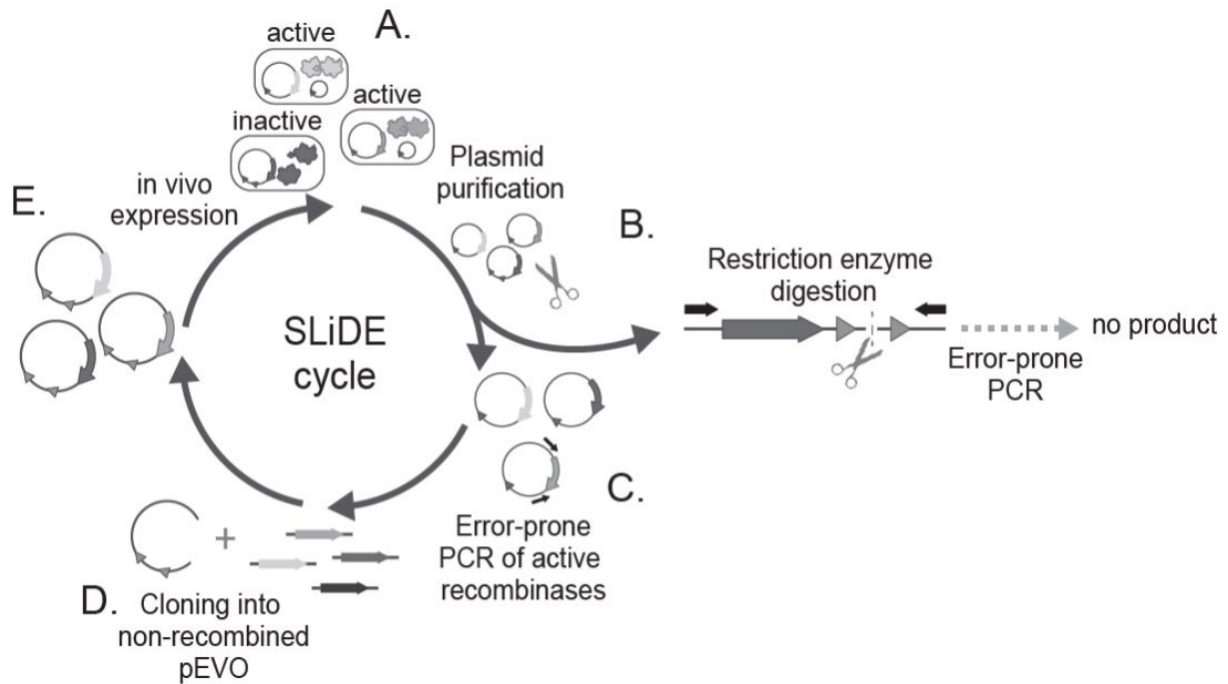
The previously mentioned recombinases were developed using a directed evolution protocol known as Substrate Linked Directed Evolution (SLiDE). This method relies on linking the coding region of the recombinase to its excision recombination substrate (the target site). That way, active recombinases will physically alter their substrate, enabling a simple selection of active variants (Buchholz and Stewart, 2001). Since site-specific recombinases like Cre favor excision over integration (Logie and Stewart, 1995), the protocol is excision-based. The entire process relies on a single plasmid called pEVO, which houses both the recombinase gene and the target sites (Figure 5). Furthermore, the expression of the recombinase gene can be regulated by adding varying amounts of L-arabinose to the bacterial growth medium, allowing for the selection of more active variants throughout the evolution process.

SLiDE begins by introducing a pEVO plasmid containing a library of recombinases and a specified target site into bacteria. The recombinases' expression is induced overnight, and the plasmids are analyzed after purification the following day (Figure 5A). If some recombinases from the library are active on the target site, the isolated plasmids will be a mix of recombined and non-recombined ones. Only the recombined plasmids encode a recombinase gene that, when expressed, recombines the target site. This recombination reaction excises a DNA fragment containing a unique restriction site from the pEVO plasmid. As a result, a restriction digest using the corresponding restriction enzyme will only cut and linearize the non-recombined plasmids (Figure 5B).

## PROTEIN ENGINEERING BY DIRECTED EVOLUTION

Subsequently, a PCR using primers specific for the circular recombined pEVO plasmids amplifies only the active recombinases. This PCR is conducted with a low-fidelity DNA polymerase to introduce diversity into the pool of active recombinase variants (Figure 5C). The diversified active recombinase fragments are then inserted into a non-recombined pEVO backbone (Figure 5D). The transformation of the new library into bacteria and overnight expression initiates the next evolutionary cycle (Figure 5E).

During the SLIDE process, researchers can manipulate various parameters to select recombinases with the desired target site specificity and activity. Consequently, it is crucial to monitor the progress of the evolution. A snapshot of a given recombinase library's activity on a specific target site is taken during each evolutionary cycle. As previously mentioned, the isolated plasmids consist of a mix of non-recombined and recombined ones. The recombined pEVO plasmids are smaller due to the excision of a DNA fragment, and this size difference can be resolved using a standard agarose gel after a restriction digestion. This digest is subsequently referred to as a test digest (Figure 5F).



**Figure 5. Substrate linked directed evolution (SLiDE).**

**A.** A starting library of recombinase genes is cloned in the non-recombined pEVO vector and expression is induced overnight. The pEVO vector contains *loxP*-like target sites as substrates for the recombinase (black triangle). **B.** The selection for active recombinases is based on the restriction digestion of non-recombined plasmids (black scissors). This is allowed by a unique restriction site (small dashed line) located between the *loxP*-like sites. Recombined pEVO plasmids contain an active recombinase gene for the given target sites and are 'immune' to the restriction digest since the site was excised. **C.** Amplification of the gene coding for the active recombinases by error-prone PCR diversifies the active recombinase library. New variants are indicated by different shades. **D.** The diversified library is cloned into a new non-recombined pEVO plasmid backbone. **E.** Transformation of the new library based on active recombinases for a given target sites starts a new cycle of evolution.

At this point, researchers can choose to reduce the expression level or expose the recombinase library to a different target site. Gradually decreasing the recombinase expression allows for the selection of more active variants. Once a particular activity level is reached, as monitored by the test digestion, the recombinases can either be exposed to another target site or the researcher can determine that the evolution is complete.

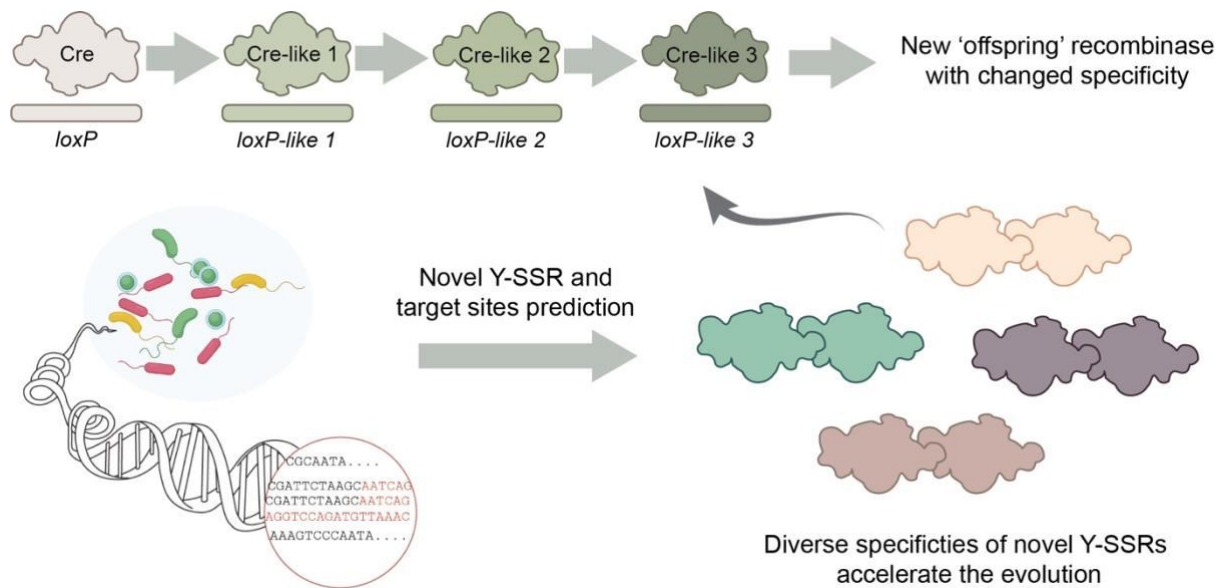
### 3.3 Challenges and optimization of directed evolution protocols for development of designer recombinases

SLiDE and other directed molecular evolution strategies are powerful tools to produce Y-SSRs with desirable properties for research or potential therapeutic applications. So far most of evolution efforts were focused on Cre recombinase. Similar to natural evolution, in order to evolve a Cre-like recombinase to recognize a different target site, the evolution must be guided through intermediate steps, known as subsites, to gradually change Cre's DNA specificity towards any other target site (Figure 6). By employing SLiDE in conjunction with gradual guidance through subsites, it should be possible to evolve Cre-like enzymes with altered DNA specificity for any target site. These novel enzymes can then be utilized to either excise or invert a specific DNA fragment flanked by two *lox*-like target sites naturally occurring in the human genome.

Even though these methods considerably ease the application of Y-SSRs, their development is still very time-consuming and laborious, especially if the novel target sequence differs substantially from the original target site (Meinke et al., 2016). For the generation of Brec1, for example, 145 evolution cycles and 12 subsites were required starting from *Cre/loxP* (Karpinski et al., 2016). Subsequently, optimization of SLiDE protocol and use of different libraries with specificity closer to the new desired target sites contributed to speeding up the design efforts of these Y-SSRs. One way to further accelerate this process is to expand the Y-SSR toolbox with novel Y-SSRs that recognize different target sites preferably having more similarities to the desired target. A collection of various Y-SSRs increases the likelihood of already finding a recombinase with the desired property or a similar recombinase that can serve as a starting point for a new evolution project (Figure 6). Furthermore, extending the range of available Y-SSRs can improve our knowledge of their DNA binding specificity and target recognition (Meinke et al., 2016), ultimately enabling a more rational design of Y-SSRs. Recent advances in sequencing technologies and the availability of already sequenced environmental samples, serve as a vast platform for the discovery of more novel Y-SSR systems.



## PROTEIN ENGINEERING BY DIRECTED EVOLUTION



**Figure 6. Stepwise evolution of recombinases and its optimization.**

Similar as natural evolution, the directed evolution of recombinases needs to be guided through different intermediate steps, so called subsites. These subsites are designed to direct the evolution of recombinase specificity towards the target site of choice, and usually differ in only a few base pairs between each other. In theory, this stepwise process allows to evolve recombinases for any target site. Having a repertoire of related Y-SSRs but with diverse properties and target site specificities can accelerate the evolution process.

## Chapter 4 GENOMIC INTEGRATION – SSRs VS NUCLEASES

Over the years, the field of transgene introduction has seen significant advancements. Exogenous DNA introduction into the genome can be broadly classified into two categories: random or targeted integration. Random integration, often achieved through viral vectors and transposases (e.g. Sleeping Beauty or PiggyBac transposase), has been widely used but presents limitations such as gene silencing and disruption of endogenous gene function. In contrast, targeted integration, which includes techniques like homologous recombination (Mansour et al., 1988), recombineering technology (Muyrers et al., 2001), and recombinase-mediated integration, offers a safer and more controllable approach for mammalian chromosomal integration. However, the low frequency of naturally occurring homologous recombination events limited its application until programmable nucleases, including ZFNs, TALENs, and CRISPR/Cas9, emerged (Capecchi 1989; Bibikova et al. 2001; Cermak et al. 2011; Mali et al. 2013).

The concept of altering mammalian genomes through HR between chromosomal DNA and exogenous foreign DNA was first proposed in the 1980s but was limited by the low frequency of naturally occurring HR events in mammalian cells (Capecchi, 1989). The discovery that meganucleases, such as I-SceI, could dramatically enhance the rate of HR (Rouet et al., 1994) spurred the development of novel, programmable nucleases, including ZFNs (Bibikova et al., 2001), TALENs (Cermak et al., 2011), and the CRISPR/Cas9 system (Mali et al., 2013). By leveraging these programmable nucleases to introduce DSBs at desired genomic loci, tailored donor DNA cassettes can be efficiently inserted through induced cellular repair of the DSB. Recently, natural CRISPR-associated transposases and engineered Cas-domain-fused transposase systems have demonstrated the potential to integrate genomic cargos *in vitro* and into bacterial genomes (Chen and Wang, 2019; Klompe et al., 2019; Strecker et al., 2019). Although CRISPR-targeted transposases have shown promising results *in vitro* and in bacterial cells, their efficacy in mammalian cells remains unreported.

Developing general methods for targeted integration in living cells has been a long-standing challenge in genome editing. While nuclease-mediated HDR can insert genetic payloads, their application is limited to actively dividing cells. Furthermore, another drawback of HDR-based insertion is its sensitivity to the size of the donor, with efficiency dropping as the donor size increases (Li et al., 2014). Moreover, these techniques are relying on cells intrinsic mechanism of DSB repair which can vary dramatically between cell types and introduces some uncertainty of the editing outcome. Site-specific recombinase (SSR)-mediated genome

## GENOMIC INTEGRATION – SSRs VS NUCLEASES

editing represents a complementary approach that typically does not rely on endogenous cellular DNA repair machinery, thus avoiding error-prone repair processes like nonhomologous end joining (NHEJ) (Turan and Bode, 2011; Meinke et al., 2016). Due to their independence from the cell's own repair machinery, undesirable sequence alterations such as indels are avoided and the enzymes can be functional in nearly any cell type and cycle stage (Grindley et al., 2006). Recombinases don't exhibit cargo size limitations, giving them further mechanistic advantages over nuclease-based editing (Durrant et al., 2022). However, the requirement for recognition sites in the genome limits the programmability and flexibility of SSR-based approaches.

## Chapter 5 AIMS

Innovative Cre-like recombinases hold immense potential for both established and emerging genome engineering techniques. The inherent characteristics of tyrosine recombinases can differ, and conducting a thorough analysis of activity levels and compatibility among various SSR systems can be especially advantageous for practical applications. In addition to considering efficiency and cross-specificity, researchers can explore the potential cytotoxicity associated with cryptic genomic recombination sites in the context of recombinase applicability. Consequently, this knowledge could enable the assignment of novel and known recombinases to specific uses or organisms. Furthermore, a comprehensive search for new Cre-type recombinases has not yet been conducted. The availability of a reliable and robust method for obtaining new enzymes could significantly advance the field of genome engineering.

Thus, one of the main aims of my thesis is to identify novel Cre-type recombinases in order to widen the genetic toolbox to meet the growing demand for better genome editing tools. I sought to establish a comprehensive prediction pipeline, combining the rational bioinformatical approach with the knowledge of biological functions of the recombinases, to enable high success rate and high-throughput identification of novel Y-SSRs systems.

Next, I want to molecularly characterize several putative candidates in depth, in order to assure their impact and successful integration in the future genome engineering applications. This means, I intend to define their activity in procaryotes (*E. coli*) and eucaryotes (human cell lines), and to determine their specificity in the sequence space of all known Cre-type target sites to ensure that the novel candidates contribute to the toolbox with the target sites distant enough to allow simultaneous use of multiple Y-SSR systems.

In addition to considering efficiency and cross-specificity, I want to explore the potential cytotoxicity associated with cryptic genomic recombination sites in the context of recombinase applicability. Consequently, this knowledge could enable the assignment of novel and known recombinases to specific uses or organisms.

Finally, I want to introduce novel Y-SSRs in the efforts of developing new designer recombinases for precise genome surgery and demonstrate their much-needed impact in accelerating the directed evolution process. Up to date, therapeutically relevant recombinases with altered DNA specificity were developed to perform an excision of the desired DNA sequence, as demonstrated in the case of Brec1 which is able to remove the HIV provirus from the genome of infected human cells (Sarkar et al., 2007; Karpinski et al.,

2016), or inversion of the large genomic sequence which is often found in individuals with severe Hemophilia A (Lansing et al., 2022).

In order to expand the use of recombinases for other therapeutic application I want to demonstrate, that it is possible to evolve recombinases which can successfully integrate large DNA cargos into naturally occurring *lox*-like sites in the human genome. The DNA integration into native sites in the human genome would allow to develop gene therapies (or even cell therapies) based on designer recombinases, where full-length endogenous gene expression is most beneficial, like in the cases of genes where disease-causing mutations are spread across the whole gene length, or where the dose of the gene is crucial for the therapeutic effect.

PART II

# RESULTS

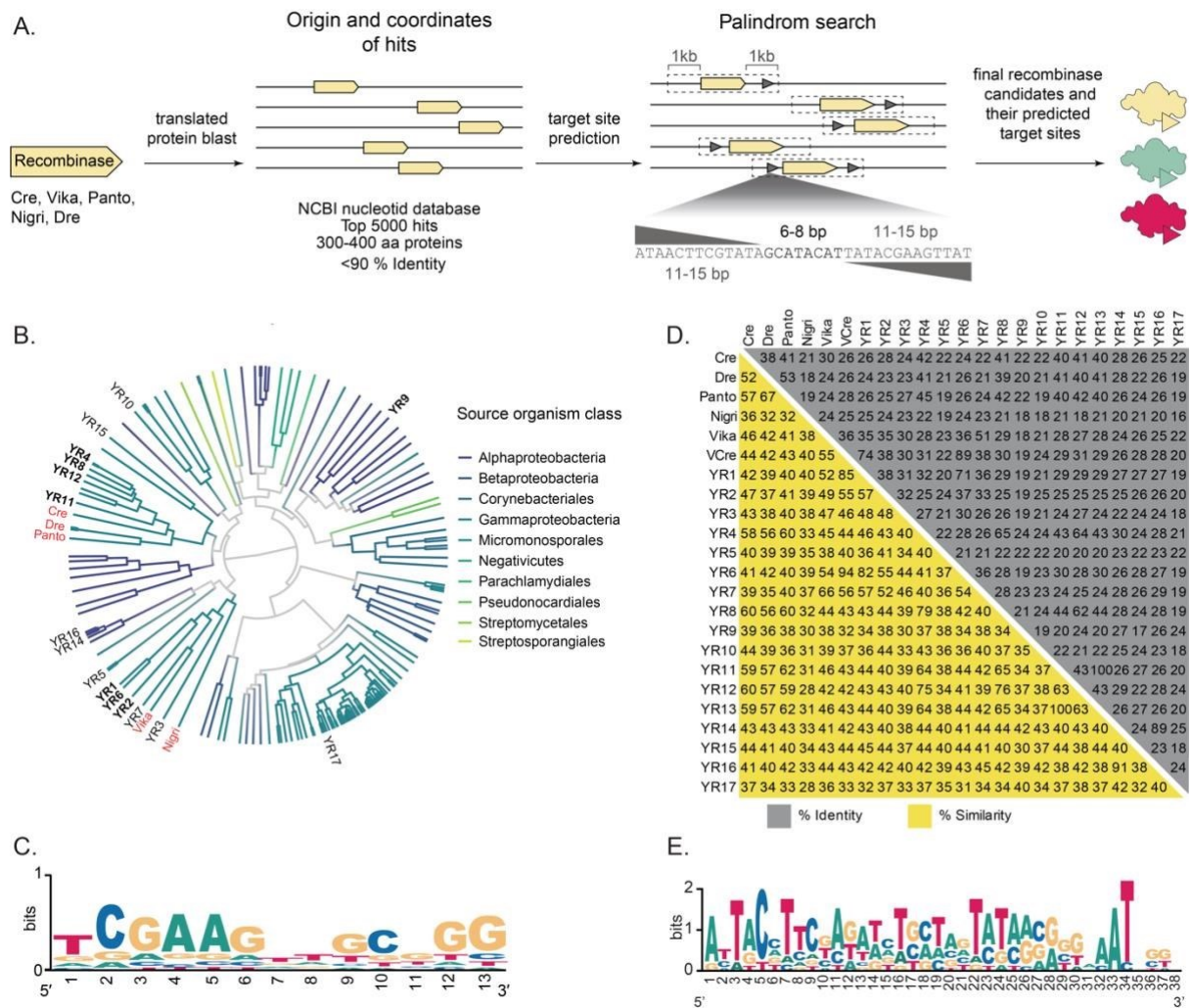
## Chapter 6 DISCOVERY AND CHARACTERIZATION OF NOVEL CRE-TYPE SSRs

### 6.1 In silico prediction of putative recombinases and their target sites

In order to identify novel Y-SSRs with diverse specificities, we searched the NCBI nucleotide collection for sequences resembling already characterized SSRs. The goal was to find sequences with similar characteristics to known Y-SSRs that might have the potential to act as novel site-specific recombinases. However, it is important to filter out sequences that are too similar to known recombinases, as these might have the same target sites and not provide much value in terms of discovering new Y-SSRs. To do this, we set a sequence identity threshold of less than 90% to the references (Cre, Vika, Dre, Panto, and Nigri), which are known Y-SSRs (Figure 7A). Additionally, we filtered for a protein sequence length of 300-400 amino acids, which is typical of Cre-type SSRs, to avoid identifying sequences that may not be Y-SSRs.

More challenging than identifying new SSRs is finding their native target site. Previous comparative genomic studies of phages suggest that target sites are typically located in close proximity to the recombinase coding sequence (Wang et al., 1995; Casjens, 2003). I therefore set the search criteria for target sites 1 kb upstream and downstream of the coding sequence of the enzyme (Figure 7A). Since Y-SSR target sites are frequently palindromic, we used the EMBOSS palindrome search tool, which looks for palindromic sequences that contain inverted complementary repeats separated by an 8 bp spacer. The search was focused on palindromic sequences that contained 13-15 base pairs of inverted complementary repeats and allowed for up to two asymmetrical mismatches, which are variations in the sequence that do not prevent the target site from being recognized by the recombinase (Figure 7A). These potential target sites were further filtered to have at least 3 mismatches (spacer included) from previously known target sites, which would increase the likelihood of discovering novel Y-SSRs. After applying these filters, a data set of approximately 500 putative recombinase/target site pairs was generated, which provided a valuable resource for experimental validation (Figure 7B and C). The hits were coming from all around the bacteria phylum, with majority coming from proteobacteria groups (Figure 7B). The putative target sites had a slight bias toward GC-rich sequences and can be seen from the DNA logo of the predicted half-sites in Figure 7C. This data set allows me to test the activity of the putative recombinases on their potential target sites and determine which of these sequences have the potential to be new Y-SSRs.

## DISCOVERY AND CHARACTERIZATION OF NOVEL CRE-TYPE SSRs



**Figure 7. Bioinformatical mining of tyrosine recombinases and their target sites.**

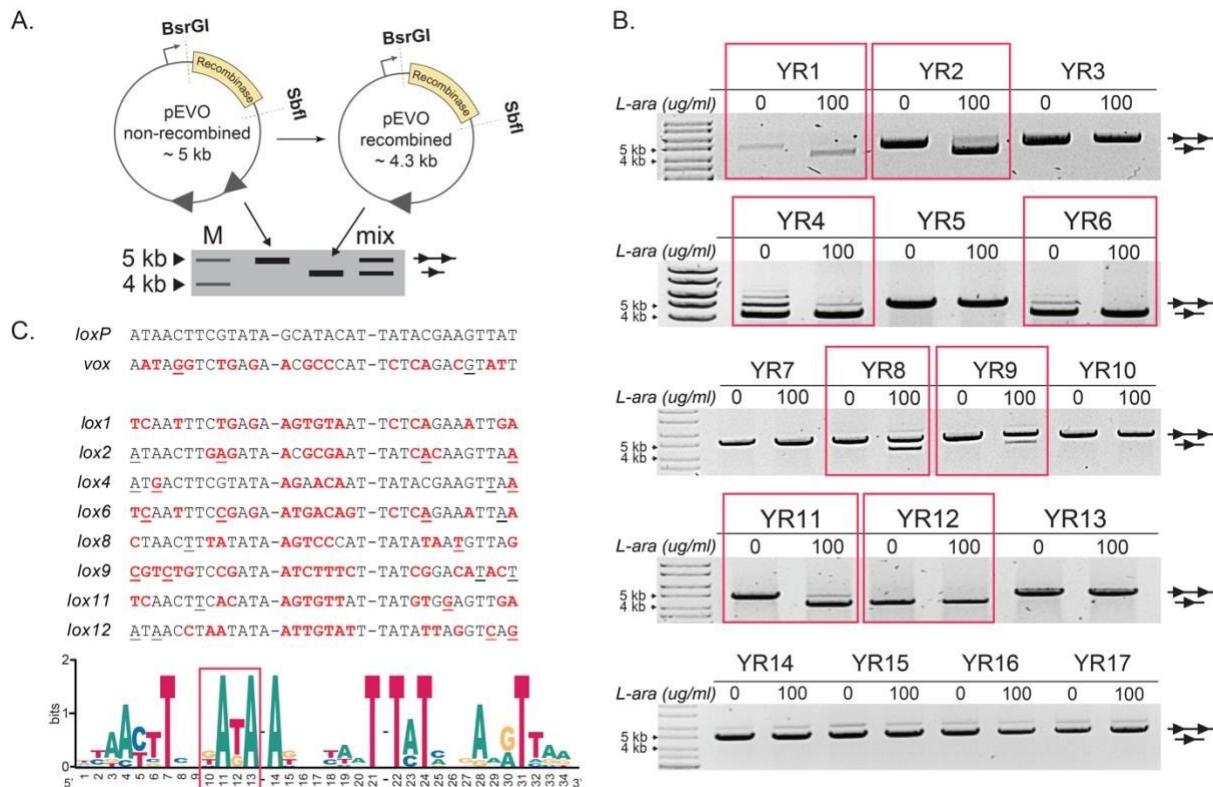
**A.** Schematic of computational workflow for prediction of novel Y-SSR candidates and their putative target sites in bacterial genomes. Dashed boxes present the palindrom search region while the triangles represent putative target site candidates. Zoom in depicts the criteria used for target site search: 11-15 bp palindromic half sites separated by 8 bp spacer with up to 2 asymmetries. **B.** Phylogenetic tree of discovered Y-SSR candidates, with branches color coded based on the class of the source organism. Candidates chosen for experimental validation are labeled as well as previously known recombinases (in red). New Y-SSRs active on their predicted target sites are depicted in bold. **C.** DNA sequence logo of 13 nucleotides left of the predicted spacers of all predicted target sites from all the predicted candidates. **D.** Protein sequence identity and similarity of 17 chosen candidates. Sequence identity (grey) and similarity (yellow) by global alignment of protein sequences was calculated and shown as a percentage for each pair of recombinases. Sequence similarity is defined as the percentage of matches between the two sequences over the reported aligned region (including any gaps in the length), where a match is defined as an alignment of two residues with a score above 1 in the BLOSUM62 matrix. **E.** DNA logos of all seventeen target sites on which recombinase activity was initially tested.



## 6.2 Novel Y-SSRs recombine their predicted target sites in bacteria

In order to experimentally test the activity of the candidate recombinases on the predicted target sequences, I randomly selected 17 candidates that best fit our criteria (Supplementary Table S3). The top candidates for experimental validation were chosen considering the sequence similarity to the reference sequences (Figure 7D), the composition of the potential target site (Figure 7E), and the conservation of catalytical residues (corresponding to R173, H289, R292, K201, W315 and Y324 in Cre) (Supplementary Figure S1). I cloned the coding sequences of the 17 prospective candidates individually into the L-arabinose inducible pEVO recombination reporter vector (Buchholz and Stewart, 2001) harboring two copies of the respective predicted target sites (Supplementary Table S1, Supplementary Figure S2). The plasmids were then transformed into *E. coli* and cultured overnight in medium containing L-arabinose to induce recombinase expression. Upon expression, successful recombination leads to excision of a ~700 bp DNA fragment from the plasmid. This size difference can be visualized by agarose electrophoresis of linearized plasmids (Figure 8A). Eight out of seventeen candidates showed activity on their predicted target sites, evident by the appearance of ~4 kb recombination bands (Figure 8B). Five of the candidates already showed efficient recombination in the samples without addition of L-arabinose to the medium (YR1, YR2, YR4, YR6 and YR12), indicating that these enzymes are active even when expressed at very low levels. Other recombinases (YR8, YR9, and YR11) recombined the plasmid only when L-arabinose was present in the growth medium, suggesting that they require a higher induction to become active in this assay. These results demonstrate that the established pipeline is able to identify novel Y-SSRs and their respective target sites at a success rate of approximately 50%. Interestingly, the target sites of the working recombinases all contained the T(G)AT(G)A motif on positions 10-13 of the half sites (positions closest to the spacer), and a conserved T7 nucleotide in the alignment of the recombined target sites (Figure 8C).

## DISCOVERY AND CHARACTERIZATION OF NOVEL CRE-TYPE SSRs



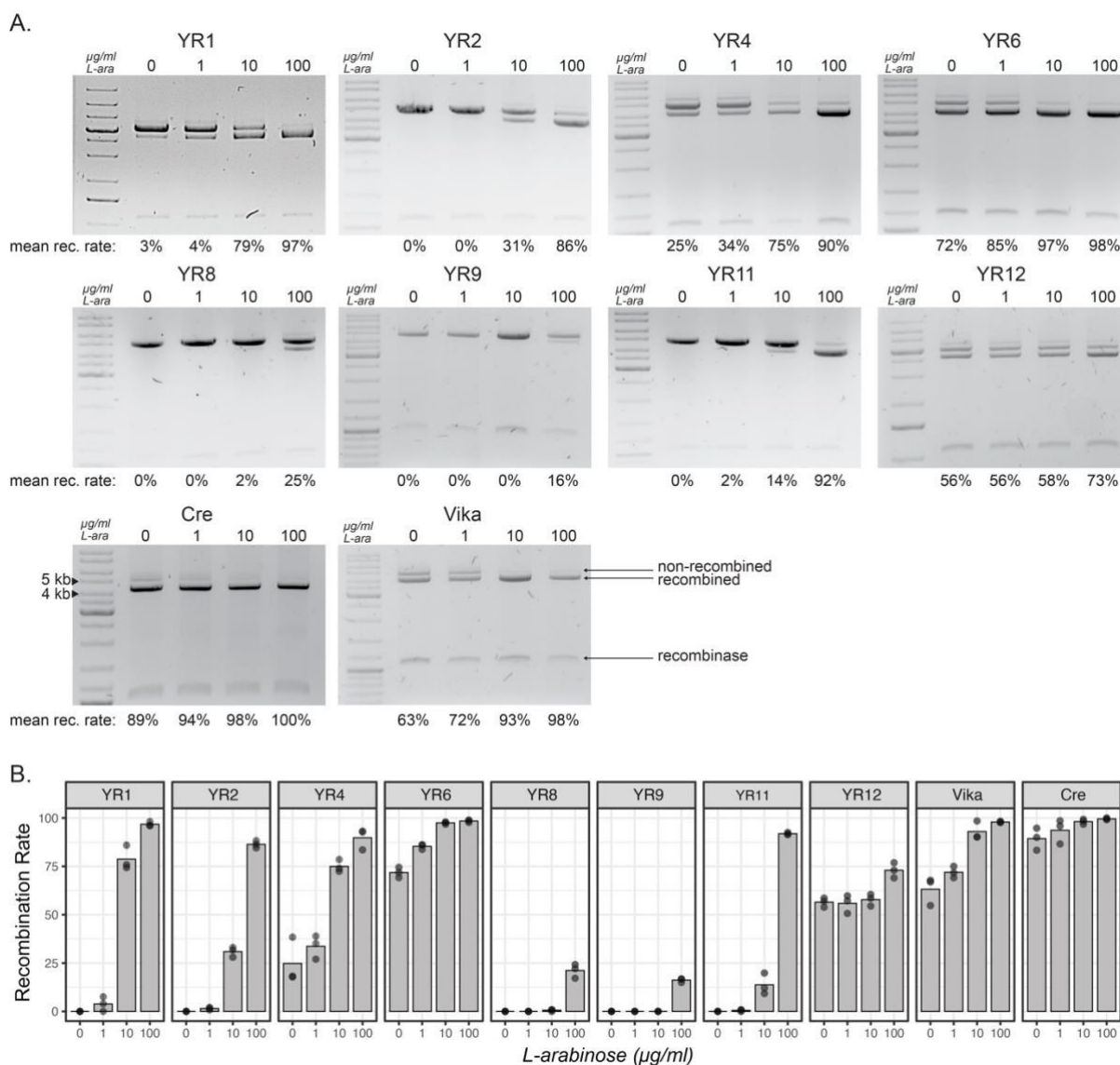
**Figure 8. Experimental validation of the chosen candidates.**

**A.** Overview of plasmid recombination assay. Important features, such as restriction sites, recombinase coding sequence and target sites (triangles) are shown. A schematic representation of expected recombination products on an agarose gel is shown below. Marker and expected sizes of recombined and unrecombined plasmids are indicated separately or together as usually seen on the gels. **B.** Recombination activity of seventeen tested putative Y-SSR/target site pairs. Each sample was tested with or without L-arabinose (100  $\mu\text{g/ml}$ ) added to the growth medium for recombinase expression, indicated with “0” or “100”. Recombination is indicated by the band aligned with the single triangle and non-recombined plasmids are indicated by two triangles. Active recombinases are highlighted in a red box. M = GeneRuler<sup>TM</sup> DNA Ladder Mix (Thermo Fisher). **C.** Nucleotide sequence alignment (up) and DNA logo (down) of the target sites that were recombined by respective putative recombinases. Full target sequences are separated into two half sites flanking the spacer sequence with a dashed line. Underlined nucleotides in the sequence represent asymmetric positions and the nucleotides in red depict differences to the loxP sequence. The common T(G)AT(G)A motif is depicted in the square box.

To characterize the active Y-SSRs in more detail, I quantified their recombination efficiencies at different expression levels in order to investigate dose response and to obtain a better side-by-side comparison of their efficiencies compared to the well-established recombinases Cre and Vika (Karimova et al., 2012). pEVO plasmids with desired recombinase/target site pairs were transformed into *E. coli* and were grown over night at different concentrations of L-arabinose to induce expression of the recombinase (Guzman et al., 1995). Extracted plasmid DNA was then assayed for recombination on agarose gels (Figure 9A). Quantification of band intensities revealed that the new recombinases have different activity profiles on their respective target sites (Figure 9B). Although, most of the recombinases were highly active when grown at high L-arabinose concentrations, showing recombination

## DISCOVERY AND CHARACTERIZATION OF NOVEL CRE-TYPE SSRs

rate between 87 and 100% (YR1, YR2, YR4, YR6 and YR11), they behaved quite differently when expressed with low L-arabinose concentrations. While YR1, YR2, YR4 and YR11 showed low recombination rates when induced at 1 or 10  $\mu\text{g/ml}$  of L-arabinose, YR6 was highly active even at these low induction levels, with its activity profile resembling Cre and Vika (Figure 9B). Interestingly, the YR12 recombinase showed a mostly constant recombination rate raging from  $\sim 50\%$  at 0  $\mu\text{g/ml}$  L-arabinose and peaking at 70% when induced with 100  $\mu\text{g/ml}$  of L-arabinose. YR8 and YR9, on the other hand, showed the weakest activity and only recombined their target sites to 25% and 17%, respectively, at the highest L-arabinose concentration.



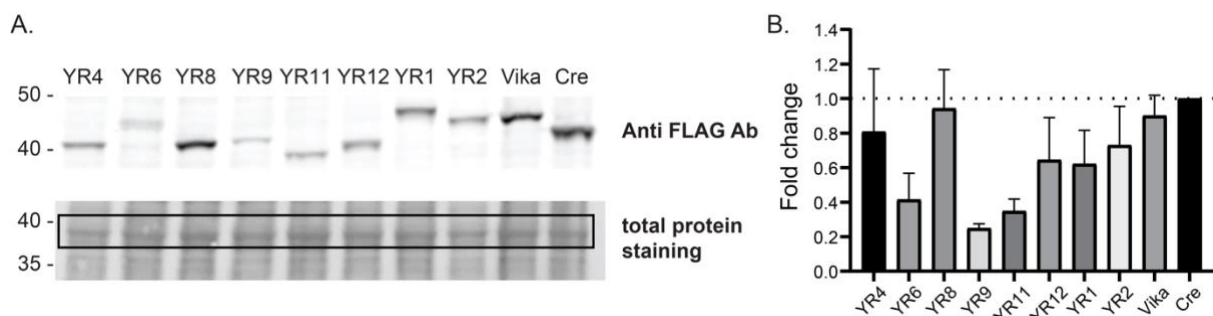
**Figure 9. Novel recombinases' activity as a function of arabinose.**

**A.** Agarose gels documenting recombination efficiencies of indicated recombinases at different L-arabinose concentrations. pEVO vectors carrying depicted recombinase – target site pairs were grown under increasing L-arabinose concentrations (0  $\mu\text{g/ml}$ , 1  $\mu\text{g/ml}$ , 10  $\mu\text{g/ml}$  and 100  $\mu\text{g/ml}$ ) to induce recombinase expression. Mean recombination efficiency is depicted as percentage for each

## DISCOVERY AND CHARACTERIZATION OF NOVEL CRE-TYPE SSRs

expression level of every recombinase. **B.** Quantification and reproducibility of recombination activity of new SSRs in *E. coli*. Recombinase expression was induced with rising concentrations ( $\mu\text{g/ml}$ ) of L-arabinose indicated along the x-axis. Vika and Cre were included as positive controls. Recombination was calculated from measuring the band intensities from agarose gels shown in A. Bacterial assays were done in triplicates ( $n = 3$ ).

One possible explanation for different activities of the examined enzymes could be different protein levels, due to varying production- and degradation-rates of the enzymes in the cells. To investigate this possibility, I measured protein concentrations in *E. coli* for each recombinase by Western blotting (Figure 10A). I cloned all the recombinases in a modified pEVO vector harboring a N-Flag tag, and transformed and expressed the recombinases at 10  $\mu\text{g/ml}$  of L-arabinose overnight. Using a Flag-antibody, I observed differential protein levels of the recombinases (Figure 10B). Cre showed the highest levels, followed by YR4 and YR8 recombinases that showed just slightly lower values. YR9 and YR11 showed weakest expression levels compared to the other recombinases (Figure 10B). Interestingly, the expression levels do not correlate well with the enzyme activities, suggesting that other factors such as catalytic activity, DNA binding affinity, dimerization or speed of isomerization also contribute to the difference in recombination efficiencies.



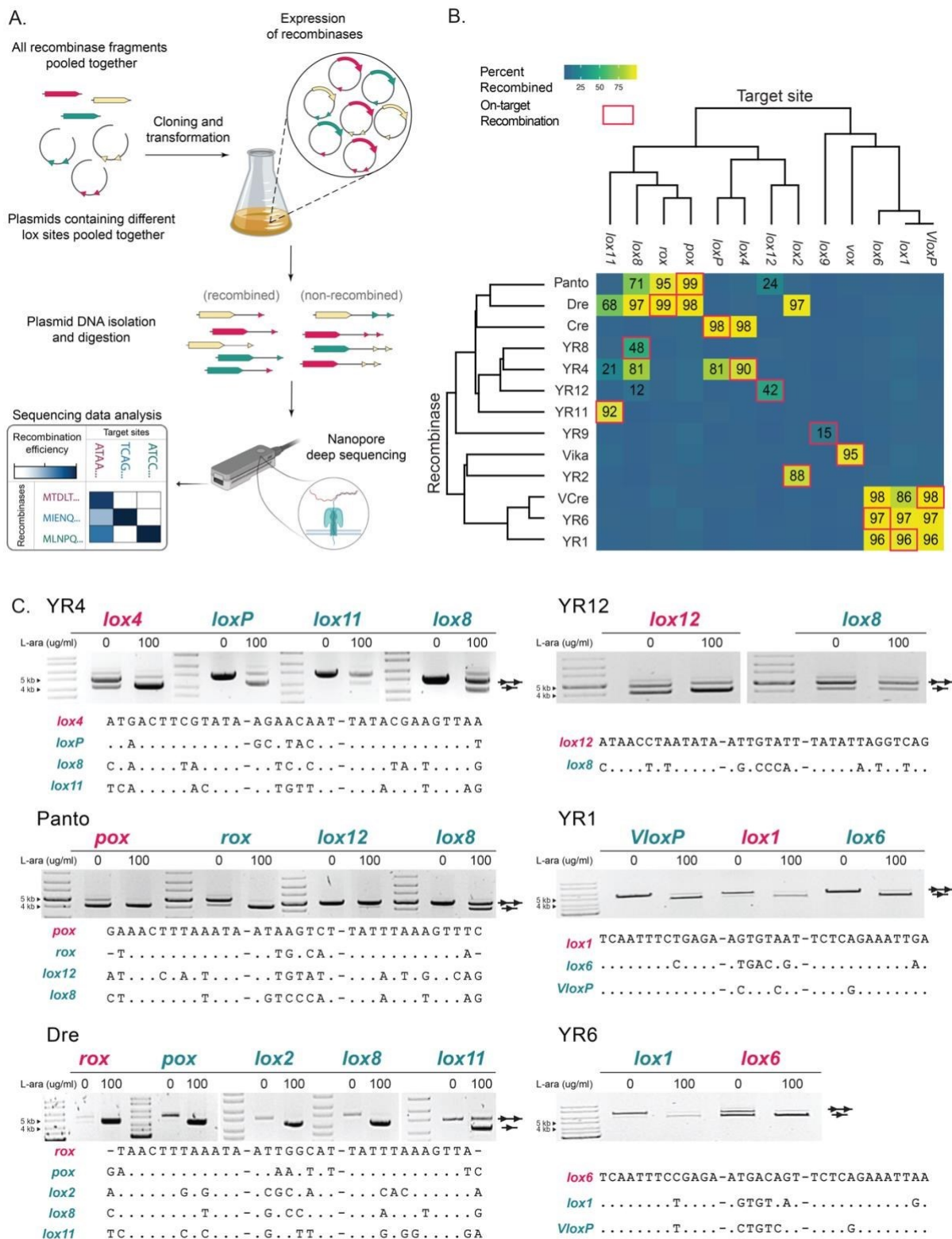
**Figure 10. Expression differences of novel Y-SSRs in *E. coli*.**

**A.** Western blot of whole-cell protein extracts from *E. coli* probed with Anti-Flag antibody (upper) or stained with total protein staining for normalization (lower). XL-1 blue *E. coli* were transformed with pEVO plasmids harboring indicated recombinases and expression of recombinases was induced with 10  $\mu\text{g/ml}$  L-arabinose. **B.** Quantitative analysis of protein expression levels. Fold change of measured protein levels of each recombinase was normalized to Cre. Bar graphs represent the mean of 3 independent biological replicates with error bars showing standard deviation of the mean (SD).

### 6.3 Profiling target-site selectivity of Cre-type recombinases

SSRs with different sequence specificity are frequently used in combination to allow sophisticated genomic or synthetic biology experiments (Feil, 2007; Livet et al., 2007; Snippert et al., 2010; Merrick et al., 2018; Sheets et al., 2020). For these experiments it is important to know the specificity of the enzymes and to consider possible cross reactivity (Fenno et al., 2014; Weinberg et al., 2017). This information may also help to provide more insight on how these enzymes recombine specific nucleotide sequences. In order to test all possible combinations, I developed a high-throughput sequencing approach where the activity of known (Panto, Dre, Cre, Vika and VCre) and new Y-SSRs (YR1, YR2, YR4, etc.) can be quantified on all target sites in a single experiment. To accomplish this, I started with a two-step cloning scheme to produce all combinations of 13 recombinases and their respective 13 target sites on 169 (13x13) individual vectors. I pooled and linearized 13 pEVO vectors harboring 13 different target site sequences to clone in a pool of the 13 recombinase coding sequences in one ligation reaction (Figure 11A). After an overnight culture and induction of recombinase expression (with 100 µg/ml of L-arabinose), I retrieved the plasmid DNA and cut out the fragments carrying the recombinase sequence on the 3'-end and target site(s) on the 5'-end (Figure 11A). Using the Oxford Nanopore Technologies' long read sequencing platform, I was able to obtain a total of 417,769 reads containing both the specified recombinases and target sites. All possible 169 combinations of Y-SSRs and target sites were identified with a minimum coverage of 224 reads (Supplementary Figure S3A). Using this data, I calculated the recombination rates for the individual recombinases on all the target sites, providing a specificity profile for each recombinase (Figure 11B).

# DISCOVERY AND CHARACTERIZATION OF NOVEL CRE-TYPE SSRs



**Figure 11. Profiling target-site selectivity of Cre-like recombinases.**

**A.** Schematic representation of the experimental workflow. Important steps are indicated by arrows. **B.** Analysis of deep sequencing results. Cross recombination events are displayed by a heatmap of recombination efficiency for each possible combination. Recombination efficiencies were calculated by dividing the number of recombined reads by the number of total reads for a given combination and expressed as a percentage. For combinations where more than 10% recombination was observed the exact number of recombination percentage is indicated. Target sites are displayed horizontally and

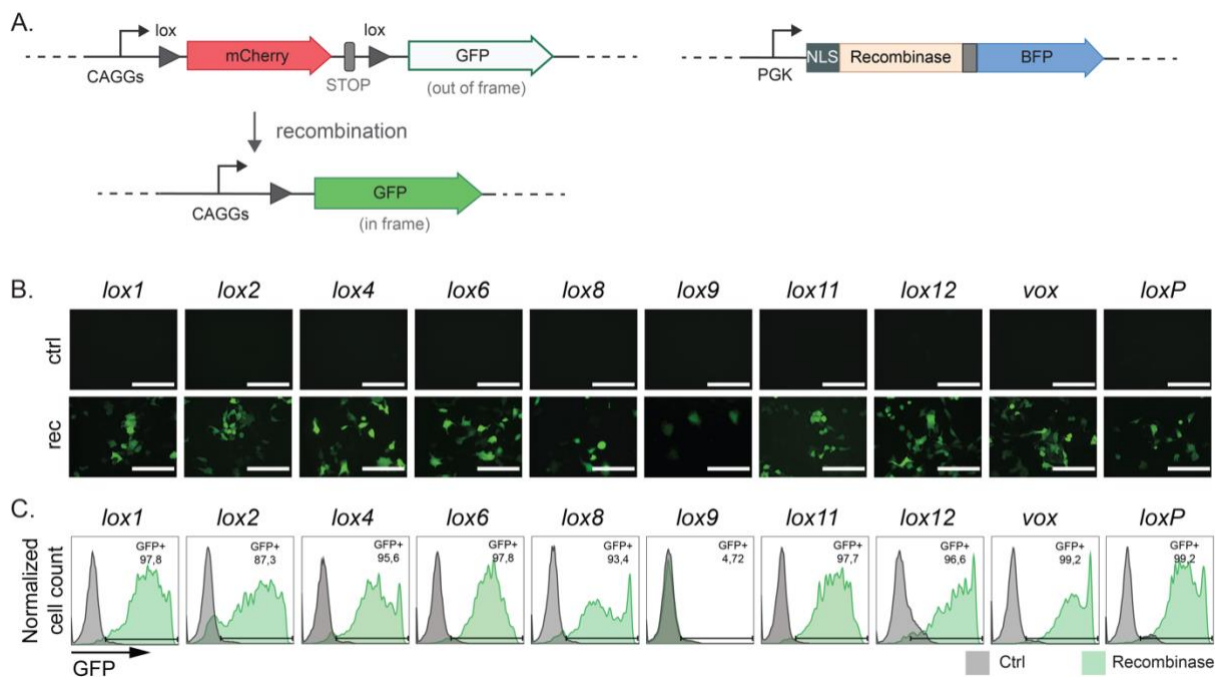
## DISCOVERY AND CHARACTERIZATION OF NOVEL CRE-TYPE SSRs

ordered based on their similarity. Recombinases tested are aligned based on their homology on the vertical axis. On-target recombination events are boxed with red squares. **C.** Validation of cross recombination events by plasmid-based recombination assay. Recombination activity of respective recombinases is shown on their on-target sites in magenta, and on off-targets in turquoise. Recombination was assessed by agarose gel electrophoresis. Each sample was tested with or without L-arabinose to induce recombinase expression, indicated with "0" or "100"  $\mu\text{g/ml}$  of L-arabinose). Recombination is indicated by the line with the single triangle, whereas a line with two triangles illustrates the non-recombined band. M = GeneRuler™ DNA Ladder Mix (Thermo Fisher). Parts of the figure were created with BioRender.com.

Interestingly, I found a wide spectrum of specificities for the different recombinases (Figure 11B). First, I identified that YR2, YR8, YR9, YR11 and Vika are highly specific recombinases, recombining only their own target sites. The other recombinases showed activity on their own, but also on varying numbers of other target sites. Analyzing this data in detail, I could make several interesting observations from the recorded cross-recombination events: i) Three-way cross-recombination between YR1, YR6 and VCre, can be explained by the high similarity of their target sites (Figure 11C, Supplementary Figure S3B) as well as relatively high protein homology (Figure 7D). ii) Two-way cross-recombination between Cre and YR4 can be explained by the similarity of their target sites, since the half-sites of *loxP* and *lox4* differ only at one nucleotide position (Supplementary Figure S3B), but, interestingly, the recombinases share only 58% similarity (42% identity, Figure 7D); iii) Dre recombinase seems to recombine a variety of target sites ranging from ones very similar to its cognate target site *rox* (*pox*, *lox8* – 2 mismatches) to *lox11* with four mismatches to the *rox* half-site (Figure 11C, Supplementary Figure S3B); iv) Three SSRs (YR4, Panto and Dre) cross-recombine targets, which are quite different to their bone-fide target sites. For example, YR4 is able to recombine *lox11* even though, its half-site differs in seven positions from *lox4* (Figure 11C, Supplementary Figure S3B). Interestingly, the reverse is not true and YR11 does not show any activity on *lox4* (Figure 11B); v) Recombinases that recombine similar target sites have different cross-recombination profiles (e.g., Dre recombines *lox11* and *lox2*, whereas Panto does not, but recombines instead *lox12*, where Dre and Panto both recombine their respective target sites, *rox* and *pox*, respectively). Altogether, these results document the complex relation of recombinase sequences and their activity on target sites, but also can serve as guidelines for simultaneous use of these recombinases in complex setups where fine spatial and temporal control of multiple target recombination is needed.

## 6.4 Activity of novel recombinases in human cells

After characterizing their activity and specificity in bacteria in depths, I wanted to investigate the potential of these new recombinases for applications in mammalian cells. To that cause, I first tested the activity of the 8 new Y-SSRs in a human cell line. I co-transfected HEK293T cells with recombinase expression plasmids alongside recombination reporter plasmids harboring the corresponding target sites (Figure 12A, Supplementary Figure S4A). In the reporter plasmids, the mCherry cassette, driven by a CAG promoter, is flanked by the target (*lox*) sites. Upon recombination, the mCherry cassette is deleted from the plasmid and the CAG promoter then drives the expression of a GFP cassette (Figure 12A). Hence, the activity of the recombinases on their predicted target sites can be visualized by fluorescent microscopy and quantified by flow cytometry. When co-transfection experiments were analyzed, all of the recombinases tested displayed GFP positive cells (Figure 12B, lower panels), demonstrating that these recombinases are active in HEK293T cells, whereas no GFP-positive cells were observed when co-transfections were done with an “empty” expression vector, lacking recombinase coding sequences (Figure 12B, upper panels). Flow cytometry analyses revealed that in almost all samples more than 90% of the cells that were co-transfected with the recombinase expression plasmid and the reporter (cells that were double, BFP and mCherry positive) were also GFP positive, indicating that these recombinases are highly active in this setting (Figure 12C, Supplementary Figure S4B and S4C). These results indicate that newly discovered recombinases can be utilized for genome engineering in mammalian cells.





**Figure 12. Activity of newly discovered SSRs in mammalian cells.**

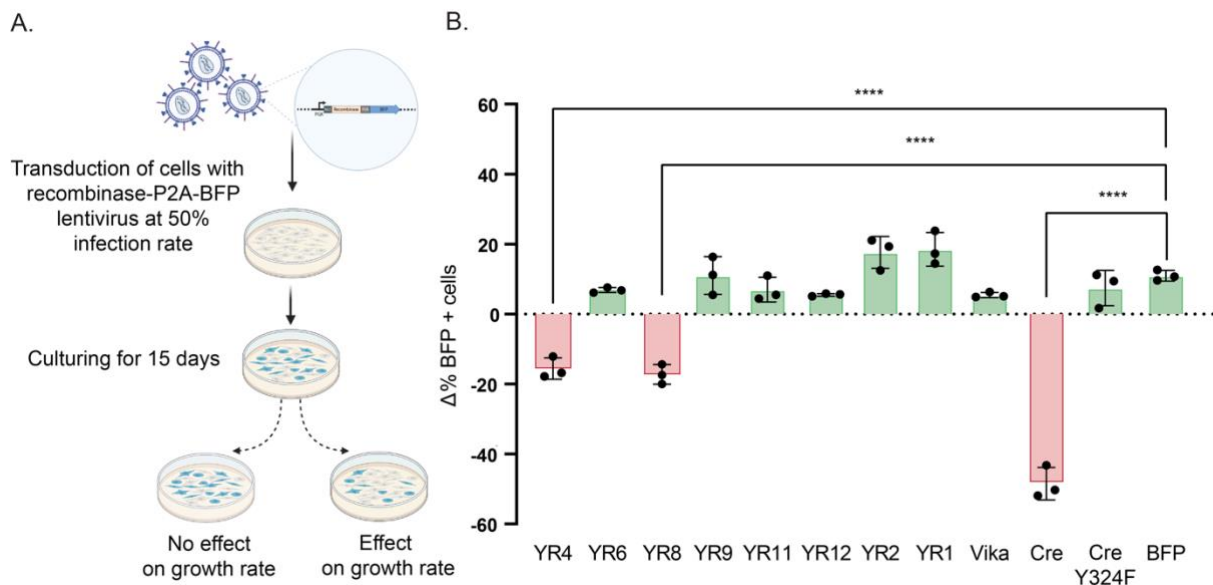
**A.** Graphical representation of the mammalian recombination reporter and expression constructs. Important features are marked in the reporter and expression vectors. Upon recombination of the reporter vector the mCherry cassette will be excised allowing for the expression of GFP (green) from the pCAG promoter (arrow). Black triangles represent different *lox* sites for each corresponding recombinase. NLS, nuclear localization signal. **B.** Fluorescence microscopy analysis in HEK293T cells. Transfected cells with the empty expression plasmids and non-recombined reporter plasmids harboring eight new target sites (top panel), or co-transfection of the reporter with the expression plasmids carrying respective recombinases (lower panel) are shown. *Vika/vox* and *Cre/loxP* were included as positive controls. Ctrl, negative controls; Rec, recombinase. **C.** FACS analysis of the samples shown in B. Grey histograms depict control samples where non-recombined reporter plasmids were co-transfected with 'empty' expression plasmids, while the green histograms show samples transfected with corresponding recombinases.

**6.5 Influence of recombinase expression on cell proliferation**

Investigations of SSRs in heterologous hosts are important to define their applied properties. Recombinases could recognize cryptic (pseudo) recombination sites in a genome that might be recombined and lead to off-target effects, potentially resulting in growth arrest or apoptosis. Indeed, active pseudo-*loxP* sites have been described in the human and mouse genome (Thyagarajan et al., 2000). Consequently, impairment of cell proliferation may occur when Cre is overexpressed in human or mouse cells (Schmidt et al., 2000; Loonstra et al., 2001; Pugach et al., 2015). Furthermore, overexpression of any DNA binding enzyme could lead to impaired regulation of genome expression or stability which could lead to additional indirect toxic effects.

To test for potential effects on cell proliferation of the newly identified Y-SSRs when they are overexpressed, I produced lentiviral particles, with the lentiviral vectors, constructed to allow co-expression of the recombinases and tagBFP (Figure 13A). As controls, I also produced viral particles for overexpression of either Cre, an inactive Cre variant (CreY324F), *Vika* and tagBFP alone. NIH3T3 cells were then infected at approx. 50% rate and the change in percentage of BFP-positive cells was monitored for 15 days. A reduction in BFP-positive cells over time would indicate a negative effect on cell proliferation due to recombinase overexpression (Schmidt et al., 2000; Pugach et al., 2015) (Figure 13A). Indeed, the number of BFP-positive cells progressively dropped when Cre recombinase was tested in this assay, while the catalytically inactive version of Cre, as well as tagBFP alone had no effect (Figure 13B). In comparison to Cre, YR4 and YR8 showed a less pronounced decrease in the percentage of BFP positive cells (~15%;  $p < 0.0001$ , and ~17%;  $p < 0.0001$ , respectively), suggesting that overexpression of these recombinases slightly inhibits cell proliferation (Figure 13B). In contrast, the percentage of BFP-positive cells did not significantly change in cells expressing the other recombinases (Figure 13B), indicating that overexpression of these recombinases is well tolerated in the cells. I conclude that 6 out of the 8 newly

identified Y-SSRs can be expressed for extended period of time without compromising cell proliferation, supporting their utility in mammalian cells.



**Figure 13. Effect on cell growth upon overexpression of Y-SSRs in mammalian cells.**

**A.** Overview of the experimental setup. Important steps are indicated by arrows. Cells were transduced at a rate of ca. 50% with a bicistronic lentivirus expression construct, where the expression of respective recombinases was linked via P2A to BFP expression. Cells were analyzed every 72 h by flow cytometry and the percentage of BFP-positive cells was recorded over the course of 2 weeks. Declining percentages of BFP-positive cells are indicative of a proliferation disadvantage of infected cells. **B.** Analysis of growth rates. Difference in the percentage of BFP-positive cells between day 3 and day 15 are plotted (biological replicates are shown as dots, n = 3). Error bars represent standard deviation of the mean (SD). Statistical significance relative to BFP control was calculated by a 1-way ANOVA test. (\*\*\*\*): P ≤ 0.0001. Parts of the figure were created with BioRender.com.

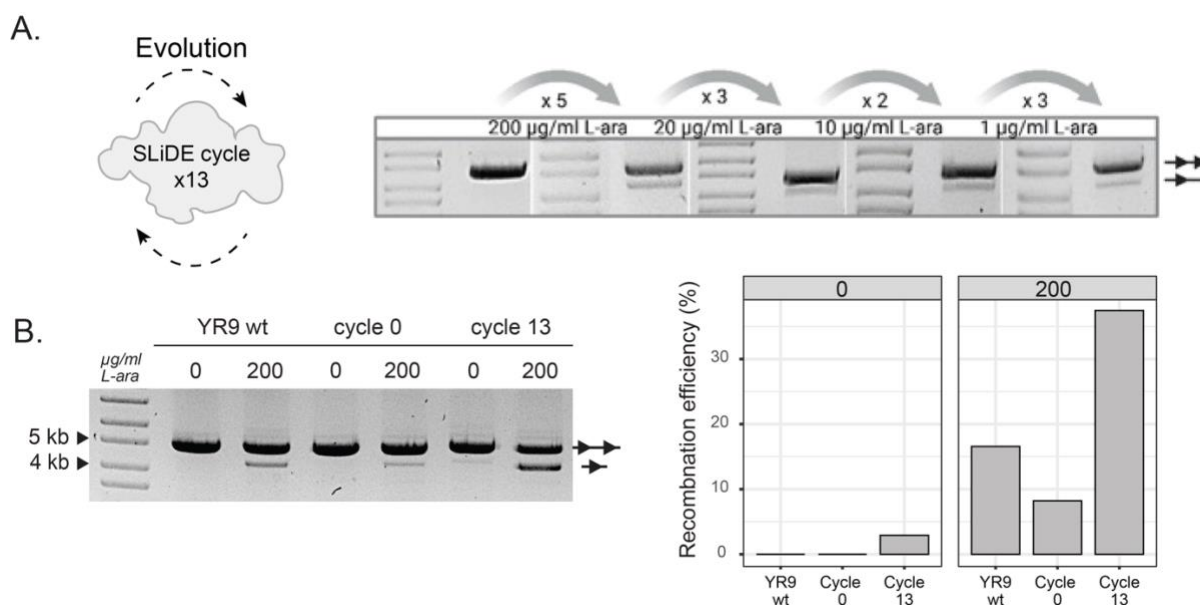
## 6.6 Directed evolution of YR9 recombinase

As demonstrated above, one of the identified recombinases, the YR9 recombinase, recombined its putative *lox9* target site with relatively low efficiency (17% at 100 μg/ml of L-arabinose in *E. coli*, and negligible recombination in HEK293T cells) (Figure 9 and Figure 12). Nevertheless, several aspects make the YR9 recombinase an interesting Y-SSR system worth investigating further: First, the YR9 recombinase seems to be one of the highly specific recombinases, recombining only the *lox9* target site (Figure 11B). Secondly, the YR9 recombinase is slightly smaller than most recombinases, and the *lox9* target site is the most unique target site among the other identified ones when compared to the *loxP* target site of Cre. This is particularly interesting when seeking to establish a diverse library of

## DISCOVERY AND CHARACTERIZATION OF NOVEL CRE-TYPE SSRs

recombinases. Thus, I sought to enhance the recombination efficiency of this recombinase through directed molecular evolution.

In order to obtain versions of YR9 that recombine *lox9* with increased efficiency, an initial library of approximately 350,000 YR9 clones was generated by error-prone PCR (see Methods section). The library was then subjected to iterative rounds of the SLiDE procedure starting with 200  $\mu\text{g/ml}$  L-arabinose. During the evolution, the recombinase expression level was gradually lowered by reducing the L-arabinose concentration in a stepwise manner in order to increase the selection pressure once the increase in the activity would be observed (Figure 14A). In total, 13 cycles of SLiDE were needed to develop a final library that showed a marked increase in recombination activity, indicating that YR9 versions with improved activity at the *lox9* target sites had evolved (see Figure 14B).



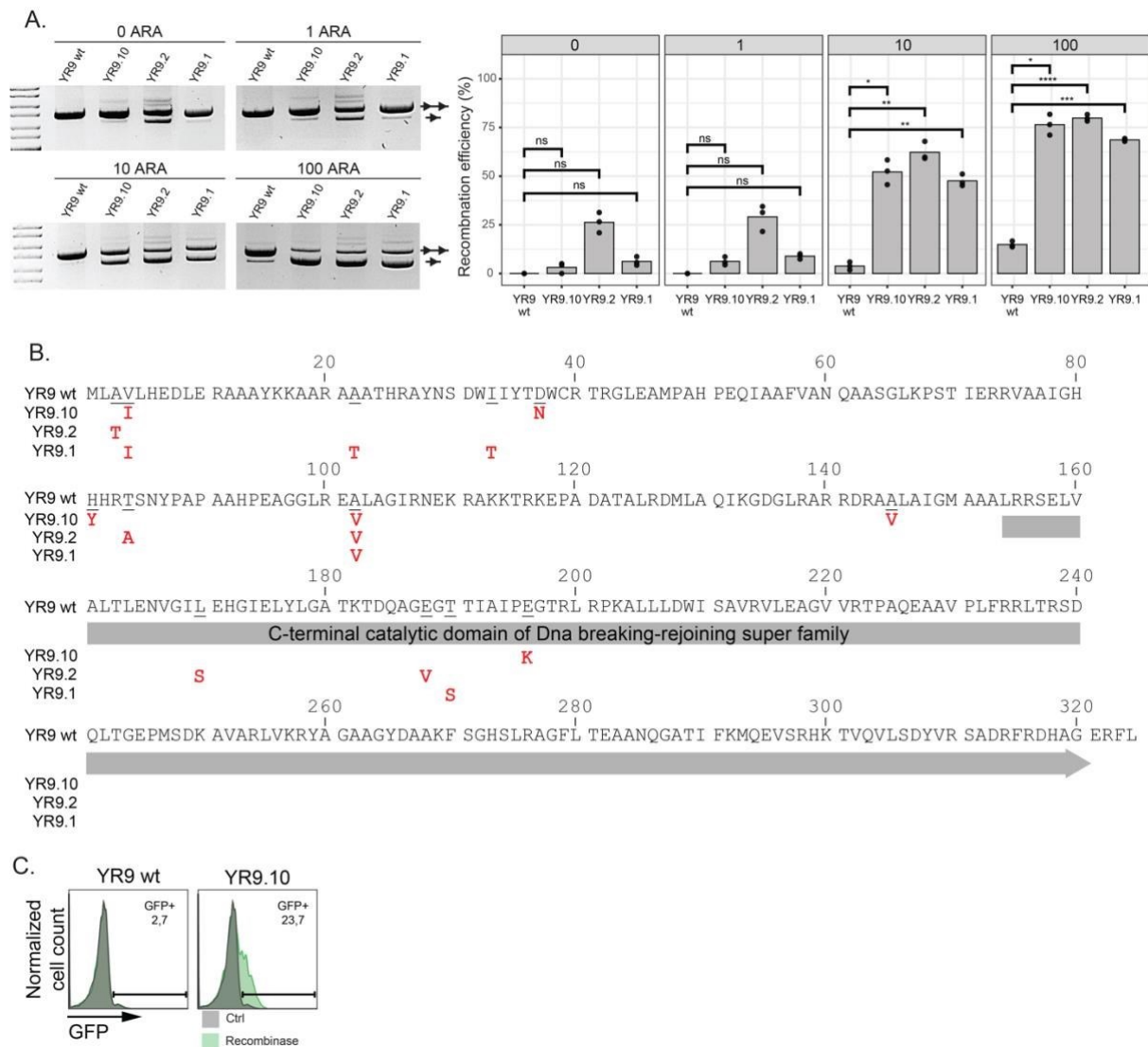
**Figure 14. Directed molecular evolution of YR9 recombinase on *lox9*.**

**A.** Representation of the SLiDE progress using 13 cycles in total. The number of cycles per L-arabinose concentration is shown below the arrows. A test digest demonstrating the recombination activity of the first and last cycle of each L-arabinose level is presented. The unrecombined and the recombined bands are denoted as a line with two triangles and one triangle, respectively. **B.** Comparison of the recombination activities of wild type R6, initial and final cycle. All samples were grown with and without L-arabinose (200  $\mu\text{g/ml}$ ) in the medium. The agarose gel image on the left shows the test digests and was used to calculate the ratio between intensities of recombined (lower band, illustrated by a line with one triangle) and unrecombined (upper band, depicted as a line with two triangles) bands for each sample. The resulting quantifications of the recombination efficiencies are shown on the right.

As final improved variants, clones 1, 2, and 10 were chosen for more detailed characterization. In order to investigate dose-response and to obtain a side-by-side comparison of their efficiencies to the wild type YR9, recombinase expression was induced with increasing concentrations of L-arabinose (0, 1, 10, and 100  $\mu\text{g/ml}$ ). All three

## DISCOVERY AND CHARACTERIZATION OF NOVEL CRE-TYPE SSRs

recombinases were highly active at 100  $\mu\text{g/ml}$  and showed significant improvement compared to the wild type YR9. Recombination rates ranged from 69% for clone 1 to 77% for clone 10 and 80% for clone 2 (Figure 15A). This means, that all clones showed about 5-fold improvement in activity compared to wild type YR9 in *E. coli*.



**Figure 15. Characterization of evolved YR9 variants.**

**A.** Plasmid based activity assay of wt YR9 recombinase and three randomly picked clones. Agarose gel pictures to the right show the best clones recombining *lox9* in comparison to wt YR9 at four different expression levels (0  $\mu\text{g/ml}$ , 1  $\mu\text{g/ml}$ , 10  $\mu\text{g/ml}$  and 100  $\mu\text{g/ml}$ ). To the right, quantification and reproducibility of recombination is shown. Recombination efficiency was calculated by comparing the ratio of recombined and non-recombined band intensities from the gels to the left. Experiments were done in triplicates ( $n = 3$ ). Comparison to YR9 wt was done with a t-test and the p-values were adjusted for multiple comparisons using the Bonferroni method. Significance: (ns)  $p > 0.05$ , (\*)  $p \leq 0.05$ , (\*\*)  $p \leq 0.01$ , (\*\*\*)  $p \leq 0.001$ , (\*\*\*\*)  $p \leq 0.0001$  **B.** Mutation analysis of the YR9 improved clones (one-letter code). The amino acid sequence of wt YR9 recombinase is shown as a reference. Red letters represent changes found in the clones YR9.10, YR9.2 and YR9.1 from top to bottom. Predicted C-terminal catalytic domain of DNA breaking-rejoining super family is labeled. **C.** Activity of the most active clone in mammalian cells. FACS analysis of wt YR9 and YR9.2 recombinase are shown. Grey

histograms depict control samples where non-recombined reporter plasmids were co-transfected with 'empty' expression plasmids, while the green histograms show samples transfected with corresponding recombinases.

The overall comparison of amino acid sequences of the improved clones to the wild type YR9 revealed only five amino acid changes for clone 1 and clone 2 and six changes for clone 10, suggesting that only minor changes were necessary to achieve high activity on the *lox9* target sites. Interestingly, two regions attract attention: The alanine residue at position 102 was mutated to valine in all three variants. Also, a further accumulation of mutations can be observed at the N-terminus of the sequence, more precisely at positions 3 and 4 (V4I for clone 10 and 1 and A3T for clone 2) (Figure 15B).

For their applied use in higher organisms, the activity of the improved YR9 variants was investigated in mammalian cells. Unexpectedly, only one variant YR9.10, demonstrated a recombination efficiency of around 24%, as deduced from flow cytometry data (see Figure 15C). Although this recombination rate is not as prominent as that of Cre, it still represents a 20-fold increase over wild type YR9. Since being highly active in *E. coli*, these results show that the obtained effects for the improved variants cannot automatically be transferred to the mammalian system.

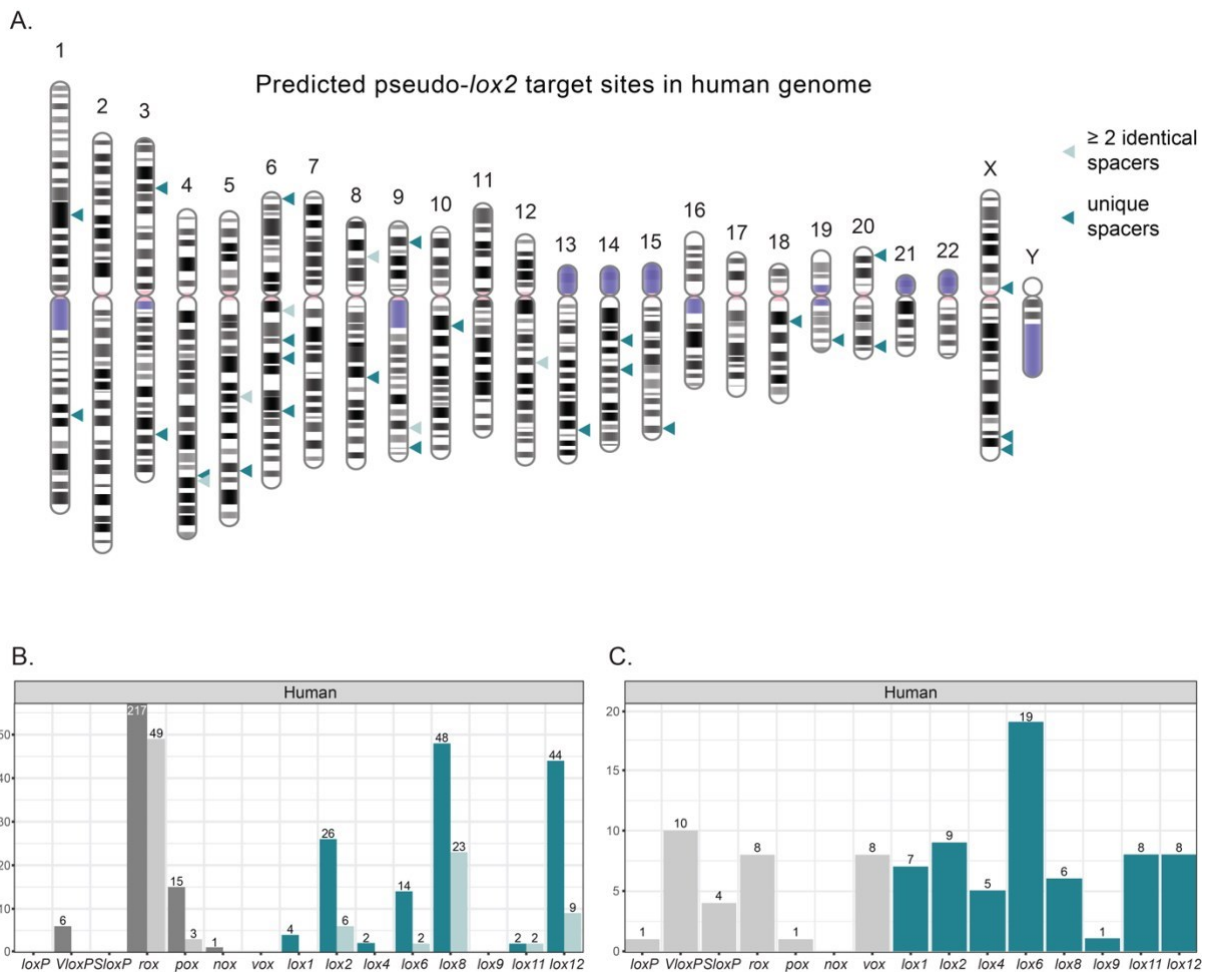
### **6.7 Prediction of possible recombination target sites in human and mouse genomes**

I sought to identify pseudo-*lox* sites of the investigated SSRs in human and mouse genomes, as these could serve as potential endogenous targets for a variety of genome engineering experiments, but also could lead to unwanted genome rearrangements (Schmidt et al., 2000; Thyagarajan et al., 2000). Hence, I searched for highly homologous target sites with unique or same putative spacers to address both possibilities (Figure 16, Supplementary Figure S5 and Supplementary Figure S6). Criteria that I used were based on sequence homology against putative target sites identified together with the recombinases, filtering for sequences carrying no more than two mismatches per half-site and leaving the spacer region entirely flexible. With the presumption that two target sites could be recombined by a single recombinase only if they possess matching spacer sequences, I searched for target sites with unique spacer to serve as landing pads for integration, while also screening for sites that share identical spacer sequence with at least one other target site to account for potential undesired inter- or intra-chromosomal rearrangements.

My search showed that the pseudo-sites with unique spacers were distributed to all the chromosomes of both genomes (Figure 16A, Supplementary Figure S5 and Supplementary Figure S6A). I found pseudo-rox sites of Dre recombinase to be the most common in both of

## DISCOVERY AND CHARACTERIZATION OF NOVEL CRE-TYPE SSRs

the genomes with 217 in human and 123 target sites in mouse genome (Figure 16B, Supplementary Figure S6B). Out of these, 49 and 22 genomic sequences, respectively, represent the subsets where at least two share the same spacer, whose recombination could lead to potential inter- or intra-chromosomal rearrangements. Pseudo-sites with identical spacers, possibly recombined by all other recombinases, were drastically less represented in the search making them more suitable for precise genomic manipulations (Figure 16B, Supplementary Figure S6B).



**Figure 16. Prediction of pseudo-recombination target sites in human genome.**

**A.** Example of chromosomal distribution of human pseudo-*lox2* sequences. Positions marked with triangles indicate sequences having no more than two mismatches per half-site of the matching target site sequence. Dark-colored triangles refer to a subset of genomic sequences, where each has a unique spacer (potential integration sites). Light-colored triangles represent a subset of genomic sequences, where at least two share the same spacer (sites for potential inter- or intra-chromosomal rearrangements). **B.** Counts of human genomic sequences with high similarity, up to two mismatches per half-site, towards already described (grey) and new (colored) Y-SSR target sites. Dark-colored bars refer to a subset of genomic sequences, where each has a unique spacer (potential integration sites). Light-colored bars refer to a subset of genomic sequences, where at least two share the same spacer (sites for potential inter- or intra-chromosomal rearrangements). **C.** Graphs depicting the number of unique sequences in the human genome with at least one half-site highly similar, with only one mismatch allowed, to target sites of the new (colored) and already described (grey) Y-SSRs.

Y-SSRs have the ability to perform recombination bidirectionally, with excision being more thermodynamically favored. Hence, their target sites have to be specifically engineered to increase the efficiency of the integration events (Araki et al., 1997, 2010). As described previously, one of the most efficient ways to shift the equilibrium of the reaction towards integration has been shown to be using the RE/LE (RE - right element, LE - left element) strategy (Thomson et al., 2003). Single mutant *loxP* sites can be modified in opposite half sites to ensure that upon recombination double mutants would be generated that inhibit reverse excision reaction and thereby stabilize the integration product. In order to show the potential utilization of new Y-SSRs for this purpose I also screened the human and mouse genomes for genomic sequences having only one half-site highly similar to the sequence of putative target sites, with only one mismatch allowed (Figure 16C, Supplementary Figure S6C). We found numerous possible entry points with our search, with the pseudo-*lox6* site possibly recombined by YR6 being most represented with 19 and 15 possible genomic sequences in both genomes (Figure 16C, Supplementary Figure S6C). By designing the compatible donor target sites multiple specifically targeted transgene insertions could possibly be realized by these Y-SSRs.

## Chapter 7 EVOLUTION OF VIKA RECOMBINASE FOR RECOMBINASE-MEDIATED INTEGRATION IN POTENTIAL SAFE HARBOR LOCUS

### 7.1 Preface

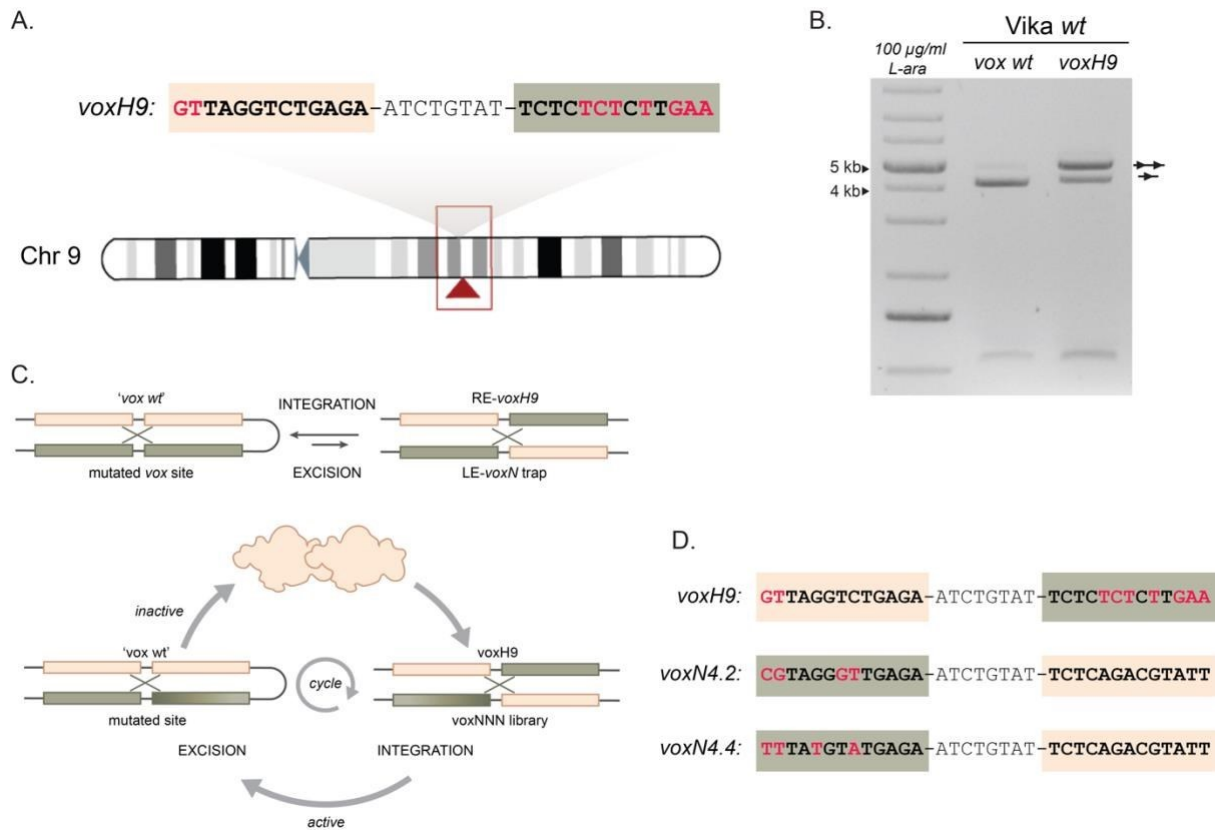
Previously, I presented evidence showing that a wide range of genomic sequences can potentially serve as entry points for site-specific transgene integration into both human and mouse genomes (refer to Figure 16 and Supplementary Figure S6 for an overview). As previously mentioned, these sequences were identified based on their high similarity to only one half-site of the putative target sites of different naturally occurring Y-SSRs (as depicted in Figure 16C). Interestingly, during my research, I found that one of the sequences that emerged from the comparison with the *vox* target site sequence had been previously observed in the Buchholz lab.

In earlier research conducted in the Buchholz lab, the *Vika/vox* system had been discovered and characterized (Karimova et al., 2012). During that time, several *vox*-like target sites were identified in the human genome. One of these target sites, named *voxH9*, is located on chromosome 9 in an area that could qualify as a safe harbor locus. This target site contains a left half-site that is highly similar to the *vox wt* site (with only two mismatches at the first two positions furthest away from the spacer), while the right half-site contains seven mismatches to the *vox* site, making it a promising candidate for a potential entry point (Figure 17A). Four positions closest to the spacer were conserved in both half-sites when compared to the *vox* site, suggesting a high likelihood that *Vika* might still have activity on this site. Indeed, when cloned into a pEVO reporter vector, this site was successfully recombined upon expression of *Vika* recombinase (Figure 17B). This site was, therefore, selected as the target for *Vika*-mediated integration.

In order to increase the efficiency of integration and decrease the likelihood of re-excision, a screen of possible 'trap sites' was designed based on the RE/LE strategy (as illustrated in the scheme shown in Figure 17C). This process led to the identification of two prospective candidates for successful integration with a lock-in mechanism: *voxN4.2* and *voxN4.4* (refer to Figure 17D for their sequences). Despite these promising findings, successful integration in the endogenous human H9 locus had not been detected before the start of my work.



## EVOLUTION OF VIKA RECOMBINASE FOR RECOMBINASE-MEDIATED INTEGRATION IN POTENTIAL SAFE HARBOR LOCUS



**Figure 17. Vika activity on pseudo-*vox* site (*voxH9*) and design of the lock-in integration strategy.**

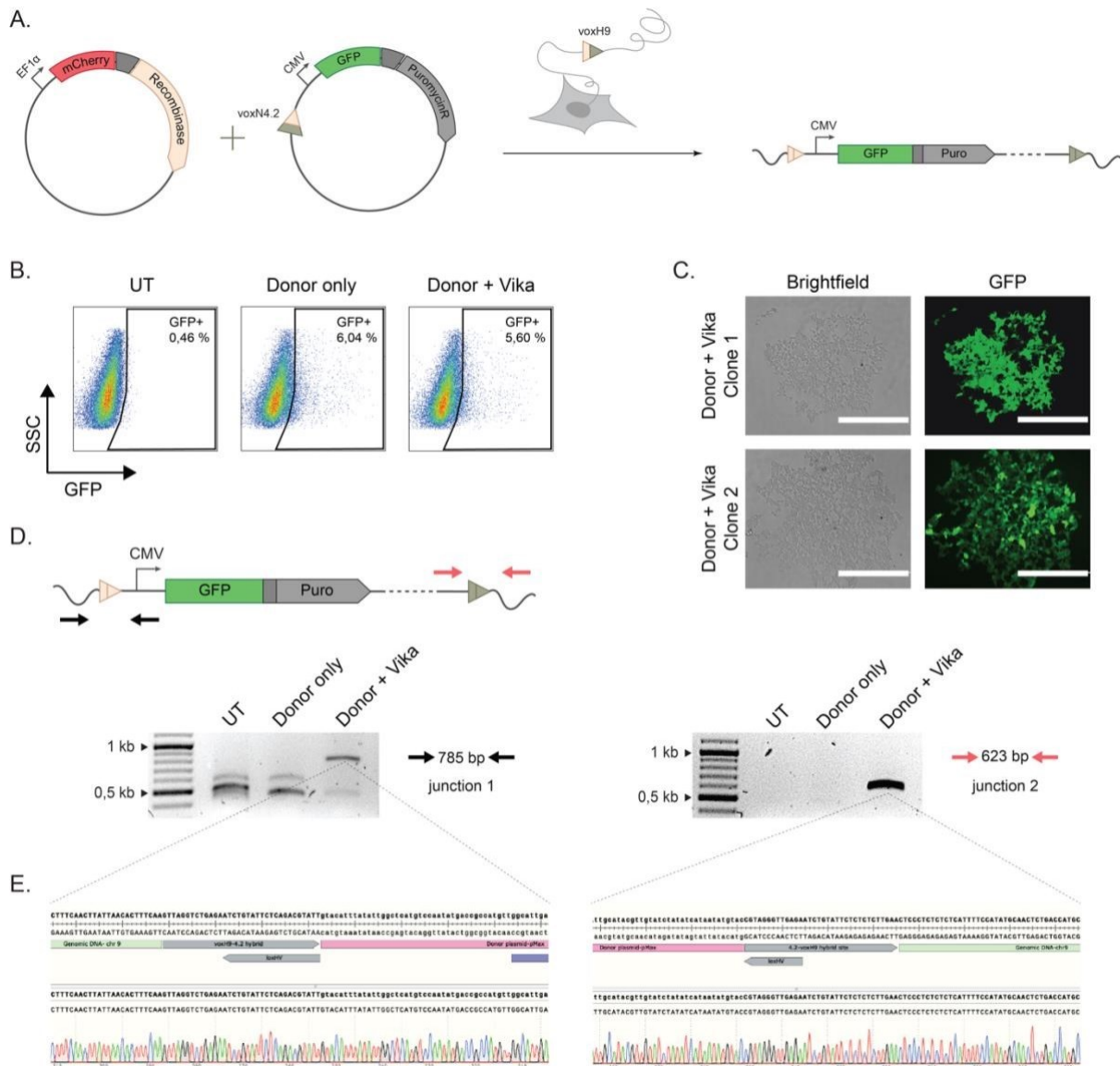
**A.** Schematic depiction of *voxH9* target site position and architecture. Nucleotides that are different from the *wt vox* site are depicted in red. Left half site is marked with the lighter shaded box as it is more similar to the *wt vox* half-site, whereas darker shaded square represents the highly mutated right half site. **B.** Comparison of the recombination activity of Vika on *wt vox* or *voxH9* target sites. The recombination activity was assessed by gel electrophoresis of the test digest. Both samples were grown in the medium containing 100 µg/ml L-arabinose to induce recombinase expression. Recombination is indicated by the line with the single triangle, whereas a line with two triangles illustrates the non-recombined band. M = GeneRuler™ DNA Ladder Mix (Thermo Fisher). **C.** The RE/LE strategy (up) and the screening pipeline for finding the trapping LE target sites (down). Library of target sites containing randomized mutations in the left half-site was subjected to rounds of positive selection for high integration efficiency and negative selection against re-excision. **D.** Sequences of the *vox* variants used for integration. The LE-*voxN* trap sites that emerged as best candidates from the selection are named *voxN4.2* and *voxN4.4*. Opposite to the *voxH9* these are LE sites meaning that mutations are on the left half-sites, depicted with the dark box.

### 7.2 Vika *wt* mediates integration into endogenous *voxH9* locus

To investigate whether Vika could facilitate the integration of a donor plasmid into the endogenous *voxH9* locus, I co-transfected a Vika expression plasmid with a donor plasmid containing the designed trap site (*voxN4.2*), CMV promoter, and GFP-P2A-PuroR expression cassette into HEK293T cells (as depicted in Figure 18A). The hypothesis was that only through integration into the genome could persistent GFP expression and puromycin resistance be achieved. However, flow cytometry analysis conducted ten days post-transfection did not reveal higher levels of GFP positive cells compared to controls

## EVOLUTION OF VIKA RECOMBINASE FOR RECOMBINASE-MEDIATED INTEGRATION IN POTENTIAL SAFE HARBOR LOCUS

where only the donor plasmid was transfected (Figure 18B). I then introduced puromycin into the growth media to initiate selection, and after one week, green colonies were observed (Figure 18C). While control samples where only the donor plasmid was transfected produced several colonies, the number of colonies on plates where the Vika expression plasmid was co-transfected was slightly higher.



**Figure 18. Vika wt can integrate donor DNA into human endogenous *voxH9* locus.**

**A.** Overview of the experiment. Recombinase expression plasmid and donor plasmid were co-transfected into HEK293T cells. Only upon the integration of the donor plasmid into endogenous *voxH9* locus, a constitutive expression of GFP-P2a-PuroR cassette can be achieved, allowing for the quantification and selection. **B.** FACS plots of un-transfected control sample, sample where only donor was transfected and the of the sample where both plasmids are transfected measured 10 days after transfection. **C.** Microscopy pictures of two examples of clones growing after puromycin selection. Both are coming from the sample where both plasmids were co-transfected. Scale bar is 400 μm. **D.** Junction PCRs performed to validate the expected integration outcome. Black arrows represent sense and antisense primers for the 5' junction (junction 1), while red arrows represent primers for the 3' junction (junction 2). Gel pictures show that the specific bands of 785 bp for junction

## EVOLUTION OF VIKA RECOMBINASE FOR RECOMBINASE-MEDIATED INTEGRATION IN POTENTIAL SAFE HARBOR LOCUS

1 and 623 bp for junction 2 are visible only in the sample containing the Vika expression plasmid. **E.** Screenshot of the Sanger sequencing results of the junction PCRs, aligned to the expected integration product map. Parts of the sequence belonging to the genomic DNA, target site and donor plasmid are depicted and show that the integration happened as expected.

To differentiate random integration events from Vika-mediated integration via the *voxH9* target site, I extracted genomic DNA from the bulk and performed a PCR over the expected junctions on the upstream and downstream borders of the integrated construct and genomic DNA (as illustrated in Figure 18D). The integration PCR band over both junctions was apparent only in the sample where the Vika expression plasmid was co-transfected with the donor plasmid, while no correct size band was evident in control samples (Figure 18E). The isolated, right-sized bands were then sent for Sanger sequencing to verify the expected recombination product (Figure 18F). This finding confirmed the successful detection of Vika-mediated integration into an endogenous human locus for the first time, and it could make a contribution in the genome engineering field, as it offers novel tool for transgene integration and stable expression from a potential safe harbor locus in the human genome. However, the integration efficiency, as estimated from the percentage difference of GFP+ cells in Vika and control samples in this experiment is relatively low for applications where selection of integrated clones is not an option, as could potentially be in a therapeutic setting. This prompted me to explore ways to enhance Vika's efficiency through evolution.

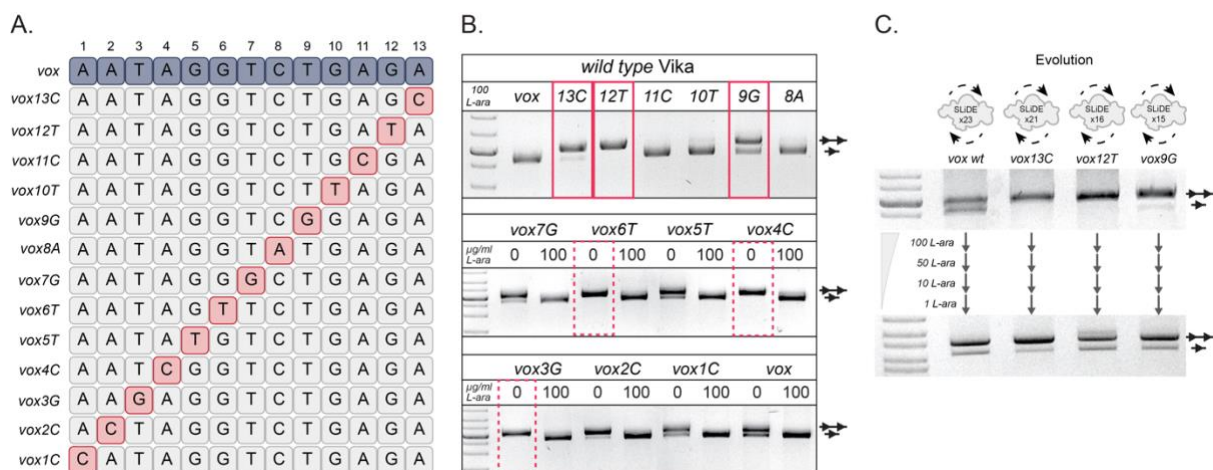
### 7.3 Establishing Vika libraries

As demonstrated previously, Vika recombinase can very efficiently and specifically recombine its predicted target site, *vox* (Karimova et al., 2012). However, lack of detailed structural data of protein-DNA interface, which is available for Cre//*loxP* system for example, precludes the application of rational design approaches. To fill this gap, I have designed thirteen single mutants of the *vox* target site, each with a single mutation in 13-bp half site on different position (Figure 19A). The nucleotides were exchanged purine for pyrimidine and vice versa and chosen so that the base pair is always switched in order to introduce the most significant change possible. I then tested the effect of these mutations on *wt* Vika's activity, in hope to reveal positions in the half-sites that are adding more weight to the complex relationship between the target sites and the recombinase that define the specificity and efficiency of recombination. As expected, with high induced levels of Vika expression most of mutations alone didn't have effect on activity except changes on 13th, 12th and 9th position (Figure 19B). These results indicated that those positions are likely where specific protein-DNA interactions are occurring. As expected, positions further away from the spacer were less influenced by the mutations, so I tested Vika activity on those without addition of L-arabinose as well. Under this condition, I could detect decreased activity on positions 6,4

## EVOLUTION OF VIKA RECOMBINASE FOR RECOMBINASE-MEDIATED INTEGRATION IN POTENTIAL SAFE HARBOR LOCUS

and 3 as well (Figure 19B). This finding supports data acquired from Cre-based evolutions, further indicating that Vika is probably functioning in a Cre-like manner.

From there, I sought to create initial Vika-based libraries. Efficiency of evolution largely depends on the size and diversity of starting libraries. Theoretical studies have shown that there is an optimal mutation rate depending on protocol and protein. Low mutation rates severely limit the sequence space that is covered by a library. Although a high mutation rate introduces many destabilizing and inactivating mutations, and strongly decreases the proportion of active enzymes, it often leads to improved variants (Drummond et al., 2005). Thus, I have generated initial library by using error-prone PCR with conditions that have been reported to increase mutational rate up to  $10^4$  (Wong et al., 2006), such as increased magnesium and manganese concentrations as well as unbalanced concentrations of nucleotides (See Methods section for details). Upon cloning of error-prone PCR product in the pEVO reporter vector initial library produced around 500,000 clones.



**Figure 19. Establishment of the Vika libraries.**

**A.** Single mutant-vox target site design. Thirteen target sites, each carrying single mutation on a different position were created in order to test the effect of isolated position change on wt Vika activity. **B.** Agarose gels showing the wt Vika activity on single mutant vox target sites. Mutants containing the changes at the positions closer to the spacer (8-13) were tested only at 100 µg/ml L-arabinose (upper gel), whereas the mutants of the further positions were examined also when no L-arabinose was added to the medium (marked with '0' in the middle and lower gels). The target sites where Vika lost most of its activity when expressed at 100 µg/ml L-arabinose are marked with the red squares, which were also chosen for the evolution. Dashed red squares depict target sites where Vika showed reduced activity but only when expressed very low from the leaky arabinose promoter without the arabinose induction. **C.** Representation of the SLiDE progress on four different target sites (wt vox, vox13C, vox12T and vox9G). The number of cycles for each target site is shown in the picture. The decreasing arabinose levels used during the evolutions are depicted next to the arrows. A test digest demonstrating the recombination activity of the first cycle at 100 µg/ml L-arabinose and last cycle of 1 µg/ml L-arabinose is presented. The unrecombined and the recombined bands are denoted as a line with two triangles and one triangle, respectively.

## EVOLUTION OF VIKa RECOMBINASE FOR RECOMBINASE-MEDIATED INTEGRATION IN POTENTIAL SAFE HARBOR LOCUS

To further diversify the libraries, I have initiated the SLiDE protocol on three target sites where Vika had exhibited reduced activity (*vox13C*, *vox12T*, *vox9G*) and also on the *wt vox* site so that I can distinguish changes that are contributing to specificity switch from ones that are improving overall efficiency, independent of the nucleotide sequence of the target site. Upon expression starting library didn't show any activity with test digestion on *vox13C* and *12T* sites similar to Vika, and showed impaired activity on *vox wt* and *vox9G* compared to *wt* enzyme (Figure 19C). After approx. 25 cycles of evolution, I acquired measurable activity of each library upon induction with 1 µg/ml L-arabinose (Figure 19C).

The four final libraries were sent for deep sequencing and differences between them were analyzed (Supplementary Figure S7). Although many mutations were detected, libraries as a whole did not diverge significantly from one another. This is likely because the selective pressure when changing only one nucleotide in the target site is not strong enough and only a few changes that lead to relaxed specificity are enough to be selected and enriched. Nevertheless, several clearly enriched mutations emerged after 25 cycles of evolution, such as W56L, Q62R, E150K, H240R and H348L (Supplementary Figure S7). Comparative sequence analysis between Vika and Cre revealed that some of these changes correspond to commonly changed positions in Cre-based evolution. However, since crystal structure of the Vika protein is still not solved, I didn't investigate these changes further. Nonetheless, these libraries present the first libraries established in our laboratory that are not Cre-based and were used as starting points for further evolution campaigns to different targets, including the *voxH9* in this work.

### 7.4 Integration-based SLiDE

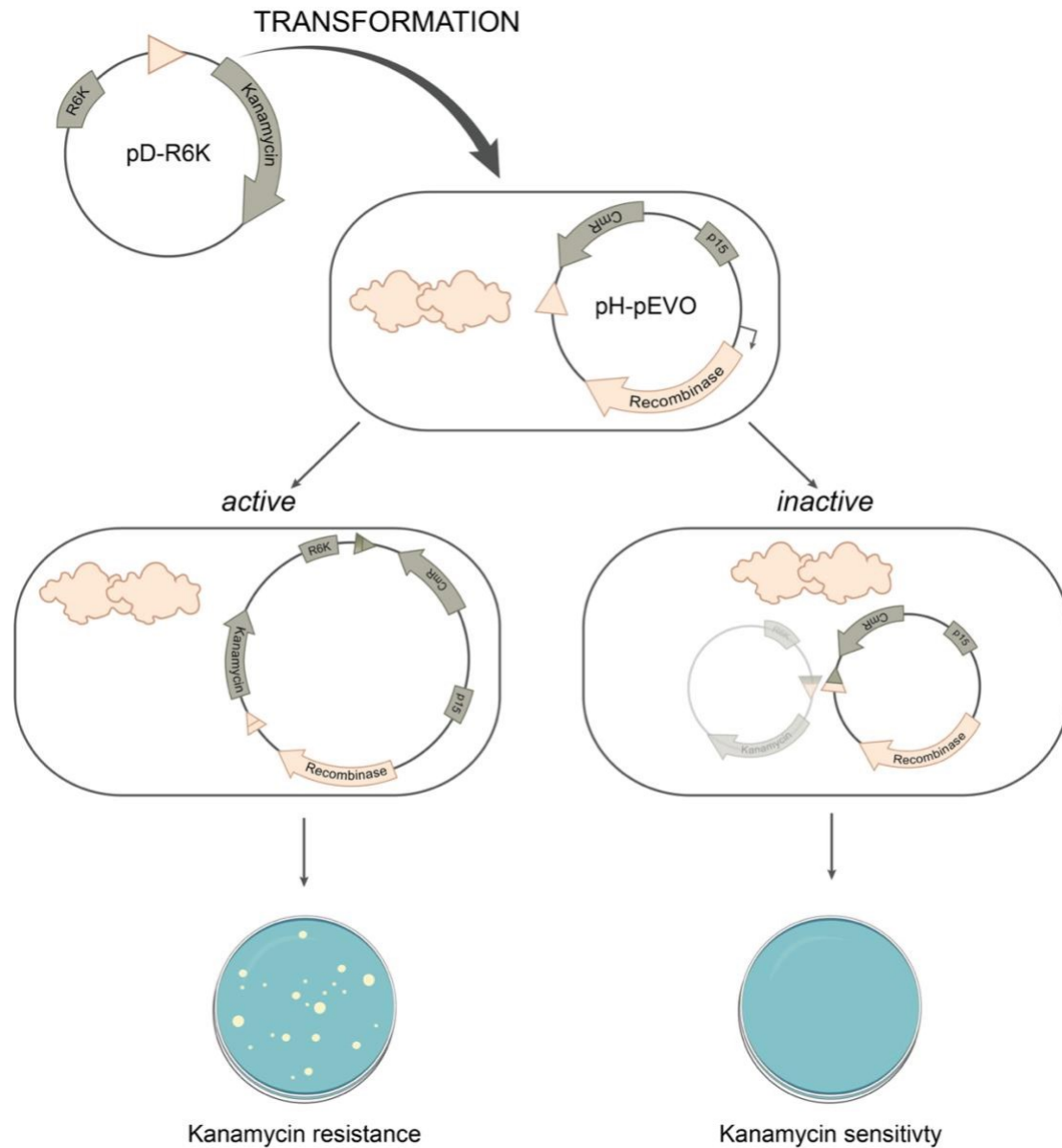
Vika was shown to be very active recombinase when challenged in the excision assay, reaching recombination efficiency on its native *vox* target site as high as Cre on the *loxP* (See Figure 9A) (Karimova et al., 2012). Furthermore, it showed over 40% excision efficiency on *voxH9* target site (Figure 17B). Nevertheless, efficiency of integration into human genome seems to be quite low. Even though, excision and integration are mechanically same recombination reactions with only difference that the first is happening between the target sites on the same DNA molecule while the latter between two DNA molecules, some recombinases still seem to compare differently when tested for integration or excision (Karimova et al., 2016). Thus, in order to provide high enough selection pressure and to select specifically for increased integration efficiency, I decided to optimize the SLiDE protocol to be based on the plasmid-integration assay.

## EVOLUTION OF VIKA RECOMBINASE FOR RECOMBINASE-MEDIATED INTEGRATION IN POTENTIAL SAFE HARBOR LOCUS

The assay is designed with plasmids based on two types of origin of replication: on p15A for the host pH plasmid (chloramphenicol resistant - pEVO plasmid), and on R6K for the donor pD- plasmid (kanamycin resistant) (Figure 20). Replication of the R6K-plasmid is pi-protein dependent and therefore this plasmid cannot replicate in a pir-negative bacterial strains (Stalker et al., 1983). Hence, kanamycin resistance conveyed by the donor plasmid depends on its integration into a host plasmid through a *lox* site to form the co-integrant and consequent replication of the kanamycin resistance gene in pir-negative cells.

To initiate the evolution, I mixed the four Vika libraries (refer to Chapter Establishing Vika libraries) and cloned them in modified pEVO vector carrying just one copy of *voxH9* target site (see Figure 21A). During each round of evolution, this plasmid was co-transformed together with the pD plasmid (R6K plasmid) carrying either *voxN4.2* or *voxN4.4* designed LE trap sites (Figure 21B). Recombinase library expression was induced for first 2h after transformation during the recovery and then the samples were grown and selected overnight in chloramphenicol and kanamycin media (Cm+Kan) (Figure 21C). Only the active variants that were able to mediate the integration of the plasmids could grow in the culture and were selected for the next cycle (Figure 21D). Next day, DNA was extracted from the culture and test digest was performed for each cycle to confirm the presence of integration product (Figure 21E). Furthermore, plasmid DNA was subjected to PCR with primers specific for the co-integration products pEVO-R6K plasmids that will amplify only the active recombinases (Figure 21F). This PCR is performed with a low-fidelity DNA polymerase to add diversity to the pool of active recombinase variants. The diversified active recombinase fragments are then re-cloned into a non-recombined pEVO backbone which marks the beginning of the new cycle.

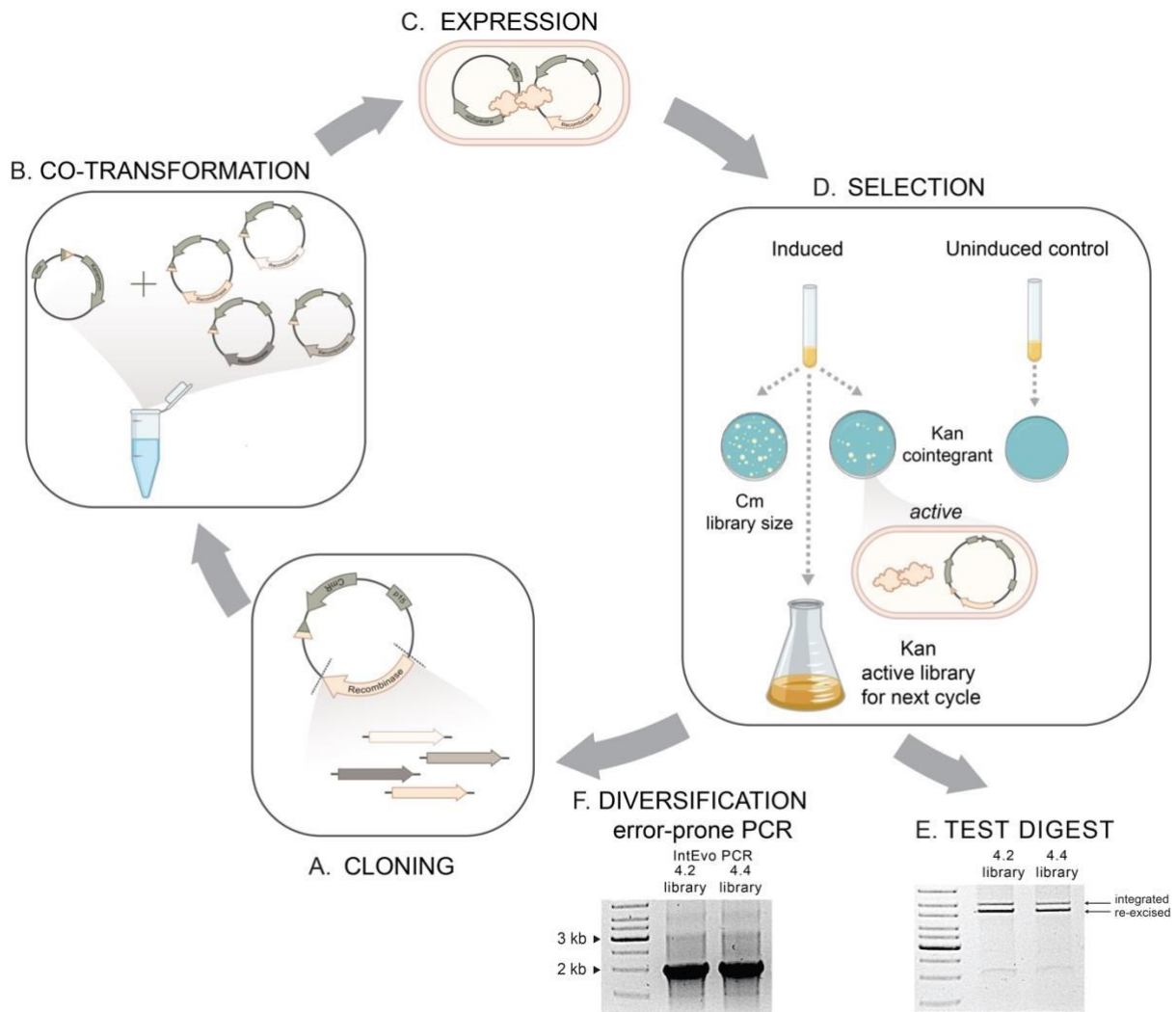
EVOLUTION OF VIKa RECOMBINASE FOR RECOMBINASE-MEDIATED INTEGRATION IN POTENTIAL SAFE HARBOR LOCUS



**Figure 20. Integration based assay.**

The assay is designed with plasmids based on two types of origin of replication: on p15A for the host pH plasmid (chloramphenicol resistant - pEVO plasmid), and on R6K for the donor pD- plasmid (kanamycin resistant). Donor plasmid is transformed into E. coli strain carrying the pEVO vector and expressing the recombinase. If recombinase is not able to perform the integration the bacteria remain sensitive to kanamycin while replication of the R6K-plasmid is pi-protein dependent and therefore this plasmid cannot replicate in a pir-negative bacterial strains. Only upon successful integration into a host plasmid through a *lox* site, the co-integrand product can ensure replication of the kanamycin resistance gene in pir-negative cells, allowing for the kanamycin resistant colonies to grow on the plate.

## EVOLUTION OF VIKA RECOMBINASE FOR RECOMBINASE-MEDIATED INTEGRATION IN POTENTIAL SAFE HARBOR LOCUS



**Figure 21. Integration-based Substrate-linked directed evolution (IntSLiDE).**

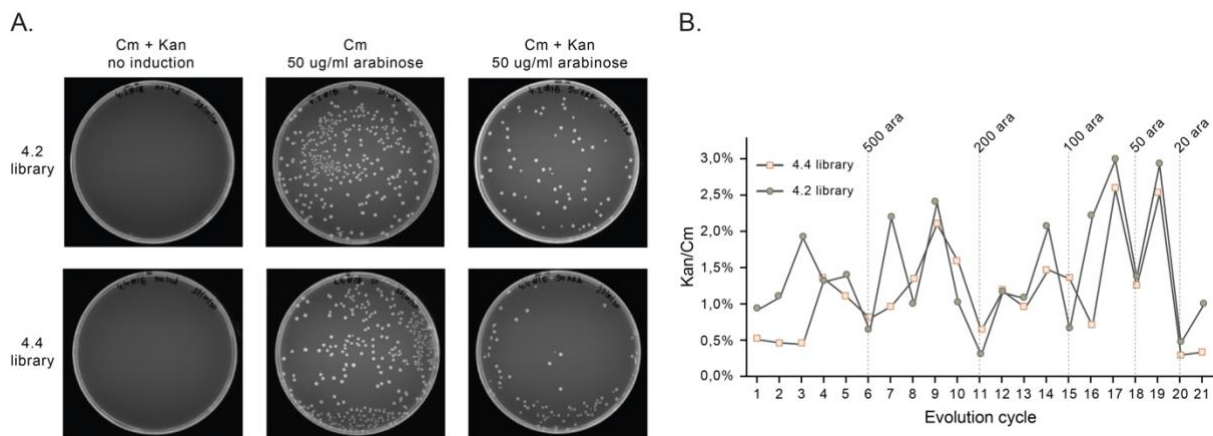
**A.** Library of recombinase variants is cloned into pEVO-ΔvoxH9. **B.** This plasmid was then co-transformed together with the pD plasmid (R6K plasmid) carrying either *voxN4.2* or *voxN4.4* designed LE trap sites into *E. coli*. **C.** In order to express the recombinase libraries, arabinose was added to the recovery medium, and the samples were recovered for 2h to ensure enough recombinase expression before the start of selection. **D.** Induced recovery was used to start Cm+Kan overnight culture to select for active variants. Additionally small amount was used to plate on Cm or CM+Kan plates to estimate the recombinase library size, or size of the active fraction, respectively. **E.** DNA was prepped and used for test digest in order to confirm the presence of the integration product, by detecting the approx. 6 kb large band on the agarose gels. **F.** Plasmid DNA was also subjected to the selective PCR that amplifies only the recombinases from the integration product plasmid. PCR is introducing novel mutations and is used to clone the library in the fresh pEVO-ΔvoxH9 plasmid in order to start the next cycle.

Over the course of 24 evolution cycles that were performed, the recombinase expression level was stepwise lowered by adding decreasing amounts of L-arabinose. The decrease in recombinase expression creates a selection pressure that allows only very active recombinases to remain in the library and propagate. Contrary to standard excision-based SLiDE, where one can easily follow the progress of evolution by regularly doing the standard test digestion (previously described in the Introduction chapter - section Directed evolution of



## EVOLUTION OF VIKA RECOMBINASE FOR RECOMBINASE-MEDIATED INTEGRATION IN POTENTIAL SAFE HARBOR LOCUS

recombinases), in this setting, no such readout exists. The test digest in this case could just serve as quality control to make sure that co-integrant band is present in the sample (Figure 21E). The only way to get an idea of the progress is to follow the increase of the number of colonies on Cm+Kan plates. In order to account for differential transformation efficiency from one cycle to the next, the evolution samples were also plated on Cm plates. Only then the ratio between number of colonies on Cm+Kan plates and the number on Cm plates could serve for estimating the relative size of active fraction of the library (Figure 22A). Since the method is sensitive to experimental variation for independent samples, the data still could not provide clear picture of progress, but the upward trend was observed when looking at the ratio throughout the evolution cycles (Figure 22B). Interestingly, the most significant drops in the ratio values coincided with decreasing of the arabinose concentration.



**Figure 22. The IntSLiDE evolution progress.**

**A.** Representative pictures of plates with the libraries after 18 rounds of evolution. Recoveries were split into two and one part was induced with arabinose whereas other one served as a non-induced control, where no kanamycin resistant colonies could be detected. By determining the number of colonies on Cm and Cm+Kan plates and calculating the ratio between the two, the evolution progress could be quantified. **B.** A graph showing the quantification of evolution progress. After each cycle, the number of colonies on Cm and Cm+Kan plates was determined, the ratio between the two was calculated and shown as the percentage on the Y axis. Beige squares represent the ratios for each cycle of 4.4 library evolution, whereas green circle represents 4.2 library. Dash vertical lines label the cycles where the L-arabinose concentration was changed in order to decrease the recombinase expression

In summary, the evolution process employed a series of 24 cycles, leveraging co-transformation, selection, and diversification to enhance recombinase activity while adjusting selection pressure by gradually lowering recombinase expression levels. Although monitoring the progress proved challenging due to the lack of a clear readout, an upward trend in the active fraction of the library was observed, indicating successful evolution under the imposed selective pressures.

## EVOLUTION OF VIKA RECOMBINASE FOR RECOMBINASE-MEDIATED INTEGRATION IN POTENTIAL SAFE HARBOR LOCUS

**Table 2. Mutations observed in consensus sequencing of the libraries.**

Mutations that were already observed in starting libraries are labeled in green. Common mutations for both IntSLiDE libraries are depicted in pink, while library specific changes are depicted in beige.

4.2 library	4.4 library
W56L	W56L
A59T	
Q62R	Q62R
E150L	E150L
Q151R	
N193S	
Q234R	
N237S	
H240R	H240R
	V260A
	K261R
K265E	
	N273D
T274A	
L275P	
N276S	N276S
E300G	E300G
	Q316L
H348L	
Q356L	

Moving on, I wanted to examine the mutations that emerged as a result of this integration-based molecular evolution in the two libraries. Consensus sequencing of the final libraries revealed that major changes occurred at sixteen positions in 4.2 library and ten positions in 4.4 library (Table 2). The four changes most pronounced in the starting libraries remained conserved in the final libraries even after evolution on different target site and selection for integration (W56L, Q62R, E150L and H240R). Interestingly, changes on two positions N276S and E300G emerged independently in both libraries only after this integration-based evolution, as those weren't previously documented in starting libraries (Table 2). Additionally, I could observe library specific mutations as follows: in 4.2 library – A59T, Q151R, N193S,

## EVOLUTION OF VIKA RECOMBINASE FOR RECOMBINASE-MEDIATED INTEGRATION IN POTENTIAL SAFE HARBOR LOCUS

Q234R, N237S, K265E, T274A, L275P, H348L, Q356L; in 4.4 library – V260A, K261R, N273D, Q316L (Table 2). In conclusion, the consensus sequencing analysis revealed significant positional changes in both 4.2 and 4.4 libraries after integration-based evolution, with some mutations conserved across libraries and others being library-specific, highlighting the dynamic nature of molecular evolution in response to distinct target sites and selection pressures.

### 7.5 Screening of best performing clones in bacteria

After the evolution of these two different libraries specific for integration of two distinguished donor target sites (*voxN4.2* and *voxN4.4*) into *voxH9* target site, I wanted to confirm that the evolution yielded more active variants than Vika. In order to accomplish that, I randomly picked 3 clones from each library, and did a one-day integration assay in parallel to *wt* Vika (see Figure 23). The six clones that were randomly picked, were sequenced and the alignment to *wt* Vika revealed that all of them were indeed different variants, having from 14 to 25 mutations compared to the *wild-type* reference (Figure 23A). Each clone and *wt* Vika were separately cloned into pEVO-*voxH9* vector and co-transformed with R6K-donor vector harboring appropriate *voxN*-trap site. The number of colonies on double selective plates with both chloramphenicol and kanamycin antibiotics (Kan+Cm) was counted and compared between the clones and *wt* Vika (Figure 23B). Since the number of colonies on plates with only chloramphenicol was comparable between all of the samples (as can be seen in the Figure 23B), this time I wasn't calculating the ratio between the number of colonies on two different plates. Noticeably, all the clones showed increased number of colonies in this test compared to Vika *wt* (Figure 23C). Both 4.4 library and Vika seem to perform more efficient integration once exposed to *voxN4.4* donors compared to 4.2 library and Vika with *voxN4.2* donor. Clone 3 from 4.2 library and clones 2 and 3 from 4.4 library showed significant increase in number of colonies, up to 5-fold in the case of 4.4 clone 2 (Figure 23C). These results demonstrated the success of integrative evolution which was one of the crucial milestones of my work. Next, I sought to screen the libraries as to nominate the best performing candidates to test in the human cells.



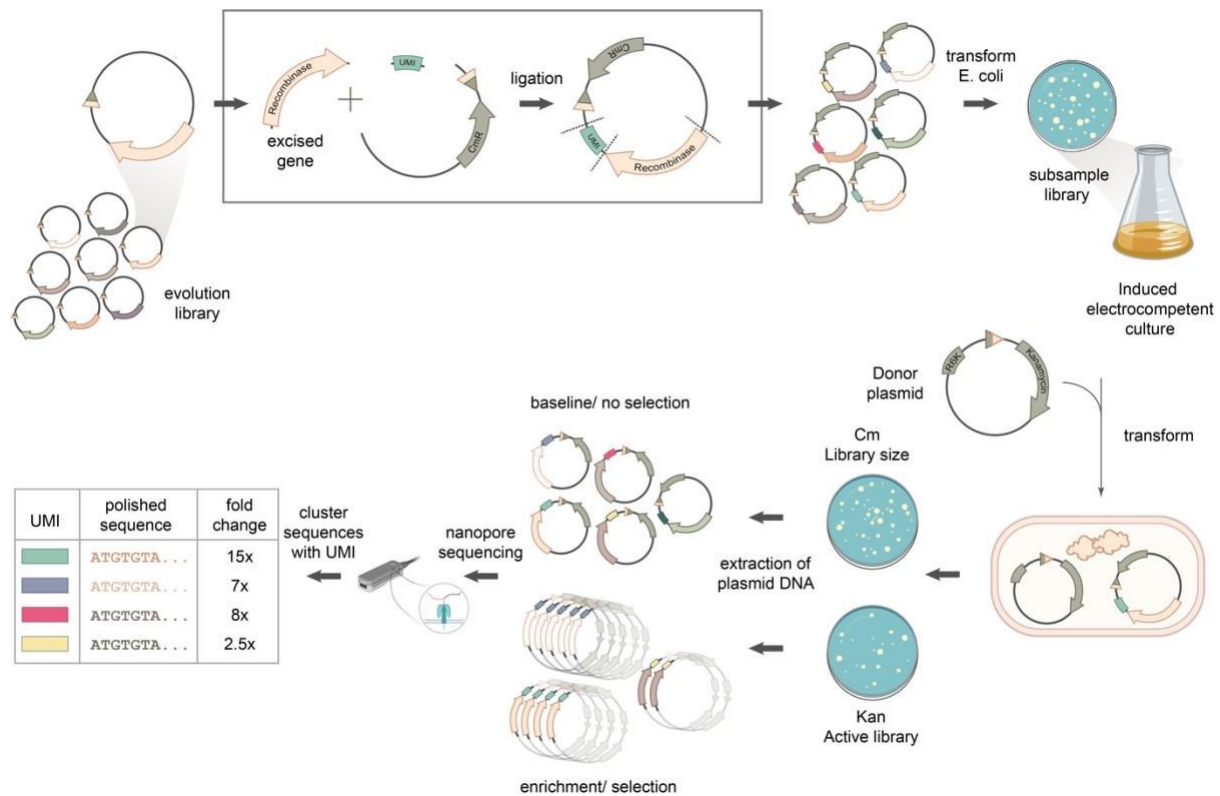
## EVOLUTION OF VIKA RECOMBINASE FOR RECOMBINASE-MEDIATED INTEGRATION IN POTENTIAL SAFE HARBOR LOCUS

without the guarantee that the best clone in the library is chosen. Recently, in our lab we developed DEQSeq (DNA Editing Quantification Sequencing), a high-throughput screening platform that enables characterization of thousands of DNA editing enzyme variants on multiple target sites. The method utilizes nanopore technology for sequencing full-length enzyme variants at fast turn-around times. Through clustering of unique molecular identifiers (UMI), a highly accurate consensus sequence is generated (Zurek et al., 2020; Karst et al., 2021). By additionally capturing the target site with the enzyme sequence, the DNA editing rate for each enzyme variant can be quantified. DEQSeq is excision based, so I decided to optimize this protocol to screen for the variants that mediate integration the most efficiently, and from now on I will be referring to this method as IntDEQSeq (Integration-based DNA Editing Quantification Sequencing).

### 7.5.1 IntDEQSeq screen

The optimization of the method mainly consisted in defining the parameters for the quantification of integration efficiency. In excision-based assay this is quite straight forward, since the recombined and unrecombined fragments are sequenced at the same time and the ratio between the number of reads for each can be used as direct measurement of recombination efficiency. This ratio could be biased by the size difference between recombined and unrecombined plasmids, since smaller plasmids generally have an advantage in replication. This doesn't pose a big problem in excision-based assay since the size difference is small (~700bp), however, same is not true for integration assay where the recombined product is 3kb bigger than the unrecombined plasmid. This means that if not selected for, the integration event is highly unlikely to be detected, while on the other hand, selection is fully shifting the bias towards integration. Only way to make a quantitative setting where integration efficiency could be determined is to establish the relative representation of the clones in the sample before and after selection. Thus, the fold change of frequency of each clone in the library after selection can serve as an estimate of integration efficiency (Figure 24). Electrocompetent XL-1 Blue *E. coli* cells expressing subfraction of around 2000 variants from both 4.2 and 4.4 libraries, were electroporated with R6K donor plasmids in triplicates to ensure the reproducibility of the results. The samples were then grown in Cm and Cm+Kan media in order to compare the representations of each clone before and after selection. As a control I included wt Vika as well.

## EVOLUTION OF VIKA RECOMBINASE FOR RECOMBINASE-MEDIATED INTEGRATION IN POTENTIAL SAFE HARBOR LOCUS



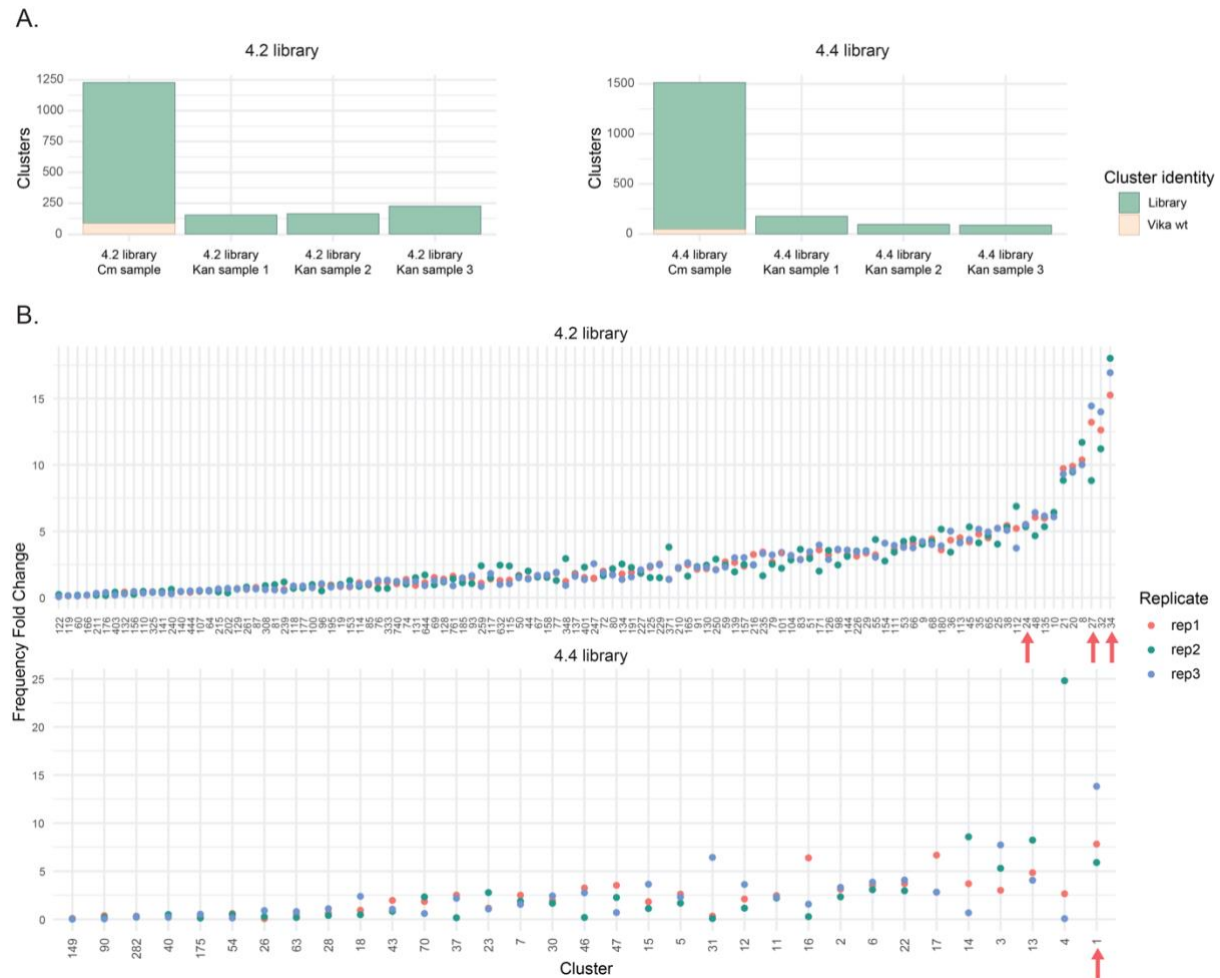
**Figure 24. Integration-based DNA Editing Quantification Sequencing (IntDEQSeq) screen workflow.**

Evolved recombinases are cloned together with a unique molecular identifier (UMI) into a vector containing single voxH9 site (pEVO- $\Delta$ voxH9). A defined number of transformed bacteria are cultured to express the enzymes and brought to electro competency. Appropriate donor plasmids are transformed into the *E. coli* expressing the recombinase library. The samples are then split and grown overnight in LB media containing chloramphenicol and kanamycin, and chloramphenicol only separately. This ensures representation of the whole library and the selected fraction that is active to perform the integration action. From all plasmids the region of interest is excised and sequenced by nanopore sequencing. Clustering of the UMIs allows consensus sequence polishing and counting of the recombinase variants. The frequency of reads of each variant is then calculated and compared between selected and non-selected samples. The frequency fold change serves as the estimate of the integration efficiency.

In total, screen yielded approx. 1200 UMI-clusters in 4.2 library and 1500 UMI-clusters in 4.4 library as could be seen in Cm samples (Figure 25A). From these I identified 120 clusters to be Vika control in 4.2 sample and 50 in 4.4 sample. As expected, a big reduction in number of UMI-clusters was observed in selected samples that were grown in presence of both of antibiotics (Cm+Kan). Interestingly, not a single UMI-cluster presenting Vika wt control could be detected in Cm+Kan samples transfected with either of the two donors, suggesting that Vika is just not efficient enough in comparison to the other clones and was massively outcompeted (Figure 25A). Furthermore, it should be noted that 4.4 library seemed to lose much larger fraction of the clusters during selection compared to 4.2 library (Figure 25A). As to identify the most successful clones, I filtered for clusters that exist in all replicates of the sample (Cm included), calculated the read frequency of each replicate (Cm included), and

## EVOLUTION OF VIKA RECOMBINASE FOR RECOMBINASE-MEDIATED INTEGRATION IN POTENTIAL SAFE HARBOR LOCUS

then I removed all the clusters that don't have at least 50 reads in all of the replicates. Finally, I calculated the fold change of the Cm+Kan samples to the Cm samples (Figure 25B). This yielded clusters whose representation increased up to 17-fold compared to the non-selected sample in the 4.2 library and 15-fold in 4.4 library (Figure 25B).



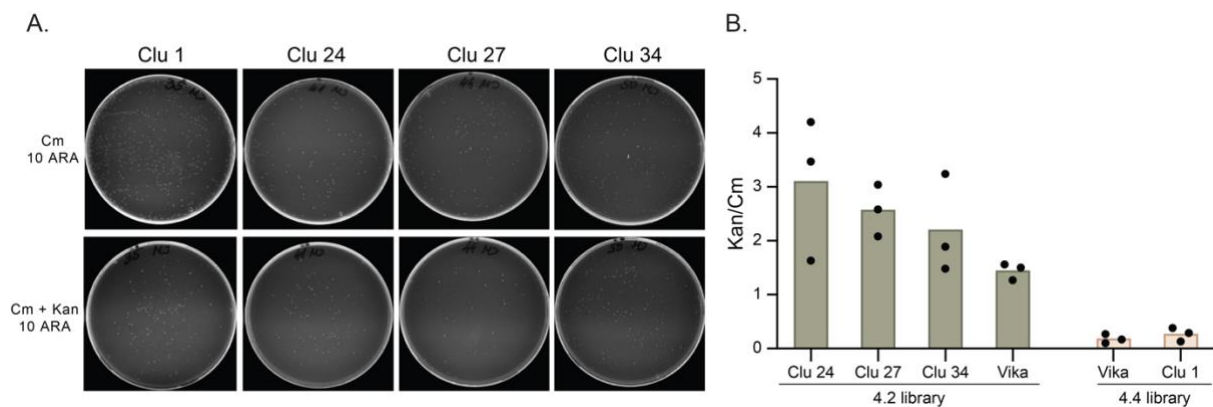
**Figure 25. The results of IntDEQSeq screen.**

**A.** Graphs depicting the number of clusters in 4.2 (left) and 4.4 (right) library samples. Each library was transformed with the donor plasmids three times, and one of the replicates was split into Cm and Cm+Kan media, while the other two were just grown in Cm+ Kan media. In Cm samples one can distinguish between the Clusters connected to the Vika wt sequence (depicted in light beige), whereas no Vika clusters could be detected in Kan samples. Generally, less clusters were found in Kan samples which was expected after the selection. **B.** Frequency fold change of the best performing clusters from 4.2 (up) and 4.4 library (down). Each Kan replicate was compared to Cm sample in order to calculate the frequency fold change, and represented as a dot of different color – as depicted in the legend.

To verify the screening results, I selected several top-performing clones and compared their integration capabilities to Vika in a one-day integration assay. However, the results were inconclusive, and reproducibility between replicates was limited (data not shown). Despite these inconsistencies, I proceeded to examine three promising clones from the 4.2 library (Clusters 24, 27, and 34) and the most promising candidate from the 4.4 library, Cluster 1, in

## EVOLUTION OF VIKA RECOMBINASE FOR RECOMBINASE-MEDIATED INTEGRATION IN POTENTIAL SAFE HARBOR LOCUS

a follow-up assay (Figure 26A). The quantification of the active fraction ratio on Cm+Kan and Cm plates largely corroborated the screening outcomes, albeit with a few discrepancies. Contrary to the screening data, Cluster 24 demonstrated the best performance, followed by Cluster 27. Cluster 34, despite having the highest enrichment in the screen, displayed the lowest Cm+Kan to Cm ratio when tested individually (Figure 26B). All clones still exhibited greater activity than Vika wt. Interestingly, in contrast to previous findings from random picking of the clones, Cluster 1 from the 4.4 library and Vika wt with voxN4.4 donor exhibited lower activity than the 4.2 clones and Vika wt with voxN4.2 donor. Nonetheless, Cluster 1 appeared to display enhanced activity compared to Vika wt (Figure 26B). In conclusion, while the integration assay results contained some inconsistencies, the selected clones generally demonstrated improved activity over Vika wt, indicating potential advancements in the recombinase libraries.



**Figure 26. Validation of the IntDEQSeq clones.**

**A.** Cm and Cm+Kan plates of all tested clusters used for colony counting and quantification of integration efficiency. **B.** Quantification of integration efficiency by determining Kan to Cm colony number ratios. Independent replicates are marked as dots, while the box height represents the mean. Bars representing the clones from 4.2 library are labeled with green, whereas orange bars represent 4.4 library clones.

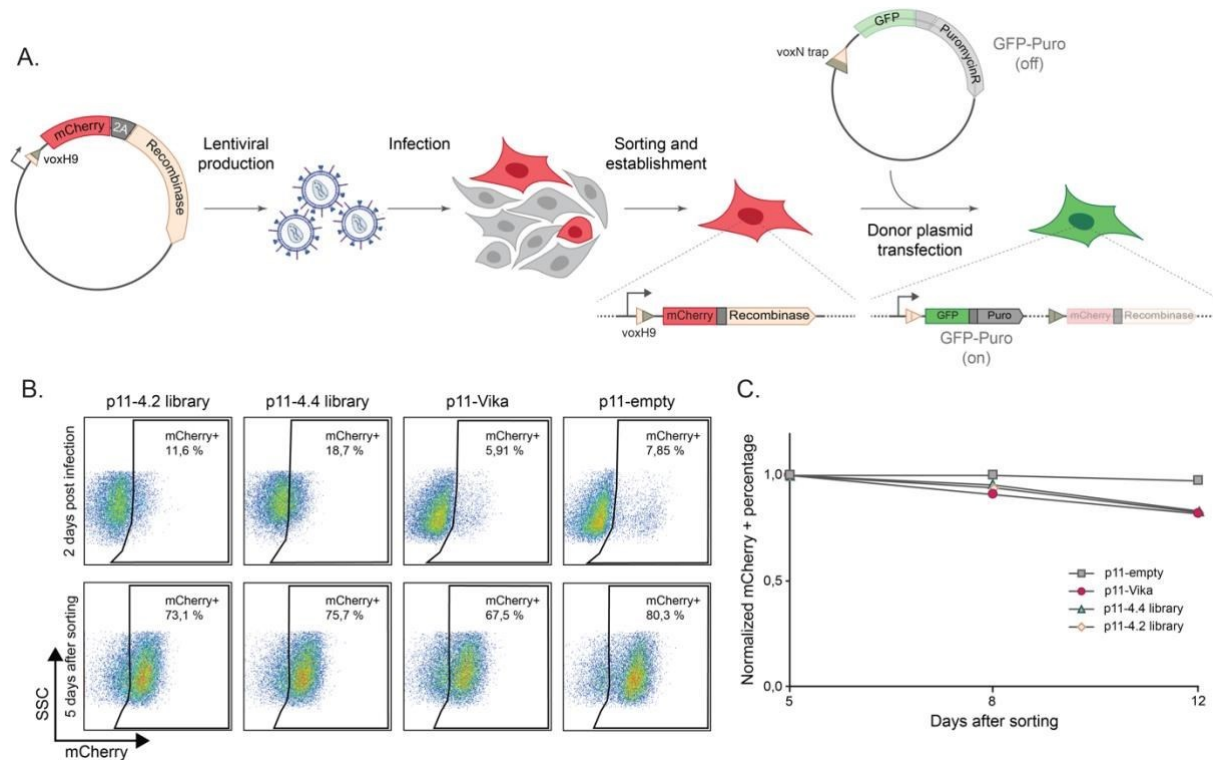
### 7.6 Screening of best performing clones in human cells

Building upon the previous findings in bacterial systems, it is important to note that the success of SLiDE evolution and DEQSeq, our novel clone screening method, does not always directly translate to mammalian cells due to differences in expression levels, folding, post-translational modifications, and degradation signals. To address this discrepancy, several evolution and screening methods have been established for mammalian systems (Berman et al., 2018; English et al., 2019; Hendel and Shoulders, 2021; Klenk et al., 2023). In pursuit of identifying the most effective clone for mediating integration in human cells, I opted to screen the libraries using a reporter cell line.



## EVOLUTION OF VIKA RECOMBINASE FOR RECOMBINASE-MEDIATED INTEGRATION IN POTENTIAL SAFE HARBOR LOCUS

To this end, I created landing pad cell lines in HEK293T cells by integrating a pEF-1 $\alpha$ -voxH9-mCherry-T2A-Recombinase cassette into the genome using a low multiplicity of lentiviral infection (MOI) (Figure 27A, B). I then expanded mCherry+ clones after sorting and assessed their clonal mCherry stability (Figure 27A, C). In that way, I established four different cell lines expressing 4.2 library, 4.4 library, wt Vika, or mCherry only (p11-empty) and observed that the cell lines with recombinases appeared somewhat less stable than the p11-empty landing pad cell line (Figure 27C).



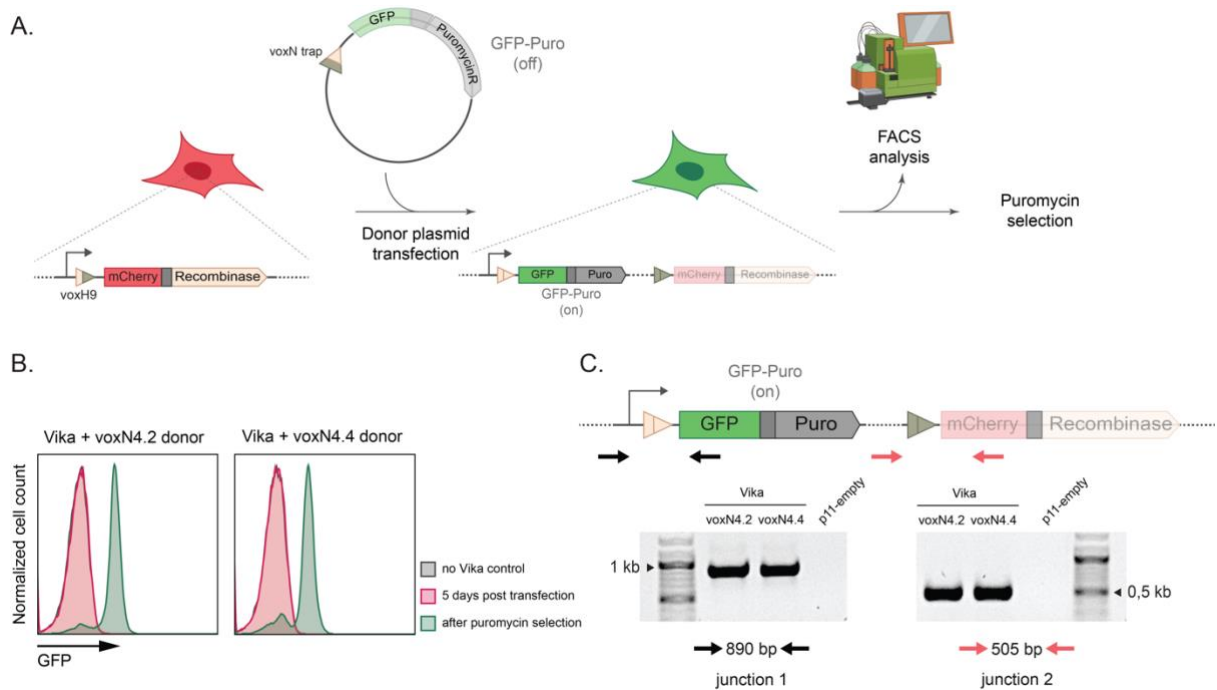
**Figure 27. Clone screening in human cells.**

**A.** Schematic of genomic landing pad assay. An EF-1 $\alpha$  promoter, voxH9 and mCherry-P2A-Recombinase cassette are integrated via lentivirus. Upon voxN donor transfection and successful integration into the landing pad, GFP and Puro are expressed, and the Recombinase and mCherry are displaced and knocked out. **B.** FACS plots depicting the landing pad cell lines establishment. The cells were infected with low MOI, as can be seen from the FACS analysis 2-day post infection (upper panel), and then sorted for mCherry expression to establish the cell line (lower panel). **C.** Expression of mCherry was followed over time in order to determine the stability of the construct and toxicity of the recombinases. The initial percentage of mCherry+ cells of each sample was normalized to 1, and then the following time points were presented relatively to the initial point.

This design allows for a promoter trapping approach, where successful recombination of the 3-kb vox donor plasmid results in a gain of GFP and puromycin N-acetyl-transferase (PAC – depicted as Puro or PuroR in the schemes and further text) expression while losing recombinase and mCherry expression (Figure 28A). The quantification can then be done via flow cytometry and puromycin selection can allow for detection of successful clones (Figure 28A). Initial testing with wt Vika showed no significant increase in GFP+ cells, but a clear

## EVOLUTION OF VIKA RECOMBINASE FOR RECOMBINASE-MEDIATED INTEGRATION IN POTENTIAL SAFE HARBOR LOCUS

GFP<sup>+</sup> population emerged after puromycin selection (Figure 28B). Successful integration was confirmed with PCR over upstream and downstream junctions and sequence verification (Figure 28C, Supplementary Figure S11).



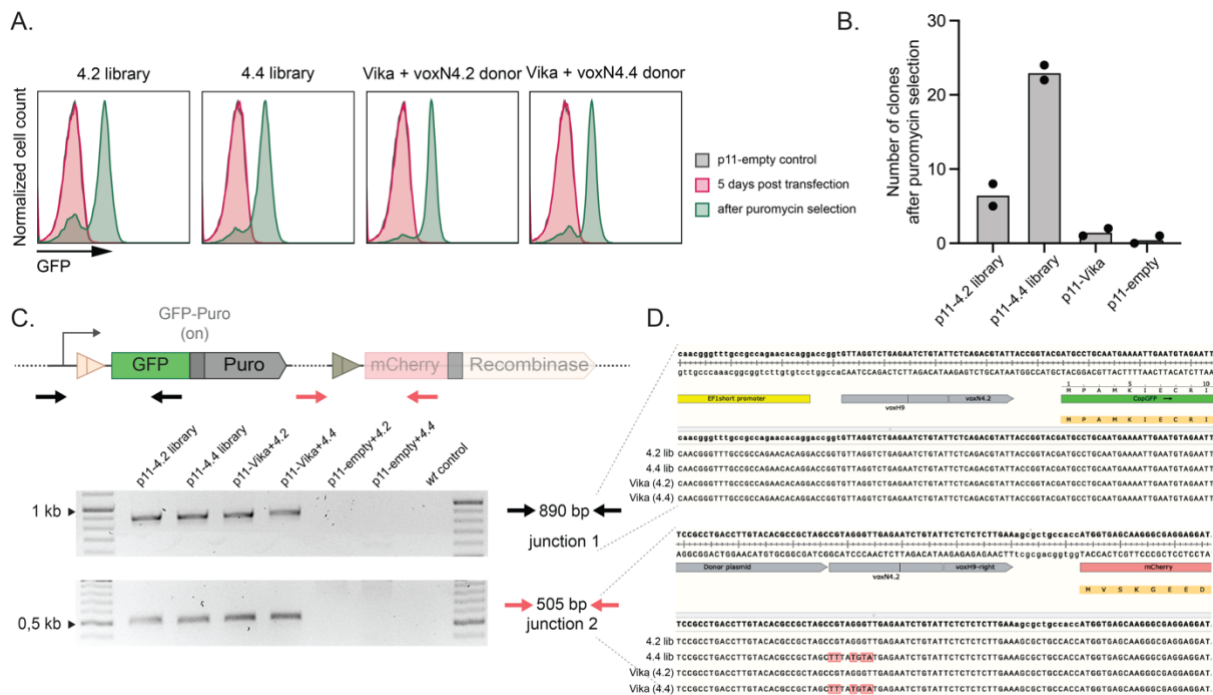
**Figure 28. Testing the landing pad construct with wild type Vika.**

**A.** Schematic of genomic landing pad assay. The HEK293T<sup>p11-Vika</sup> or HEK293T<sup>p11-empty</sup> cell line was transfected with either voxN4.2 or voxN4.4 donor plasmids. The integration outcome was identified by FACS analysis of GFP expression pre and post puromycin selection. **B.** FACS plots showing the cell populations from three samples: p11-empty, p11-Vika 5 days after transfection and, p11-Vika after puromycin selection examined for the GFP expression. The samples are labeled as described in the legend. **C.** Junction PCRs performed to validate the expected integration outcome into the landing pad reporter. Black arrows represent sense and antisense primers for the 5' junction (junction 1), while red arrows represent primers for the 3' junction (junction 2). Gel pictures show that the specific bands of 890 bp for junction 1 and 505 bp for junction 2 are visible only in the samples where HEK293T<sup>p11-Vika</sup> was transfected with the donors.

Subsequently, I conducted experiments with cell lines expressing the two libraries. Although the percentage of GFP<sup>+</sup> cells five days after donor transfection did not reveal a clear advantage between samples, both libraries demonstrated a distinct GFP<sup>+</sup> population after puromycin selection (Figure 29A). These cells were also mCherry<sup>-</sup>, consistent with the desired outcome of a GFP donor integrating into the landing pad while simultaneously displacing and knocking out the mCherry-Recombinase cassette, unlike the p11-empty sample (Supplementary Figure S12). Interestingly, in contrast to the IntDEQSeq results, the number of surviving clones after puromycin selection indicated a clear advantage for the 4.4 library in two independent experiments (Figure 29B). Integration was confirmed through

## EVOLUTION OF VIKA RECOMBINASE FOR RECOMBINASE-MEDIATED INTEGRATION IN POTENTIAL SAFE HARBOR LOCUS

junction PCRs for each puromycin-selected clone pool, and the sequences were verified (Figure 29C and D).



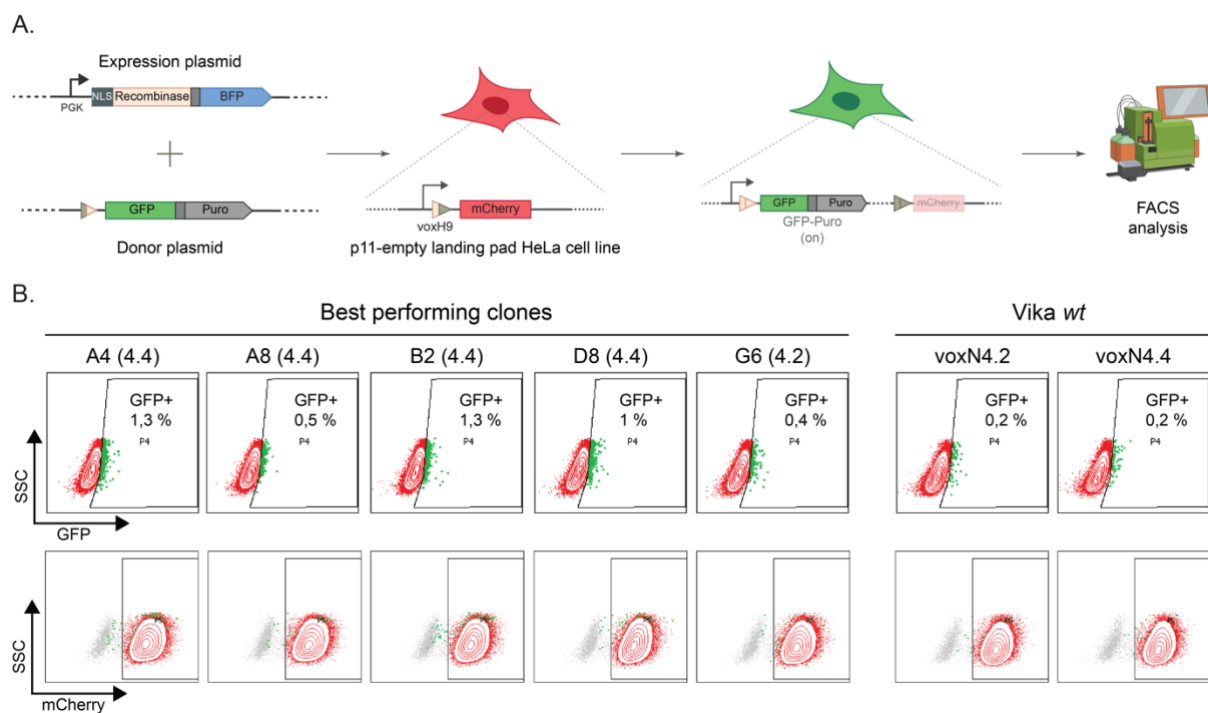
**Figure 29. Screening of the libraries with the landing pad integration assay.**

**A.** FACS plots showing the cell populations of the landing pad cell lines HEK293T<sup>p11-Vika</sup>, HEK293T<sup>p11-4.2library</sup>, HEK293T<sup>p11-4.4library</sup> or HEK293T<sup>p11-empty</sup> when transfected with appropriate donor plasmids. The GFP expression was analyzed 5 days post transfection and after puro selection. Puro selected population display clear shift demonstrating GFP expression in all the samples except for HEK293T<sup>p11-empty</sup>. **B.** Number of clones detected to survive the puromycin selection in 10-cm dishes. Two independent experiments were performed. **C.** Junction PCRs performed to validate the expected integration outcome into the landing pad reporters from the puromycin selected pull. Black arrows represent sense and antisense primers for the 5' junction (junction 1), while red arrows represent primers for the 3' junction (junction 2). Gel pictures show that the specific bands of 890 bp for junction 1 and 505 bp for junction 2 are visible only in the HEK293T<sup>p11-Vika</sup>, HEK293T<sup>p11-4.2library</sup>, HEK293T<sup>p11-4.4library</sup> samples transfected with the corresponding donors. **D.** Screenshot of the Sanger sequencing results of the junction PCRs, aligned to the expected integration product map. Parts of the sequence belonging to the landing pad construct, target site and donor plasmid are depicted and show that the integration happened as expected. For the expected integration product map, the voxN4.2 donor was used, thus, the samples were voxN4.4 donor was used display mismatches in the position where these two target sites differ.

Transitioning from the experiments with cell lines, I sought to retrieve active clones from the puromycin selected libraries, by performing another PCR reaction that amplifies the recombinase cassette. I subsequently cloned these PCR products into the PGK-Recombinase-p2A-BFP expression plasmid to facilitate parallel quantification of the efficiency of these clones (Supplementary Figure S13). To achieve this, I selected 96 clones, verified their sequences, and prepped the DNA for co-transfection with their respective donor plasmids into p11-empty reporter cell lines (Figure 30A).

## EVOLUTION OF VIKA RECOMBINASE FOR RECOMBINASE-MEDIATED INTEGRATION IN POTENTIAL SAFE HARBOR LOCUS

Samples were analyzed by FACS at 3- and 7-days post-transfection to estimate the transfection and integration efficiency for each clone. Regrettably, none of the clones exhibited a significant increase in integration efficiency, as estimated from the percentage of GFP+ cells compared to Vika wt, with the observed maximum remaining around 1% (data not shown). Despite this, I proceeded with the five clones that demonstrated the highest integration efficiency and conducted a follow-up assay on a larger scale. Nine days after co-transfecting recombinase expression and donor plasmids, FACS analysis revealed a modest increase in GFP+ cells for three of the five clones compared to Vika wt samples (Figure 30B). However, the highest integration efficiency did not surpass approximately 1.3% in the case of the A4 and B2 clones (Figure 30B).



**Figure 30. Validation of the best performing clones in the landing pad integration assay.**

**A.** Scheme representing the experiment workflow. Active recombinases were cloned into pPGK-NLS-P2A-BFP expression plasmid and co-transfected together with the donor plasmids into the HeLa<sup>p11-empty</sup> cell line. The integration outcome was quantified by the flow cytometry analysis as the percentage of GFP positive cells. **B.** FACS plots depicting the integration assay results for 5 best performing clones and Vika wt. Depicted GFP percentages were corrected with the background fluorescence detected in the samples where no recombinase was transfected.

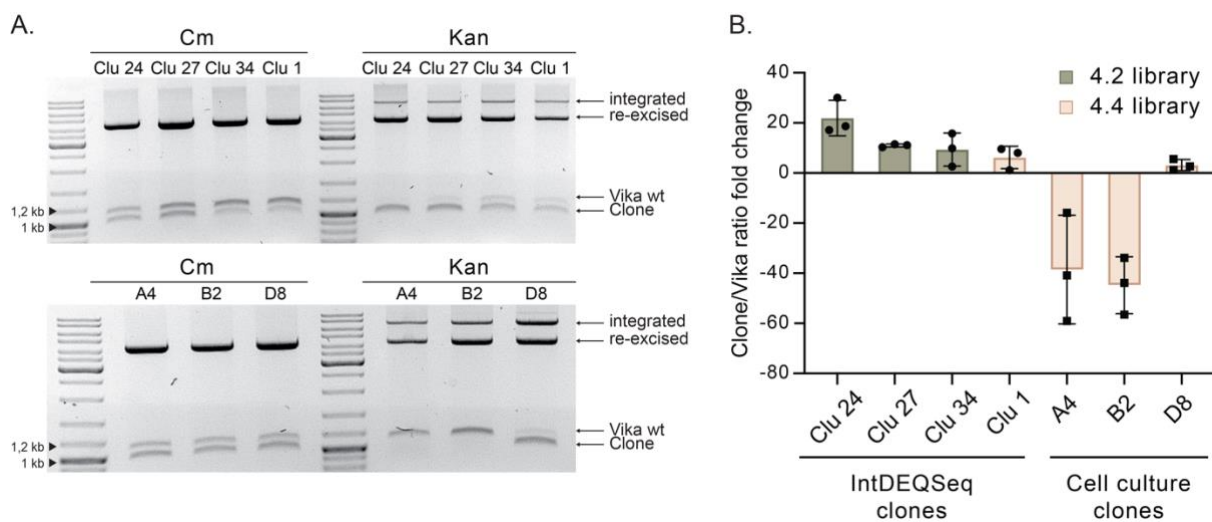
### 7.7 Competition assay

Although a clear advantage of the clones selected in cell culture was not apparent, I sought to determine if all the clones could be quantitatively described as more efficient than wt Vika. Inspired by the IntDEQSeq data where clusters of Vika wt were completely depleted after selection, I designed a competition assay, allowing me to evaluate the advantage of each

## EVOLUTION OF VIKA RECOMBINASE FOR RECOMBINASE-MEDIATED INTEGRATION IN POTENTIAL SAFE HARBOR LOCUS

clone compared to Vika wt individually. I transformed XL-1 Blue *E. coli* cells with wt Vika and seven clones (four from the IntDEQSeq screen and three from cell culture) in parallel, then mixed bacteria with wt Vika and each clone separately in a 1:1 ratio. Following overnight growth, I created electrocompetent cells to transform each Vika/clone mix with pD-voxN4.2 or 4.4 plasmids. The recombinases were induced with 10 µg/ml L-arabinose over 3h during the preparation of electrocompetent cells.

After growing the transformations in Cm and Cm+Kan media, I isolated plasmid DNA and conducted a test digest to compare selected and unselected samples. The test digest, designed to distinguish between Vika-derived bands and clone-derived bands after agarose gel electrophoresis (Figure 31A), revealed an approximately 1:1 ratio of Vika wt and respective clone bands in Cm samples. However, in Cm+Kan samples with selection pressure for integration, a clear enrichment of either clones (in most cases) or wt Vika bands (in A4 and B2 samples) was observed (Figure 31A).



**Figure 31. Competition assay.**

**A.** Agarose gels showing the test digest done to discriminate between Vika wt and clones' fraction of the plasmid mix. On the left side of the gels are the samples grown in chloramphenicol where no selection for the integration outcome was done, whereas to the right are samples grown in kanamycin media in order to select for the integration. Upper panel - the clones coming from IntDEQSeq screen; lower panel - clones selected in the cell culture. **B.** Quantification of the integration. Gels from A. were used to quantify Vika and each clone derived bands and the fold change of the ratio between these two bands in Cm and Kan samples is plotted as the estimation of integration efficiency. Clone coming from 4.2 and 4.4 libraries are depicted with different colors, and clones coming from IntDEQSeq screen and cell culture are labeled.

To quantify the difference, I calculated the ratio of the two bands in both Cm and Cm+Kan samples and determined the fold change between them. This quantification confirmed the validation results for Clusters 24, 27, and 34 derived from the IntDEQSeq screen, with Cluster 24 emerging as the most successful clone, enriched 22-fold compared to Vika wt

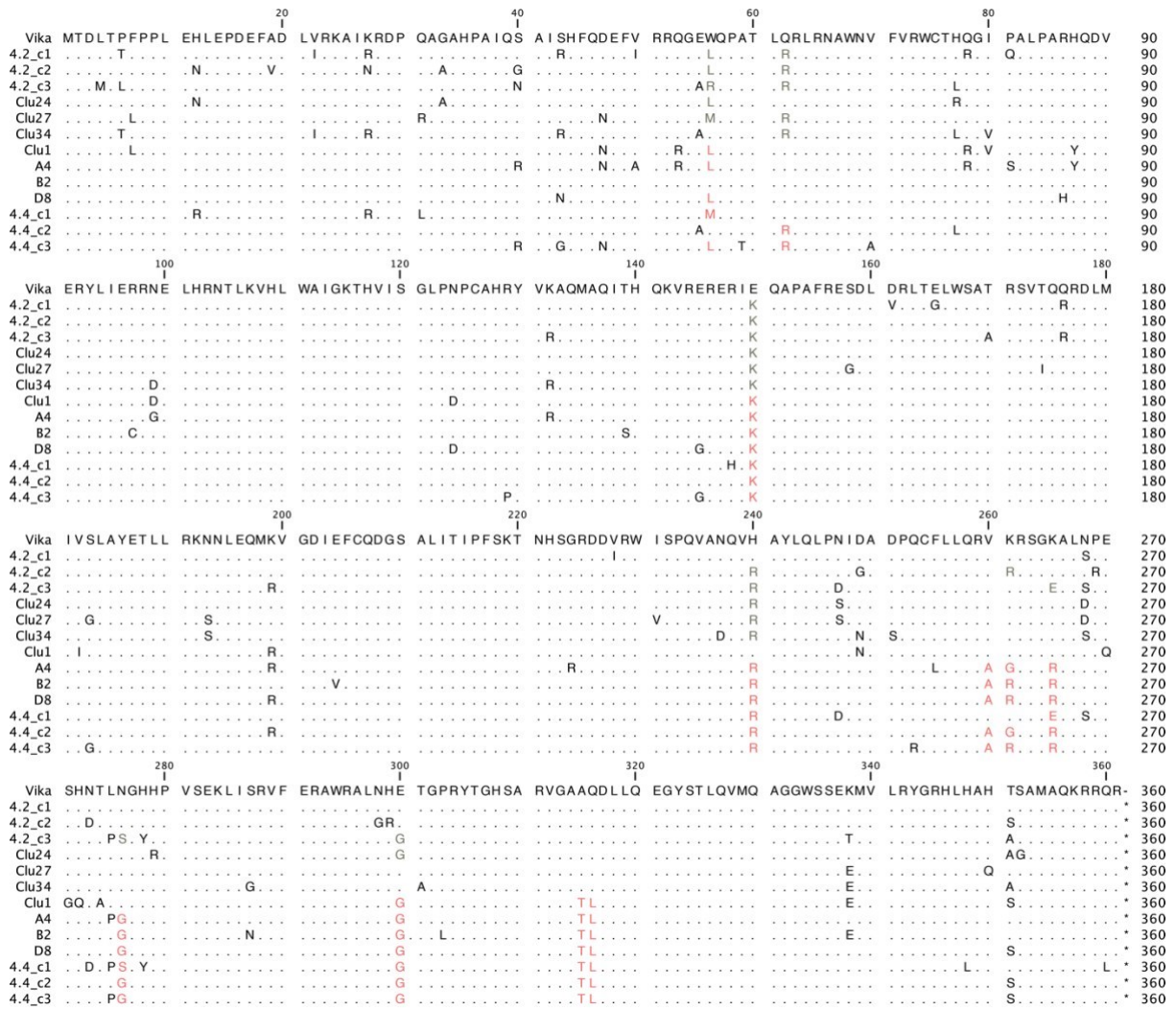
## EVOLUTION OF VIKA RECOMBINASE FOR RECOMBINASE-MEDIATED INTEGRATION IN POTENTIAL SAFE HARBOR LOCUS

(Figure 31B). Conversely, clones selected from human cell assays did not exhibit the same trend as in cell culture. Intriguingly, the seemingly best clones, A4 and B2, were outcompeted by Vika wt (up to 60-fold change in favor of Vika), while the D8 clone maintained a 5-fold change advantage over Vika (Figure 31B). The reasons behind these discrepancies in results could lie in the fact that different clones really do behave differently across heterologous hosts and warrant further investigation. In conclusion, while the competition assay highlighted certain successful clones, inconsistencies across different experimental conditions might indicate the need for additional research to fully understand the underlying factors affecting recombinase efficiency.

### 7.8 Mutational analysis of the clones

In order to narrow down the reason for this discrepancy between the clones' activity in bacteria and cell culture, I performed mutational analysis of all of the tested clones. The positions that emerged in the starting libraries: 56, 62, 150 and 240 still showed high conservation of the mutations that happened in the initial excision-based evolutions (50%, 54%, 100% and 85% respectively) (Figure 32). Surprisingly, the two positions that displayed conserved N276S and E300G changes in the consensus sequence of both libraries after integration-based evolution, were slightly changed in the clones. The 276th position was reverted to the wt residue (asparagine) in most of the clones coming from the 4.2 library, whereas all the clones except Cluster 1 had a glutamine conserved at that position (N276G) (Figure 32). Only randomly picked clone 3 from 4.2 library and random clone 1 from 4.4 library contained the N276S change detected in the consensus sequence. The E300G change was also more conserved in 4.4 library clones compared to 4.2 library, where only two clones contained the mutation (Figure 32). Interestingly, mutational analysis didn't reveal any obvious differences between clones picked from bacteria and the ones from the cell culture. Nevertheless, clear distinction between the clones coming from the different two libraries could be made, thanks to several positions that testify their diverging evolutions. For example, the clearest distinctions could be made when looking at the 315th and 316th position. In all the clones from 4.4 library these two positions were changed from alanine to threonine (A315T) and from glutamine to leucine (Q316L) (Figure 32). Several other positions that are characteristic for 4.4 library with more than 85% conservation among the clones were: V260A, K265R, N276G, and the clones from 4.2 library displayed a higher conservation of the Q62R change, contrary to 4.4 clones. In general, clones from 4.4 library showed more common features, depicted in highly conserved regions in comparison to the clones from 4.2 library.

# EVOLUTION OF VIKA RECOMBINASE FOR RECOMBINASE-MEDIATED INTEGRATION IN POTENTIAL SAFE HARBOR LOCUS



**Figure 32. Mutational analysis of the clones.**

Sequence alignment of all tested clones. Vika wt was used as a reference. Matching residues are depicted as dots. Changes addressed in the text are colored as follows: green – for clones that are derived from 4.2 library; orange – clones derived from 4.4 library.

PART III  
**DISCUSSION**



## Chapter 8 NOVEL CRE-TYPE RECOMBINASES FOR MANIPULATION OF THE GENOMES

### 8.1 Mining for novel Y-SSRs and other genome editing enzymes

**D**iverse microbial organisms and their gene pool represent an almost inexhaustible source for the identification of new DNA editing enzymes (Harrington et al., 2018; Pausch et al., 2020; Karvelis et al., 2021; Durrant et al., 2022). The discovery and characterization of some of these enzymes has enabled the development of molecular biology techniques for advanced genome engineering. Nevertheless, the multitude of publications in recent years shows that there is a growing need for new and better genome engineering tools.

For example, the race for finding the smallest possible CRISPR/Cas systems that still preserves editing efficiency achieved with Cas9 gave rise to countless new Cas proteins (Pausch et al., 2020; Bigelyte et al., 2021; Xu et al., 2021). Furthermore, targeting the genome with Cas proteins is still limited by the requirement for usually small DNA sequence named PAM (NGG in the case of spCas9) right next to the target sequence. Recently, a study interrogated a massively expanded dataset of metagenome and virome assemblies for accurate and comprehensive PAM predictions, in order to identify novel Cas9s selected for their PAM requirements (Ciciani et al., 2022). This PAM prediction pipeline can be instrumental to generate a Cas9 nuclease repertoire responding to any PAM requirement leading towards a natural PAM-free genome editing toolbox, as an alternative to the engineering efforts to generate PAM-less spCas9 variants (Hu et al., 2018). Additionally, novel CRISPR-associated transposases are being described at a high rate, in an attempt to find the system that can successfully integrate extracellular DNA into mammalian genomes (Klompe et al., 2019; Karvelis et al., 2021; Rybarski et al., 2021).

The research efforts mining for SSRs are not falling behind either. It is becoming more and more clear that fully programmable SSRs could pose as the most efficient, versatile and safe genome editing tool. Recent success in large-scale identification of novel large serine recombinases led to defining minimal set of recombinases needed to target over 85% of the human genome (Durrant et al., 2022) which could diminish the need for extensive engineering. Furthermore, by marrying the easy programmability of CRISPR/Cas system, ability to insert small pieces of DNA by coupled reverse transcription and high and specific integration of large cargo via S-SSRs, PASTE system has the potential to expand the scope of genome editing and enable new applications across basic biology and therapeutics

(Yarnall et al., 2022). Novel Cre like recombinases discovered so far, already found a multitude of applications as well (Minorikawa and Nakayama, 2011; Yoshimura et al., 2018; Phillips et al., 2021). Hand-in-hand, engineering and large-scale mining of new editing proteins are pushing the frontiers of genome editing field and opening the doors to new sophisticated solutions for genome manipulation.

Owing to the ability of SSR to carry out three basic catalytic reactions (excision, inversion and integration), complex schemes can be designed by combining several recombinases and virtually no design should be unimaginable. Several elegant techniques were developed based on multiple recombinases (independent rearrangements on two alleles, double gene knock-outs, RMCE). Although the canonical use designates SSR to the conditional knockout strategies where one to two SSR enable the experiment, novel multiplexed engineering technologies now contribute to the need of broader spectrum of recombinases. Moreover, modern genome engineering aims at “clean” rearrangements, which means avoidance of co-introducing prokaryotic vector parts or leaving behind a selection marker. “Clean” DNA modifications mostly are achieved via recombinase-mediated excision and this imposes an additional necessity of multiple recombinases available for more complex step-wise rearrangements within one genome. Thus, increasing sophistication and precision of the DNA modifications *in vivo* dictate the need for novel heterospecific SSRs.

In the first part of my thesis, with help I performed a bioinformatic-guided genome-wide search to identify novel Cre-type recombinases and their associated target sequences. Based on the bioinformatic search, we identified over 500 putative Y-SSRs candidates of which a selection of 17 candidates were experimentally tested and 8 novel Y-SSRs systems were molecularly characterized in depth. The results show that the pipeline is able to predict new Y-SSRs and their target sites at a success rate of about 50%, indicating that many more active Y-SSR systems should be retrievable from the list.

The success of prediction still proves to be largely limited by the nomination of the appropriate target site. My search was focused on scanning regions 1kb upstream and downstream of the putative recombinase gene as this genomic organization was reported for previously described phage-related Y-SSRs (Sternberg and Hamilton, 1981; Casjens, 2003; Karimova et al., 2012; Casjens and Hendrix, 2015). Nevertheless, some of these SSRs come from temperate phages that integrated their genomes into bacterial chromosomes and commonly undergo a complex decay process consisting of inactivating point mutations, genome rearrangements, modular exchanges, invasion by further mobile DNA elements, and DNA deletions. These events could cause the disturbance of the expected organization and lead to misplacement of the native target sites thus hindering the search process.

## 8.2 Molecular characterization of novel recombinases

The determination of the *in vivo* recombinase reactivity and maximum reachable recombination rate may be crucial for planning an experiment. Sophisticated genetic schemes that include simultaneous modifications (e.g. double knock-outs) will require recombinases of comparable activity, whereas for sequential DNA rearrangements (e.g. RMCE) it may be more important to design the experiment in a manner that intermediate recombination events could be selected for. Notably, lower catalysis rate may also confer an advantage for recombinases over the very efficient ones for certain applications. For instance, less active recombinases perform better in integration rather than very active enzymes. Because excision and integration are reversible reactions, the integrated molecules flanked by wild-type lox sites can readily be excised in the presence of Cre-like highly efficient recombinases. Therefore, for the applications where both excision and integration are fulfilled by SSR (like RMCE), the presence of both very efficient and moderate or low active recombinases is needed. By describing and comparing Y-SSRs that have different recombination activities, I offer new tools for both of these cases.

In this study, the eight newly characterized Y-SSRs showed varying activity on their predicted target sites in bacteria, from over 90 % in the case of YR1, YR4, YR6 and YR11 to 82% and 75% for YR2 and YR12, respectively, and going as low as around 15% in the case of YR8 and YR9 (Figure 8). The observed low activity of the latter is possibly due to suboptimal expression of the protein (Figure 10), but the need for additional cofactors or optimal temperature for more efficient recombination of the target sites could also play a role. Furthermore, the annotated sequence may contain errors for the target site or for the recombinase, which can come from prophage that accumulated mutations over time, influencing the activity of the enzyme.

I profiled the specificity of Cre-type recombinases for the first time on a large scale, providing a valuable overview of possible cross recombination events on all target sites. I found that five recombinases showed high specificity on the tested target sites (Vika, YR2, YR8, YR9 and YR11) (Figure 11B). Hence, these recombinases likely represent good candidates for experiments where high specificity of recombination is desired. On the other hand, different cross recombination properties described in this work could have important implications for the design of various complex experiments requiring the use of multiple Y-SSRs.

Interestingly, my results reveal that the correlation between amino acid sequence homology of the Y-SSRs and nucleotide sequence homology of their target sites that they recombine is not straightforward. In addition, cross recombination is not always reciprocal (e.g., YR4 and

YR11 as well as YR11, YR2, YR8 and Dre, Figure 11), suggesting that there are yet unknown determinants that render these recombinases more or less tolerant to the sequence variations of the target sites. Indeed, the quality and the quantity of the nucleotide changes, as well as positional effect all contribute to the different outcomes of the recombination, and in-depth selectivity profiling would be needed to uncover all of these mechanisms for each of these Y-SSRs. Nevertheless, this data should be useful to investigate key specificity determinants of these enzymes. Indeed, in previous work, detailed comparative analyses of Y-SSRs have identified several amino acids as key players in target site distinction. For example, it was shown that K43, R259, and G263 of Cre are critical residues for the discrimination between the *loxP* and *rox* sites (Karimova et al., 2016), and that nearest-neighbor amino acids of these residues influence the activity of Cre-type recombinases (Soni et al., 2020). The data I provide here, could inspire more of such studies, which will ultimately bring us one step closer to elucidate the mysterious mechanisms behind the complex protein-DNA substrate relationships of these SSRs.

For their applied use in higher organisms, I tested the new Y-SSRs for their activity and compatibility in mammalian cells. All the recombinases except for YR9 recombinase showed high activity in a plasmid-based assay in HEK293T cells (Figure 12). Interestingly, YR8 showed high recombination rates in mammalian cells, whereas this recombinase had only weak activity in bacteria, suggesting that activity profiles of recombinases can vary in heterologous hosts.

When applying recombinases in heterologous cells, it is crucial to consider potential adverse effects, not only from off-target recombination on pseudo-sites within the host genome but also from other mechanisms, such as potential impacts on DNA replication or gene expression. These factors could lead to impaired cell growth. Although most recombinases did not exhibit this effect upon overexpression, some illegitimate recombination events might cause 'silent' effects that do not affect cell growth (Figure 13). Consequently, I performed a bioinformatic screening of the human and mouse genome for *lox*-like sites for all known Cre-type recombinases (Figure 16 and S7). This information is valuable in two ways: i) it estimates potential off-target sites that could compromise an experiment, and ii) it identifies potential endogenous target sites that could be used for genome engineering exercises, such as targeted delivery of DNA cargo into a safe harbor locus. Investigating the activity of these recombinases on these sites could potentially serve as starting point for multiple genome manipulation strategies.

### 8.3 Limitations

The success of prediction still proves to be largely limited by the nomination of the appropriate target site. My search was focused on scanning regions 1kb upstream and downstream of the putative recombinase gene as this genomic organization was reported for previously described phage-related Y-SSRs (Sternberg and Hamilton, 1981; Casjens, 2003; Karimova et al., 2012; Casjens and Hendrix, 2015). Nevertheless, a lot of these SSRs come from temperate phages that integrated their genomes into bacterial chromosomes and commonly undergo a complex decay process consisting of inactivating point mutations, genome rearrangements, modular exchanges, invasion by further mobile DNA elements, and DNA deletions. These events could cause the disturbance of the expected organization and lead to misplacement of the native target sites. Furthermore, *Cre/loxP* is naturally not a frequently occurring system because of its biological function. It is a part of the genome stability maintaining system of P1 phage that exists as single copy plasmid. Replication of the P1 circular genome can produce dimers and *Cre/loxP* is dedicated to resolve the genome into monomeric state. The preference of P1 phage not to integrate into host genome during lysogenic stage but to exist as a plasmid is a unique feature, compared to other temperate phages that “sleep” in integrated form. Therefore, search for such a system is restricted to finding the rare class of phages harboring recombinase system that function in a similar manner, but are still different enough so that the target site specificity is distant enough to bring new useful tools to the table.

Given the various mechanisms involved and the diversity within the Y-SSR family, my search criteria were designed to increase the likelihood of successful identification, but also limited the number of putative recombinases that we characterized. Several recently reported methods for identification of prophage elements in bacterial genomes may improve the annotation of new Y-SSRs and their target sites in the future (Houdt et al., 2012; Smyshlyayev et al., 2021; Durrant et al., 2022).

Additionally, identifying large numbers of potential recombinases is one thing, but experimental validation and characterization is very laborious and rate limiting. However, a more high-throughput pipeline for initial experimental validation would greatly improve the efficiency and speed of this process. For example, the pipeline I developed for testing numerous possible cross recombination events, presented in the Figure 11 could potentially serve as a starting point for the development of a high-throughput method for experimental validation. Such a pipeline would allow for the simultaneous testing of multiple recombinases and their activity on various target sites, reducing the time and resources needed for validation. By streamlining the validation process, this high-throughput method could

accelerate the identification and characterization of novel recombinases, ultimately expanding the toolbox of genetic engineering tools further.

Another limitation of this study is that the recombinases were only evaluated for their efficiency in an excision-based assay. For example, it has been previously documented that even bidirectional recombinases may exhibit a preference for DNA inversions between inverted sites over deletions between directly repeated sites (Han et al., 2021). Moreover, it has been suggested that less active recombinases in excision assays could demonstrate higher integration efficiency, since the re-excision step tends to be less pronounced (Karimova et al., 2016). Therefore, it would be advantageous to compare these recombinases in integration- or inversion-based assays to reveal differences in their efficiency when mediating various recombination outcomes. This would further guide the application of these recombinases in complex genome engineering contexts, but also help to further elucidate the mechanism of target site recognition.

Lastly, while the present work aims to provide a more diverse collection of starting points for recombinase optimization, the lack of molecular mechanisms underlying their specificities and target site recognition limits rational design approaches. Even though all tyrosine recombinases have highly conserved 3D structure on a large scale (Meinke et al., 2016), the crystal structures of Cre and its evolved progeny Tre reveal that the 'grip' of a recombinase on DNA can shift depending on the target sequence, meaning that the same residues can interact with different nucleotides (Meinke et al., 2017). It would therefore be helpful to obtain additional structural data for these new Y-SSR-systems to gain insights into key structural features and amino acid residues that contribute to target site recognition and specificity.

### 8.4 What comes next

Although several key milestones were achieved during my thesis work, there is still a lot of ground to be covered. The novel Cre-type Y-SSRs still have to show their impact and potential as highly efficient genome engineering tools. The Y-SSRs described here as independent tools, can enable sophisticated genetic strategies. Apart from the classical conditional gene knockout and selection marker deletions, there are immense perspectives for the multi-allele designs that imply multiple recombinase-mediated sequential rearrangements of the genetic material *in vivo*. Until now the presence of several selectable markers, lack of the enzymes to target and modify several alleles simultaneously was a limitation in the genome engineering (Glaser et al., 2005; Anastassiadis et al., 2010). This can be overcome by adding one or more recombinases to the geneticist's toolbox. Hence, novel Y-SSRs and extensive data of their specificity in the sequence space of so far described target sites increase the repertoire of the experimental genetic designs.

Moreover, these recombinases might become an additional tool or even a valuable substitute for experimental settings where Cre is not applicable. For example, in the cell lines or organisms where Cre showed high cytotoxicity, recombinases that I discovered could be used instead, considering their better toxicity profile and relatively similar efficiency. Furthermore, I have identified a lot of potential pseudo-*lox* sites in the human and mouse genomes, and activity of novel recombinases on these sites should be addressed in order to estimate the real risk of unwanted genome rearrangements when using any of these recombinases in human or mouse cells.

Following steps to integrate novel recombinases into the genome engineering of higher organisms can be creation of the inducible expression systems established to work very well with Cre so far, controlled by the known ligand activated protein (e.g., estrogen- or progesterone receptor). Generation of mouse reporters or mice expressing these Y-SSRs could be a helpful step further, as was shown with single reporter mouse line that has been developed for Vika, Flp, Dre, and Cre-recombination, providing a versatile tool for a wide range of applications (Karimova et al., 2018). More complex models could now be integrated with the novel Y-SSRs described here.

Finally, the discovery and characterization of different naturally occurring recombinases should be useful to accelerate the development of novel enzymes via directed evolution in two ways: i) the detailed investigation of DNA binding should lead to a better understanding how these enzymes specifically recognize their targets. With this knowledge, rational design of enzymes with novel specificities might become possible. By resolving more crystal structures of Cre-like Y-SSRs or by performing mutational analysis of the new targets sites, or Y-SSR proteins could give a novel insight into the specificity paradigm of the tyrosine recombinases; ii) It has been demonstrated that shuffling of related genes can speed up the evolution process (Cramer et al., 1998). Hence, new recombinases with desired properties might become available in shorter time through family shuffling. In addition, by using the wide variety of new naturally occurring recombinases with activity on various target sites, one could start different evolution campaigns depending on similarity to the desired target site.

This knowledge could, in turn, accelerate the engineering of new designer recombinases with desired specificities and functionalities. For example, by identifying the critical amino acid residues or structural motifs that contribute to target site recognition in Cre-type Y-SSRs, we can potentially engineer new recombinases with altered specificities or improved performance. Furthermore, this data could also improve machine learning algorithms developed to predict recombinase activity on desired target sites (Schmitt et al., 2022). By

## NOVEL CRE-TYPE RECOMBINASES FOR MANIPULATION OF THE GENOMES

using data from evolved Y-SSRs, in combination with data from naturally occurring SSR systems, I expect that a smarter design of synthetic SSRs can be generated to rapidly target novel loci in desired genomes. In addition, by understanding the underlying principles of target site recognition, we can also design more efficient and specific assays for characterizing recombinase activity and specificity, which can aid in the discovery and validation of new recombinases.

Overall, the collection of Cre-type Y-SSRs that I characterized in this work, coupled with additional structural data, has the potential to advance our understanding of recombinase target site recognition and accelerate the development of new and improved recombinases for various applications in biotechnology, medicine, and beyond.



## Chapter 9 NOVEL Y-SSRS AS PLATFORMS FOR EXTENSIVE DEVELOPMENT OF DESIGNER RECOMBINASES

With just few exceptions, all molecular evolution campaigns to develop designer recombinases were Cre-based. Although in theory, one can evolve Cre to any target site, this can prove to be very tedious and time consuming. For the generation of Brec1, for example, 145 evolution cycles and 12 subsites were required starting from Cre/*loxP* (Karpinski et al., 2016). Through these experiences it was shown that differences in nucleotide positions in the desired target sites and starting sites have different effect, with some requiring much more effort to adapt Cre for. For example, changing the 12<sup>th</sup> or 13<sup>th</sup> position in the half-sites was generally shown to be very challenging in several evolution campaigns. By introducing different recombinases with diverse target site specificity one can cover much more sequence space, ultimately accelerating the evolution process. Furthermore, with the work of Lansing et al. we showed that evolving a functional heterodimer, we can now address the asymmetrical target sites which was a huge limitation for designer recombinases applications (Lansing et al., 2019, 2022). Now, with libraries based on novel Y-SSRs one can combine their different specificity in order to ease the targeting of asymmetrical target sites. Moreover, one can explore even the development of functional hetero-tetrameric recombinases for targeting of two different asymmetrical target sites which will almost completely remove restrictions for the choice of target site identification in the human genomes.

In this work I applied directed evolution methods to already two novel Y-SSRs systems, thus establishing novel recombinase libraries that are not Cre-based. As proof-of-concept I evolved YR9 recombinase, as to increase its activity on the predicted *lox9* target site. Furthermore, I established novel Vika libraries that served as starting point for development of designer recombinases for DNA integration.

One of the eight newly identified recombinases, the YR9 recombinase, recombined its putative *lox9* target site with relatively low efficiency (17% at 100 µg/ml of L-arabinose in *E. coli*, and close to 0 in HEK293T, see Figure 8 and Figure 12). Nevertheless, several aspects make the YR9 recombinase an interesting Y-SSR system worth investigating further: First, the YR9 recombinase was shown to be a highly specific recombinase, recombining only the *lox9* target site (Figure 11B). Secondly, the YR9 recombinase is slightly smaller than most recombinases, and the *lox9* target site is the most unique target site among the other

identified ones. This is particularly interesting when seeking to establish a diverse library of recombinases. Thus, I decided to increase its activity through directed evolution.

Overall, 13 rounds of substrate-linked directed evolution (SLiDE) were enough in order to achieve 4.5-fold increase in the library activity (Figure 14). The results demonstrated a qualitative difference between the recombination activity of the selected YR9 variants and the wild type YR9, showing that the activity of weaker enzymes can be significantly improved with directed molecular evolution. Remarkably, sequence analysis revealed that only 5-6 mutations per variant led to this enhancement in activity (Figure 15B). However, only one variant, YR9.10, demonstrated increased activity also in mammalian cells, showing that the activity of recombinases can vary in heterologous hosts (Figure 15C). Since the overall recombination efficiency of YR9.10 was still low, the use of this improved YR9 variant is also very limited in mammalian cells and further optimization of expression could increase the potential of this recombinase system. Nevertheless, this enhanced version can now be used in mammalian system since lower-performing recombinases could be useful for sequential DNA rearrangements such as RMCE.

## Chapter 10 Y-SSRS AS TOOLS FOR EFFICIENT TARGETED DNA INTEGRATION

The postgenomic era marked the rapid development of genome editing techniques. In particular, targeted insertion of exogenous DNA sequences has opened up new venues for fundamental biological studies as well as industrial and therapeutic applications (Zhang et al., 2021). However, efficient and precise integration of foreign DNA into the mammalian genome remains challenging, particularly for large DNA fragments. Two mechanistically distinct systems, CRISPR/Cas9 and SSRs, have greatly expanded our capabilities to perform targeted chromosomal insertion, but they are also facing various limitations. Despite the great programmability offered by CRISPR/Cas9, its applications have been haunted by the off-target activity and undesired editing outcomes associated with the DSB repair. Although SSR-mediated integration is capable of large DNA integration, highly specific, and does not invoke cellular DSB repair, its applications are often limited by the rigid requirement of substrate DNA sequences. A win-win solution to overcome the bottlenecks in both of these tools is to develop programmable SSRs that can operate on the endogenous mammalian genome. Encouraging progress has been made over the past decade in this direction, such as the discovery of naturally existing programmable integrases, or the directed evolution of recombinases in order to change their substrate specificity.

In this work I wanted to explore deeper the potential of using the Y-SSRs for precise and efficient genomic integration. To that intent, I described pseudo-*lox* target sites with potential LE/RE trapping architecture in the human and mouse genomes as described previously, as entry points for genomic integration (Figure 16C). Out of that data set, I selected the voxH9 target site located on chromosome 9 as a target for DNA integration (Figure 17A). This site, already observed in previous work in the Buchholz lab, is particularly interesting as it fulfills most of the safe harbor criteria (Papapetrou et al., 2011). For instance, it doesn't overlap with any ORFs, enhancers or promoters. It overlaps only with a non-coding transcript that seems to be lncRNAs (long non-coding RNAs) so far reported to be expressed only in spermatogonia. The nearest neighboring genes on the 5' or 3' side are 270 and 700 kb away, respectively. Thus, I hypothesized that a locus like this should pose minimal risk for perturbation of genome regulation and gene expression while still enabling open chromatin structure necessary for successful and prolonged expression of the desired transgene.

Furthermore, previous work demonstrated Vika wt activity on voxH9 in an integration-based assay in bacteria, which I also confirmed by the excision-based assay (Figure 17B). Most importantly, I detected successful Vika-mediated integration of a 4 kb DNA cargo into the

endogenous voxH9 in HEK293T human cell line. However, the observed efficiency was very low, detectable only after selecting for the integration outcome (Figure 18). While these results marked an exciting success in the effort to use Y-SSRs for genomic integration, I had to increase the efficiency of this process to achieve the goal of developing Y-SSR-based tools for therapeutic transgene delivery where selection is usually not an option.

In order to do so, I established a novel evolution scheme based on the integration assay. This enabled me to produce highly active variants after only 23 cycles of evolution, demonstrated by the different screening methods in bacteria. Random picking and the IntDEQSeq screen resulted in variants that appeared to be 5-15 times more active than wild-type Vika (Figure 23 and Figure 25). Concurrently, I screened the libraries of active clones in cell culture to compare the two approaches and ensure that any discrepancies between bacterial and human cell results would not hinder the identification of successful clones. Although the overall efficiency of DNA integration into the human genome remained relatively low, some clones exhibited higher efficiency than *wild-type* Vika (up to a 5-fold increase) (Figure 30). Interestingly, all successful clones in cell culture originated from the 4.4 library, which also consistently displayed more clones surviving after puromycin selection (Figure 29B), contrary to the results of the clones from the IntDEQSeq screen, where the 4.2 library had a larger proportion of clones in the selection sample (Cm+Kan). These clones were also more enriched compared to clones from 4.4 library (Figure 25).

Additionally, I observed intriguing dynamics among the variants in the IntDEQSeq screen. In fact, all variants seemingly outperformed Vika in the selective culture (Cm+Kan), even though Vika clusters were numerous in the pre-selection sample (Cm) (Figure 25A). This observation led me to develop a strategy to characterize each clone by its ability to outcompete Vika when simultaneously challenged in an integration assay. By measuring the potency of each clone to outcompete Vika in establishing integration plasmids providing kanamycin resistance and growing in selective media, I could quantify the integration efficiency. The competition assay confirmed the differing performance of clones selected from cell culture and those selected from bacterial screens, as illustrated by the A4 and B2 clones which were significantly outperformed by Vika in the competition assay in bacteria, despite being the best-performing candidates in cell culture (Figure 31). Given the low integration efficiency of these clones in cell culture, it is important to verify these results and investigate the source of this difference. A side-by-side comparison of all clones in cell culture should be conducted again to clarify the discrepancy between the clones. This may reveal that clones selected in bacteria are not optimal for successful integration in human cells due to subtle differences in the reaction mechanism, varying expression levels, or

stability of the variants. If this is the case, a better screening or even selection method in cell culture should be established to enrich variants that perform well in human cells.

The clones from bacteria and cell culture did not exhibit distinct mutations that could explain the observed discrepancies between the results in cell culture and bacteria. However, the clones displayed a clear distinction in mutational patterns depending on the library they were derived from (Figure 32). This observation is particularly interesting, as the two libraries were evolved on very similar target sites, which could help elucidate the mechanism of how recombinases mediate integration, especially in the context of the RE/LE strategy. The hypothesis suggests that the wild-type half-sites are where monomers first achieve an active conformation due to better binding and the initial strand exchange occurs on the 5' side of the spacer. The second pair of monomers then acquire an active conformation through isomerization, tolerating slightly suboptimal binding on the mutated half-sites, leading to the second strand exchange and resolution. Once integrated the full mutant target site consist of both half-sites where suboptimal binding of the monomers is happening which disables them to achieve active conformation, preventing the re-excision reaction. In theory, these two libraries were evolved to mediate integration on target sites that differ only in one of the mutant half-sites, voxN4.2 and voxN4.4, which differ by five positions. It would be worthwhile to further explore these clones, for instance, comparing their performance when provided with a donor target site they were not evolved on, in order to better understand the selection pressures that shaped the distinct outcomes of these two evolutions.

### 10.1 Caveats of using Y-SSRs for genomic integration

In this research, my aim was to investigate the potential of Y-SSRs for genome integration, as no designer recombinase has been developed for this purpose thus far. However, the use of tyrosine site-specific recombinases for genomic integration presents several challenges, particularly in the directed evolution process and the inherent nature of these enzymes. One concern with any directed evolution experiment is that "you get what you screen for." This concept highlights the fact that a system adapts to the constraints it encounters, and sometimes the outcome deviates from the experimenter's intentions. In the classical SLIDE, for example, the predominant (positive) selection pressure is on activity, and not specificity. Indeed, it has been frequently observed that evolved recombinases exhibit a relaxed specificity: While gaining activity on a new target, they retain activity on *loxP* (Buchholz and Stewart, 2001; Santoro and Schultz, 2002; Sarkar et al., 2007; Karpinski et al., 2016).

In the context of the directed evolution strategy employed in this study, the lack of concrete mechanistic data supporting the hypothesized mechanism of integration via RE/LE target

sites is an issue. As the integration process is performed in a plasmid-based assay, various complex mechanisms for regulating plasmid persistence in cells could be implicated in ways that are difficult to predict. These evolutionary mechanisms are deeply embedded, as plasmids and their persistence in cells are critical for the high adaptability of prokaryotic populations to survive under different selection pressures. Moreover, the co-transformation step of the host and donor plasmids in my evolution procedure represent a major bottleneck, while it has been reported that only up to 3% of electrocompetent bacteria are able to receive more than 1 copy of the plasmids per cell (Goldsmith et al., 2007; Velappan et al., 2007; Tomoiaga et al., 2022). This can be illustrated by the fact that measured ratios between the number of colonies between Cm and Cm+Kan plates were never recorded to be higher than 3,5-4% (See Figure 22 and Figure 26).

Furthermore, the evolution process in bacteria involves selection pressures that can ensure enrichment of desired properties in bacteria; however, these properties might not function similarly in mammalian cells and could even be counterproductive. This issue is evident from the observed discrepancy between clones screened in bacteria and those in cell culture. To overcome these constraints, it would be beneficial to develop an evolution pipeline specifically designed for enriching integration in mammalian cells. Such a pipeline would consider the unique characteristics and requirements of mammalian cells, ensuring that the selected recombinases function optimally in the intended context. By tailoring the directed evolution process to mammalian systems, researchers can more effectively bridge the gap between bacterial and mammalian cell experiments, leading to more reliable and reproducible results for genomic integration using recombinases.

Another challenge arises from the nature of tyrosine site-specific recombinases themselves. Their natural role is usually not integration of DNA sequences, but rather maintaining the phage extrachromosomal genome integrity, primarily through excision reactions. In contrast, the unidirectional nature of serine recombinases contributes to their higher integration rate, making them more suitable for DNA integration. Tyrosine recombinases, on the other hand, are generally preferred for genomic deletions.

By leveraging our extensive expertise and deeper understanding of Y-SSR specificity mechanisms gained from prior directed molecular evolution endeavors, I demonstrated that it is possible to enhance the efficiency of Y-SSR-mediated integration. However, to further advance DNA integration efforts in the human genome, it would be advantageous to include serine recombinases in the development of programmable tools for DNA integration. By doing so, it is possible to harness the inherent advantages of serine recombinases in

integration reactions and overcome some of the limitations presented by tyrosine recombinases. In summary, addressing the challenges of using tyrosine site-specific recombinases for genomic integration requires a deeper understanding of directed evolution strategies and exploring alternative enzymes like serine recombinases to achieve more efficient and targeted genome engineering.

### 10.2 Future prospects of DNA integration via voxH9 target sites

The successful integration of DNA into the human genome using Vika-based recombinases, as well as the ability to increase the efficiency of this process through directed evolution, represents a significant milestone. However, further studies are required to ensure the translation of these successful data in human cells. A deeper understanding of the differences between clones selected in bacteria and those selected in cell culture is essential for optimizing the performance of recombinases in human cells. Investigating the distinct mutational patterns of the clones derived from different libraries could offer valuable insights into the mechanisms of recombinase-mediated integration and guide the development of more efficient and targeted genome engineering tools.

Furthermore, the Vika recombinase tool can be developed further by incorporating additional DNA binding domains. This may enhance target site specificity and increase the overall efficiency of the recombinase. Alternatively, recombinase-mediated cassette exchange (RMCE) with two recombinases could be employed, as it has demonstrated high efficiency in mammalian cells (Osterwalder et al., 2010). This could involve using a Vika-based designer recombinase for the integration step via voxH9 sites, followed by a Cre-based designer recombinase for the subsequent excision of *loxP*-like sites identified downstream, ultimately completing the cassette exchange.

Secondly, the voxH9 locus must be characterized and proven to function as a safe harbor locus. This would entail demonstrating that the locus enables constitutive, uninterrupted expression of the transgene without disrupting other expression patterns in the cells. Transcriptome analysis of various cell types could be performed to confirm this, ensuring that the voxH9 locus can be employed safely and effectively in genome engineering applications (Aznauryan et al., 2022).

Lastly, integrating larger, therapeutically relevant DNA cargo into the voxH9 locus would serve as a proof of concept for gene therapy applications. Successful integration and expression of such cargo would pave the way for employing the evolved Vika recombinase and the voxH9 locus in the development of novel therapeutic strategies. This could ultimately

## Y-SSRS AS TOOLS FOR EFFICIENT TARGETED DNA INTEGRATION

lead to more effective treatments for various genetic diseases, highlighting the significant potential of this research in the field of gene therapy.



## CONCLUSION

### Chapter 11 CONCLUSION

In the recent years, it is becoming more and more clear, that fully programmable SSRs could be the most powerful editing tools, as they can modify the DNA with the nucleotide precision independent of the DNA repair machinery, all the while enabling variety of different rearrangements such as excision, integration, inversion or cassette exchange of the DNA fragments flanked by the two recognition sites. This was exemplified through recent developments of therapeutically relevant designer recombinases such as: Brec1, for excision of HIV provirus, a heterodimer recombinase RecF8 for correction of an inversion of the first exon in F8 gene of numerous Hemophilia A patients, and most recently RecHTLV for excision of HTLV virus causing an aggressive form of adult T cell leukaemia/lymphoma. Brec1 is even set out to go into clinical trials by the end of the year, presenting a revolutionary milestone.

In conclusion, this work has made significant contributions to the field of genome engineering by establishing a high-throughput method for the identification of novel Y-SSR systems. This led to the discovery of eight new Y-SSRs, which I further characterized in depth. These novel Y-SSRs have the potential to advance genome engineering practices, as demonstrated by the expansion of targetability of native human sequences and the acceleration of the evolution process.

Moreover, the successful application of previously established evolution strategies from Cre to these novel recombinases supports the hypothesis that the integration of libraries based on new Y-SSRs can indeed accelerate the development of innovative designer recombinases. This notion was further exemplified by the efforts to develop Vika-based recombinases capable of integrating DNA into the genome. For the first time, I demonstrated that wild-type Vika can integrate DNA cargo into native human pseudo-vox sequence, and I increased the efficiency of this process through directed evolution by specifically selecting for the integration outcome.

While the evolved variants exhibited a significant increase in integration efficiency in bacteria, it is essential to conduct further studies to ensure the successful translation of these findings to human cells. Ultimately, the findings presented in this work provide a solid foundation for further development and application of Y-SSR systems in genome engineering, paving the way for innovative and potentially life-changing gene therapies.

PART IV

# MATERIAL AND METHODS

## MATERIALS

### Chapter 12 MATERIALS

#### 12.1 Organisms

**Table 3. Bacterial strains and cell lines used in this work.**

<b>Organism</b>	<b>Genotype/characteristics</b>	<b>Reference</b>
<b><i>E. coli</i> strain</b>		
<i>E. coli</i> XL1-blue	<i>recA1 endA1 gyrA96 thi-1 hsdR17 supE44 relA1 lac [F proABlacIqZΔM15 Tn10(Tetr)]</i>	Laboratory stock
<i>E. coli</i> PIR1	<i>F- Δlac169 rpoS(Am) robA1 creC510 hsdR514 endA recA1 uidA(ΔMluI)::pir-116</i>	Invitrogen
<b>Mammalian cell lines</b>		
HEK293T	Wild type human embryonic kidney cells 293, contain SV40 Large T-Antigen	Laboratory stock
HEK293T <sup>p11-empty</sup>	Reporter cell line harboring EF1alpha promoter-voxH9-mCherry	This work
HEK293T <sup>p11-Vika</sup>	Reporter cell line harboring EF1alpha promoter-voxH9-mCherry-P2A-Vika	This work
HEK293T <sup>p11-4.2 library</sup>	Reporter cell line harboring EF1alpha promoter-voxH9-mCherry-P2A-4.2 library of Vika variants	This work
HEK293T <sup>p11-Vika</sup>	Reporter cell line harboring EF1alpha promoter-voxH9-mCherry-P2A-4.4 library of Vika variants	This work
HeLa Kyoto	Human papillomavirus-related endocervical adenocarcinoma cells	Laboratory stock
HeLa <sup>p11-empty</sup>	Reporter cell line harboring EF1alpha promoter-voxH9-mCherry	This work

#### 12.2 Synthetic oligonucleotides

Oligonucleotides were synthesised by Sigma-Aldrich (standard cloning), IDT (UMI oligos).

List of all oligonucleotides used in this work can be found in Appendix (Supplementary Table S2).

## MATERIALS

### 12.3 Molecular biological products

**Table 4. Molecular biological products and materials used in this work.**

<b>Name</b>	<b>Manufacturer</b>
<b>Enzymes</b>	
Herculase II Fusion DNA Polymerase	Agilent
BsrGI-HF	New England Biolabs
SbfI-HF	New England Biolabs
XbaI	New England Biolabs
AgeI-HF	New England Biolabs
EcoRI-HF	New England Biolabs
BglII	New England Biolabs
NdeI	New England Biolabs
AvrII	New England Biolabs
ScaI	New England Biolabs
HindIII-HF	New England Biolabs
XhoI	New England Biolabs
NheI-HF	New England Biolabs
SacI-HF	New England Biolabs
T4 Ligase	New England Biolabs
BIOTAQ™ DNA Polymerase	Meridian Bioscience
MyTaq™ Polymerase	Meridian Bioscience
Q5 high-fidelity DNA polymerase	New England Biolabs
<b>Molecular biological reagents and chemicals</b>	
5x Herculase II Reaction Buffer	Agilent
5x Q5 reaction buffer	New England Biolabs
100 nM dNTP Mix (25 nM each)	Agilent
6X Orange DNA Loading Dye	Thermo Fisher Scientific
Gel Loading Dye Purple (6X)	New England Biolabs
Tris Pufferan > 99.9%, p.A.	Roth
Boric acid	VWR
EDTA	VWR
RedSafe (20000x)	Intron Biotechnology
GeneRuler DNA Ladder Mix	Thermo Fisher Scientific
GeneRuler 1kb Plus DNA Ladder	Thermo Fisher Scientific
UltraPure TRIS-Borat-EDTA (TBE)	Invitrogen
L-(+)-Arabinose	Sigma-Aldrich
Chloramphenicol	Sigma-Aldrich
10X CutSmart®-Buffer	New England Biolabs
10X NEBuffer r3.1	New England Biolabs
10X NH4 Reaction Buffer	Meridian Bioscience
MgCl <sub>2</sub>	BIOTEC media kitchen
MnCl <sub>2</sub>	BIOTEC media kitchen
5x MyTaq™ Reaction Buffer Red	Meridian Bioscience
T4 DNA Ligase Reaction Buffer	New England Biolabs
Glucose 20%	BIOTEC media kitchen
Distilled water, autoclaved	BIOTEC media kitchen
1% penicillin-streptomycin	Thermo Fisher Scientific
GlutaMAX™	Gibco
10% fetal bovine serum	Gibco

## MATERIALS

Trypsin	Gibco
D-PBS	BIOTEC media kitchen
Chloroquine	Sigma-Aldrich
Polyethylenimine (PEI)	Sigma-Aldrich
Lipofectamine® 2000 Transfection Reagent	Invitrogen
Glycine	AppliChem Panreac
Sodium dodecyl sulfate (SDS)	Merck Millipore
Glycerol	VWR
Dithiothreitol (DTT)	Thermo Fisher Scientific
Bromphenol Blue	Sigma-Aldrich
Amido black	Merck Millipore
Methanol	VWR
Acetic acid	Merck Millipore
Nonfat dried milk powder	AppliChem Panreac
Tween® 20	SERVA
Trichloroacetic acid	Sigma-Aldrich
<b>Antibodies</b>	
Monoclonal ANTI-FLAG® M2 mouse antibody	Sigma-Aldrich
IRDye® 800CW Donkey anti-Mouse IgG Secondary Antibody	LI-COR
<b>Molecular biological kits</b>	
GeneJET Plasmid Miniprep Kit	Thermo Fisher Scientific
NucleoBond Xtra Plasmid Maxi Kit	Macherey-Nagel
ISOLATE II PCR and Gel Kit	Bioline
Cold Fusion Cloning Kit	System Biosciences
QIAamp DNA Blood Mini Kit	Qiagen
<b>Media</b>	
Lysogeny broth (LB) medium	BIOTEC media kitchen
LB agar plates containing chloramphenicol (15 µg/ml)	BIOTEC media kitchen
LB agar plates containing chloramphenicol (15 µg/ml) and kanamycin (15 µg/ml)	BIOTEC media kitchen
LB agar plates containing kanamycin (30 µg/ml)	BIOTEC media kitchen
LB agar plates containing ampicillin (100 µg/ml)	BIOTEC media kitchen
SOC medium	BIOTEC media kitchen
Dulbecco's modified Eagle's medium	Gibco
Opti-MEM medium	Thermo Fisher Scientific

## METHODS

### Chapter 13 METHODS

#### 13.1 Bacterial growth conditions

Bacterial cultures were grown at 37°C in L-broth or SOC medium, cells were grown on L-broth agar plates as described. Ampicillin was added to L-broth and L-Broth agar plates at concentration of 100µg/ml. Kanamycin was added to L-broth and L-Broth agar plates at concentration of 15 µg/ml. Chloramphenicol was added to L-broth at concentration of 25 µg/ml and to L-Broth agar plates at concentration of 15 µg/ml.

#### 13.2 Recombinant DNA techniques

##### 13.2.1 DNA purification

Plasmid DNA was purified in large scale using the “NucleoBond Xtra Plasmid Maxi Kit by MACHEREY-NAGEL” (referred to as “maxiprep”) or in a small scale using the “GeneJET Plasmid Miniprep Kit” (referred to as “miniprep”) according to manufacturer's instructions. DNA concentration and purity were assessed using the NanoDrop 8000 Spectrophotometer.

Genomic DNA was isolated with QIAamp DNA Blood Mini Kit (Qiagen) according to the manufacturer's manual.

##### 13.2.2 High-fidelity polymerase chain reaction (PCR)

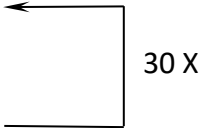
To amplify DNA with high accuracy, PCR was performed in a final volume of 50 µl and carried out in a PeqSTAR 2X thermocycler (Peqlab, USA). The following two reaction mixtures and cycling programs were used:

**Table 5. Reaction mixture for Herculase II Fusion high-fidelity DNA polymerase chain reaction.**

Component	Volume (50 µl in total)	Final concentration
5X Herculase II reaction buffer	10 µl	1X
100 mM dNTPs	0.5 µl	1 mM
12.5 µM primers	1 µl each	0.25 µM each
Template DNA	1 µl	~ 30 ng
5 U/µl Herculase II Fusion DNA polymerase	0.5 µl	0.05 U/µl
Ultrapure water	36 µl	-

## METHODS

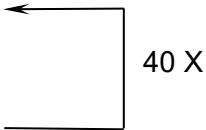
### Thermocycler program high-fidelity PCR

1. Initial denaturation	95°C	3 minutes	
2.	95°C	20 seconds	
3.	54°C	20 seconds	
4.	72°C	1 minute	
5. Final extension	72°C	5 minutes	
6. Hold	8°C	∞	

**Table 6. Reaction mixture for Q5 high-fidelity DNA polymerase chain reaction.**

Component	Volume (50 µl in total)	Final concentration
5X Q5 reaction buffer	10 µl	1X
10 mM dNTPs	1 µl	200 µM
12.5 µM primers	2 µl each	0.5 µM each
Template DNA	1 µl	~ 30 ng of plasmid 100- 200 ng of genomic
2000 u/ml Q5 High-Fidelity DNA polymerase	0.5 µl	0.02 U/µl
Ultrapure water	36 µl	-

### Thermocycler program high-fidelity PCR

1. Initial denaturation	98°C	30 seconds	
2.	98°C	20 seconds	
3.	65°C	20 seconds	
4.	72°C	1 minute	
5. Final extension	72°C	5 minutes	
6. Hold	8°C	∞	

PCR products were purified using the “ISOLATE II PCR and Gel Kit” according to the manufacturer's instructions and eluted in water. DNA concentration and purity were assessed using the NanoDrop 8000 Spectrophotometer.

## METHODS

### 13.2.3 Restriction enzyme digestion

Purified PCR products or plasmid DNA were cut via restriction enzyme digestion according to Table 7. To digest purified PCR products, the total elution volume was used. Standard digestions and PCR product digestions were typically incubated at 37 °C for 0.5 - 1 hour while backbone digestions were incubated at 37 °C overnight.

**Table 7. Different reaction mixtures used for restriction enzyme digestion.**

<b>Component</b>	<b>Test digests</b>	<b>PCR product digest</b>	<b>Backbone digest</b>
10X CutSmart buffer	2 µl	4 µl	5 µl
DNA	500 ng	33 µl elution	5 µg
Restriction enzyme 1	1 µl	1.5 µl	2.5 µl
Restriction enzyme 2	1 µl	1.5 µl	2.5 µl
Ultrapure water	Up to 20 µl	-	Up to 50 µl

### 13.2.4 Gel electrophoresis

PCR products were analyzed via analytical gel electrophoresis on 0.8% - 1% UltraPure agarose gels with 1X TBE buffer (see Buffer and staining solutions) and 1X Red Safe. Narrow 12-lane combs were used. 5 µl sample were mixed with 1 µl 6X loading dye and loaded into the wells. For high-fidelity PCR products, Orange DNA Loading Dye was used and Gel Loading Dye Purple otherwise. Fragment sizes were determined using GeneRuler DNA Ladder Mix or GeneRuler 1kb Plus DNA Ladder as a size standard. The agarose gel electrophoresis ran at around 80 V in TBE buffer for 40 - 90 minutes. Afterwards, the gel was viewed with the Infinity VX2-3026 (Vilber) using the Infinity Capt software. If the separated DNA was needed further, preparative gel electrophoresis was performed. Therefore, 4-lane combs were used and the whole sample was loaded with the respective amount of 6X loading dye into the wells. After separation, the desired bands were excised from the gel using the Safelmager™ 2.0 (Invitrogen) and extracted using the "ISOLATE II PCR and Gel Kit" according to the manufacturer's instructions and eluted in water. DNA concentration and purity were assessed using the NanoDrop 8000 Spectrophotometer.

### 13.2.5 Ligation

Ligations were carried out by mixing vector and insert with a molar insert-to-vector ratio of 3:1 according to Table 8. The mixture was incubated at room temperature for 0.5 – 1 hour and then heat-inactivated at 65 °C for 10 minutes. For library cloning, the sample was



## METHODS

additionally purified on a membrane (MF-Millipore™ Membrane Filter, 0.025 µm pore size, Merck Millipore) by diffusion. Therefore, a 6-well plate was filled with ultrapure water and the membrane was placed on the water surface. The ligation sample was transferred onto the membrane for 20 - 30 minutes.

**Table 8. Reaction mixture of a typical ligation reaction.**

Component	Volume (20 µl in total)	Final concentration
10X T4 DNA Ligase buffer	2 µl	1X
Insert	variable	30 ng or 60 ng
Vector	variable	60 ng or 120 ng
400,000 U/ml T4 DNA Ligase	1 µl	20,000 U/ml
Ultrapure water	Up to 20 µl	-

### 13.2.6 Transformation

For transformations, electroporation of electrocompetent *E. coli* XL1-Blue cells was performed. Electrocompetence was achieved by serial washes in ultrapure water and 10% glycerol after being grown to a mid-logarithmic phase of growth. 0.8 – 1 µl ligation sample for standard cloning or retransformations and 4 µl of purified ligation sample for library cloning, respectively, were added to ~ 55 µl of thawed electrocompetent cells. The mixture was transferred to an electroporation cuvette (Gene Pulse Cuvette, BioRad) and electroporated at 1700 V/cm using the Eppendorf Eporator® (Eppendorf). When pEVO plasmid was used, the cells were recovered in 950 µl of SOC medium at 37 °C 1 hour, under constant shaking at 200 rpm. The recovered cells were used for subsequent plating on LB plates containing chloramphenicol (overnight at 37°C) or for liquid culture containing chloramphenicol (overnight at 200 rpm and 37°C). When plasmids for mammalian expression were used, the sample was directly plated on LB plates containing ampicillin. In case of R6K donor plasmids, pir+ top10 electrocompetent *E. coli* strain was used. After transformation, the cells were recovered for 1h at 37 °C in SOC medium, and then plated on LB plates containing kanamycin to grow overnight. Single colony was picked, sequence verified and grown in liquid culture as to maxi prep the plasmid DNA the next day as described above.

### 13.2.7 Sequencing

To verify successful cloning, plasmid DNA or *E. coli* colonies were sequenced using Microsynth Seqlab's (Göttingen, Germany) Sanger sequencing service. Plasmid DNA was diluted according to Microsynth's recommendations (DNA concentration 40-100 ng/µl).

## METHODS

Bacterial colonies were picked from the agar plate and transferred into the liquid inside the provided tubes from Microsynth Seqlab.

### 13.2.8 Test digest

To assess the recombination activity of a single recombinase or a library on a target site, a test digest was performed. Therefore, pEVO plasmid harboring the respective target sites, and a single recombinase or library was transformed into *E. coli* (XL-1 Blue) and grown overnight in LB medium containing chloramphenicol and a desired amount of L(+)-arabinose. In the pEVO construct, the recombinase is expressed from the L-arabinose promoter, which allows inducible expression of the enzyme by addition of L(+)-arabinose. After incubation, purified plasmid DNA, containing a mixture of recombined and non-recombined target sequences, was digested with BsrGI-HF and XbaI in a 20 µl reaction (see Restriction enzyme digestion). That way, if non-recombined, the plasmid is cut into fragments of size ~1kb and ~5kb and if recombined, the plasmid is cut into fragments of size ~1kb and ~4.2kb. The recombinase-mediated excision events could then be detected through the different fragment sizes using analytical gel electrophoresis (see Gel electrophoresis). To quantify the recombination efficiency, the intensities of the recombined and unrecombined bands on the agarose gel were measured using GelAnalyzer 19.1 Software, and the ratio of the intensities of both bands was calculated.

Test digest was also done after integration assays to confirm integration events. In this case, HindIII restriction enzyme was used and the co-integrant specific band could be detected at the height of 6058 bp, while re-excised pEVO band could be detected as 5342 bp band.

## 13.3 Bioinformatics

### 13.3.1 Identification of putative recombinases and native target sites

Potential new Y-SSRs were identified by using tblastn from BLAST+ 2.10.1 (<https://www.ncbi.nlm.nih.gov/books/NBK131777/>) with the protein sequences of Cre, Vika (Karimova et al., 2012), Nigri, and Panto (Karimova et al., 2016) as references to search the NCBI nucleotide collection database (v5). The results were filtered for below 90% identity and a sequence length of 300 to 400 amino acids with GNU awk. Protein sequences were acquired with efetch (<https://dataguide.nlm.nih.gov/edirect/efetch.html>). Full genome sequences of the potential SSRs were gathered using bastdbcmd (part of BLAST+). Potential target sites were identified by searching the genome sequences 1000 bp upstream and downstream of the potential Y-SSRs for palindromes. Palindrome search was performed with EMBOSS palindrome (v6.6.0.0) (Rice et al., 2000) with a minimum palindromic length of

## METHODS

13 to 15 base pairs and a gap limit of 8 with one mismatch allowed. The program output was then converted to a tabular format with GNU awk and combined with the protein data in R with the dplyr package. Potential SSRs with a Levenshtein distance (stringdist R package, <https://journal.r-project.org/archive/2014/RJ-2014-011/index.html>) below 10 to the references were removed. Potential SSRs were clustered with complete hierarchical clustering (base R) based on Levenshtein distances and cluster groups were formed with a cut-off distance of 11. The same clustering method was also used on the potential half-sites of the palindromic sequences, here the cluster cut-off was a distance of 2. The top candidates for testing were chosen considering the clustering, their organism they were found in and their distance to the reference recombinases (Supplementary Table S3, Supplementary Figure S1).

The phylogeny tree of known and putative recombinases was generated by performing an all-against-all pairwise sequence alignment of the protein sequences using EMBOSS needle (Needleman and Wunsch, 1970), followed by complete hierarchical clustering of the sequence dissimilarities. Visualization of the tree was done with R packages tidygraph and gggraph. For brevity, the tree was cut at the 98% sequence similarity, to represent almost identical proteins as a single node.

### 13.3.2 Identification of pseudo-*lox* sites in human and mouse genomes

Human and mouse genomic sequences with high similarity to potential target sites were identified using PatMaN (Prüfer et al., 2008). The search was performed on half-sites only, allowing for up to 2 mismatches. If two genomic sequences matching the same half-site were found to be located on opposite strands, with a distance of 8 bp between them, they were called a potential target site of the respective recombinase. Genomic coordinates manipulation and sequence extraction steps were performed with the BEDTools suite (Quinlan and Hall, 2010).

### 13.4 Plasmids

List of all plasmids used in this work can be found in the Supplementary Table S1 of Appendix section. Plasmid maps of the main constructs generated in this work can be found in Supplemental figures 2, 8, 9 and 10.

## METHODS

### 13.5 Plasmid construction

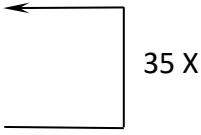
#### 13.5.1 Expression of recombinases

For expression in *E. coli*, codon-optimized DNA sequences of 17 predicted candidates were synthesized by Twist Biosciences amplified with pTWIST fw and pTWIST rev primers (See Supplementary Table S2 for primer sequences) and cloned into pEVO vector via BsrGI and SbfI restriction sites (Supplementary Figure S2) (Buchholz and Stewart, 2001). Respective target sites were introduced via pEVO-target PCRs (Table 9) with primers that were designed to carry the desired target sites and an overlap with the pEVO vector (Supplementary Table S2). The PCR fragment, that was generated when using the pEVO-*loxP* vector as template, was then cloned via Cold Fusion into a new, BglII digested pEVO backbone (System Biosciences) (reaction mixture for backbone digestion is described in Table 7).

**Table 9. Reaction mixture for pEVO-target PCR.**

Component	Volume (50 $\mu$ l in total)	Final concentration
5X Herculase II reaction buffer	10 $\mu$ l	1X
100 mM dNTPs	0.5 $\mu$ l	1 mM
12.5 $\mu$ M primer mix	1 $\mu$ l	0.25 $\mu$ M
Template DNA	1 $\mu$ l	~ 30 ng
5 U/ $\mu$ l Herculase II Fusion DNA polymerase	0.5 $\mu$ l	0.05 U/ $\mu$ l
Ultrapure water	37 $\mu$ l	-

#### Thermocycler program pEVO-target PCR

1. Initial denaturation	95°C	2 minutes	
2.	95°C	20 seconds	
3.	60°C	20 seconds	
4.	72°C	30 seconds	
5. Final extension	72°C	3 minutes	
6. Hold	8°C	$\infty$	

The FLAG-tagged versions of novel recombinases were constructed by amplifying all the novel recombinases with forward primers containing FLAG tag as to tag the recombinases on the N-terminal end (primers listed in Supplementary Table S2). The tagged recombinase gene fragments were then cloned into pEVO backbones via SacI and XbaI restriction sites.

## METHODS

For expression of recombinases in mammalian cells, two vectors were used: a lentiviral p81 - PGK-NLS-BFP plasmid (Supplementary Figure S4A), or a version of p11 - pECVW (EF1-alpha-mCherry-P2A-Vika) (Supplementary Figure S8). For cloning into p81, recombinase sequences were amplified via high-fidelity PCR with primers specific for each recombinase from pEVO vectors and cloned via BsrGI and XbaI restriction sites (Supplementary Table S2). The reverse primers were designed so, as to remove stop codons from the recombinase gene as it was cloned as a first ORF in P2A operon. The before mentioned lentiviral vector, harboring recombinases was either transfected into HEK293T or HeLa cells for transient expression, or used for virus production and infection for continuous expression. The pECVW vector was used just in one experiment, in an attempt to detect donor DNA integration in endogenous voxH9 target site. The p11-pECCW (EF1alpha-mCherry-p2A-Cre) was linearized via AgeI and XhoI restriction digestion and used as backbone to clone in the mCherry-P2A-Vika cassette which was generated by overlap PCR using p11 as a template for mCherry and pEVO-Vika as template for Vika (primers used can be found in Supplementary Table S2).

### 13.5.2 Recombination reporters

For the integration assay, the host plasmid carrying voxH9 target site (pEVO- $\Delta$ voxH9) was constructed by cloning the annealed oligonucleotides coding for the site (oMJ49 and oMJ50, Supplementary Table S2) in the pEVO-vox vector that was digested by BglII restriction enzyme. Donor R6K plasmids carrying voxN4.2 or 4.4 target sites were cloned by amplifying KanR gene with the forward primers containing the target sites (oMJ53 and oMJ54, respectively) and oMJ55 reverse primer (Supplementary Table S2). The PCR product was then digested with XbaI and XhoI restriction enzymes and cloned into the compatibly linearized R6K-neo based backbone.

For the construction of the excision recombination reporters in mammalian cells, pCAG-*loxP*-mCherry-*loxP*-GFP 'traffic light' vector was used (Supplementary Figure S4A) (Karpinski et al., 2016). Oligonucleotides containing respective target sites were used to amplify mCherry cassette and the fragment was later ligated to the pCAG plasmid via NheI and HindIII restriction sites (Supplementary Table S2).

For the mammalian integration reporter, the stock of p11-pECCW (EF1alpha-mCherry-p2A-Cre) lentiviral vector available in the laboratory was used as a backbone. First, mCherry cassette was amplified via high-fidelity PCR by using sense primer harboring voxH9 target site and AgeI restriction site upstream for subsequent cloning (oMJ100). Two overlapping reverse primers were designed as to remove BsrGI site from mCherry gene, and to exclude

## METHODS

Cre gene in order to replace it with BsrGI and XbaI cloning site for the cloning of recombinases (oMJ101 and oMJ102, Supplementary Table S2). This PCR was then introduced into p11 backbone by ligation via AgeI and XhoI restriction sites. The generated vector p11-empty was then either used as it is to produce lentiviral particles to generate the reporter cell lines HEK293T<sup>p11-empty</sup> and HeLa<sup>p11-empty</sup> (Figure 27B), or used as a backbone in order to clone wt Vika or Vika libraries via BsrGI and XbaI restriction sites. These plasmids were also used to produce respective cell line reporters: HEK293T<sup>p11-Vika</sup>, HEK293T<sup>p11-4.2library</sup>, HEK293T<sup>p11-4.4library</sup>, expressing either wt Vika or Vika libraries.

Donor plasmids for mammalian integration are based on pmaxGFP plasmid by Lonza (Shagin et al., 2004). First, puromycin N-acetyl-transferase (PAC) gene together with P2A polypeptide upstream, was introduced after turboGFP gene in order to ensure puromycin resistance upon integration (turboGFP-P2A-PAC cassette) (oMJ103 and oMJ104). For endogenous integration, a version with CMV promoter was generated by introducing voxN4.2 or voxN4.4 sites as oligonucleotides via BsrGI restriction site upstream of the CMV promoter (oMJ105-108, Supplementary Table S2). For landing pad integration, versions without CMV promoter and voxN target sites right upstream from the beginning of copGFP-P2A-PAC cassette were synthesized by Twist Biosciences as clonal genes.

### 13.6 Cell culture

#### 13.6.1 Cell culture maintenance

HEK293T or HeLa cells were cultured in different formats using T75 flasks, 10 cm dishes, 6-well or 24-well plates (Corning). Cells were cultured in Dulbecco's modified Eagle's medium (DMEM) with 10% fetal bovine serum, 1% penicillin-streptomycin (10 000 U/ml) and GlutaMAX™ (hereinafter referred to as DMEM+), at 37 °C, 5% CO<sub>2</sub> in HERAcCell Incubator 240i (Thermo Fischer Scientific, USA).

Cells were split when they reached 90-100% confluency or when needed for an experiment. Therefore, the cells were washed once with PBS and then detached using Trypsin. When the cells started to lift off, trypsin was inactivated by adding DMEM+, and rinsing multiple times by pipetting up and down. Depending on how many cells were needed for a subsequent experiment, the desired amount of the cell suspension was combined with the respective amount of fresh DMEM+ to reach the total volume for each format (see Table 10). To determine the accurate cell count, cells were counted using the Countess 3 FL Automated Cell Counter (Thermo Fisher).

## METHODS

**Table 10. Pipetting scheme for splitting cells for different formats.**

	<b>T75 flask</b>	<b>10 cm dish</b>	<b>6-well</b>	<b>24-well</b>
PBS	7 $\mu$ l	4 $\mu$ l	1.5 $\mu$ l	0.5 $\mu$ l
Trypsin	2 $\mu$ l	1 $\mu$ l	0.4 $\mu$ l	0.1 $\mu$ l
DMEM+ for Trypsin inactivation	8 $\mu$ l	4 $\mu$ l	1.6 $\mu$ l	0.4 $\mu$ l
DMEM+ to reach total volume	Up to 14 $\mu$ l	Up to 8 $\mu$ l	Up to 3 $\mu$ l	Up to 1 $\mu$ l

### 13.6.2 Fluorescent activated cell analysis

Cells were washed once with PBS and then detached using Trypsin (Gibco). Next, the cells were transferred to a Falcon® 5 mL Round Bottom Polystyrene Test Tube with Cell Strainer Snap Cap or to a 96-well round bottom plate in their respective cell culture medium. Cells were analyzed with the MACSQuant® VYB Flow Cytometer (Miltenyi) or with BD LSR Fortessa Cell Analyzer (BD Biosciences). Analysis of the data was performed using FlowJo™ 10 (BD). Generally, the gating hierarchy was following: Population of interest -> Single cells -> transfection fluorescence -> reporter fluorescence (Supplementary Figure S4B), except in the case of integration assays when cells were analyzed 7 or 10 days after transfection. In that case, reporter fluorescence was gated directly from single cells.

### 13.6.3 Plasmid transfection

HEK293T or Hela TDS cells were transfected 24h after seeding. The appropriate amount (see Table 11) of plasmid and Opti-MEM I Reduced Serum Medium (ThermoFisher) was prepared each in a 1.5 ml tube for each well/dish. Next, the appropriate volumes of Lipofectamine 2000 (ThermoFisher) were added to the 1.5 ml tube with Opti-MEM I Reduced Serum Medium and mixed by pipetting. This mixture was then added to the 1.5 ml tube containing the plasmid, vortexed for 10s at maximum speed and incubated for 20 min at RT. During this time the medium of the cells was replaced with fresh medium. The transfection mixture was slowly added to the cells. After 24 h the medium was replaced with fresh medium. Cells were analyzed earliest 48 h post transfection.

## METHODS

**Table 11. Pipetting scheme for plasmid transfections for different formats and reagents.**

	<b>Seeding density HEK293T</b>	<b>Seeding density HeLa</b>	<b>DNA</b>	<b>Lipofectamine 2000</b>	<b>PEI</b>	<b>Opti-MEM</b>
24-well	0.2 x 10 <sup>6</sup>	0.09 x 10 <sup>6</sup>	0.8 µg	2 µl	-	2 x 50 µl
6-well	1 x 10 <sup>6</sup>	0.4 x 10 <sup>6</sup>	4 µg	8 µl	-	2 x 250 µl
10-cm	5 x 10 <sup>6</sup>	3.5 x 10 <sup>6</sup>	8 µg	-	40 ul	2 x 700 µl

### 13.6.4 Lentiviral particle production

Lentiviral particles were used to stably integrate the pPGK-NLS-Recombinase-BFP construct randomly into the NIH/3T3 genome, or p11-empty (EF1alpha-voxH9-mCherry) and its derivatives into HEK293T cells for generation of landing pad reporter cell lines. In both cases, the first step comprised of using HEK293T cells to produce the infectious lentiviral particles. Therefore, cells were plated at a density of 5 x 10<sup>6</sup> cells in 10 cm dishes and incubated overnight as described above (day 1). On day 2, 12 µl chloroquine was added to each dish as an endosomal acidification and autophagy inhibitor. A DNA solution with a total of 15.2 µg plasmid DNA was prepared according to the Table 12 as well as a PEI solution consisting of 45 µl PEI and 655 µl Opti-MEM medium. Both solutions were combined, vortexed well, and incubated at room temperature for around 30 minutes.

**Table 12. Reaction mixture to prepare DNA solution for virus production in a 10 cm dish.**

<b>Component</b>	<b>Amount</b>
Opti-MEM medium	700 µl
Desired plasmid DNA	6.4 µg
psPAX2 (gag/pol)	5.6 µg
pMD2.G (Env)	3.2 µg

After incubation, the mixture was added dropwise to the HEK293T cells. The cells were then incubated overnight as described above and after one day (day 3), the transfection medium was replaced with fresh DMEM+. The viral supernatants were then collected on day 4, filtered through a 0.45 µl Filtropur cell strainer (Sarstedt) and 1% protamine-sulfate was added. The virus supernatants were, then, either frozen or used directly for infection.



## METHODS

### 13.6.5 Generation of landing pad cell lines

Different concentrations of viral particles produced with p11-EF1alpha-voxH9-mCherry empty and its derivatives expressing Vika wt, 4.2 or 4.4 library were used for the transduction of the freshly seeded HEK293T to estimate the MOI (Multiplicity of infection) using the MACSQuant® VYB Flow Cytometer (Miltenyi) 2 days after infection (Figure 27B - upper). The viral load that resulted in a transduction efficiency of less than 20% was used to establish the reporter cell lines. The mCherry<sup>+</sup> cells were then enriched by BD FACSAria III Sorter (BD Biosciences) in order to establish the cell lines HEK293T<sup>p11-Vika</sup>, HEK293T<sup>p11-4.2library</sup>, HEK293T<sup>p11-4.4library</sup> or HEK293T<sup>p11-empty</sup> (Figure 27B – lower panel). The stability of the reporter cell lines was followed over the course of 12 days by measuring % of mCherry<sup>+</sup> cells at three different time points (day 5, 8 and 12 after sorting) using the MACSQuant® VYB Flow Cytometer (Miltenyi) (Figure 27C).

### 13.7 Biochemistry

#### 13.7.1 Preparation of the protein extract

Sample preparation for western blot analysis was performed by Trichloroacetic acid (TCA) precipitation. pEVO-N-FLAG plasmids carrying 8 newly described recombinases and Vika and Cre as controls were transformed in *E. coli* and grown ON with 10 µg/ml L-arabinose to induce recombinase expression. Next day, the cultures were diluted and grown until OD<sub>600</sub>= 0.5 twice in order to synchronize most of the cells in the mid-exponential phase, to ensure that equal number of cells is used for protein extraction. 900 µl of cultures was then incubated with 100 µl of ice-cold TCA for 30 minutes and then the pellets were centrifugated for 5 min at 13 000 rpm. Pellets were then washed with cold acetone and centrifugated again. The acetone was discarded and the samples were incubated at 37°C for 5 minutes so that the remaining acetone evaporates.

#### 13.7.2 SDS-PAGE

Remaining pellets were resuspended in 2X Laemmli Buffer and incubated at 95°C for 10 minutes. Samples were stored at -20°C until needed. 20 µl of each sample were separated by electrophoresis on NuPAGE™ Novex 4 - 12% Bis-Tris Gel (Invitrogen) with 1 X NuPAGE™ MOPS SDS Running Buffer (20 X) (Invitrogen) in an Invitrogen™ XCell SureLock™ Mini-Cell and XCell II™ Blot Modul (Invitrogen) at 120 V for 80 minutes.

## METHODS

### 13.7.3 Western blotting

After separation, the samples were transferred from the gel to a 0.45 µm Nitrocellulose Blotting Membrane (Amersham™ Protran®) in a TE77 Large SemiPhor Semi-Dry Transfer Unit (Hofer) using Extra Thick Blot Filter Paper (BioRad) immersed in 1 X Blotting Buffer (see Buffer and staining solutions) at 50 mA for 90 min. For both separation and blotting, EPS-601 Electrophoresis Power Supply (Amersham Pharmacia) was used. The Nitrocellulose Blotting Membrane was stained using 1 X Amido black solution to visualize total protein fraction and then photographed as to use it for loading control. For the detection of recombinase expression monoclonal ANTI-FLAG® M2 mouse antibody (1:5000, Sigma-Aldrich) was used. The membranes were then incubated with IRDye® 800CW Donkey anti-Mouse IgG Secondary Antibody (1:10,000, LI-COR Biosciences) and protein signal was then detected using the Odyssey® Classic Imaging System (LI-COR) and Image Studio™ Software (LI-COR). The band intensities were quantified by using Fiji-ImageJ and visualized with GraphPad Prism. The quantification was done from 3 independent biological replicates.

### 13.8 Recombination assays

#### 13.8.1 Recombination assay based on pEVO vector- excision

To visualize the recombination activity of novel recombinases on their predicted target sites, a plasmid-based assay was used as previously described (Karimova et al., 2012, 2016)(Figure 8A). In short, expression of the recombinases from the pBAD promoter was induced with L-arabinose (Sigma-Aldrich Chemie GmbH). Single clones containing the pEVO plasmid with the recombinase and recombination target sites were cultured overnight in 6 ml LB medium with 25 µg/ml Cm and either 0 or 100 µg/ml L-arabinose at 37°C and 200 rpm. The recombinase mediated excision event was detected by agarose gel electrophoresis after digestion with BsrGI and SbfI restriction enzymes. The recombined plasmid is smaller in size compared to the non-recombined plasmid. Therefore, after gel electrophoresis a slower migrating non-recombined band (~5.0 kb), and a faster migrating (~4.3 kb) band for recombined plasmids can be seen (Figure 8A).

To compare the recombination efficiency of the novel recombinases on their native target sites, recombinase expression was induced with increasing concentrations of L-arabinose (0, 1, 10 or 100 µg/ml medium) overnight in 6 ml culture volume. The test digest (See section test digest) was done for each induction level and recombination efficiency was estimated by agarose gel electrophoresis. To quantify the recombinase activity, the ratio of band intensities was determined using Fiji-ImageJ for image processing. The quantified

## METHODS

recombination was plotted in R 4.0.3 with dplyr v1.0.7 and visualized with ggplot2 v3.3.5. All test digests were done in triplicates (n = 3).

### 13.8.2 Recombination assay based on pEVO vector- integration assay

Integration assay is a two-plasmid assay that reports integrative recombination (Buchholz & Stewart 2001). Here the integration property of the recombinase is tested by its ability to utilize two target sites located on different plasmids in order to merge them into one plasmid through the site-specific recombination reaction. The assay is designed with vectors based on two types of origin of replication: on p15A for the host pH plasmid (chloramphenicol resistant), and on R6K for the donor pD- plasmid (kanamycin resistant) (Figure 20). Replication of the R6K-plasmid is *pi*-protein dependent and therefore this plasmid cannot replicate in a *pir*-negative bacterial strains. Hence kanamycin resistance conveyed by the donor plasmid depends on its integration into a host plasmid through a *lox* site to form the co-integrand and consequent replication of the kanamycin resistance gene in *pir*-negative cells.

In a two-day protocol, host plasmid (pEvo- $\Delta$ voxH9) was electroporated into *E. coli* XL1-Blue on the first day. Cells were grown overnight in Cm. On day two, a fresh 50ml culture of cells harboring the host plasmids were grown for 3.5 hours with Cm and desired amount of  $\mu$ g/ml L- arabinose to induce recombinase expression. Cells were brought to competent state and electroporated with 50 ng of the donor plasmid DNA. After 1-hour recovery cells were plated on kanamycin plates. Control experiments were conducted in absence of arabinose induction with no growth on Km plates observed.

In a one-day protocol, both pH (pEvo- $\Delta$ voxH9) and pD (R6K) plasmids were co-transformed into electrocompetent XL1-Blue *E. coli* in 1:2 ratio (24,3 ng pEVO and 25,7 ng R6K). The samples were then split in two and recovered for 2 hours in LB medium with or without addition of arabinose. After 2-hour recovery cells were plated on chloramphenicol (Cm) and kanamycin plates (Cm+Kan). The samples where arabinose was not added during recovery lacked induction of recombinase expression and no growth on Cm+Kan plates was observed.

### 13.8.3 Mammalian recombination assays

For the mammalian excision reporter assay, HEK293T cells were plated at a density of  $2 \times 10^5$  cells per well in 24-well plates and cultured in glucose Dulbecco's Modified Eagle's Medium (DMEM, Gibco®), supplemented with 10% fetal bovine serum (Invitrogen), 1% Penicillin-Streptomycin (10,000 U/ml, Thermo Fisher). At a confluency of 70-80%, cells were

## METHODS

co-transfected with pPGK-NLS-Recombinase-P2A-BFP plasmids expressing a recombinase and pCAG-*lox*-mCherry-*lox*-GFP traffic light reporters using Lipofectamine® 2000 Transfection Reagent (Invitrogen) according to manufacturer's instructions. Per well 0.5 µg of DNA (0.25 µg of each plasmid) and 2.5 µl of Lipofectamine® 2000 reagent diluted in 100 µl Opti-MEM® Reduced Serum Media each were used. On the next day, the media was changed and the cells were further cultured at 37 °C and 5% CO<sub>2</sub>. Upon recombination between the target sites, the mCherry cassette is excised and CAG promoter starts driving the expression of downstream green fluorescent protein (GFP). The cells were analyzed 2 days after transfection with fluorescent activated cell analysis and were then imaged with a fluorescent microscope (EVOS FL imaging system; Thermo Fisher Scientific).

For the mammalian integration assay, HEK293T, HEK293T<sup>p11-Vika</sup>, HEK293T<sup>p11-4.2library</sup>, HEK293T<sup>p11-4.4library</sup> or HEK293T<sup>p11-empty</sup> cell lines were seeded at a density of 5 x 10<sup>6</sup> in a 10 cm dishes and cultured in glucose Dulbecco's Modified Eagle's Medium (DMEM, Gibco®), supplemented with 10% fetal bovine serum (Invitrogen), 1% Penicillin-Streptomycin (10,000 U/ml, Thermo Fisher). At a confluency of 70-80%, cells were transfected with donor plasmids (CMV versions in the case of endogenous integration or promoterless versions for landing pad integration) by using cationic polymer polyethylenimine (PEI). Per dish 8 µg of DNA and 40 µl of PEI diluted in 700 µl Opti-MEM® Reduced Serum Media each were used. On the next day, the media was changed and the cells were further cultured at 37 °C and 5% CO<sub>2</sub>. Upon integration, the GFP-P2A-PAC cassette gets constitutively expressed by either, EF1-alpha promoter in the landing pad cell lines or by CMV promoter of the integrated plasmid from the endogenous *voxH9* locus of HEK293T wt cells. The cells were, therefore, analyzed for GFP expression, 7 or 10 days (with CMV donors) after transfection with fluorescent activated cell analysis (FACS). After FACS analysis, puromycin selection was started by using 1 µg/ml of puromycin at first and then increasing to 2 µg/ml until selection was done. The clones were analyzed individually by microscopy (EVOS FL imaging system; Thermo Fisher Scientific), or in bulk with FACS. Genomic DNA was isolated with QIAamp DNA Blood Mini Kit (Qiagen) according to the manufacturer's manual, and subjected to the PCR amplification specific for the donor plasmid/genome upstream and downstream junctions.

For validation of the individual clones in mammalian cells, the integration assay was slightly modified. Potentially active variants of the two libraries were amplified from the puromycin selected pull, by high-fidelity PCR with upstream primer binding into the donor plasmid ensuring once more that only active variants are amplified (primers oMJ121 and oMJ74). The PCR product of active recombinase pulls was digested with BsrGI and XbaI and cloned into pPGK-NLS-Recombinase-P2A-BFP expression plasmids described previously. HeLa<sup>p11-</sup>

## METHODS

<sup>empty</sup> cell line was produced in the same way as HEK293T<sup>p11-empty</sup>, and was used for these experiments. HeLa<sup>p11-empty</sup> cells were seeded at a density of  $3 \times 10^5$  in 6-well plates and cultured in glucose Dulbecco's Modified Eagle's Medium (DMEM, Gibco®), supplemented with 10% fetal bovine serum (Invitrogen), 1% Penicillin-Streptomycin (10,000 U/ml, Thermo Fisher). At a confluency of ~80%, cells were co-transfected with pPGK-NLS-Recombinase-P2A-BFP expression, and pmaxGFP-PAC-voxN4.2 or pmaxGFP-PAC-voxN4.4 donor plasmids using Lipofectamine® 2000 Transfection Reagent (Invitrogen) according to manufacturer's instructions. Per well 4 µg of DNA (2 µg of each plasmid) and 8 µl of Lipofectamine® 2000 reagent diluted in 250 µl Opti-MEM® Reduced Serum Media each were used. On the next day, the media was changed and the cells were further cultured at 37 °C and 5% CO<sub>2</sub>. Samples were analyzed by FACS at 3 and 7 days post-transfection to estimate the transfection and integration efficiency for each clone. Vika wt and empty pPGK-NLS-BFP vector were used as controls in these experiments.

### 13.9 Cross-recombination assay: Nanopore sequencing.

Eight new recombinases and Cre, Vika, Panto, Dre (Anastassiadis et al., 2009), VCre (Suzuki and Nakayama, 2011) were amplified from pEVO vectors, cleaned using the Isolate II PCR and Gel Cleanup Kit (Bioline) and mixed together in a 1:1 ratio. All respective target sites were cloned into the pEVO vectors with Cold Fusion Cloning kit as previously described and the resulting vectors were also mixed with equal molar ratio. Both, the mix of recombinases and pEVO backbones were digested with BsrGI and SbfI and ligated in a single reaction, thus creating a library of different recombinase/ target site pairs. Plasmids were transformed in XL1-Blue electrocompetent *E. coli* cells and grown overnight with 100 µg/ml L-arabinose to induce recombinase expression. On the next day, plasmids were linearized with BsrGI and Scal and fragments carrying the recombinase sequence and target sites were isolated by agarose gel excision using the Isolate II PCR and Gel Cleanup Kit (Bioline). These DNA fragments were then prepared for nanopore sequencing with the SQK-LSK110 Kit according to the "Amplicons by Ligation" protocol on a MinION R9.4.1 Flow Cell (Oxford Nanopore Technologies). Base calling of the sequence data was performed with guppy v5.0.7 on the high accuracy model (Oxford Nanopore Technologies). The sequence reads were then filtered for a read length of at least 1800 bp and a minimum mean phred score of 10 with filtlong (<https://github.com/rrwick/Filtlong>). To identify the recombinases, the reads were aligned to the reference recombinase sequences with minimap2 v2.17 (Li, 2018). The target sites were identified using exonerate v2.2.0 using the affine:local model (<https://www.ebi.ac.uk/about/vertebrate-genomics/software/exonerate>). The read ID and the

## METHODS

matching references were then extracted from both alignments and combined in R with the `dplyr` package. Visualization of the data was performed with the R package `ggplot2`.

### 13.10 Overexpression studies in mammalian cells

For the viral delivery, pPGK-Recombinase-P2A-BFP vectors were used to produce lentiviral particles as described previously. NIH/3T3 mouse fibroblasts were seeded at a density of  $4 \times 10^4$  cells per well in 24-well plates and grown at 37 °C and 5% CO<sub>2</sub>. The next day, fibroblasts were transduced with the different lentiviruses with a MOI of 0,5 in order to achieve about 50% of infection rate. The percentage of BFP expressing cells from at least  $2 \times 10^4$  cells was tracked over the course of 15 days using MACSQuant® VYB Flow Cytometer (Miltenyi). The difference in the percentage of BFP cells at the last time point (day 15) and first time point was calculated and visualized with GraphPad Prism and statistical significance relative to BFP control was calculated by doing 1-way ANOVA test with 95% CI.

### 13.11 Substrate-linked directed evolution (SLiDE)

The SLiDE protocol was used to generate diverse libraries of the Vika recombinase with slightly different specificities in a stepwise manner. First, a diverse starting library of randomly mutated recombinases was generated via error-prone PCR using pEVO-Vika as a template and oMJ109 and oMJ110 as primers (see Table 13 and Supplementary Table S2). The PCR is characterized by (1) an imbalance of dNTPs increasing the likelihood of wrong incorporation of bases, (2) the addition of manganese chloride, which reduces the base pair specificity and (3) an increased amount of magnesium chloride, which stabilizes non-complementary base pairing. Therefore, the following reaction mixture and cycling program was used:

**Table 13. Reaction mixture for error-prone PCR to generate the starting library.**

Component	Volume (50 µl in total)	Final concentration
10X NH <sub>4</sub> Reaction Buffer	5 µl	1X
50 mM MgCl <sub>2</sub>	7 µl	7 mM
5 mM MnCl <sub>2</sub>	1 µl	0.1 mM
100 mM dNTPs (A=G, C=T)	0.1 µl, 0.5 µl	0.2 mM, 1mM
12.5 µM primers	1 µl each	0.25 µM
Template DNA	1 µl	25 ng
5 U/µl BIOTAQ™ DNA Polymerase	0.5 µl	0.05 U/µl

## METHODS

---

Ultrapure water	32.3 $\mu$ l	-
-----------------	--------------	---

---

### Thermocycler program error-prone PCR

1. Initial denaturation	94°C	90 seconds	
2.	94°C	15 seconds	← 26 X
3.	54.5°C	15 seconds	
4.	72°C	40 seconds	
5. Final extension	72°C	3 minutes	
6. Hold	8°C	$\infty$	

The resulting starting library was cloned into the evolution vectors pEVO-vox, pEVO-vox13C, pEVO-vox12T or pEVO-vox9G via BsrGI-HF and XbaI-HF restriction sites as described. Since library cloning was performed, the ligation sample was purified (see Ligation) and 4  $\mu$ l were used for transformation via electroporation into *E. coli* XL1-blue (see Transformation). The transformation efficiency (library size) was determined by plating 2  $\mu$ l (1:500) of the recovery on a fresh LB plate containing chloramphenicol (15  $\mu$ g/ml). The remaining recovery was grown overnight in LB medium containing chloramphenicol (25  $\mu$ g/ml) and 200  $\mu$ g/ml L(+)- arabinose to induce recombinase expression. Active recombinase variants recombine the pEVO plasmid and thereby excise the intervening region containing NdeI and AvrII restriction sites, while inactive variants do not recombine their target site. The plasmids were subsequently purified from the overnight culture ("miniprep", see Plasmid DNA purification). To perform selection for excision, the purified plasmids were digested with NdeI and AvrII in a 20  $\mu$ l reaction (see Restriction enzyme digestion) to linearize all non-recombined, and therefore inactive variants. Since active variants excised their NdeI and AvrII restriction sites, they remain circular. To retrieve these active variants, evolution PCR was performed using the NdeI/AvrII digestion as a template and oMJ109 and oMJ110 as primers (see Table 14). To fine-tune the mutation rate to avoid incorporating too many destabilizing and inactivating mutations, this PCR constitutes an adapted version of the previously described error-prone PCR, characterized by the use of the low-fidelity MyTaq™ Polymerase. The following reaction mixture and cycling program was used:

## METHODS

**Table 14. Reaction mixture for evolution PCR.**

Component	Volume (50 $\mu$ l in total)	Final concentration
5x MyTaq™ Reaction Buffer Red (contains dNTPs)	10 $\mu$ l	1X
12.5 $\mu$ M primers	1 $\mu$ l each	0.25 $\mu$ M
Template DNA	1 $\mu$ l	~ 25 ng
5 U/ $\mu$ l MyTaq™ Polymerase	1 $\mu$ l	0.1 U/ $\mu$ l
Ultrapure water	36 $\mu$ l	-

### Thermocycler program evolution PCR

1. Initial denaturation	94°C	90 seconds	
2.	94°C	15 seconds	
3.	54.5°C	15 seconds	
4.	72°C	40 seconds	
5. Final extension	72°C	3 minutes	
6. Hold	8°C	$\infty$	

The amplification of the active recombinase library (~ 1.8 kb PCR fragment) was confirmed by analytical gel electrophoresis (see Gel electrophoresis). The selected variants were then carried on to the next cycle by cloning into fresh pEVO-vox, pEVO-vox13C, pEVO-vox12T or pEVO-vox9G vectors via BsrGI and XbaI restriction sites and transformed into *E. coli* (XL1-Blue) as described above. *E. coli* was grown overnight in LB medium containing chloramphenicol (25  $\mu$ g/ml) and the desired amount of L(+)- arabinose to induce recombinase expression. To monitor the library size, transformation efficiency was determined for every evolution cycle by plating 2  $\mu$ l (1:500) of the recovery on a fresh LB plate containing chloramphenicol (15  $\mu$ g/ml). The library size was determined the next day based on the colony counts of the plate and the dilution factor (colony number x 500). Only if the library size exceeded 100,000 clones it was continued with the next SLiDE cycle. Then, a miniprep was performed from 10 ml of the liquid culture (see Plasmid DNA purification) and the evolution cycle was started again. For every cycle, recombination efficiency was monitored by performing a test digest of the miniprep in parallel to the NdeI/AvrII restriction enzyme digestion (see Test digest). The whole procedure was repeated until an appropriate recombination activity of the library was observed in the test digest. Over the course of the



## METHODS

evolution cycles, the amount of L(+)-arabinose was decreased in a stepwise manner, allowing only very active recombinases to remain in the library and propagate.

### 13.12 Integration-based SLiDE

The integration-based SLiDE protocol is based on one-day integration assay protocol. To initiate the evolution, the four Vika libraries generated during standard SLiDE (refer to Chapter 8.3) were mixed and cloned in via BsrGI and XbaI restriction sites into modified pEVO vector - pEVO- $\Delta$ voxH9. Since library cloning was performed, the ligation sample was purified (see Ligation) and 5,4  $\mu$ l (equal to approx. 24,3 ng of pEVO DNA) was co-transformed with 1  $\mu$ l of R6K-voxN4.2 or R6K-voxN4.4 (approx. 25,7 ng R6K, for 1:2 ratio) via electroporation into XL1-Blue E. coli. The samples were then split in two and recovered for 2 hours in LB medium with or without addition of arabinose, as to induce library expression. The transformation efficiency (library size) was determined by plating 1  $\mu$ l of recovery (1:1000) of the recovery on a fresh LB plate containing chloramphenicol (15  $\mu$ g/ml). Additional 10  $\mu$ l (1:100) was plated on double selective plates containing both chloramphenicol (15  $\mu$ g/ml) and kanamycin (15  $\mu$ g/ml) (Cm+Kan), as to follow the progress of evolution by calculating the ratio between number of colonies on Cm+Kan plates and the number on Cm plates accounting also for the dilution factor (colony number on Cm+Kan plates x 100/colony number on Cm plates x 1000). The rest of the recovery was used to inoculate fresh 100 ml of LB media containing both antibiotics. Only the active variants that were able to mediate the integration of the plasmids could grow in the culture and were selected for the next cycle. Next day, DNA was purified from the culture (see DNA purification) and subjected to PCR with primers specific for the co-integration products pEVO-R6K plasmids that will amplify only the active recombinases (depicted as IntEvo PCR in the Figure 21F, Table 14). This PCR is performed with a low-fidelity DNA polymerase to add diversity to the pool of active recombinase variants. The diversified active recombinase fragments are then re-cloned into a non-recombined pEVO backbone which marks the beginning of the new cycle. For every cycle, successful integration was confirmed by performing a test digest of the miniprep in parallel to the selective PCR. Minipreps contain a mix of co-integrated vectors, re-excised pEVO vectors and might contain R6K plasmids in traces. Digestion of the plasmid DNA via HindIII restriction enzyme was done as described previously (see Table 7), allowing the distinction of bands specific for all three scenarios, but couldn't serve as a quantitative measure of the integration efficiency (Figure 21E). The whole procedure was repeated until an appropriate recombination activity of the library was deduced from the ratio between Cm and Cm+Kan plates. Over the course of the evolution

## METHODS

cycles, the amount of L(+)-arabinose was decreased in a stepwise manner, allowing only very active recombinases to remain in the library and propagate (Figure 22B).

### 13.13 PCR-based genomic integration detection

The integration into mammalian genome was undoubtedly confirmed by performing the PCR amplification of the upstream and downstream junctions between the integrated plasmid DNA and targeted genomic locus. For integration into endogenous *voxH9* locus, the junction PCRs were performed with following primers: oMJ114 and oMJ117 for junction 1; oMJ118 and oMJ116 for junction 2 (Supplementary Table S2). If the integration was successful the 785 bp band for upstream junction and 623 bp band for the downstream junction could be detected on the agarose gels (Figure 18D). On the other hand, integration into integrated landing pad was detected with the following primers: oMJ119 and 120 for the 5' junction (j1); oMJ121 and oMJ122 for the 3' junction (j2) (Supplementary Table S2). Successful integration could be visualized on agarose gels by the presence of the 890 bp band for upstream junction and 505 bp band for the downstream junction (Figure 28C). All PCRs were performed with Q5 high-fidelity polymerase as described in the Table 6. All PCR products were cleaned up with the Isolate II PCR and Gel Kit (Bioline) and sent for Sanger sequencing to confirm the expected plasmid/genomic DNA border.

### 13.14 IntDEQSeq screen

#### 1.1.1 UMI fragment preparation

To make these oligonucleotides double stranded a 50  $\mu$ l PCR was performed with 20  $\mu$ M of the primers oMJ124 and oMJ125 (Supplementary Table S2), 10  $\mu$ M of the UMI oligonucleotide (oMJ123), 10  $\mu$ l of 5x MyTaq buffer and 1  $\mu$ l MyTaq polymerase (Bioline). The PCR-cycler was set 94  $^{\circ}$ C for 90 seconds, followed by 10 cycles of 15 seconds at 94  $^{\circ}$ C, 15 seconds at 54  $^{\circ}$ C and 15 seconds at 72  $^{\circ}$ C. The resulting PCR product was digested with XbaI and HindIII. The digest was then again cleaned up with the Isolate II PCR and Gel Kit (Bioline) and measured with a Qubit HS dsDNA Kit on a Qubit 2.0 (Thermo Fisher Scientific).

#### 13.14.1 Recombinase variant barcoding and integration assay

The evolved libraries were acquired from pEVO plasmids by digesting with XbaI and BsrGI-HF. The libraries and Vika wt control were ligated in a ratio of 60 ng of recombinase gene fragment, 4.8 ng UMI fragment and 100 ng of BsrGI-HF and HindIII digested pEVO- $\Delta$ voxH9 plasmid. The ligated plasmids were desalted with MF-Millipore membrane filters (Merck) on

## METHODS

distilled water for 30 minutes and transformed into XL-1 Blue E. coli via electroporation. The transformed bacteria were cultured in SOC medium for 1 hour at 37 °C. 2 µl of this culture was spread on agarose plates with 15 mg/ml chloramphenicol and incubated overnight at 37 °C. The number of colonies on the plates were counted to calculate the number of transformed bacteria present per µl of SOC culture.

To nominate the number of variants for the screen, an amount of the SOC culture equal to the desired number of variants was cultured overnight in 100 ml LB medium with 25 mg/ml chloramphenicol. From each library around 2000 transformed bacteria were culture together with around 150 transformed bacteria of Vika wt control. Next day, 500 µl of overnight cultures was used to inoculate 50 ml of fresh LB media. 50 ml culture of cells harboring the UMI-barcoded recombinase genes were grown for 3.5 hours with Cm and 10 µg/ml L-arabinose to induce recombinase expression. Cells were brought to competent state and electroporated with 50 ng of the pD plasmids (R6K-voxN4.2 or R6K-voxN4.4) in triplicates to ensure the reproducibility of the results. The samples were then grown in Cm and Cm+Kan media in order to compare the representations of each cluster before and after selection. Next day, the plasmid DNA of these cultures was extracted with the GeneJet Plasmid Miniprep Kit (Thermo Fisher Scientific). The barcoded plasmid extracts from the libraries were digested with BsrGI-HF and Scal in the case of Cm samples, or BsrGI and XhoI in the case of Cm+Kan samples. The resulting fragments containing the evolved gene, the UMI, and the target site products were isolated via agarose gel excision with the Isolate II PCR and Gel Kit (Bioline).

### 13.14.2 Nanopore sequencing and processing of screen libraries

The concentration of DNA fragments was measured with a Qubit dsDNA HS Assay Kit on a Qubit 2.0 Fluorometer (Thermo Fisher Scientific). Nanopore sequencing library preparation was performed according to the "Native Barcoding Kit 24 V12 (SQK-NBD112.24)". The pulled sample library was then loaded onto a MinION FLO-MIN110 flow cell with r10.4 pores (Oxford Nanopore Technologies). Sequencing was performed for 72 hours.

Basecalling of the sequence data was performed on guppy version 6.0.1 with the high accuracy model. Reads were first filtered with Filtlong v0.2.1 (<https://github.com/rrwick/Filtlong>) to be at least 2000 bp long for the Cm samples and 2500 bp long for the Cm+Kan samples. The sequences were then aligned with minimap2 (Li, 2018) to a reference sequence containing Vika wt and the UMI of 50 'N' on a pEVO non-integrated fragment or pEVO-R6K co-integrated fragment.

From the filtered alignment the UMIs were then extracted with the `stackStringsFromBam` function from the R package `GenomicAlignments` (Lawrence et al., 2013). The UMIs were

## METHODS

then clustered with VSEARCH4 (Rognes et al., 2016) with a cluster\_identity value of 0.7. Sequence reads from clusters with a minimum size of 50 reads were then transferred to separate files and aligned to the gene-UMI reference sequence. These separate read files and alignments were used to construct consensus sequences with racon (Vaser et al., 2017) followed by further polishing with medaka (<https://github.com/nanoporetech/medaka>), both with standard settings. The polishing process was run in parallel with GNU (<https://doi.org/10.5281/zenodo.1146014>). Finally, gene sequences were extracted with the R package GenomicAlignments and translated to amino acids.

The clusters that exist in all replicates of one or the other library were filtered (Cm included), and the read frequency of each replicate (Cm included) was calculated. All the clusters that don't have at least 50 reads in all of the replicates were then removed. Finally, for each cluster, frequency fold change of the Cm+Kan samples to the Cm samples was calculated. The results from the screen were then combined. All further data processing and visualization was performed in R with the tidyverse and stringdist packages (Loo)

### 13.15 Competition assay

In order to directly compare the integration efficiency mediated by the individual clones to wt Vika, a competition assay imitating the conditions in the IntDEQSeq screen was designed. XL-1 Blue E. coli cells were transformed with wt Vika and the individual clones in parallel, and grown over night on chloramphenicol selective LB plates (15 µg/ml of Cm). Next day, a colony from Vika wt and clones' plates was picked and grown until mid-exponential phase ( $OD_{600}=0.5$ ). 100 µl of Vika wt over-day culture was added to the 100 µl of over-day culture of each clone (1:1 ratio) to ensure that bacterial mix contains equal amounts of bacteria carrying Vika wt or clone recombinase gene. Total of 200 µl of the mix was then used to inoculate 20 ml of LB-Cm media and grown overnight.

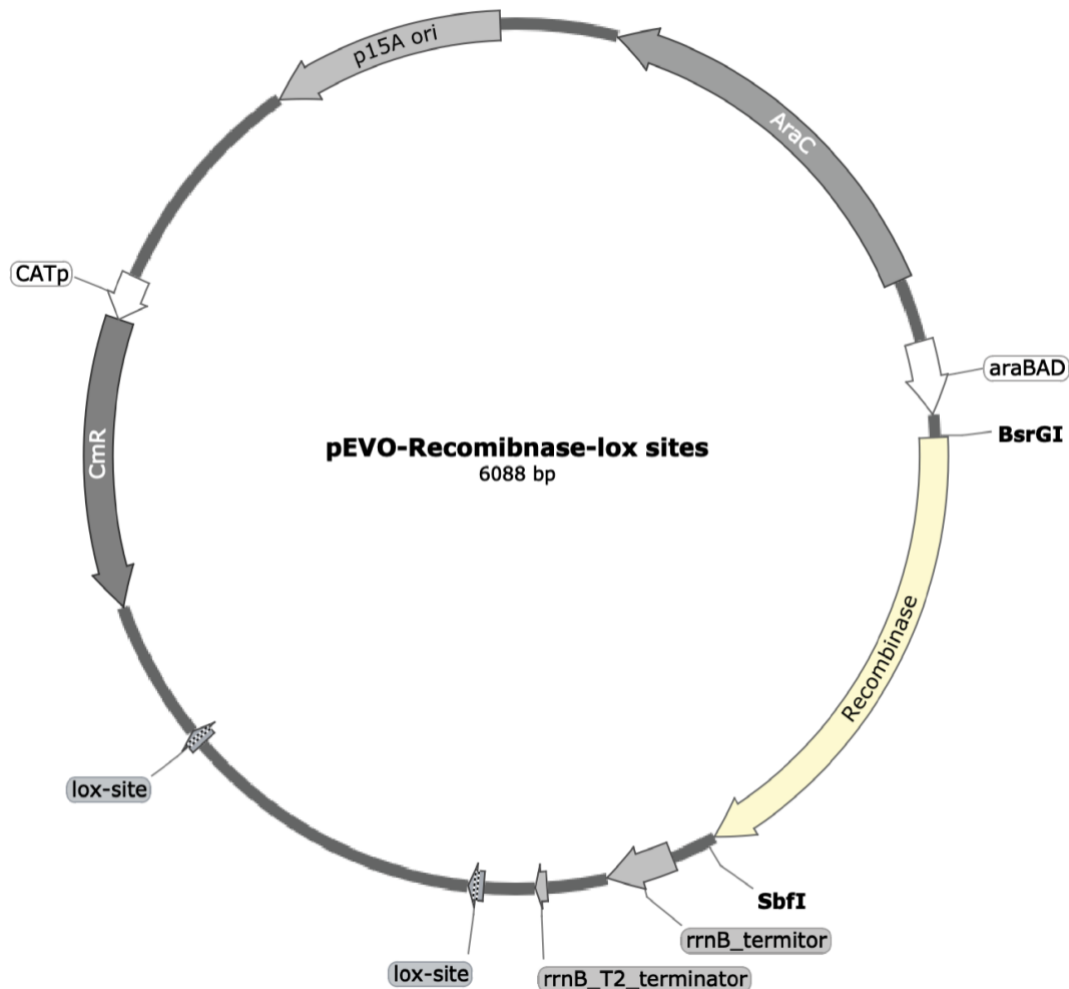
Next day, cells were made electrocompetent as described before and transformed with appropriate R6K donors. After growing the transformations in Cm and Cm+Kan media, plasmid DNA was isolated and a test digest to compare selected and unselected samples was conducted. For the test digest, plasmid DNA was digested with BsrGI and SbfI restriction enzymes in order to distinguish the Vika wt specific band from the band specific for the clones. This was possible since all the tested clones contained SbfI restriction site while the site was destroyed by codon-optimization in Vika wt gene. Thus, the difference of ~200 bp could be visualized on the agarose gels after electrophoresis. Quantification of each clone mediated integration efficiency was calculated relatively to Vika wt. Based on the difference between the ratio of Vika wt and clone specific bands in Cm and Cm+Kan samples, a fold change was calculated representing the relative efficiency of each clone.

PART V  
**APPENDIX**



**Supplementary Figure S1. Clustal Omega amino acid sequence alignment of the seventeen recombinases chosen to be experimentally validated.**

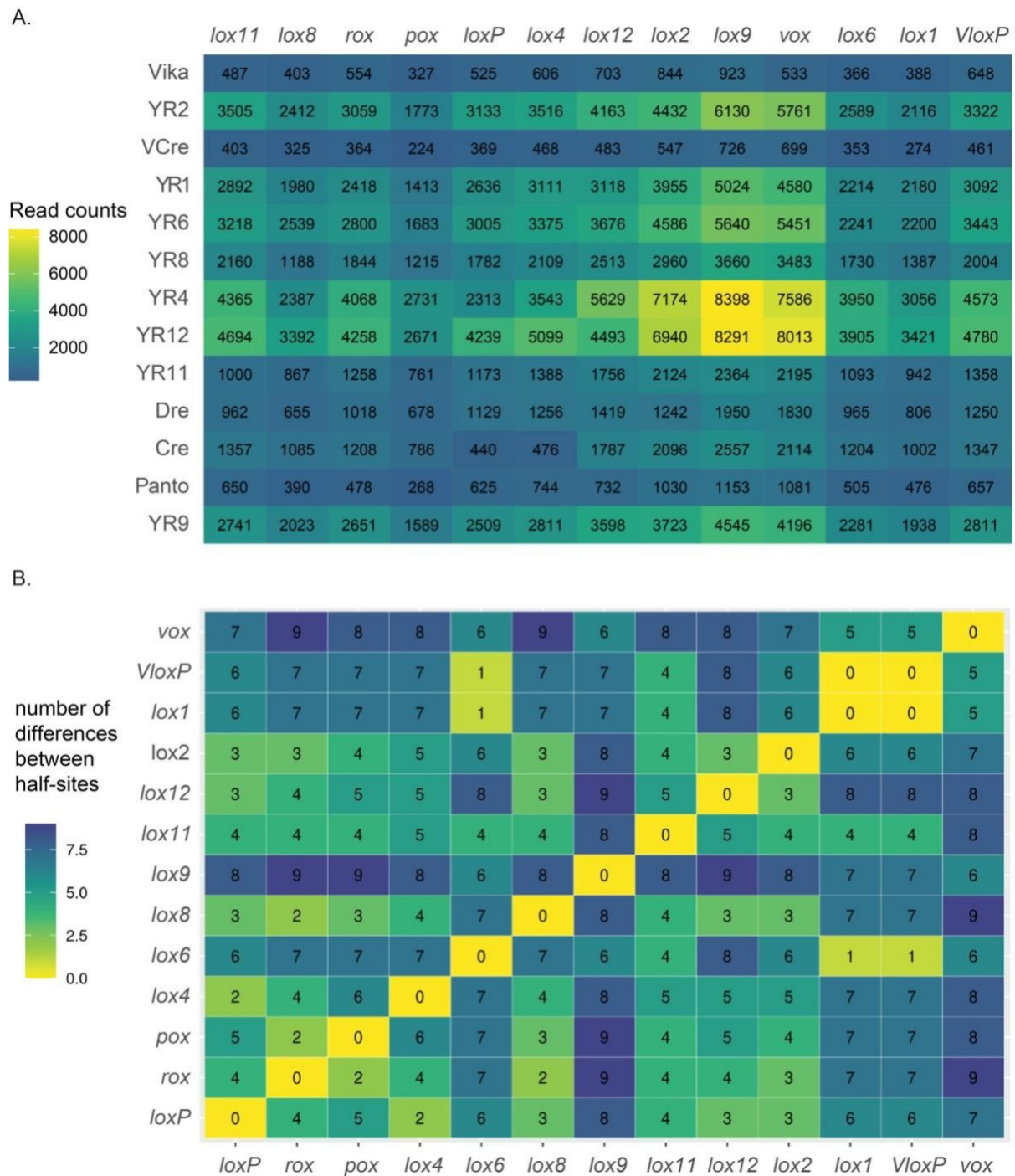
Recombinases are aligned together with Cre as a reference and the residues are shaded based on the conservation score. Secondary structure elements of Cre are depicted. Catalytic residues are highlighted with black arrows, and the nucleophilic tyrosine is marked by a red arrow.



**Supplementary Figure S2. Plasmid map of pEVO recombination reporter.**

Protein coding genes, the origin of replication (oriP15A) and the pBAD promoter are depicted as arrows. The protein coding genes include the chloramphenicol resistance gene (cmR), the arabinose regulatory protein (araC) and the genes encoding for the recombinases of interest. Expression of the recombinase is driven by pBAD promoter upon addition of arabinose. Recombination between two lox sites leads to excision of ~ 700 bp stuffer sequence. BsrGI and SbfI restriction enzymes are used for cloning of the recombinases as well as for linearization of the plasmids for test digest.

SUPPLEMENTARY FIGURES

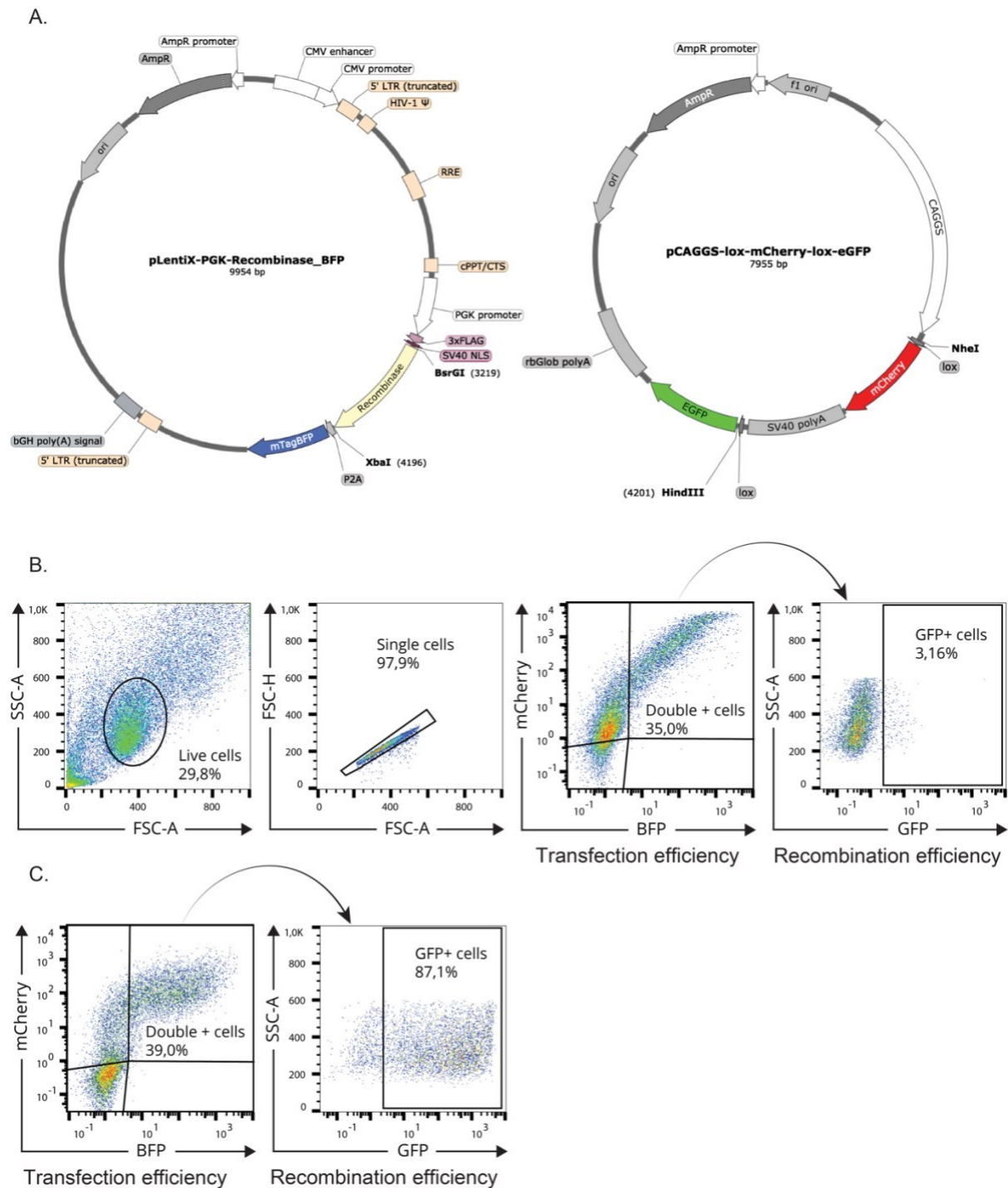


**Supplementary Figure S3. Profiling target site selectivity of Y-SSRs.**

**A.** Heat map depicting number of reads recovered for each recombinase/ target site combination found in the screen after filtering for size and reference identity. Library of new and already known Y-SSRs was cloned into a mixture of pEVO vectors carrying respective target sites. Fragments with the recombinases and target sites were recovered and sequenced by Oxford Nanopore long read sequencing platform to retrieve recombination rates for each combination. **B.** Heat map representing the number of differences between half-sites of all tested target sites in the cross-recombination screen. Identical half-sites are depicted with yellow.



SUPPLEMENTARY FIGURES



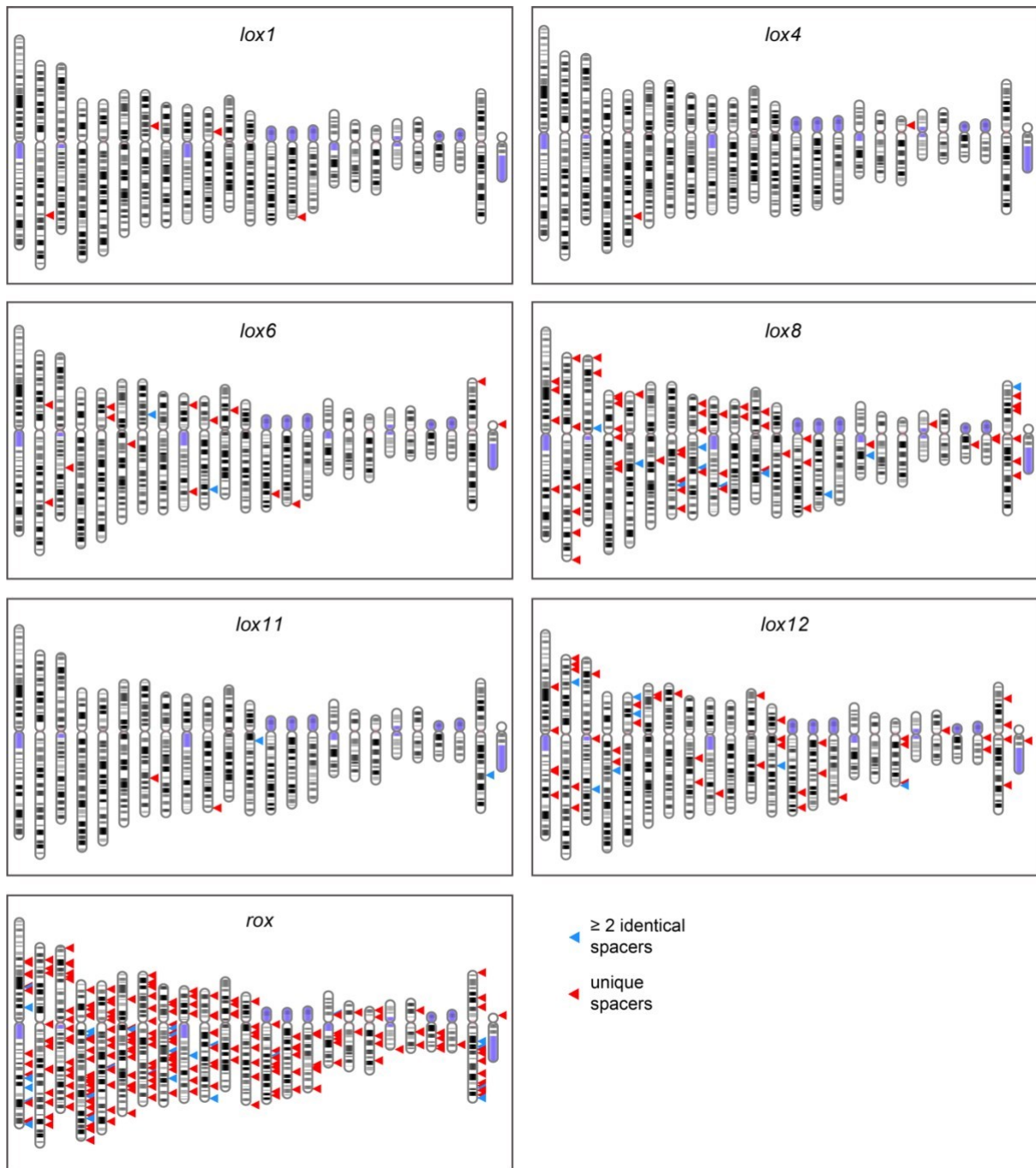
**Supplementary Figure S4. Gating strategy for recombination assay in HEK293T cell line.**

**A.** Plasmid maps of expression lentiviral vector (left) and reporter vector (right). CMV, PGK and CAG promoters for mammalian expression of viral RNA, Recombinase-P2A-BFP cassette and mCherry cassette, respectively, are shown as white arrows. Features for viral production on pLentiX vector are depicted in orange. The lentiviral vector was either transfected into HEK293T cells for recombination assay, or used for virus production and infection as to test the effect of continuous expression of the recombinases. BsrGI and XbaI restriction sites used for cloning of the recombinases are marked on the expression vector while NheI and HindIII sites labeled on pCAG-lox-mCherry-lox-GFP reporter plasmid were used for cloning of all the target sites. Ori – ColE1 origin of replication in bacteria; AmpR – ampicillin resistance gene; bGH poly (A) signal – bovine growth hormone polyadenylation signal; rbGlob-polyA – rabbit  $\beta$ -globin polyadenylation signal. **B.** Single cells were gated out the live population and then gated for mCherry and BFP to assess the transfection efficiency. Double positive population was then examined for GPR expression and gated in FITC channel. Here, the control

## SUPPLEMENTARY FIGURES

sample where empty expression vector was co-transfected with pCAG-lox2 reporter and the basal GFP expression is shown. **C.** Gating example for samples where recombinase was transfected. Again, gating for mCherry and BFP shows transfection efficiency and double positive cells are then gated for GFP expression.

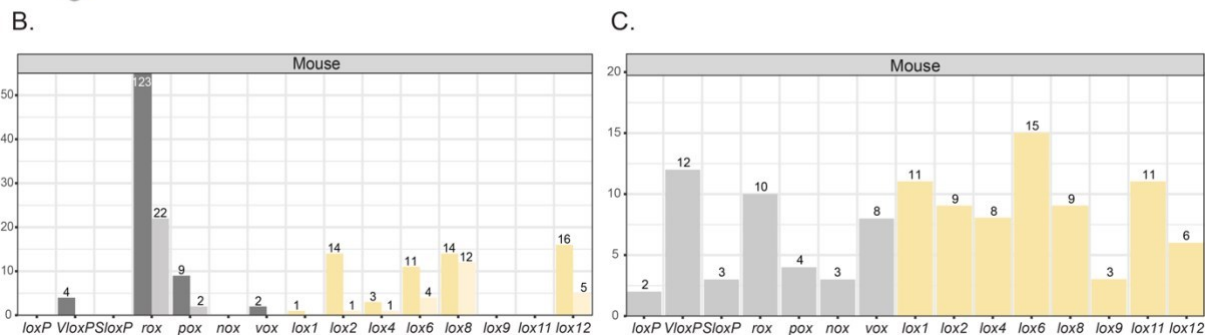
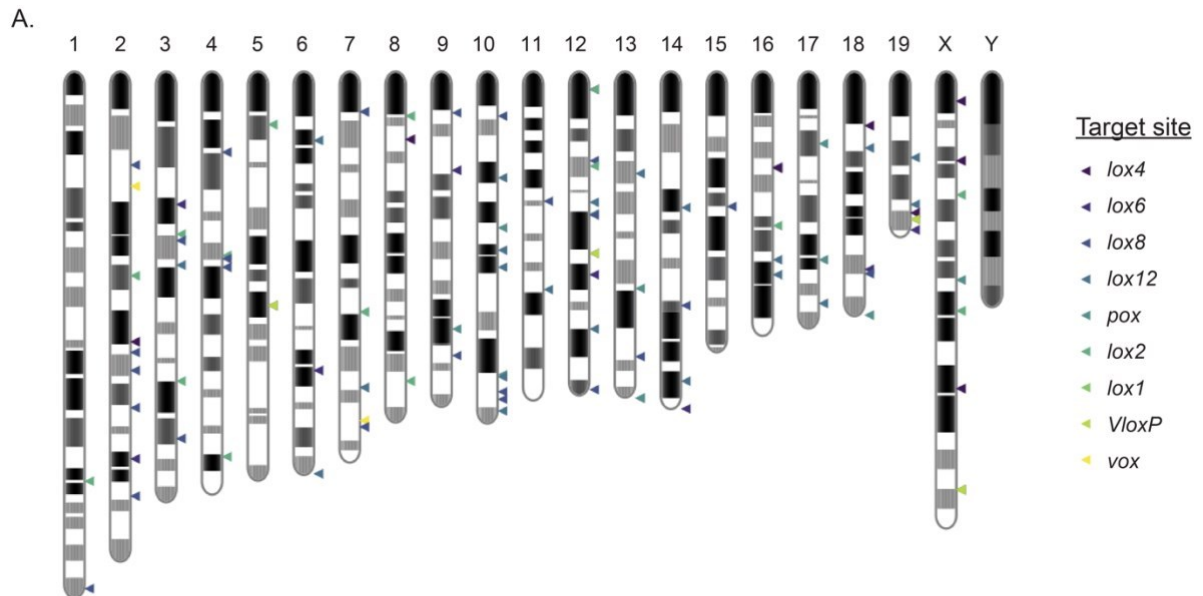
SUPPLEMENTARY FIGURES



**Supplementary Figure S5. Prediction of pseudo-recombination target sites in the human genome.**

Panels of chromosomal distribution of human genomic sequences with high similarity towards remaining newly described Cre-type Y-SSR target sites. Pseudo-rox sites were included as well, considering their high abundance. Positions marked with triangles indicate sequences having no more than two mismatches per half-site of the matching target site sequence. Red triangles refer to a subset of genomic sequences, where each has a unique spacer (potential integration sites). Blue triangles represent a subset of genomic sequences, where at least two share the same spacer (sites for potential inter- or intra-chromosomal rearrangements).

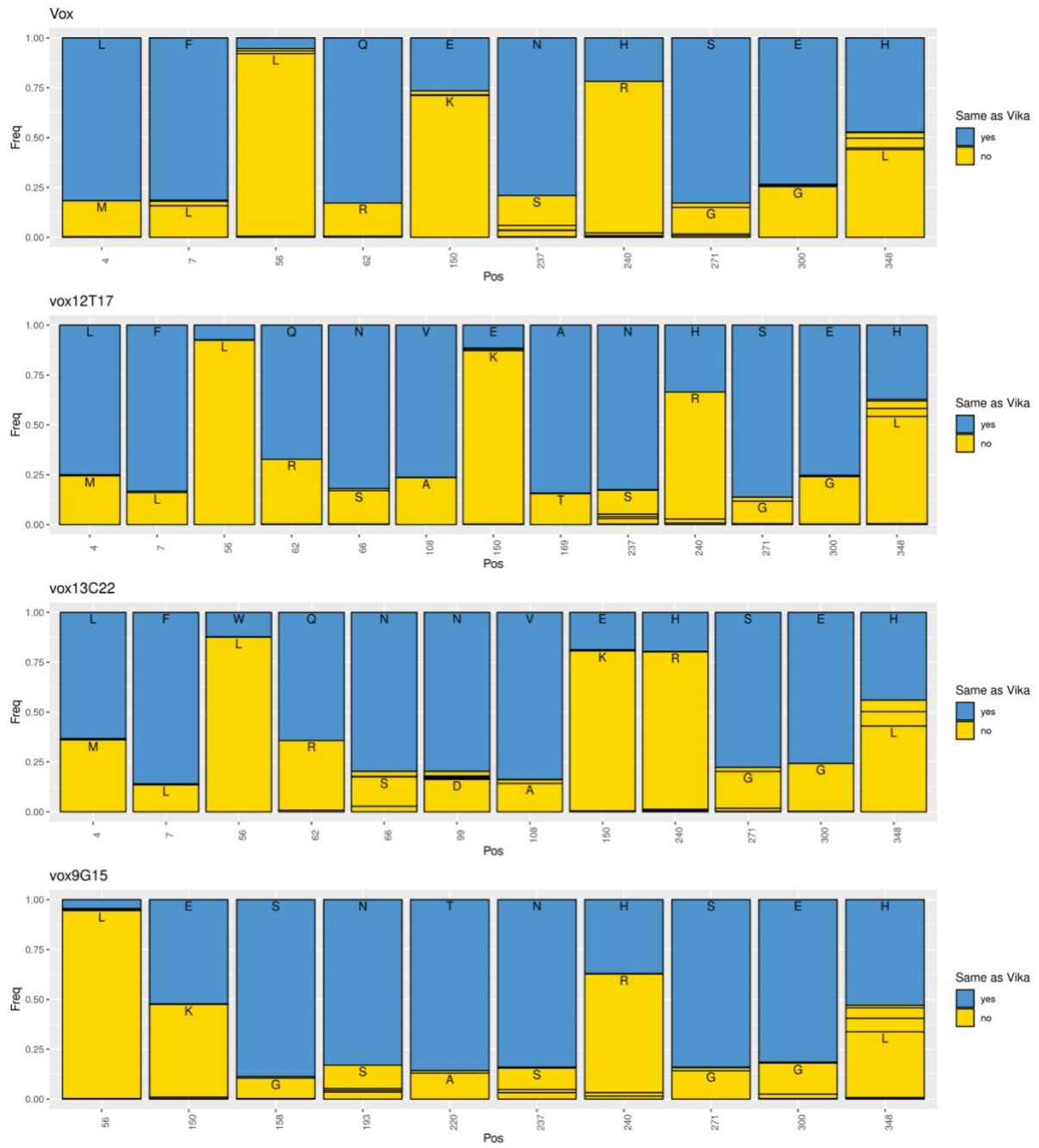
SUPPLEMENTARY FIGURES



**Supplementary Figure S6. Prediction of pseudo-recombination target sites in the mouse genome.**

**A.** Chromosomal distribution of mouse genomic sequences with high similarity towards the described Cre-type Y-SSR target sites. Positions marked with triangles indicate sequences having no more than two mismatches per half-site of the matching target site sequence (shown by different colors), allowing any composition of the 8 bp spacer region. **B.** Counts of mouse genomic sequences with high similarity, up to two mismatches per half-site, towards already described (grey) and new (colored) Y-SSR target sites. Dark-colored bars refer to a subset of genomic sequences, where each has a unique spacer (potential integration sites). Light-colored bars refer to a subset of genomic sequences, where at least two share the same spacer (sites for potential inter- or intra-chromosomal rearrangements). **C.** Graphs depicting the number of unique sequences in the mouse genome with at least one half-site highly similar, with only one mismatch allowed, to target sites of the new (colored) and already described (grey) Y-SSRs.

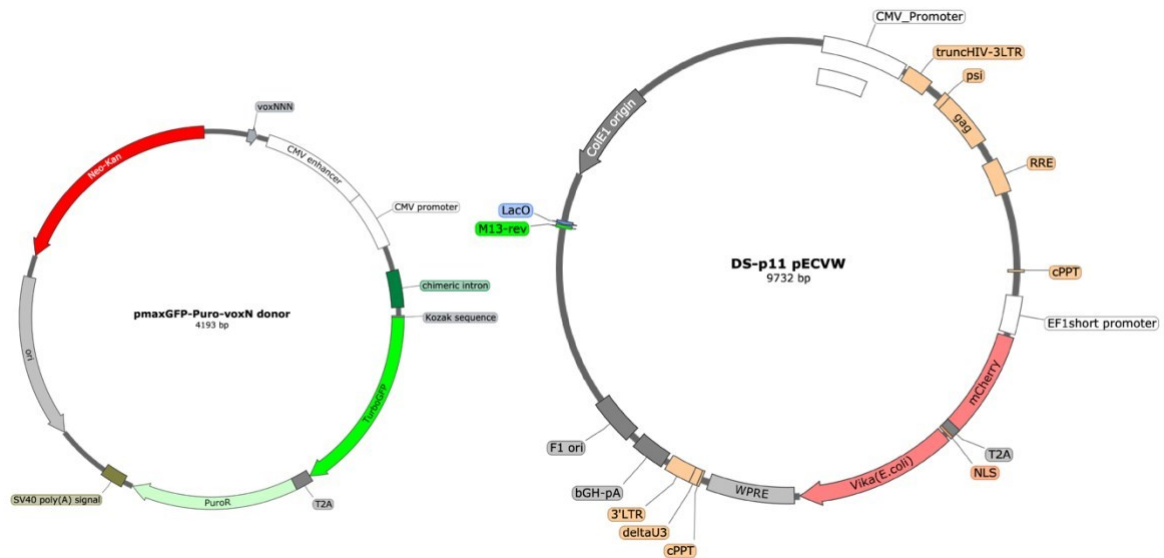
SUPPLEMENTARY FIGURES



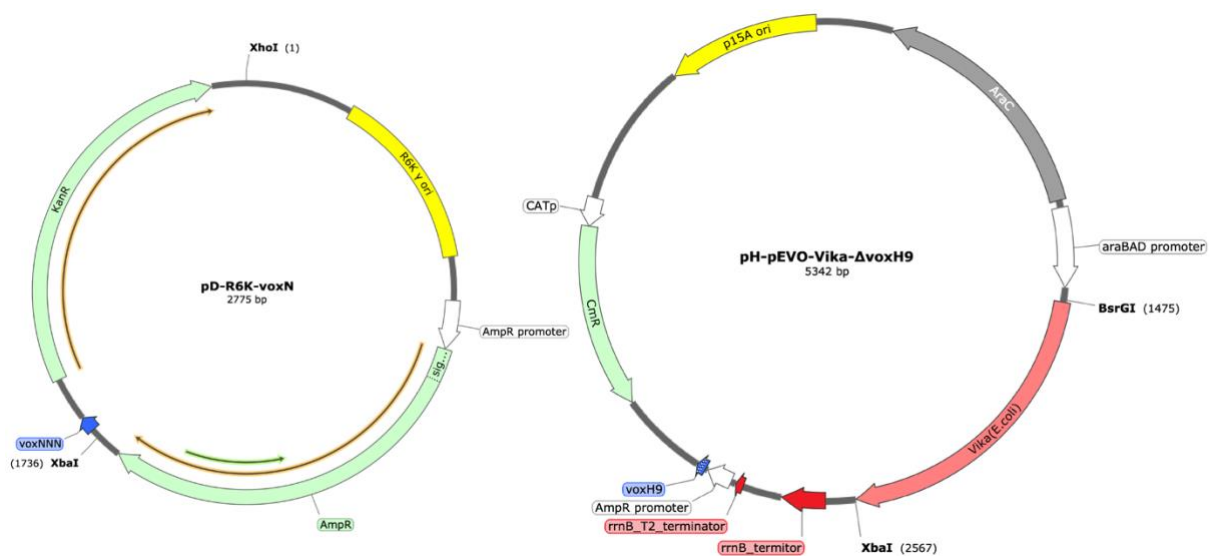
**Supplementary Figure S7. Mutational analysis of Vika libraries.**

Four Vika libraries that were evolved on vox wt, vox12T, vox13C and vox9G target sites, were sent for deep sequencing. The pronounced changes compared to Vika wt are shown.

SUPPLEMENTARY FIGURES

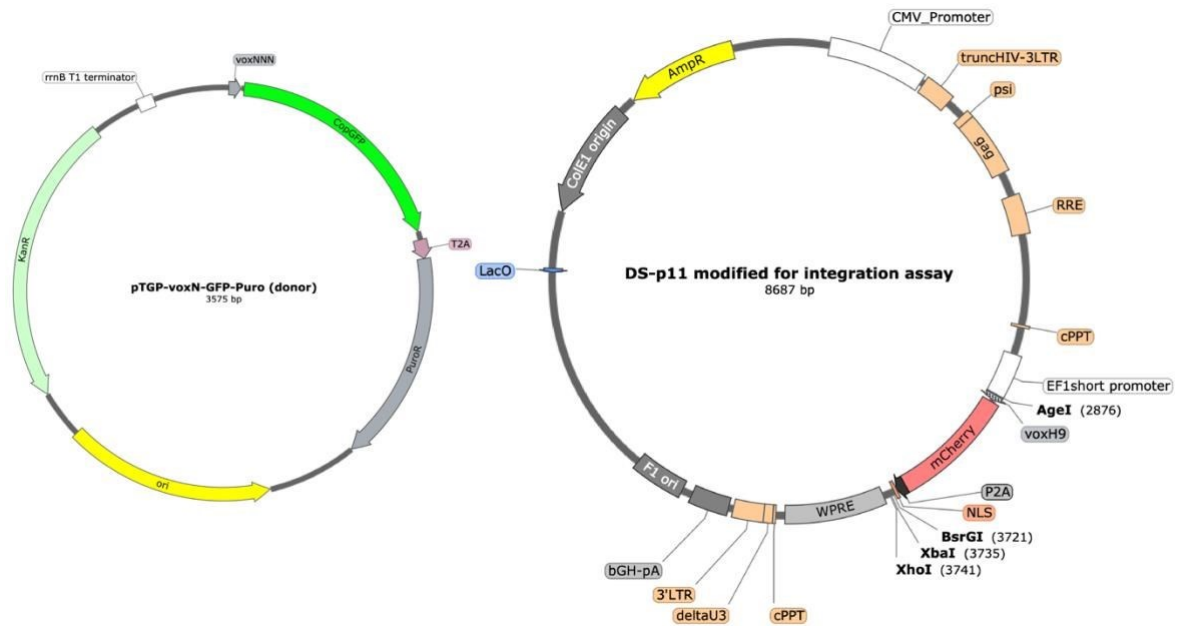


**Supplementary Figure S8. Plasmid maps of donor and expression vectors used to test integration into endogenous *voxH9* site in human genome.**  
 Donor plasmid (left) contains the trap site and GFP-P2A-PuroR cassette expressed by CMV promoter. Expression plasmid (right) harbors EF-1 alpha promoter, which drives expression of mCherry-P2A-Vika cassette.



**Supplementary Figure S9. Plasmid maps of donor and host vectors for integration assay in *E. coli*.**  
 Recombinase is expressed from arabinose promoter of pEVO plasmid (right) that harbors just a single *voxH9* target site which serves as substrate for integrative reaction. Donor plasmid (left) contains the trap site and downstream kanamycin resistance gene. R6K origin cannot secure replication of this plasmid in XL-1 Blue, as it lacks *pir* protein necessary for replication initiation.

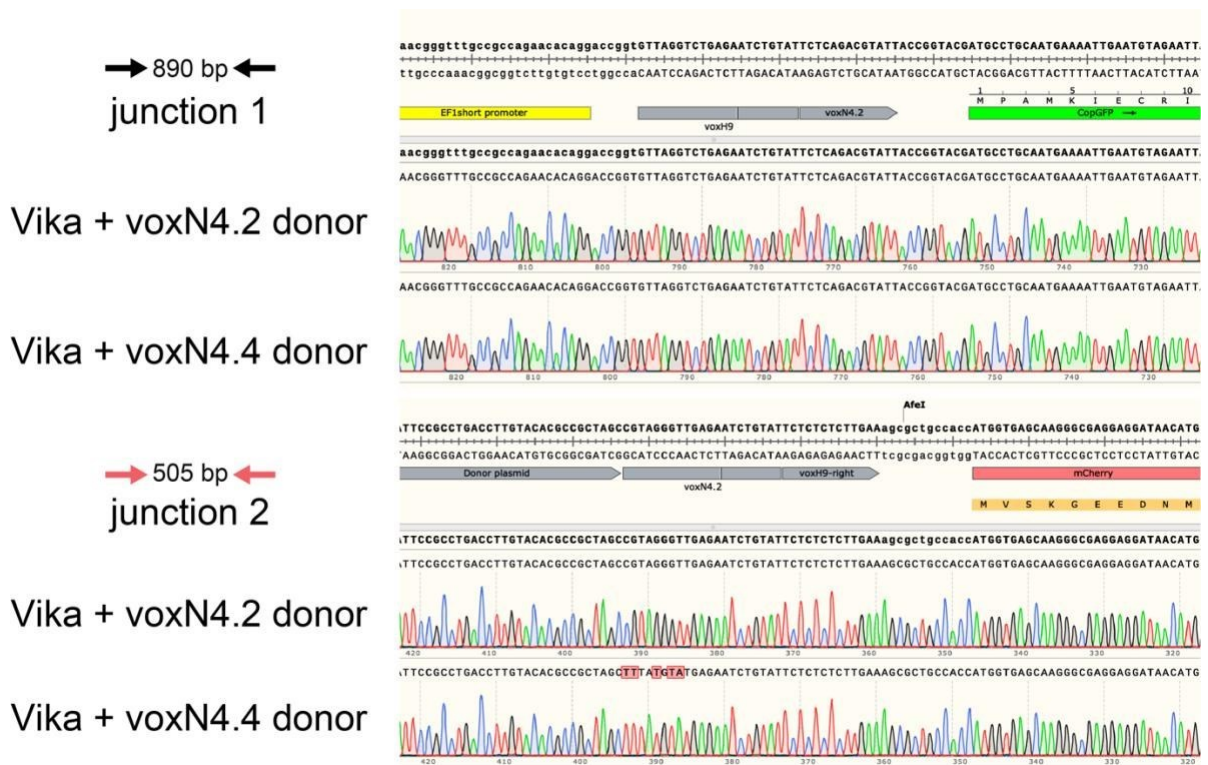
SUPPLEMENTARY FIGURES



**Supplementary Figure S10. Plasmid maps of vectors used for screening of the libraries with the landing pad integration assay.**

Donor (left) harbors trap site and a GFP-P2A-PuroR cassette without any promoter, so that the cassette can't be transiently expressed from the plasmid. To the right is the map of the landing pad vector, constructed to serve as a reporter for successful integration of the donor via voxH9 sites. The landing pad vector contains EF-1 alpha promoter followed by the voxH9 site and mCherry-P2A-cassette. BsrGI and XbaI sites for straightforward cloning of the recombinases in frame with mCherry are found downstream.

SUPPLEMENTARY FIGURES

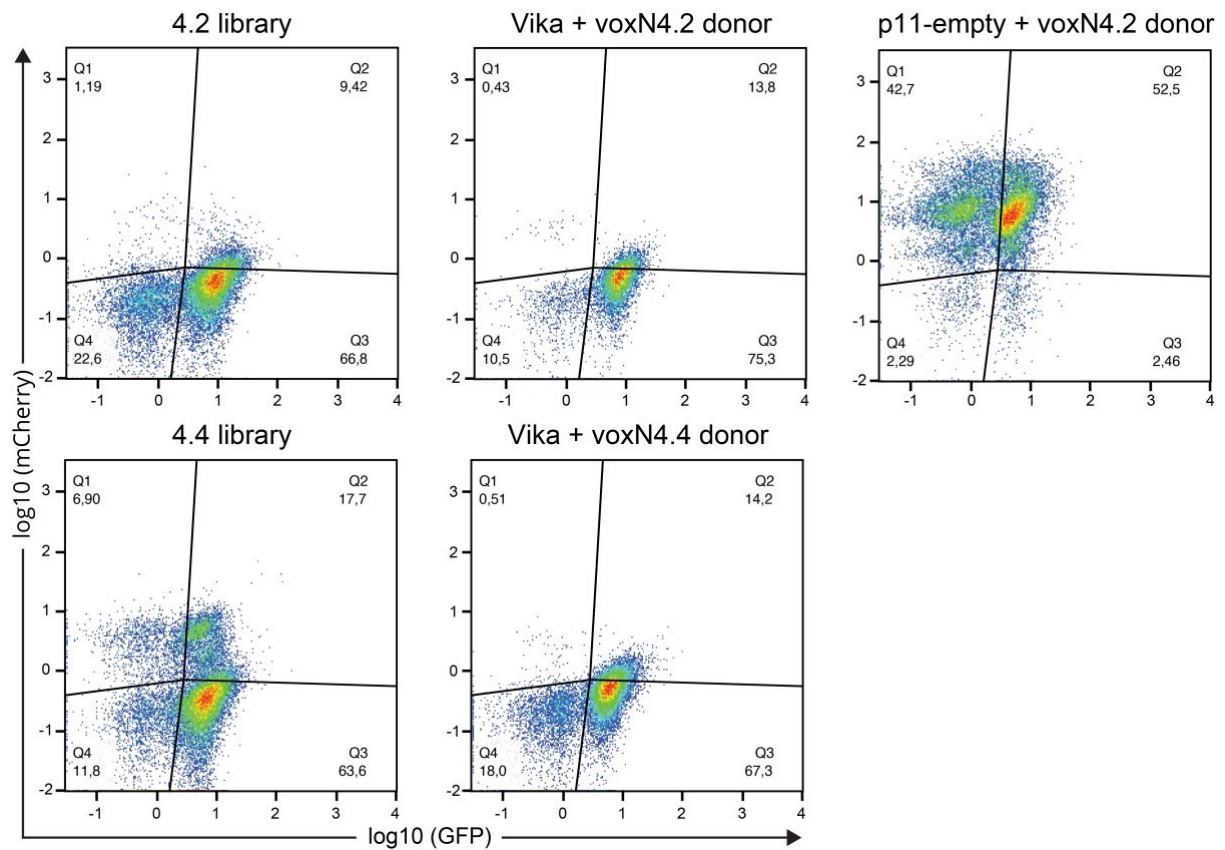


**Supplementary Figure S11. Sanger sequencing of 5' and 3' junctions.**

Amplification of the border region between the donor and landing pad locus can confirm successful integration product. The two PCRs were sent for Sanger sequencing and the results were aligned to the expected recombination product map.

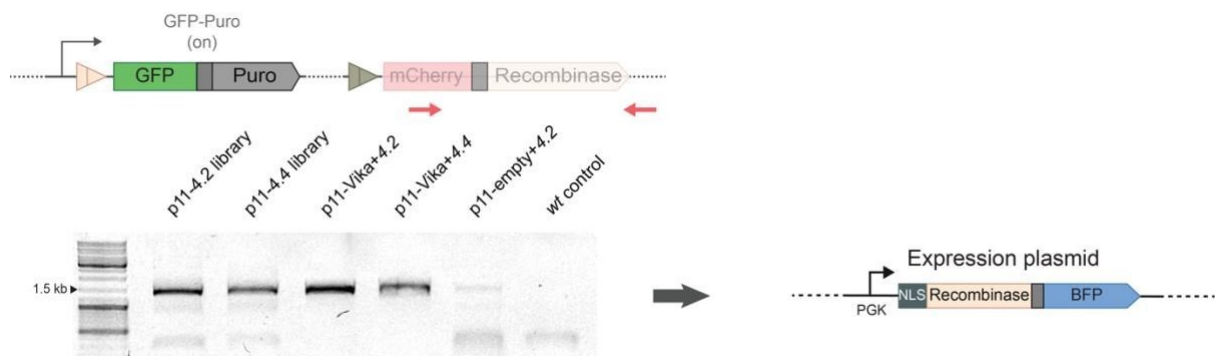


SUPPLEMENTARY FIGURES



**Supplementary Figure S12. Gating strategy for confirming integration into landing pad locus.**

FACS plots of puro selected samples depicting the mCherry and GFP expression. Upon voxN donor transfection and successful integration into the landing pad, GFP and PuroR are expressed, and the Recombinase and mCherry are displaced and knocked out. Only the sample where recombinase was not present has a double positive population (p11-empty + voxN4.2 donor), suggesting that the GFP expression is coming from the random integration of the donor.



**Supplementary Figure S13. PCR and cloning scheme of active clone pulls for testing the individual clones.**

## SUPPLEMENTARY TABLES

**Supplementary Table S1. Plasmids used in this work.**

Name	Description	Reference
pEVO-loxP	Evolution vector harboring loxP target sites, <i>cm<sup>r</sup></i>	Laboratory stock; Based on pEVO vector (Buchholz and Stewart, 2001)
pEVO-Cre	Evolution vector harboring Cre recombinase without target sites, <i>cm<sup>r</sup></i>	Laboratory stock; Based on pEVO vector (Buchholz and Stewart, 2001)
pEVO-vox	Evolution vector harboring vox target sites, <i>cm<sup>r</sup></i>	Laboratory stock; Based on pEVO vector (Buchholz and Stewart, 2001)
pEVO-Vika	Evolution vector harboring Vika recombinase without target sites, <i>cm<sup>r</sup></i>	Laboratory stock; Based on pEVO vector (Buchholz and Stewart, 2001)
pEVO-lox1	Evolution vector harboring lox1 target sites, <i>cm<sup>r</sup></i>	This work
pEVO-YR1	Evolution vector harboring YR1 recombinase, <i>cm<sup>r</sup></i>	This work
pEVO-lox2	Evolution vector harboring lox2 target sites, <i>cm<sup>r</sup></i>	This work
pEVO-YR2	Evolution vector harboring YR2 recombinase, <i>cm<sup>r</sup></i>	This work
pEVO-lox4	Evolution vector harboring lox4 target sites, <i>cm<sup>r</sup></i>	This work
pEVO-YR4	Evolution vector harboring YR4 recombinase, <i>cm<sup>r</sup></i>	This work
pEVO-lox6	Evolution vector harboring lox6 target sites, <i>cm<sup>r</sup></i>	This work
pEVO-YR6	Evolution vector harboring YR6 recombinase, <i>cm<sup>r</sup></i>	This work
pEVO-lox8	Evolution vector harboring lox8 target sites, <i>cm<sup>r</sup></i>	This work
pEVO-YR8	Evolution vector harboring YR8 recombinase, <i>cm<sup>r</sup></i>	This work
pEVO-lox9	Evolution vector harboring lox9 target sites, <i>cm<sup>r</sup></i>	This work
pEVO-YR9	Evolution vector harboring YR9 recombinase, <i>cm<sup>r</sup></i>	This work
pEVO-YR9.1	Evolution vector harboring YR9.1 variant recombinase, <i>cm<sup>r</sup></i>	This work
pEVO-YR9.10	Evolution vector harboring YR9.10 variant recombinase, <i>cm<sup>r</sup></i>	This work
pEVO-YR9.12	Evolution vector harboring YR9.12 variant recombinase, <i>cm<sup>r</sup></i>	This work
pEVO-lox11	Evolution vector harboring lox11 target sites, <i>cm<sup>r</sup></i>	This work
pEVO-YR11	Evolution vector harboring YR11	This work

SUPPLEMENTARY TABLES

	recombinase, <i>cm<sup>r</sup></i>	
pEVO-lox12	Evolution vector harboring lox12 target sites, <i>cm<sup>r</sup></i>	This work
pEVO-YR12	Evolution vector harboring YR12 recombinase, <i>cm<sup>r</sup></i>	This work
pEVO-pox	Evolution vector harboring pox target sites used for Oxford Nanopore sequencing-based cross-reactivity experiment, <i>cm<sup>r</sup></i>	Laboratory stock; Based on pEVO vector (Buchholz and Stewart, 2001)
pEVO-Panto	Evolution vector harboring Panto recombinase, <i>cm<sup>r</sup></i>	Laboratory stock; Based on pEVO vector (Buchholz and Stewart, 2001)
pEVO-rox	Evolution vector harboring rox target sites used for Oxford Nanopore sequencing-based cross-reactivity experiment, <i>cm<sup>r</sup></i>	Laboratory stock; Based on pEVO vector (Buchholz and Stewart, 2001)
pEVO-Dre	Evolution vector harboring Dre recombinase, <i>cm<sup>r</sup></i>	Laboratory stock; Based on pEVO vector (Buchholz and Stewart, 2001)
pEVO-VloxP	Evolution vector harboring VloxP target sites used for Oxford Nanopore sequencing-based cross-reactivity experiment, <i>cm<sup>r</sup></i>	Laboratory stock; Based on pEVO vector (Buchholz and Stewart, 2001)
pEVO-VCre	Evolution vector harboring VCre recombinase, <i>cm<sup>r</sup></i>	Laboratory stock; Based on pEVO vector (Buchholz and Stewart, 2001)
pPGK-NLS-BFP	Lentiviral expression vector, <i>amp<sup>r</sup></i>	Laboratory stock
pPGK-NLS-Cre-P2A-BFP	Lentiviral expression vector harboring Cre recombinase with N-terminal tagged NLS and FLAG for transient or constitutive expression, <i>amp<sup>r</sup></i>	This work
pPGK-NLS-YR1-P2A-BFP	Lentiviral expression vector harboring YR1 recombinase with N-terminal tagged NLS and FLAG for transient or constitutive expression, <i>amp<sup>r</sup></i>	This work
pPGK-NLS-YR2-P2A-BFP	Lentiviral expression vector harboring YR2 recombinase with N-terminal tagged NLS and FLAG for transient or constitutive expression, <i>amp<sup>r</sup></i>	This work
pPGK-NLS-YR4-P2A-BFP	Lentiviral expression vector harboring YR4 recombinase with N-terminal tagged NLS and FLAG for transient or constitutive expression, <i>amp<sup>r</sup></i>	This work
pPGK-NLS-YR6-P2A-BFP	Lentiviral expression vector harboring YR6 recombinase with N-terminal tagged NLS and FLAG for transient or constitutive expression, <i>amp<sup>r</sup></i>	This work
pPGK-NLS-YR8-P2A-BFP	Lentiviral expression vector harboring YR8 recombinase with N-terminal tagged NLS and FLAG for transient or constitutive expression, <i>amp<sup>r</sup></i>	This work

SUPPLEMENTARY TABLES

pPGK-NLS-YR9-P2A-BFP	Lentiviral expression vector harboring YR9 recombinase with N-terminal tagged NLS and FLAG for transient or constitutive expression, <i>amp<sup>r</sup></i>	This work
pPGK-NLS-YR11-P2A-BFP	Lentiviral expression vector harboring YR11 recombinase with N-terminal tagged NLS and FLAG for transient or constitutive expression, <i>amp<sup>r</sup></i>	This work
pPGK-NLS-YR12-P2A-BFP	Lentiviral expression vector harboring YR12 recombinase with N-terminal tagged NLS and FLAG for transient or constitutive expression, <i>amp<sup>r</sup></i>	This work
pPGK-NLS-YR9.1-P2A-BFP	Lentiviral expression vector harboring YR9.1 enhanced variant with N-terminal tagged NLS and FLAG for transient or constitutive expression, <i>amp<sup>r</sup></i>	This work
pPGK-NLS-YR9.10-P2A-BFP	Lentiviral expression vector harboring YR9.10 enhanced variant with N-terminal tagged NLS and FLAG for transient or constitutive expression, <i>amp<sup>r</sup></i>	This work
pPGK-NLS-YR9.12-P2A-BFP	Lentiviral expression vector harboring YR9.12 enhanced variant with N-terminal tagged NLS and FLAG for transient or constitutive expression, <i>amp<sup>r</sup></i>	This work
pPGK-NLS-Vika-P2A-BFP	Lentiviral expression vector harboring Vika recombinase with N-terminal tagged NLS and FLAG for transient or constitutive expression, <i>amp<sup>r</sup></i>	This work
pPGK-NLS-A4-P2A-BFP	Lentiviral expression vector harboring A4 Vika variant with N-terminal tagged NLS and FLAG for transient or constitutive expression, <i>amp<sup>r</sup></i>	This work
pPGK-NLS-A8-P2A-BFP	Lentiviral expression vector harboring A8 Vika variant with N-terminal tagged NLS and FLAG for transient or constitutive expression, <i>amp<sup>r</sup></i>	This work
pPGK-NLS-B2-P2A-BFP	Lentiviral expression vector harboring B2 Vika variant with N-terminal tagged NLS and FLAG for transient or constitutive expression, <i>amp<sup>r</sup></i>	This work
pPGK-NLS-D8-P2A-BFP	Lentiviral expression vector harboring D8 Vika variant with N-terminal tagged NLS and FLAG for transient or constitutive expression, <i>amp<sup>r</sup></i>	This work
pPGK-NLS-G6-P2A-BFP	Lentiviral expression vector harboring G6 Vika variant with N-terminal tagged NLS and FLAG for transient or constitutive expression, <i>amp<sup>r</sup></i>	This work
pPGK-NLS-Clu1-	Lentiviral expression vector harboring	This work

SUPPLEMENTARY TABLES

P2A-BFP	Cluster 1 Vika variant with N-terminal tagged NLS and FLAG for transient or constitutive expression, <i>amp<sup>r</sup></i>	
pPGK-NLS-Clu24-P2A-BFP	Lentiviral expression vector harboring Cluster 24 Vika variant with N-terminal tagged NLS and FLAG for transient or constitutive expression, <i>amp<sup>r</sup></i>	This work
pPGK-NLS-Clu27-P2A-BFP	Lentiviral expression vector harboring Cluster 27 Vika variant with N-terminal tagged NLS and FLAG for transient or constitutive expression, <i>amp<sup>r</sup></i>	This work
pPGK-NLS-Clu34-P2A-BFP	Lentiviral expression vector harboring Cluster 34 Vika variant with N-terminal tagged NLS and FLAG for transient or constitutive expression, <i>amp<sup>r</sup></i>	This work
pCAGGS-loxP-mCherry-loxP-EGFP	Reporter vector harboring loxP target sites used for plasmid excision assay, <i>amp<sup>r</sup></i>	Laboratory stock; based on pCAGGS-IRES-puro (Karpinski et al., 2016)
pCAGGS-vox-mCherry-vox-EGFP	Reporter vector harboring vox target sites used for plasmid excision assay, <i>amp<sup>r</sup></i>	Laboratory stock; based on pCAGGS-IRES-puro (Karpinski et al., 2016)
pCAGGS-lox1-mCherry-lox1-EGFP	Reporter vector harboring lox1 target sites used for plasmid excision assay, <i>amp<sup>r</sup></i>	Laboratory stock; based on pCAGGS-IRES-puro (Karpinski et al., 2016)
pCAGGS-lox2-mCherry-lox2-EGFP	Reporter vector harboring lox2 target sites used for plasmid excision assay, <i>amp<sup>r</sup></i>	Laboratory stock; based on pCAGGS-IRES-puro (Karpinski et al., 2016)
pCAGGS-lox4-mCherry-lox4-EGFP	Reporter vector harboring lox4 target sites used for plasmid excision assay, <i>amp<sup>r</sup></i>	Laboratory stock; based on pCAGGS-IRES-puro (Karpinski et al., 2016)
pCAGGS-lox6-mCherry-lox6-EGFP	Reporter vector harboring lox6 target sites used for plasmid excision assay, <i>amp<sup>r</sup></i>	Laboratory stock; based on pCAGGS-IRES-puro (Karpinski et al., 2016)
pCAGGS-lox8-mCherry-lox8-EGFP	Reporter vector harboring lox8 target sites used for plasmid excision assay, <i>amp<sup>r</sup></i>	Laboratory stock; based on pCAGGS-IRES-puro (Karpinski et al., 2016)
pCAGGS-lox9-mCherry-lox9-EGFP	Reporter vector harboring lox9 target sites used for plasmid excision assay, <i>amp<sup>r</sup></i>	Laboratory stock; based on pCAGGS-IRES-puro (Karpinski et al., 2016)
pCAGGS-lox11-mCherry-lox11-EGFP	Reporter vector harboring lox11 target sites used for plasmid excision assay, <i>amp<sup>r</sup></i>	Laboratory stock; based on pCAGGS-IRES-puro (Karpinski et al., 2016)
pCAGGS-lox12-mCherry-lox12-EGFP	Reporter vector harboring lox12 target sites used for plasmid excision assay, <i>amp<sup>r</sup></i>	Laboratory stock; based on pCAGGS-IRES-puro (Karpinski et al., 2016)
pEVO-voxH9	Evolution vector harboring two voxH9 target sites for excision assay in bacteria, <i>cm<sup>r</sup></i>	This work
pEVO-ΔvoxH9	Evolution vector harboring single voxH9 target site used for integration assays in bacteria, <i>cm<sup>r</sup></i>	This work
pR6K-voxN4.2	Donor vector with R6K <i>ori</i> harboring	Laboratory stock; based on

SUPPLEMENTARY TABLES

	voxN4.2 trap target site used for integration assays in bacteria, kan <sup>r</sup>	pR6K-vox (Karimova et al., 2012)
pR6K-voxN4.4	Donor vector with R6K <i>ori</i> harboring voxN4.4 trap target site used for integration assays in bacteria, kan <sup>r</sup>	Laboratory stock; based on pR6K-vox (Karimova et al., 2012)
pEVO-Vika_WT- $\Delta$ voxH9	Evolution vector harboring Vika wt recombinase and single voxH9 target site used for integration assays in bacteria, <i>cm<sup>r</sup></i>	This work
pEVO-4.2_Clone1- $\Delta$ voxH9	Evolution vector harboring randomly picked Clone1 from 4.2 library and single voxH9 target site used for integration assays in bacteria, <i>cm<sup>r</sup></i>	This work
pEVO-4.2_Clone2- $\Delta$ voxH9	Evolution vector harboring randomly picked Clone2 from 4.2 library and single voxH9 target site used for integration assays in bacteria, <i>cm<sup>r</sup></i>	This work
pEVO-4.2_Clone3- $\Delta$ voxH9	Evolution vector harboring randomly picked Clone3 from 4.2 library and single voxH9 target site used for integration assays in bacteria, <i>cm<sup>r</sup></i>	This work
pEVO-4.4_Clone1- $\Delta$ voxH9	Evolution vector harboring randomly picked Clone1 from 4.4 library and single voxH9 target site used for integration assays in bacteria, <i>cm<sup>r</sup></i>	This work
pEVO-4.4_Clone2- $\Delta$ voxH9	Evolution vector harboring randomly picked Clone2 from 4.4 library and single voxH9 target site used for integration assays in bacteria, <i>cm<sup>r</sup></i>	This work
pEVO-4.4_Clone3- $\Delta$ voxH9	Evolution vector harboring randomly picked Clone3 from 4.4 library and single voxH9 target site used for integration assays in bacteria, <i>cm<sup>r</sup></i>	This work
pEVO-Clu1- $\Delta$ voxH9	Evolution vector harboring Cluster 1 from 4.4 library and single voxH9 target site used for integration assays in bacteria, <i>cm<sup>r</sup></i>	This work
pEVO-Clu24- $\Delta$ voxH9	Evolution vector harboring Cluster 24 from 4.2 library and single voxH9 target site used for integration assays in bacteria, <i>cm<sup>r</sup></i>	This work
pEVO-Clu27- $\Delta$ voxH9	Evolution vector harboring Cluster 27 from 4.2 library and single voxH9 target site used for integration assays in bacteria, <i>cm<sup>r</sup></i>	This work
pEVO-Clu34- $\Delta$ voxH9	Evolution vector harboring Cluster 34 from 4.2 library and single voxH9 target site used for integration assays in bacteria, <i>cm<sup>r</sup></i>	This work
pEVO-A4- $\Delta$ voxH9	Evolution vector harboring A4 Vika variant and single voxH9 target site used for integration assays in bacteria,	This work

SUPPLEMENTARY TABLES

	<i>cm<sup>r</sup></i>	
pEVO-B2-ΔvoxH9	Evolution vector harboring B2 Vika variant and single voxH9 target site used for integration assays in bacteria, <i>cm<sup>r</sup></i>	This work
pEVO-D8-ΔvoxH9	Evolution vector harboring D8 Vika variant and single voxH9 target site used for integration assays in bacteria, <i>cm<sup>r</sup></i>	This work
pEF-1α-mCherry-P2A-NLS-Vika	p11-pECVW lentiviral vector for expression of recombinases, <i>amp<sup>r</sup></i>	This work; Based on p11-pECCW laboratory stock
pEF-1α-voxH9-mCherry-P2A-NLS	p11-pECVW lentiviral vector for generation of landing pad reporter cell lines, <i>amp<sup>r</sup></i>	This work; Based on p11-pECCW laboratory stock
pEF-1α-voxH9-mCherry-P2A-NLS-Vika	p11-pECVW lentiviral vector for generation of landing pad reporter cell lines and simultaneous expression of Vika recombinases, <i>amp<sup>r</sup></i>	This work; Based on p11-pECCW laboratory stock
pEF-1α-voxH9-mCherry-P2A-NLS-4.2 Vika variant library	p11-pECVW lentiviral vector for generation of landing pad reporter cell lines and simultaneous expression of 4.2 Vika variant recombinase library, <i>amp<sup>r</sup></i>	This work; Based on p11-pECCW laboratory stock
pEF-1α-voxH9-mCherry-P2A-NLS-4.4 Vika variant library	p11-pECVW lentiviral vector for generation of landing pad reporter cell lines and simultaneous expression of 4.4 Vika variant recombinase library, <i>amp<sup>r</sup></i>	This work; Based on p11-pECCW laboratory stock
pmaxTurboGFP-P2A-PuroR-voxN4.2	Donor vector harboring CMV-turboGFP-P2A-PuroR cassette and voxN4.2 target site used for integration into human endogenous voxH9 locus, <i>kan<sup>r</sup></i>	Laboratory stock; Based on pmaxGFP (Lonza)
pmaxTurboGFP-P2A-PuroR-voxN4.4	Donor vector harboring CMV-turboGFP-P2A-PuroR cassette and voxN4.4 target site used for integration into human endogenous voxH9 locus, <i>kan<sup>r</sup></i>	Laboratory stock; Based on pmaxGFP (Lonza)
pTwist-copGFP-P2A-PuroR-voxN4.2	Donor vector harboring voxN4.2-copGFP-P2A-PuroR cassette used for integration into landing pad reporter, <i>kan<sup>r</sup></i>	Ordered from Twist Bioscience
pTwist-copGFP-P2A-PuroR-voxN4.4	Donor vector harboring voxN4.4-copGFP-P2A-PuroR cassette used for integration into landing pad reporter, <i>kan<sup>r</sup></i>	Ordered from Twist Bioscience
psPAX2 (gag/pol)	Packaging plasmid for the production of lentiviral particles	Addgene plasmid # 12260; <a href="http://n2t.net/addgene:12260">http://n2t.net/addgene:12260</a> ; RRID:Addgene_12260
pMD2.G (Env)	Helper plasmid expressing VSV-G envelope for the production of lentiviral particles	Addgene plasmid # 12259; <a href="http://n2t.net/addgene:12259">http://n2t.net/addgene:12259</a> ; RRID:Addgene_12259

SUPPLEMENTARY TABLES

*amp<sup>r</sup>*, ampicillin resistance; *cm<sup>r</sup>*, chloramphenicol resistance; *kan<sup>r</sup>*, kanamycin resistance; *puro<sup>r</sup>*, puromycin resistance  
**Supplementary Table S2. Primers used in this work.**

oMJ	Primer name	Nucleotide sequence 5' → 3'
	<b>Construction of pEVO vectors</b>	
1	pEVO-lox1 fw	TGAAAAGGAAGAGTATGAGATCTTCAATTTCTGAGAAGTGAATTCTCAGAAATTGAaagcttgcacg
2	pEVO-lox1 rev	GGCGACACCGAAATGTTGAGATCTTCAATTTCTGAGAATTACACTTCTCAGAAATTGAtcgaactgtacc
3	pEVO-lox2 fw	TGAAAAGGAAGAGTATGAGATCTTAACTTGTGATAATTCGCGTTATCTCAAGTTATtcgaactgtacc
4	pEVO-lox2 rev	GGCGACACCGAAATGTTGAGATCTTAACTTGTGATAATTCGCGTTATCTCAAGTTATtcgaactgtacc
5	pEVO-lox3 fw	TGAAAAGGAAGAGTATGAGATCTTAACTTGTGATAATTCGCGTTATCTCAAGTTATtcgaactgtacc
6	pEVO-lox3 rev	GGCGACACCGAAATGTTGAGATCTTAGGATGCTTATGACCGGCCATCGAGAACATAGTtcgaactgtacc
7	pEVO-lox4 fw	TGAAAAGGAAGAGTATGAGATCTTAACTTGTGATAATTCGCGTTATCTCAAGTTATtcgaactgtacc
8	pEVO-lox4 rev	GGCGACACCGAAATGTTGAGATCTTAACTTGTGATAATTCGCGTTATCTCAAGTTATtcgaactgtacc
9	pEVO-lox5 fw	TGAAAAGGAAGAGTATGAGATCTTAACTTGTGATAATTCGCGTTATCTCAAGTTATtcgaactgtacc
10	pEVO-lox5 rev	GGCGACACCGAAATGTTGAGATCTTAACTTGTGATAATTCGCGTTATCTCAAGTTATtcgaactgtacc
11	pEVO-lox6 fw	TGAAAAGGAAGAGTATGAGATCTTAACTTGTGATAATTCGCGTTATCTCAAGTTATtcgaactgtacc
12	pEVO-lox6 rev	GGCGACACCGAAATGTTGAGATCTTAACTTGTGATAATTCGCGTTATCTCAAGTTATtcgaactgtacc
13	pEVO-lox7 fw	TGAAAAGGAAGAGTATGAGATCTTAACTTGTGATAATTCGCGTTATCTCAAGTTATtcgaactgtacc
14	pEVO-lox7 rev	GGCGACACCGAAATGTTGAGATCTTAACTTGTGATAATTCGCGTTATCTCAAGTTATtcgaactgtacc
15	pEVO-lox8 fw	TGAAAAGGAAGAGTATGAGATCTTAACTTGTGATAATTCGCGTTATCTCAAGTTATtcgaactgtacc
16	pEVO-lox8 rev	GGCGACACCGAAATGTTGAGATCTTAACTTGTGATAATTCGCGTTATCTCAAGTTATtcgaactgtacc
17	pEVO-lox9 fw	TGAAAAGGAAGAGTATGAGATCTTAACTTGTGATAATTCGCGTTATCTCAAGTTATtcgaactgtacc
18	pEVO-lox9 rev	GGCGACACCGAAATGTTGAGATCTTAACTTGTGATAATTCGCGTTATCTCAAGTTATtcgaactgtacc
19	pEVO-lox10 fw	TGAAAAGGAAGAGTATGAGATCTTAACTTGTGATAATTCGCGTTATCTCAAGTTATtcgaactgtacc
20	pEVO-lox10 rev	GGCGACACCGAAATGTTGAGATCTTAACTTGTGATAATTCGCGTTATCTCAAGTTATtcgaactgtacc
21	pEVO-lox11 fw	TGAAAAGGAAGAGTATGAGATCTTAACTTGTGATAATTCGCGTTATCTCAAGTTATtcgaactgtacc
22	pEVO-lox11 rev	GGCGACACCGAAATGTTGAGATCTTAACTTGTGATAATTCGCGTTATCTCAAGTTATtcgaactgtacc
23	pEVO-lox12 fw	TGAAAAGGAAGAGTATGAGATCTTAACTTGTGATAATTCGCGTTATCTCAAGTTATtcgaactgtacc
24	pEVO-lox12 rev	GGCGACACCGAAATGTTGAGATCTTAACTTGTGATAATTCGCGTTATCTCAAGTTATtcgaactgtacc
25	pEVO-lox13 fw	TGAAAAGGAAGAGTATGAGATCTTAACTTGTGATAATTCGCGTTATCTCAAGTTATtcgaactgtacc
26	pEVO-lox13 rev	GGCGACACCGAAATGTTGAGATCTTAACTTGTGATAATTCGCGTTATCTCAAGTTATtcgaactgtacc
27	pEVO-lox14 fw	TGAAAAGGAAGAGTATGAGATCTTAACTTGTGATAATTCGCGTTATCTCAAGTTATtcgaactgtacc
28	pEVO-lox14 rev	GGCGACACCGAAATGTTGAGATCTTAACTTGTGATAATTCGCGTTATCTCAAGTTATtcgaactgtacc
29	pEVO-lox15 fw	TGAAAAGGAAGAGTATGAGATCTTAACTTGTGATAATTCGCGTTATCTCAAGTTATtcgaactgtacc
30	pEVO-lox15 rev	GGCGACACCGAAATGTTGAGATCTTAACTTGTGATAATTCGCGTTATCTCAAGTTATtcgaactgtacc
31	pEVO-lox16 fw	TGAAAAGGAAGAGTATGAGATCTTAACTTGTGATAATTCGCGTTATCTCAAGTTATtcgaactgtacc
32	pEVO-lox16 rev	GGCGACACCGAAATGTTGAGATCTTAACTTGTGATAATTCGCGTTATCTCAAGTTATtcgaactgtacc
33	pEVO-lox17 fw	TGAAAAGGAAGAGTATGAGATCTTAACTTGTGATAATTCGCGTTATCTCAAGTTATtcgaactgtacc
APPENDIX		132



SUPPLEMENTARY TABLES

34	pEVO-lox17 rev	GGCGACACGGAAATGTTGAGATCTCCGGCTTCTTCGATCAGCACGTCGAAGTTGCGGGtcgaactgtacc ggttgtagtga
35	pEVO-vox fw	TGAAAAGGAAGAGTATGAGATCTAATAGGTCTGAGAcgccatTCTCAGACGTATTAagcttgcagtc ctgcagatcgag
36	pEVO-vox rev	GGCGACACGGAAATGTTGAGATCTAATACGTCTGAGAAATGGCGTCTCAGACCTATTTcgaactgtacc ggttgtagtga
37	pEVO-YR1-N-FLAG fw	GAATTCGAGCTCATGgactacaaggatcatgatattgattacaaagacgatgacgataagGGTGGTAGCG GTGGTTCacTGTACATGATTGAAAACAGCTGAGCCTGC
38	pEVO-YR2-N-FLAG fw	GAATTCGAGCTCATGgactacaaggatcatgatattgattacaaagacgatgacgataagGGTGGTAGCG GTGGTTCacTGTACATGAACAACGAAATTATTCATAGCACCAGC
39	pEVO-YR4-N-FLAG fw	GAATTCGAGCTCATGgactacaaggatcatgatattgattacaaagacgatgacgataagGGTGGTAGCG GTGGTTCacTGTACATGAGTGAACCTCTGCCAC
40	pEVO-YR6-N-FLAG fw	GAATTCGAGCTCATGgactacaaggatcatgatattgattacaaagacgatgacgataagGGTGGTAGCG GTGGTTCacTGTACATGATCGAGAATCAACTTCTCTCTGG
41	pEVO-YR8-N-FLAG fw	GAATTCGAGCTCATGgactacaaggatcatgatattgattacaaagacgatgacgataagGGTGGTAGCG GTGGTTCacTGTACATGGGTAAGCTTCCCTTACGAACC
42	pEVO-YR9-N-FLAG fw	GAATTCGAGCTCATGgactacaaggatcatgatattgattacaaagacgatgacgataagGGTGGTAGCG GTGGTTCacTGTACATGTTGGCGGTTTACATGAGGAC
43	pEVO-YR11-N-FLAG fw	GAATTCGAGCTCATGgactacaaggatcatgatattgattacaaagacgatgacgataagGGTGGTAGCG GTGGTTCacTGTACATGGTTGGTGGCATGAGTTTCGTTCG
44	pEVO-YR12-N-FLAG fw	GAATTCGAGCTCATGgactacaaggatcatgatattgattacaaagacgatgacgataagGGTGGTAGCG GTGGTTCacTGTACATGACCGAGATGATTGTCGCAAACC
45	pEVO-Vika-N-FLAG fw	GAATTCGAGCTCATGgactacaaggatcatgatattgattacaaagacgatgacgataagGGTGGTAGCG GTGGTTCacTGTACATGACCGATCTGACCCCGTTTCC
46	pEVO-Cre-N-FLAG fw	GAATTCGAGCTCATGgactacaaggatcatgatattgattacaaagacgatgacgataagGGTGGTAGCG GTGGTTCacTGTACATGTCCAATTTACTGACCGTACACC
47	TWIST adapter fw	GAAGTGCCATTCGCGCTGACCT
48	TWIST adapter rev	CACTGAGCCTCCACCTAGCCT
49	ΔvoxH9 oligo fw	GATCTGTTAGGTCTGAGAATCTGTATTCTCTCTCTTGA
50	ΔvoxH9 oligo rev	GATCTTTCAAGAGAGAGAATACAGATTCTCAGACCTAACA
51	pEVO-voxH9 fw	TGAAAAGGAAGAGTATGAGATCTGTTAGGTCTGAGAACTGTATTCTCTCTCTTGAaagcttgcagtc ctgcagatcgag
52	pEVO-voxH9 rev	GGCGACACGGAAATGTTGAGATCTTTCAAGAGAGAGAATACAGATTCTCAGACCTAACTcgaactgtacc ggttgtagtga
	<b>Construction of R6K-voxNNN vectors</b>	
53	R6K-Km TS cloning-voxN4.2 fw	AACTTCTAGACGTAGGGTTGAGAATCTGTATTCTCAGACGTATTggtctgacgctcagtggaacg
54	R6K-Km TS cloning-voxN4.4 fw	AACTTCTAGATTTATGTATGAGAATCTGTATTCTCAGACGTATTggtctgacgctcagtggaacg
55	R6K-Km TS cloning rev	actCTCGAGGAAATGTGCGCGAACC
	<b>Construction of pCAGGS reporter vectors</b>	
56	SV40pA-vox rev	TGGCGAAGCTTAGATCTAATACGTCTGAGAAATGGCGTCTCAGACCTATTCTGCATCGACCTAGACTA GC
57	lox1-mCherry fw	ATTCCGCTAGCTCAATTTCTGAGAAGTGAATTCTCAGAAATTGACGCCACCATGGTGAGCAAGG
58	SV40pA-lox1 rev	TGGCGAAGCTTAGATCTTCAATTTCTGAGAAATACACTTCTCAGAAATTGACCTGCATCGACCTAGACTA GC
59	lox2-mCherry fw	ATTCCGCTAGCATAACTTGAATAACCGAATTATCACAAGTTAACGCCACCATGGTGAGCAAGG
60	SV40pA-lox2 rev	TGGCGAAGCTTAGATCTTTAACTTGTGATAATTGCGGTTATCTCAAGTTATCTGCATCGACCTAGACTA GC
61	lox4-mCherry	ATTCCGCTAGCTGACTTCGTATAAGAACAATTATACGAAGTTACGCCACCATGGTGAGCAAGG

SUPPLEMENTARY TABLES

	fw	
62	SV40pA-lox4 rev	TGGCGAAGCTTAGATCTTAACTTCGTATAATTTGTTCTTATACGAAGTCACCTGCATCGACCTAGACTAGC
63	lox6-mCherry fw	ATTCCGCTAGCTCAATTTCCGAGAATGACAGTTCTCAGAAATTAACGCCACCATGGTGAGCAAGG
64	SV40pA-lox6 rev	TGGCGAAGCTTAGATCTTAAATTTCTGAGAACTGTCATTCTCGAAATTGACCTGCATCGACCTAGACTAGC
65	lox8-mCherry fw	ATTCCGCTAGCCTAACTTTATATAAGTCCCATTATATAATGTTAGCGCCACCATGGTGAGCAAGG
66	SV40pA-lox8 rev	TGGCGAAGCTTAGATCTCTAACATTATATAATGGGACTTATATAAAGTTAGCCTGCATCGACCTAGACTAGC
67	lox9-mCherry fw	ATTCCGCTAGCCGTCTGTCCGATAATCTTTCTTATCGGACATACTCGCCACCATGGTGAGCAAGG
68	SV40pA-lox9 rev	TGGCGAAGCTTAGATCTAGTATGTCCGATAAGAAAGATTATCGGACAGACGCCTGCATCGACCTAGACTAGC
69	lox11-mCherry fw	ATTCCGCTAGCGTCAACTTTCACATAAGTGTATTATGTGGAGTTGACCGCCACCATGGTGAGCAAGG
70	SV40pA-lox11 rev	TGGCGAAGCTTAGATCTGTCAACTCCACATAATAACACTTATGTGAAGTTGACCTGCATCGACCTAGACTAGC
71	lox12-mCherry fw	ATTCCGCTAGCATAACCTAATATAATTGTATTTATATTAGGTCAGCGCCACCATGGTGAGCAAGG
72	SV40pA-lox12 rev	TGGCGAAGCTTAGATCTCTGACCTAATATAAATACAATTATATTAGGTTATCCTGCATCGACCTAGACTAGC
	<b>Construction of mammalian expression vectors</b>	
73	Vika BsrGI fw	TCAGGAATTGTACATGACCGATCTGACCCCGTTTCC
74	Vika no stop rev	GTCGACTCTAGAACGCTGACGACGTTTCTGTGCC
75	YR1 BsrGI fw	AGGAATTGTACATGATTGAAAACCAGCTGAGCCTGC
76	YR1 no stop rev	GTCGACTCTAGAGCGTTTGCTGCCCCAGCTCAGATCAATG
77	YR2 BsrGI fw	AGGAATTGTACATGAACAACGAAATATTTCATAGCACCAGC
78	YR2 no stop rev	GTCGACTCTAGAGTTGCGCGGTTTCTGGC
79	YR4- BsrGI fw	AGGAATTGTACATGAGTGAACCTCCTGCCAC
80	YR4-no stop rev	GAACATCTAGAGCCTGGATAATCAGGATCC
81	YR6-BsrGI fw	AGGAATTGTACATGATCGAGAATCAACTTTCTCTTCTGG
82	YR6-no stop rev	GTCGACTCTAGAAATGGTGTCTTGTGTACC
83	YR8-BsrGI fw	AGGAATTGTACATGGGTAAGCTTTCCCTACGAACC
84	YR8-no stop rev	GTCGACTCTAGACTTAGTGTGTTGTGGTCCAGC
85	YR9-BsrGI fw	AGGAATTGTACATGTTGGCGGTTTTACATGAGGAC
86	YR9-no stop rev	GTCGACTCTAGACAGAAAGCGCTCACCGCGGTG
87	YR11-BsrGI fw	AGGAATTGTACATGGTTGGTGGCATGAGTTTCGTTTCG
88	YR11-no stop rev	GTCGACTCTAGAACCAACGAGGTCAATCATCACGC
89	YR12-BsrGI fw	AGGAATTGTACATGACCGAGATGATTGTGCGCAAACC
90	YR12-no stop rev	GTCGACTCTAGATTTCATCCTCTCCATAAGTCCG
91	R6.c1-opt BamHI-Kozak fw	tgacgcgggatccgccaccATGTTGGCAATCCTGCACGAAG
92	R6.c10-opt BamHI-Kozak fw	tgacgcgggatccgccaccATGCTGGCCATCTTGCACG
93	R6.c2-opt BamHI-Kozak fw	tgacgcgggatccgccaccATGCTCACAGTGTACACGAAG

SUPPLEMENTARY TABLES

94	R6.wt-opt BamHI-Kozak fw	tgacgcgggatccgccaccATGCTGGCAGTCTGCACG
95	BFP rev	CTCGGATGTGCACTTGAAGT
96	mCherry fw	gaccggttctagagcgctgc
97	mCherryVika rev	GGCGGAAACGGGGTCAGATCGGTGTACAAGGGGTCTTCTACCTTTC
98	mCherryVika fw	GAAAGGTAGAAGACCCCTTGTACACCGATCTGACCCCGTTTCCGCC
99	Vika rev	cagctggCTCGAGTTAACGCTGACGACGTTTCTGTG
	<b>Construction of mammalian integration reporter vectors</b>	
100	p11-voxH9- mChery fw	ggccacaccggtGTTAGGTCTGAGAATCTGTATTCTCTCTTGAAGcgctgccaccATGGTGAG
101	p11-voxH9- mChery-p2a1 rev	CTGTACAAGGGGTCTTCTACCTTCTCTCTTTTtaggcccgggatctcctccacgtcacctgcttgtt tgagtagtgagaagtttgttgcCTTaTAGACTCGTCCATGCCGCCG
102	p11-voxH9- mChery-NLS2 rev	agctggCTCGAGTCTAGAATCTGAACTGTACAAGGGGTCTTCTACCTTTCTC
103	pmaxGFP- P2A-PuroR fw	gatctcgagctcgaGGAAGCGGAGAGGGCAGAG
104	pmaxGFP- P2A-PuroR rev	ctcatcgagctcTTAGGCACCGGGCTTGCG
105	voxN4.2- BsrGI fw	gtacAATACGTCTGAGAATACAGATTCTCAACCCTACG
106	voxN4.2- BsrGI rev	gtacCGTAGGGTTGAGAATCTGTATTCTCAGACGTATT
107	voxN4.4- BsrGI fw	gtacTTTATGTATGAGAATCTGTATTCTCAGACGTATT
108	voxN4.4- BsrGI rev	gtacAATACGTCTGAGAATACAGATTCTCATACATAAA
	<b>Primers for SLiDE protocol</b>	
109	EvoPCR-rec fw	CGGCGTCACACTTTGCTATG
110	EvoPCR rev	CGTGTGCGCCCTTATTCCTT
111	pEvoseq rev	CCCAACTGATCTTCAGCATC
112	pEvo fw	ctcttctcgctaaccacaaaccg
113	intSLiDE PCR rev	GACGTTTCCCGTTGAATATGGC
	<b>Primers for genomic PCRs</b>	
114	H9-gDNA j1 fw	TGATCTCGTGATCTGCCCTCC
115	copGFP-pTGP- j1 rev	TCCGTCCTTGTCTGGTGTTC
116	H9-gDNA-j2 rev	GCTTCATAAGCCACACTCAGAGC
117	pMax-CMV-j1 rev	ccgtcattgacgtcaatagggg
118	KanR-pMax- GFP-j2 fw	agtgacaacgtcgagcacag
119	p11-EF1a-j1 fw	taagtgcagtagtcgcccgtg
120	pTGP-Puro-j1 rev	CTCTTGTAGCAAGTCTGACTGTAGG



SUPPLEMENTARY TABLES

**Supplementary Table S3. Y-SSR candidates chosen for validation**

Label	GenBank Accession number Amino acid sequence	Protein size	Putative target site	Origin
YR1	<p>WP_080176046.1  MIENQLSLLGDFTDVVRPSDVKTAI  EKAQKKGVVVAEDHVFQAAINHLL  NEFKKREDRYSPTLRRLESAGWC  FVEWCLDNKRHSLPASPDTEKFL  IYKAESVHRNTLSIYKWAISRVRH  VAGCPNPCNDVFVEDRYKALVRVK  VQSGEAIKQASPFNELHLNALVEK  WKQHERVLERRNLALLGVAYESML  RAAELANIKLSDIELAGDGTAILT  IPITKTNHSGDPDTCILSHDVVGL  IMDYIEAGELHLKQDGYLFTGVSK  HNKCTKPKVDKETGEVITYKPIITK  TVEGIFKAAWSELELGRQGVKPF  GHSARVGATQDILLRKGYNLQIQQ  SGRWSSEVMVARYGRAILARESAM  AQSRVKTKNIDLSWGSKR*</p>	378	<p>TCAATTTCTGAGA  AGTGTAAT  TCTCAGAAATTGA</p>	Photobacterium toruni
YR2	<p>WP_051647626.1  MNNEIIHSTSNTSLSQYPAEHIQK  ALANGDIPTDShLFQSAADHLIN  YRSREGLAENTFLALDTGWSLFVD  WCVEHNRVSLPASSKTVEDYVKS  SKVLRNNTIRVRKWAITKIHKICG  LPNPFDFSEFVTQTISGIYKKKLHE  DEITEQASPFNETHLEALELLYAD  STLKKRRDMLMMTIAYESLLRSSE  LCNIKLRRLRIGKEIHITIPVTK  TNHSGNPDVVALSEHATNQVLEYL  NDHSMKLSGDGYLFRRLRRNGLAY  PSTKQAMSNQSVIDVFNSVHNDLG  GSDVLHCEPFTSHSCRVGGAQDLL  AAGYSILQVQQAGRWDPSMVYRY  GRGIFAAKSAMAHFRRNRQKPRN*</p>	359	<p>ATAACTTGAGATA  ACGCGAAT  TATCACAAGTTAA</p>	Vibrio shiloi
YR3	<p>WP_128342862.1  MKSLLTLAENNTPHLFPNQTNHS  DPRVDAIIAGIDFDVITPSDQAE</p>	394	<p>ACTATGTTCTCGA  TGGCCCGG  TCATAAGCATCCT</p>	Aeromonas caviae

SUPPLEMENTARY TABLES

	<p>IRLSRMRGIDITTTSLALRQSIDRF  IADFEKRIGADAGNQAQSNTVRS  LSNWKLFARWCNENNVAGPLPAPM  ATVEQYLMYRFEKGLSKHSLVMDQ  WAIRRFHLEGGCPDPTSEERIKGL  VAKLKRDRVFIHNDLTKQATAMRK  STLKRLVELWGGPESTLKQKRDLA  MMVIAYCTLLRGSELARIKLEHFS  VKSDGRGAVLMI PVSKTNHSGSPD  AVFLKDRQM QHVYRYLAADGRSIN  DSGYLLGAVTANGRKP IRRVDPLT  VQTVRDTFRRAWDATAQPGSNERP  FSAHSARVGAAQDLRLKDGVKIQD  IMHAGRWSNESMVLRYTRNVDAEE  TSAVMIQDDF</p>			
YR4	<p>CP031651.1  MSELLPLTPLTVDRNSDITERLRQ  FVQDKEAFSPNTWRQLLSVMRICN  RWSEDNQRSFLPMSADDLRDYLSF  LAESGRASSTVTSHAALISMLHRN  AGLPVPNVSPLVFRTMKKINRVAV  INGERAGQAVPFRLSDLLALDEEW  SGSDNLQALRDALFLHVAYATLLR  ISELSRLRVRDVMRAGDGRIILDV  AWTKTIVQTGGLIKALSARSTQRL  EEWIEASGLSSQPDWLFTAVHRS  GRPLIAEKPMSTRALEQIFSRWR  TAGKEGAVKANKNRYTGWSGHSAR  VGAAQDMADKGYPIARIMQEGTWK  KPETLMRYIRHVDAHKGAMVEFME  QYGDPPDYPG*</p>	345	<p>TGACTTCGTATA  AGAACAAT  TATACGAAGTTA</p>	P. agglomerans
YR5	<p>LR136958.1  MSSIFKPGSNTQLSIVDGDSTNTE  RAEYVDAYFRDIFTKLPRNTQRAY  ISDFNDFAI FCQSERIDSFSDDFS  HNEYAIKRYVEELCKSPAYRTIK  RRLSALS KFLGIAKLPNPIIQSVY  LRDFVRLSLIENRKFQLSHNQAVP  LTIDLLDEINNKIIPDTLLEMRDL  TIMNLMFDALLRADELVRVCAEHI  SRRNNAVLVVTSKSDQSGKGS HRF  ISTSTINMVDEYIAEANFN AKTQS  ERLVDDLRRINRGILFRRVSNRGH  ALLPYDEQANTHQHSP IHESSILD  YSSVYRIWKRVAARAEIKENITPH  SGRVGGAVSLAENGATLPELQLAG  GWQSPMPGHYTQQANVKRGGMAK  LSEKFKR*</p>	368	<p>ACATTTTAACTGT  TGTTATAT  ACAGTTAAATTGG</p>	Alteromonas sp. 76-1

SUPPLEMENTARY TABLES

<p>YR6</p>	<p>CP010077.1 MIENQLSLLGDFSGVRPDDVKA AVQAAQKKGINVAENEQFKAVFDHLL GEFKKREERYSPNTLRRLES AWTCFVDWCLAHHRHSLPATPDTVEAFF IERSETLHRNTLSVYRWAI SRVHRVAGCPDPCLDIYVEDRLKAI SRKKVREGETVKQASPFNEQHLLKLTSL WYLSDKLLRRNLALLAVAYESML RAAELANIRVSDLELSGDGTAVLT IPITKTNHSGEPDTCILSQDVVSL LMDYTEAGRLDMRADGYL FVGI SKHNTCINPKRDADTGECLHKPITTK TVEGVFYSAWQALELERQGVKPF T AHSARVGA AQDLLKKGYN TLQIQQSGRWSSGTMVARYGRAIL ARDGAM AHSRVKTRNVSIDWGS GGSKNTI*</p>	<p>383</p>	<p>TCAATTTCCGAGA ATGACAGT TCTCAGAAATTAA</p>	<p><i>V. anguillarum</i></p>
<p>YR7</p>	<p>CP009356.1 MTTSLVSEVPFERLLPHEFAEGL AAAQRAGEALEGHPLVEAAITHYQ GEFFRAERLQPASLVRLKSAWAT FVAWCCEQDRCALPASQTV EAYLIAEQDRLHRNTLKVQLWAI GKTHQISGCPDPCHNDYVKAQLQ QIHHRKVRQREVIRQAV ALRESHLNALADLWDRPEASL TECRDLLIVSMLYETL LRKSNLETLRVGDVDWQADG SGLIKVFVTKTDKSGDV KYSYVSPSTMDLLARYLGHAD IVDNPEAFLIQRVKLSSQQLK GSARTQA AISPVSAKLI GRVCAKAAKTLGLSTDRPFT GHSA RVGATQDLLAEGFSSLQV QQAGWSSERMVLRYGGSV LASESAMAQRRQRKSPK*</p>	<p>367</p>	<p>AATACGTCCTAGA ACTGCCAT TCTAGGACGTATG</p>	<p><i>V.tubiashii</i> ATCC 19109</p>
<p>YR8</p>	<p>CP003406.1 MGKLSPTNQTLPAIQAEEDV LARLKEFVQDKEAFSPNTWR QLMSVMRI</p>	<p>349</p>	<p>CTAACTTTATATA AGTCCCAT TATATAATGTTAG</p>	<p><i>R. aquatilis</i> HX2</p>

SUPPLEMENTARY TABLES

	<p>CHRWSIENSRSFLPMLPADLRDYL          NWLQESGRASSTIATHGSLISMLH          RNAGLIPPNTSPLVFRVAVKKINRV          AVVTGERTGQAVPFRLEDLLELDA          LWSDSISLRHKRDLAFLHVAYSTL          LRISLARLRVRDISRATDGRIIL          NVSYTKTIVQTGGLIKSLSSQSSR          RLTEWMSVSGINAEPDAFLFCPVH          RSGSATLSVTRPLSTPAIESIFAQ          AWLTIGAGEPIIPNKGRYTAWTGH          SARVGAAQDMAGRGYAVAQIMQEG          TWKKPETLMRYIRNLQAHEGAMTD          IMEKSTLDHNNTK*</p>			
YR9	<p>CP018821.1          MLAVLHEDLERAAAYKKAARAAAT          HRAYNSDWIITYDWCRTGLEAMP          AHPEQIAAFVANQAASGLKPSTIE          RRVAATIGHHRTSNYPAPAAHPEA          GGLREALAGIRNEKRAKKTREPA          DATAALDMLAQIKGDGLRARRDRA          ALAIGMAAALRRSELVALTLENVG          ILEHGIELYLGATKTDQAGEGTTI          AIPGTRLRPKALLLDWISAVRVL          EAGVVRTPAQEAAVPLFRRLTRSD          QLTGEPMSDKAVARLVKRYAGAAG          YDAAKFSGHS LRAGFLTEAANQGA          TIFKMQEVSRRHKTQVLSDYVRS          DRFRDHAGERFL*</p>	324	<p>CGTCTGTCCGATA          ATCTTTCT          TATCGGACATACT</p>	S. koreensis
YR10	<p>CP031536.1          MKKNIPLITDPMILKQKVTEFSDT          VCFEQFQHLTHWQYSKNSALAMAK          DWNHFVTFCKIRCVTPLPGSTTAV          RQFIETEARVRKYATIRRYMVTIG          I IHLLSMKDPTQNRLTQLSLFRL          KGEKGDDAKQATPLTKKHLALDI          QLIHSSHKKDIRDLAIYYVMFECA          LKRSELKKLSIGQALTVDGQMKII          VGNVDYYLSEASLALKKWLQLLN          KESNIVFCSIDRHGNISNRSLNDA          SIFRILRHAGQRLGIFELRFSGQS          TRVGAAQELAKQGYKTQDIQQFGR          WLSPPAMPAQYIGKLDIAESEQMKF          KVIKPFDP*</p>	320	<p>ATAACTAGTCTGATA          ATCGAAGA          TTCATTCTAGTTAT</p>	V. anguillarum



SUPPLEMENTARY TABLES

<p>YR11</p>	<p>FO818638.1  MVGGMSEFVRRDVVVI PDNPDNLNDE  VIRNLNAFMKDREAF AENTWKQLM  MAVRLWCHWCI AKGRPYLPVDADY  LRDYLLLELHDNGLAPATISNYAAM  LNLLHRQAGLIPAGESQKVKRVLK  KISRTSIIKGETVQQAIPFRIADL  NQVDEAWEASDRLKTIRNLAFLFV  AYNTLLRISNIAHLKVKDLAFDHD  GSVMLNIGYTKTLVDGKGITKALS  PRASARVLKWLHVSGLLDHPDAYL  FCKVYRTNKASVTTDKPLTLHPLE  SIFSEAWAVIHGEKVGIKNKGRYA  TWTGHSARVGAAQDMTESGYSLAQ  IMHEGTWKAPKTVLGYTRNLEAKK  SVMIDLVG*</p>	<p>344</p>	<p>GTCAACTTCACATA  AGTGTTAT  TATGTGGAGTTGAC</p>	<p>X. bovienii</p>
<p>YR12</p>	<p>CP002436.1  MTEMIVANPLLAQFSASDDISAKL  ASFVRDREAFSSNTWRQLLSVMRI  CWRWSEENHRSFLPMAPEDLRDYL  LHLQCI GRASSTI STHAALISMLH  RNAGLVPPNVSPDVFRVVKKINRA  AVIAGERTGQAVPFCRQDLKLDLDT  AWQGS PRLQQLRDLAFMHVAYSTL  LRLSELSRLRVRDISRAADGRMIL  DVAWTKTIVQSGGIVKALSTQSSQ  RLTDWIVAAGLTGEPDAMIFCPVH  RSNRMTKKIFSPMSTPCLEDIFLR  AREAAGVAALSRTNKGRYAGWSGH  SARVGAAQDMARKGFSVAQIMQEG  TWTRTETVMRYIRMVEAHKGAMIG  LMEEDE*</p>	<p>342</p>	<p>ATAACCTAATATA  ATTGTATT  TATATTAGGTCAG</p>	<p>Pantoea sp. At-9b</p>
<p>YR13</p>	<p>CP015368.1  MVGGMSEFVRRDVVVI PDNPDNLNDE  VIRNLNAFMKDREAF AENTWKQLM</p>	<p>345</p>	<p>ATTGCCTTCGATA  ATACCAAC  TATCGCGGGAAAT</p>	<p>M.  phyllosphaerae</p>

SUPPLEMENTARY TABLES

	<p>MAVRLWCHWCIAKGRPYLPVDADY          LRDYLLELHDNGLAPATISNYAAM          LNLHRQAGLIPAGESQVKRVLK          KISRTSIIKGETVQAI PFRIADL          NQVDEAWEASDRKLTIRNLAF LFV          AYNTLLRISNIAHLKVKDLAFDHD          GSVMLNIGYTKTLVDGKGITKALS          PRASARVLKWLHVSGLLDHPDAYL          FCKVYRTNKASVTTDKPLTLHPLE          SIFSEAWAVIHGEKVG IKNKGRYA          TWTGHSARVGAAQDMTESGYS LAQ          IMHEGTWKAPKTVLGYTRNLEAKK          SVMIDLVG*</p>			
YR14	<p>JX627580.1          MPVLF TDALPPGLDLLIERLEQHA          RAAQGA FADNTVRAFAADSRIFSA          WCGQAGRAMLPAAPETVAAFIDAQ          AEIKARATVERYRSSIAALHRAAG          LSNPCADEIVRLAVKRMNRAKGRR          QKQAEPLNRTSIERMLEVKTPGRL          HRRITEAKREVPLIALRNAALVAV          AYDTLLRRSELVSLYIGDLHRGAD          GSGTVLVRRSKADQEGEGA IKYLA          PDTMAHIEAWLSAAHLESGPLFRP          LTKGGQVGTGALGGGEVARVFRDL          SMAAGLKLARLPSGHSTRVGATQD          MFAAGFELLEVMQAGSWKTPAMPA          RYGERLRAQRGAARKLATLQNR A*</p>	336	<p>ATTGCCTTCGATA          ATACCAAC          TATCGCGGGAAT</p>	<p><b>Methylobacterium          oryzae CBMB20</b></p>
YR15	<p>LS483250.1          MPKALINIRNNSIVSSEHLTEEHI          ENLIRFSDRKEQLEENTLKSLHYH          VSKFN DYCLTHNVIPLPLQDATIL          EAFLIQENERGCKAQTLRIYAVAV          SKIHSLAGLEIPYFKTIIITARLKI          ISKREVKLGVKRKQAVAFSYSHLK          FVNEQLDINSLLSIRNALLLNICY          DGLLRESEACNLFVDQIQKVNGLY          NINITNSKTDKSLDGSIVHLSKFT          SKLLPLYLQKVSAAHGHQFLFKRIT          PRGGKLEKLPKSDHSPNPHDKPIS          TKTVELVFANTWHSITDYNQSVSF          DLYVDLPERPFSGHSARVGASCDL          VSAGYDDSLVMRAGRWKTLRMVEL          YTRGVNKKFATVREFRERLDTMTD          I*</p>	362	<p>TACTATCCAGTTA          TTATCGGT          TAACTGGAAAGAA</p>	<p><b>Moritella          yayanosii</b></p>

SUPPLEMENTARY TABLES

<p>YR16</p>	<p>FP103043.1  MELVATDSAAEPQRDAFNPPVPFA  DALPPGLELLIERLEQHARAARGA  FADNTVRALAADSRIFAAWCREEG  RAMLPATPETVAAFIDAQGETKAR  ATVERYRSSIAALHRAAGLPNPCA  DEIVRLAVKRMNRARGRRQKQAEF  LNRASIERMLEVKTPGRLHRRVTE  AKRETPLIALRNAALVAVAYDTLL  RRSELVSLYIGDLHKGADGSGTVL  VRRSKADQEGEGAIKYLAPDTMAH  IEAWLSAAHLESGPLFRPLTKGGQ  VGTVALGGGEVARVFRDLATAAGL  KLARLPSGHSTRVGATQDMFAAGF  ELLEVMQAGSWKTPAMPARYGERL  RAQRGAARKLATLQNR*</p>	<p>354</p>	<p>ATTTCCCGCGATA  GATGGTGT  TATCGCAGGCAAT</p>	<p><b>Methylobacterium  extorquens AM1</b></p>
<p>YR17</p>	<p>LT629738.1  MSELD RYLHAATRDNTRRSYRAAI  EHFEVTWGGFLPATSDSVARYLVA  YAGELSINTLKLRLSALAQWHNSQ  GFVDPTKAPVVRQVFKGIRALHPA  QEKQAEPLQLQHLEQVIWLEQEA  SQARLDDNQPALLRARRDSALILL  GFWRGFRSDELCLRLRIEHVQAVAG  SGISLYLPRSKSDRDNLGKTWHTP  ALRRLCPVQAYIEWINAAALVRGP  VFRGIDRWGHLSEEGHANSVIPL  LRQALERAGVAAEQYTSHSLRRGF  ATWAHRSGWDLKSLMSYVGWKDIK  SAMRYVEASPFLGMALTQEKPVGE  *</p>	<p>313</p>	<p>CCCGCAACTTCGA  CGTGCTGA  TCGAAGAAGCCGG</p>	<p><b>Pseudomonas  chlororaphis</b></p>

## SUMMARY

### SUMMARY

Tyrosine site-specific recombinases (Y-SSRs) are DNA editing enzymes that play a valuable role for the manipulation of genomes, due to their precision and versatility. They have been widely used in biotechnology and molecular biology for various applications, and are slowly finding their spot in gene therapy in recent years. However, the limited number of available Y-SSR systems and their often narrow target specificity have hindered the full potential of these enzymes for advanced genome engineering. In this PhD thesis, I conducted a comprehensive investigation of novel Y-SSRs and their potential for advancing genome engineering. This PhD thesis aims to address the current limitations in the genetic toolbox by identifying and characterizing novel Cre-type recombinases and demonstrating their impact on the directed evolution of designer recombinases for precise genome surgery.

To achieve these aims, I developed in a collaboration a comprehensive prediction pipeline, combining a rational bioinformatical approach with knowledge of the biological functions of recombinases, to enable high success rate and high-throughput identification of novel tyrosine site-specific recombinase (Y-SSR) systems. Eight putative candidates were molecularly characterized in-depth to ensure their successful integration into future genome engineering applications. I assessed their activity in prokaryotes (*E. coli*) and eukaryotes (human cell lines), and determined their specificity in the sequence space of all known Cre-type target sites. The potential cytotoxicity associated with cryptic genomic recombination sites was also explored in the context of recombinase applicability. This approach allowed the identification of novel Y-SSRs with distinct target sites, enabling simultaneous use of multiple Y-SSR systems, and provided knowledge that will facilitate the assignment of novel and known recombinases to specific uses or organisms, ensuring their safe and effective implementation.

The introduction of these novel Y-SSRs into the genome engineering toolbox opens up new possibilities for precise genome manipulation in various applications. The broader targetability offered by these enzymes could accelerate the development of novel gene therapies, as well as advance the understanding of gene function and regulation. Moreover, these recombinases could be used to design custom genetic circuits for synthetic biology, allowing researchers to create more complex and sophisticated cellular systems.

Finally, I introduced the novel Y-SSRs into efforts aimed at developing designer recombinases for precise genome surgery, demonstrating their impact on accelerating the directed evolution process. Therapeutically relevant recombinases with altered DNA specificity have been developed for excision or inversion of specific DNA sequences.

## SUMMARY

However, the potential for evolving recombinases capable of integrating large DNA cargos into naturally occurring lox-like sites in the human genome remained untapped so far. Thus, I embarked on evolving the Vika recombinase to mediate the integration of DNA cargo into a native human sequence. I discovered that Vika could integrate DNA into the voxH9 site in the human genome, and then, I enhanced the process through directed evolution. The evolved variants of Vika displayed a marked improvement in integration efficiency in bacterial systems. However, the translation of these results into mammalian systems has not yet been entirely successful. Despite this, the study laid the groundwork for future research to optimize the efficiency and applicability of Y-SSRs for genomic integration.

In summary, this thesis made significant strides in the identification, characterization, and development of novel Y-SSRs for advanced genome engineering. The comprehensive prediction pipeline, combined with in-depth molecular characterization, has expanded the genetic toolbox to meet the growing demand for better genome editing tools. By exploring efficiency, cross-specificity, and potential cytotoxicity, this research lays the foundation for the safe and effective application of novel Y-SSRs in various therapeutic settings. Furthermore, by demonstrating the potential of these recombinases to improve efforts in creating designer recombinases through directed evolution, this research has opened new avenues for precise genome surgery. The successful development and implementation of these novel recombinases have the potential to revolutionize gene therapy, synthetic biology, and our understanding of gene function and regulation.

### ZUSAMMENFASUNG

Ortsspezifische Rekombinasen vom Tyrosin-Typ (Y-SSRs) sind DNA-editierende Enzyme, die dank ihrer Präzision und Vielseitigkeit eine wertvolle Rolle bei der Manipulation von Genomen spielen. Sie wurden bereits für eine Vielzahl von Anwendungen in der Biotechnologie und der Molekularbiologie verwendet, und finden in den letzten Jahren langsam ihren Platz in der Gentherapie. Die begrenzte Anzahl verfügbarer Y-SSR-Systeme und ihre oftmals hohe Zielspezifität haben jedoch das volle Potenzial dieser Enzyme für fortgeschrittene Genomtechnik eingeschränkt. In dieser Doktorarbeit führte ich eine umfassende Untersuchung neuer Y-SSRs und ihrer Potenziale für die Weiterentwicklung der Genomtechnik durch. Diese Doktorarbeit zielt darauf ab, die derzeitigen Einschränkungen im genetischen Werkzeugkasten durch Identifizierung und Charakterisierung neuer Cre-Typ-Rekombinasen zu beheben und ihre Auswirkungen auf die gerichtete Evolution von Designer-Rekombinasen für präzise Genomchirurgie zu demonstrieren.

Um diese Ziele zu erreichen, entwickelte ich in einer Kollaboration eine umfassende Vorhersage-Pipeline, die einen rationalen bioinformatischen Ansatz mit dem Wissen über die biologischen Funktionen von Rekombinasen kombiniert, um eine hohe Erfolgsquote und eine schnelle Identifizierung neuer Systeme von Ortsspezifischen Rekombinasen vom Tyrosin-Typ zu ermöglichen. Acht mutmaßliche Kandidaten wurden eingehend molekular charakterisiert, um ihre erfolgreiche Integration in zukünftige Genomtechnikanwendungen sicherzustellen. Ich bewertete ihre Aktivität in Prokaryoten (*E. coli*) und Eukaryoten (menschliche Zelllinien) und bestimmte ihre Spezifität im Sequenzraum aller bekannten Cre-Typ-Zielstellen. Die potenzielle Zytotoxizität im Zusammenhang mit kryptischen genomischen Rekombinationsstellen wurde ebenfalls im Kontext der Rekombinase-Anwendbarkeit untersucht. Dieser Ansatz ermöglichte die Identifizierung neuer Y-SSRs mit unterschiedlichen Zielstellen und ermöglichte den gleichzeitigen Einsatz mehrerer Y-SSR-Systeme und lieferte Wissen, welches die Zuordnung neuer und bekannter Rekombinasen zu bestimmten Verwendungen oder Organismen erleichtern wird, um ihre sichere und effektive Implementierung sicherzustellen.

Die Einführung dieser neuen Y-SSRs in den Werkzeugkasten für Genomtechnik eröffnet neue Möglichkeiten für präzise Genommanipulationen in verschiedenen Anwendungen. Die größere Reichweite an Zielen, die diese Enzyme bieten, könnte die Entwicklung neuer Gentherapien beschleunigen und das Verständnis von Genfunktion und -regulation voranbringen. Darüber hinaus könnten diese Rekombinasen verwendet werden, um

## ZUSAMMENFASSUNG

kundenspezifische genetische Schaltkreise für die synthetische Biologie zu entwerfen, wodurch Forscher komplexere und ausgefeiltere zelluläre Systeme schaffen können.

Schließlich führte ich die neuen Y-SSRs in die Bemühungen ein, Designer-Rekombinasen für präzise Genomchirurgie zu entwickeln, und demonstrierte ihre Auswirkungen auf die Beschleunigung des gerichteten Evolutionsprozesses. Therapeutisch relevante Rekombinasen mit veränderter DNA-Spezifität wurden für die Exzision oder Inversion spezifischer DNA-Sequenzen entwickelt. Das Potenzial für die Evolution von Rekombinasen, die in der Lage sind, große DNA-Cargos in natürlich vorkommende *Lox*-ähnliche Stellen im menschlichen Genom zu integrieren, blieb jedoch bisher ungenutzt. Daher begann ich mit der Evolution der Vika-Rekombinase, um die Integration von DNA-Cargos in eine native menschliche Sequenz zu vermitteln. Ich entdeckte, dass Vika DNA in die *voxH9*-Stelle im menschlichen Genom integrieren konnte und verbesserte diesen Prozess dann durch gerichtete Evolution. Die entwickelten Varianten von Vika zeigten eine deutliche Verbesserung der Integrationseffizienz in bakteriellen Systemen. Die Übertragung dieser Ergebnisse in Säugetiersysteme war jedoch noch nicht vollständig erfolgreich. Trotzdem legte die Studie den Grundstein für zukünftige Forschungen zur Optimierung der Effizienz und Anwendbarkeit von Y-SSRs in der Genomintegration.

Zusammenfassend hat diese Arbeit bedeutende Fortschritte bei der Identifizierung, Charakterisierung und Entwicklung neuer Y-SSRs für die fortgeschrittene Genomtechnik gemacht. Die umfassende Vorhersage-Pipeline in Kombination mit einer eingehenden molekularen Charakterisierung hat den genetischen Werkzeugkasten erweitert, um der wachsenden Nachfrage nach besseren Genomeditierungswerkzeugen gerecht zu werden. Durch die Untersuchung von Effizienz, Kreuzspezifität und potenzieller Zytotoxizität legt diese Forschung die Grundlage für die sichere und effektive Anwendung neuer Y-SSRs in verschiedenen therapeutischen Umgebungen. Darüber hinaus hat diese Forschung durch die Demonstration des Potenzials dieser Rekombinasen zur Verbesserung der Bemühungen um Designer-Rekombinasen durch gerichtete Evolution neue Wege für präzise Genomchirurgie eröffnet. Die erfolgreiche Entwicklung und Implementierung dieser neuen Rekombinasen hat das Potenzial, die Gentherapie, die synthetische Biologie und unser Verständnis von Genfunktion und -regulation zu revolutionieren.

## BIBLIOGRAPHY

## BIBLIOGRAPHY

Abi-Ghanem J, Chusainow J, Karimova M, Spiegel C, Hofmann-Sieber H, Hauber J, Buchholz F, Pisabarro MT. 2012. Engineering of a target site-specific recombinase by a combined evolution- and structure-guided approach. *Nucleic Acids Res* 41:2394–2403.

Albert H, Dale EC, Lee E, Ow DW. 1995. Site-specific integration of DNA into wild-type and mutant lox sites placed in the plant genome. *Plant J* 7:649–659.

Anastassiadis K, Fu J, Patsch C, Hu S, Weidlich S, Duerschke K, Buchholz F, Edenhofer F, Stewart AF. 2009. Dre recombinase, like Cre, is a highly efficient site-specific recombinase in *E. coli*, mammalian cells and mice. *Dis Model Mech* 2:508–515.

Anastassiadis K, Glaser S, Kranz A, Bernhardt K, Stewart AF. 2010. Chapter Seven A Practical Summary of Site-Specific Recombination, Conditional Mutagenesis, and Tamoxifen Induction of CreERT2. *Methods Enzymol* 477:109–123.

Anzalone AV, Randolph PB, Davis JR, Sousa AA, Koblan LW, Levy JM, Chen PJ, Wilson C, Newby GA, Raguram A, Liu DR. 2019. Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature* 576:149–157.

Araki K, Araki M, Yamamura K. 2002. Site-directed integration of the cre gene mediated by Cre recombinase using a combination of mutant lox sites. *Nucleic Acids Res* 30:e103–e103.

Araki K, Araki M, Yamamura K -i. 1997. Targeted integration of DNA using mutant lox sites in embryonic stem cells. *Nucleic Acids Res* 25:868–872.

Araki K, Okada Y, Araki M, Yamamura K. 2010. Comparative analysis of right element mutant lox sites on recombination efficiency in embryonic stem cells. *Bmc Biotechnol* 10:29.

Austin S, Ziese M, Sternberg N. 1981. A novel role for site-specific recombination in maintenance of bacterial replicons. *Cell* 25:729–736.

Aznauryan E, Yermanos A, Kinzina E, Devaux A, Kapetanovic E, Milanova D, Church GM, Reddy ST. 2022. Discovery and validation of human genomic safe harbor sites for gene and cell therapies. *Cell Reports Methods* 2:100154.

Berman CM, Papa LJ, Hendel SJ, Moore CL, Suen PH, Weickhardt AF, Doan N-D, Kumar CM, Uil TG, Butty VL, Hoeben RC, Shoulders MD. 2018. An Adaptable Platform for Directed Evolution in Human Cells. *J Am Chem Soc* 140:18093–18103.

Bianchi MM. 1992. Site-specific recombination of the circular 2 microns-like plasmid pKD1 requires integrity of the recombinase gene A and of the partitioning genes B and C. *J Bacteriol* 174:6703–6706.

Bibikova M, Carroll D, Segal DJ, Trautman JK, Smith J, Kim Y-G, Chandrasegaran S. 2001. Stimulation of Homologous Recombination through Targeted Cleavage by Chimeric Nucleases. *Mol Cell Biol* 21:289–297.



## BIBLIOGRAPHY

Bigelyte G, Young JK, Karvelis T, Budre K, Zedaveinyte R, Djukanovic V, Ginkel EV, Paulraj S, Gasior S, Jones S, Feigenbutz L, Clair GST, et al. 2021. Miniature type V-F CRISPR-Cas nucleases enable targeted DNA modification in cells. *Nat Commun* 12:6191.

Bonnet J, Yin P, Ortiz ME, Subsoontorn P, Endy D. 2013. Amplifying Genetic Logic Gates. *Science* 340:599–603.

Broach JR, Guarascio VR, Jayaram M. 1982. Recombination within the yeast plasmid 2 $\mu$  circle is site-specific. *Cell* 29:227–234.

Buchholz F, Angrand P-O, Stewart AF. 1998. Improved properties of FLP recombinase evolved by cycling mutagenesis. *Nat Biotechnol* 16:657–662.

Buchholz F, Ringrose L, Angrand P-O, Rossi F, Stewart AF. 1996. Different Thermostabilities of FLP and Cre Recombinases: Implications for Applied Site-Specific Recombination. *Nucleic Acids Res* 24:4256–4262.

Buchholz F, Stewart AF. 2001. Alteration of Cre recombinase site specificity by substrate-linked protein evolution. *Nat Biotechnol* 19:1047–1052.

Cadwell RC, Joyce GF. 1992. Randomization of genes by PCR mutagenesis. *Genome Res* 2:28–33.

Capecchi MR. 1989. Altering the Genome by Homologous Recombination. *Science* 244:1288–1292.

Casjens S. 2003. Prophages and bacterial genomics: what have we learned so far? *Mol Microbiol* 49:277–300.

Casjens SR, Hendrix RW. 2015. Bacteriophage lambda: Early pioneer and still relevant. *Virology* 479:310–330.

Cermak T, Doyle EL, Christian M, Wang L, Zhang Y, Schmidt C, Baller JA, Somia NV, Bogdanove AJ, Voytas DF. 2011. Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting. *Nucleic Acids Res* 39:7879–7879.

Chaikind B, Bessen JL, Thompson DB, Hu JH, Liu DR. 2016. A programmable Cas9-serine recombinase fusion protein that operates on DNA sequences in mammalian cells. *Nucleic Acids Res* 44:9758–9770.

Chen SP, Wang HH. 2019. An Engineered Cas-Transposon System for Programmable and Site-Directed DNA Transpositions. *Crispr J* 2:376–394.

Christian M, Cermak T, Doyle EL, Schmidt C, Zhang F, Hummel A, Bogdanove AJ, Voytas DF. 2010. Targeting DNA Double-Strand Breaks with TAL Effector Nucleases. *Genetics* 186:757–761.

Ciciani M, Demozzi M, Pedrazzoli E, Visentin E, Pezzè L, Signorini LF, Blanco-Miguez A, Zolfo M, Asnicar F, Casini A, Cereseto A, Segata N. 2022. Automated identification of sequence-tailored Cas9 proteins using massive metagenomic data. *Nat Commun* 13:6474.

Cramer A, Raillard S-A, Bermudez E, Stemmer WPC. 1998. DNA shuffling of a family of genes from diverse species accelerates directed evolution. *Nature* 391:288–291.

## BIBLIOGRAPHY

Drummond DA, Iverson BL, Georgiou G, Arnold FH. 2005. Why High-error-rate Random Mutagenesis Libraries are Enriched in Functional and Improved Proteins. *J Mol Biol* 350:806–816.

Duplus-Bottin H, Spichty M, Triqueneaux G, Place C, Mangeot PE, Ohlmann T, Vittoz F, Yvert G. 2021. A single-chain and fast-responding light-inducible Cre recombinase as a novel optogenetic switch. *Elife* 10:e61268.

Durrant MG, Fanton A, Tycko J, Hinks M, Chandrasekaran SS, Perry NT, Schaepe J, Du PP, Lotfy P, Bassik MC, Bintu L, Bhatt AS, et al. 2022. Systematic discovery of recombinases for efficient integration of large DNA sequences into the human genome. *Nat Biotechnol* 1–12.

Duyme GDV. 2001. A Structural View of Cre- loxP Site-Specific Recombination. *Annu Rev Bioph Biom* 30:87–104.

English JG, Olsen RHJ, Lansu K, Patel M, White K, Cockrell AS, Singh D, Strachan RT, Wacker D, Roth BL. 2019. VEGAS as a Platform for Facile Directed Evolution in Mammalian Cells. *Cell* 178:748-761.e17.

Ennifar E, Meyer JEW, Buchholz F, Stewart AF, Suck D. 2003. Crystal structure of a wild-type Cre recombinase– lox P synapse reveals a novel spacer conformation suggesting an alternative mechanism for DNA cleavage activation. *Nucleic Acids Res* 31:5449–5460.

Enquist LW, Kikuchi A, Weisberg RA. 1979. The Role of Integrase in Integration and Excision. *Cold Spring Harb Sym* 43:1115–1120.

Esposito D, Scocca JJ. 1997. The integrase family of tyrosine recombinases: evolution of a conserved active site domain. *Nucleic Acids Res* 25:3605–3614.

Feil R. 2007. Conditional Mutagenesis: An Approach to Disease Models. *Handb Exp Pharmacol* 3–28.

Fenno LE, Mattis J, Ramakrishnan C, Hyun M, Lee SY, He M, Tucciarone J, Selimbeyoglu A, Berndt A, Grosenick L, Zalocusky KA, Bernstein H, et al. 2014. Targeting cells with single vectors using multiple-feature Boolean logic. *Nat Methods* 11:763–772.

Gaj T, Mercer AC, Gersbach CA, Gordley RM, Barbas CF. 2011. Structure-guided reprogramming of serine recombinase DNA sequence specificity. *Proc National Acad Sci* 108:498–503.

Gaj T, Mercer AC, Sirk SJ, Smith HL, Barbas CF. 2013. A comprehensive approach to zinc-finger recombinase customization enables genomic targeting in human cells. *Nucleic Acids Res* 41:3937–3946.

Gaudelli NM, Komor AC, Rees HA, Packer MS, Badran AH, Bryson DI, Liu DR. 2017. Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage. *Nature* 551:464–471.

Glaser S, Anastassiadis K, Stewart AF. 2005. Current issues in mouse genome engineering. *Nat Genet* 37:1187–1193.

## BIBLIOGRAPHY

- Goldsmith M, Kiss C, Bradbury ARM, Tawfik DS. 2007. Avoiding and controlling double transformation artifacts. *Protein Eng Des Sel* 20:315–318.
- Gopaul DN, Duyne GDV. 1999. Structure and mechanism in site-specific recombination. *Curr Opin Struc Biol* 9:14–20.
- Grindley NDF, Whiteson KL, Rice PA. 2006. Mechanisms of Site-Specific Recombination\*. *Annu Rev Biochem* 75:567–605.
- Groth AC, Calos MP. 2004. Phage Integrases: Biology and Applications. *J Mol Biol* 335:667–678.
- Guo F, Gopaul DN, Duyne GDV. 1997. Structure of Cre recombinase complexed with DNA in a site-specific recombination synapse. *Nature* 389:40–46.
- Guo F, Gopaul DN, Duyne GDV. 1999. Asymmetric DNA bending in the Cre-loxP site-specific recombination synapse. *Proc National Acad Sci* 96:7143–7148.
- Guzman LM, Belin D, Carson MJ, Beckwith J. 1995. Tight regulation, modulation, and high-level expression by vectors containing the arabinose PBAD promoter. *J Bacteriol* 177:4121–4130.
- Han P, Ma Y, Fu Z, Guo Z, Xie J, Wu Y, Yuan Y. 2021. A DNA Inversion System in Eukaryotes Established via Laboratory Evolution. *Acs Synth Biol* 10:2222–2230.
- Hardy S, Legagneux V, Audic Y, Paillard L. 2010. Reverse genetics in eukaryotes. *Biol Cell* 102:561–580.
- Harrington LB, Burstein D, Chen JS, Paez-Espino D, Ma E, Witte IP, Cofsky JC, Kyrpides NC, Banfield JF, Doudna JA. 2018. Programmed DNA destruction by miniature CRISPR-Cas14 enzymes. *Science* 362:839–842.
- Hendel SJ, Shoulders MD. 2021. Directed evolution in mammalian cells. *Nat Methods* 18:346–357.
- Hoersten J, Ruiz-Gómez G, Lansing F, Rojo-Romanos T, Schmitt LT, Sonntag J, Pisabarro MT, Buchholz F. 2021. Pairing of single mutations yields obligate Cre-type site-specific recombinases. *Nucleic Acids Res* gkab1240-.
- Hoess RH, Wierzbicki A, Abremski K. 1986. The role of the loxP spacer region in P1 site-specific recombination. *Nucleic Acids Res* 14:2287–2300.
- Houdt RV, Leplae R, Lima-Mendez G, Mergeay M, Toussaint A. 2012. Towards a more accurate annotation of tyrosine-based site-specific recombinases in bacterial genomes. *Mobile Dna-uk* 3:6–6.
- Hu JH, Miller SM, Geurts MH, Tang W, Chen L, Sun N, Zeina CM, Gao X, Rees HA, Lin Z, Liu DR. 2018. Evolved Cas9 variants with broad PAM compatibility and high DNA specificity. *Nature* 556:57–63.
- Jin H, Liu K, Zhou B. 2021. Dual recombinases-based genetic lineage tracing for stem cell research with enhanced precision. *Sci China Life Sci* 64:2060–2072.

## BIBLIOGRAPHY

Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. 2012. A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science* 337:816–821.

Justice MJ, Siracusa LD, Stewart AF. 2011. Technical approaches for mouse models of human disease. *Dis Model Mech* 4:305–310.

Karimova M, Abi-Ghanem J, Berger N, Surendranath V, Pisabarro MT, Buchholz F. 2012. Vika/vox, a novel efficient and specific Cre/loxP-like site-specific recombination system. *Nucleic Acids Res* 41:e37–e37.

Karimova M, Baker O, Camgoz A, Naumann R, Buchholz F, Anastassiadis K. 2018. A single reporter mouse line for Vika, Flp, Dre, and Cre-recombination. *Sci Rep-uk* 8:14453.

Karimova M, Splith V, Karpinski J, Pisabarro MT, Buchholz F. 2016. Discovery of Nigri/nox and Panto/pox site-specific recombinase systems facilitates advanced genome engineering. *Sci Rep-uk* 6:30130.

Karpinski J, Hauber I, Chemnitz J, Schäfer C, Paszkowski-Rogacz M, Chakraborty D, Beschorner N, Hofmann-Sieber H, Lange UC, Grundhoff A, Hackmann K, Schrock E, et al. 2016. Directed evolution of a recombinase that excises the provirus of most HIV-1 primary isolates with high specificity. *Nat Biotechnol* 34:401–409.

Karst SM, Ziels RM, Kirkegaard RH, Sørensen EA, McDonald D, Zhu Q, Knight R, Albertsen M. 2021. High-accuracy long-read amplicon sequences using unique molecular identifiers with Nanopore or PacBio sequencing. *Nat Methods* 18:165–169.

Karvelis T, Druteika G, Bigelyte G, Budre K, Zedaveinyte R, Silanskas A, Kazlauskas D, Venclovas Č, Siksnys V. 2021. Transposon-associated TnpB is a programmable RNA-guided DNA endonuclease. *Nature* 599:692–696.

Klenk C, Scrivens M, Niederer A, Shi S, Mueller L, Gersz E, Zauderer M, Smith ES, Strohner R, Plückthun A. 2023. A Vaccinia-based system for directed evolution of GPCRs in mammalian cells. *Nat Commun* 14:1770.

Klompe SE, Vo PLH, Halpin-Healy TS, Sternberg SH. 2019. Transposon-encoded CRISPR-Cas systems direct RNA-guided DNA integration. *Nature* 571:219–225.

Kolb AF. 2002. Genome Engineering Using Site-Specific Recombinases. *Cloning Stem Cells* 4:65–80.

Komor AC, Kim YB, Packer MS, Zuris JA, Liu DR. 2016. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* 533:420–424.

Landy A. 1989. Dynamic, Structural, and Regulatory Aspects of lambda Site-Specific Recombination. *Annu Rev Biochem* 58:913–941.

Langer SJ, Ghafoori AP, Byrd M, Leinwand L. 2002. A genetic screen identifies novel non-compatible loxP sites. *Nucleic Acids Res* 30:3067–3077.

Lansing F, Mukhametzyanova L, Rojo-Romanos T, Iwasawa K, Kimura M, Paszkowski-Rogacz M, Karpinski J, Grass T, Sonntag J, Schneider PM, Günes C, Hoersten J, et al.

## BIBLIOGRAPHY

2022. Correction of a Factor VIII genomic inversion with designer-recombinases. *Nat Commun* 13:422.

Lansing F, Paszkowski-Rogacz M, Schmitt LT, Martin Schneider P, Romanos TR, Sonntag J, Buchholz F. 2019. A heterodimer of evolved designer-recombinases precisely excises a human genomic DNA locus. *Nucleic Acids Res* 48:472–485.

Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, Carey VJ. 2013. Software for Computing and Annotating Genomic Ranges. *Plos Comput Biol* 9:e1003118.

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

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34:3094–3100.

Li K, Wang G, Andersen T, Zhou P, Pu WT. 2014. Optimization of Genome Engineering Approaches with the CRISPR/Cas9 System. *Plos One* 9:e105779.

Liu K, Jin H, Zhou B. 2020. Genetic lineage tracing with multiple DNA recombinases: A user's guide for conducting more precise cell fate mapping studies. *J Biol Chem* 295:6413–6424.

Livet J, Weissman TA, Kang H, Draft RW, Lu J, Bennis RA, Sanes JR, Lichtman JW. 2007. Transgenic strategies for combinatorial expression of fluorescent proteins in the nervous system. *Nature* 450:56–62.

Lloyd JPB, Ly F, Gong P, Pflüger J, Swain T, Pflüger C, Khan MA, Kidd B, Lister R. 2022. Synthetic memory circuits for programmable cell reconfiguration in plants. *Biorxiv* 2022.02.11.480167.

Logie C, Stewart AF. 1995. Ligand-regulated site-specific recombination. *Proc National Acad Sci* 92:5940–5944.

Loo M van der. The stringdist Package for Approximate String Matching. *R Journal*.

Loonstra A, Vooijs M, Beverloo HB, Allak BA, Drunen E van, Kanaar R, Berns A, Jonkers J. 2001. Growth inhibition and DNA damage induced by Cre recombinase in mammalian cells. *Proc National Acad Sci* 98:9209–9214.

Mali P, Yang L, Esvelt KM, Aach J, Guell M, DiCarlo JE, Norville JE, Church GM. 2013. RNA-Guided Human Genome Engineering via Cas9. *Science* 339:823–826.

Mansour SL, Thomas KR, Capecchi MR. 1988. 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.

Meinke G, Bohm A, Hauber J, Pisabarro MT, Buchholz F. 2016. Cre Recombinase and Other Tyrosine Recombinases. *Chem Rev* 116:12785–12820.

Meinke G, Karpinski J, Buchholz F, Bohm A. 2017. Crystal structure of an engineered, HIV-specific recombinase for removal of integrated proviral DNA. *Nucleic Acids Res* 45:gkx603-.

## BIBLIOGRAPHY

- Mercer AC, Gaj T, Fuller RP, Barbas CF. 2012. Chimeric TALE recombinases with programmable DNA sequence specificity. *Nucleic Acids Res* 40:11163–11172.
- Merrick CA, Zhao J, Rosser SJ. 2018. Serine Integrases: Advancing Synthetic Biology. *Acs Synth Biol* 7:299–310.
- Minorikawa S, Nakayama M. 2011. Recombinase-mediated cassette exchange (RMCE) and BAC engineering via VCre/VloxP and SCre/SloxP systems. *Biotechniques* 50:235–246.
- Missirlis PI, Smailus DE, Holt RA. 2006. A high-throughput screen identifying sequence and promiscuity characteristics of the loxP spacer region in Cre-mediated recombination. *Bmc Genomics* 7:73.
- Mullard A. 2018. 2017 FDA drug approvals. *Nat Rev Drug Discov* 17:81–85.
- Muyrers JPP, Zhang Y, Stewart AF. 2001. Techniques: Recombinogenic engineering—new options for cloning and manipulating DNA. *Trends Biochem Sci* 26:325–331.
- Needleman SB, Wunsch CD. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* 48:443–453.
- Nunes-Düby SE, Kwon HJ, Tirumalai RS, Ellenberger T, Landy A. 1998. Similarities and differences among 105 members of the Int family of site-specific recombinases. *Nucleic Acids Res* 26:391–406.
- O’Flaherty R, Bergin A, Flampouri E, Mota LM, Obaidi I, Quigley A, Xie Y, Butler M. 2020. Mammalian cell culture for production of recombinant proteins: A review of the critical steps in their biomanufacturing. *Biotechnol Adv* 43:107552.
- Olorunniji FJ, Rosser SJ, Stark WM. 2016. Site-specific recombinases: molecular machines for the Genetic Revolution. *Biochem J* 473:673–684.
- Onouchi H, Yokoi K, Machida C, Matsuzaki H, Oshima Y, Matsuoka K, Nakamura K, Machida Y. 1991. Operation of an efficient site-specific recombination system of *Zygosaccharomyces rouxii* in tobacco cells. *Nucleic Acids Res* 19:6373–6378.
- Osterwalder M, Galli A, Rosen B, Skarnes WC, Zeller R, Lopez-Rios J. 2010. Dual RMCE for efficient re-engineering of mouse mutant alleles. *Nat Methods* 7:893–895.
- Packer MS, Liu DR. 2015. Methods for the directed evolution of proteins. *Nat Rev Genet* 16:379–394.
- Papapetrou EP, Lee G, Malani N, Setty M, Riviere I, Tirunagari LMS, Kadota K, Roth SL, Giardina P, Viale A, Leslie C, Bushman FD, et al. 2011. Genomic safe harbors permit high  $\beta$ -globin transgene expression in thalassemia induced pluripotent stem cells. *Nat Biotechnol* 29:73–78.
- Pausch P, Al-Shayeb B, Bisom-Rapp E, Tsuchida CA, Li Z, Cress BF, Knott GJ, Jacobsen SE, Banfield JF, Doudna JA. 2020. CRISPR-Cas $\Phi$  from huge phages is a hypercompact genome editor. *Sci New York N Y* 369:333–337.

## BIBLIOGRAPHY

Phillips S, Ramos PV, Veeraraghavan P, Young SM. 2021. VikAD, a Vika site-specific recombinase-based system for efficient and scalable helper-dependent adenovirus production. *Mol Ther Methods Clin Dev* 24:117–126.

Prüfer K, Stenzel U, Dannemann M, Green RE, Lachmann M, Kelso J. 2008. PatMaN: rapid alignment of short sequences to large databases. *Bioinformatics* 24:1530–1531.

Pugach EK, Richmond PA, Azofeifa JG, Dowell RD, Leinwand LA. 2015. Prolonged Cre expression driven by the  $\alpha$ -myosin heavy chain promoter can be cardiotoxic. *J Mol Cell Cardiol* 86:54–61.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26:841–842.

Raymond CS, Soriano P. 2007. High-Efficiency FLP and  $\Phi$ C31 Site-Specific Recombination in Mammalian Cells. *Plos One* 2:e162.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet* 16:276–277.

Rice PA, Mouw KW, Montaña SP, Boocock MR, Rowland S-J, Stark WM. 2010. Orchestrating serine resolvases. *Biochem Soc T* 38:384–387.

Ringrose L, Angrand P, Stewart AF. 1997. The Kw Recombinase, an Integrase from *Kluyveromyces Waltii*. *Eur J Biochem* 248:903–912.

Rognes T, Flouri T, Nichols B, Quince C, Mahé F. 2016. VSEARCH: a versatile open source tool for metagenomics. *Peerj* 4:e2584.

Rouet P, Smih F, Jasin M. 1994. Expression of a site-specific endonuclease stimulates homologous recombination in mammalian cells. *Proc National Acad Sci* 91:6064–6068.

Rybarski JR, Hu K, Hill AM, Wilke CO, Finkelstein IJ. 2021. Metagenomic discovery of CRISPR-associated transposons. *Proc National Acad Sci* 118:e2112279118.

Sadowski PD. 1995. The Flp recombinase of the 2-microns plasmid of *Saccharomyces cerevisiae*. *Prog Nucleic Acid Re* 51:53–91.

Santoro SW, Schultz PG. 2002. Directed evolution of the site specificity of Cre recombinase. *Proc National Acad Sci* 99:4185–4190.

Sarkar I, Hauber I, Hauber J, Buchholz F. 2007. HIV-1 proviral DNA excision using an evolved recombinase. *Sci New York N Y* 316:1912–5.

Sato T, Samori Y, Kobayashi Y. 1990. The *cisA* cistron of *Bacillus subtilis* sporulation gene *spoIVC* encodes a protein homologous to a site-specific recombinase. *J Bacteriol* 172:1092–1098.

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

## BIBLIOGRAPHY

- Saunders TL. 2010. Transgenic Mouse Methods and Protocols. *Methods Mol Biology* 693:103–115.
- Schmidt EE, Taylor DS, Prigge JR, Barnett S, Capecchi MR. 2000. Illegitimate Cre-dependent chromosome rearrangements in transgenic mouse spermatids. *Proc National Acad Sci* 97:13702–13707.
- Schmitt LT, Paszkowski-Rogacz M, Jug F, Buchholz F. 2022. Prediction of designer-recombinases for DNA editing with generative deep learning. *Nat Commun* 13:7966.
- Shagin DA, Barsova EV, Yanushevich YG, Fradkov AF, Lukyanov KA, Labas YA, Semenova TN, Ugalde JA, Meyers A, Nunez JM, Widder EA, Lukyanov SA, et al. 2004. GFP-like Proteins as Ubiquitous Metazoan Superfamily: Evolution of Functional Features and Structural Complexity. *Mol Biol Evol* 21:841–850.
- Sheets MB, Wong WW, Dunlop MJ. 2020. Light-Inducible Recombinases for Bacterial Optogenetics. *Acs Synth Biol* 9:227–235.
- Sherratt DJ, Arciszewska LK, Blakely G, Colloms S, Grant K, Leslie N, McCulloch R. 1995. Site-specific recombination and circular chromosome segregation. *Philosophical Transactions Royal Soc Lond Ser B Biological Sci* 347:37–42.
- Siuti P, Yazbek J, Lu TK. 2013. Synthetic circuits integrating logic and memory in living cells. *Nat Biotechnol* 31:448–452.
- Smyshlyaev G, Bateman A, Barabas O. 2021. Sequence analysis of tyrosine recombinases allows annotation of mobile genetic elements in prokaryotic genomes. *Mol Syst Biol* 17:e9880.
- Snippert HJ, Flier LG van der, Sato T, Es JH van, Born M van den, Kroon-Veenboer C, Barker N, Klein AM, Rheenen J van, Simons BD, Clevers H. 2010. Intestinal Crypt Homeostasis Results from Neutral Competition between Symmetrically Dividing Lgr5 Stem Cells. *Cell* 143:134–144.
- Soni A, Augsburg M, Buchholz F, Pisabarro MT. 2020. Nearest-neighbor amino acids of specificity-determining residues influence the activity of engineered Cre-type recombinases. *Sci Rep-uk* 10:13985.
- Stalker DM, Filutowicz M, Helinski DR. 1983. Release of initiation control by a mutational alteration in the R6K pi protein required for plasmid DNA replication. *Proc National Acad Sci* 80:5500–5504.
- Standage-Beier K, Brookhouser N, Balachandran P, Zhang Q, Brafman DA, Wang X. 2019. RNA-Guided Recombinase-Cas9 Fusion Targets Genomic DNA Deletion and Integration. *Crispr J* 2:209–222.
- Stemmer WPC. 1994. Rapid evolution of a protein *in vitro* by DNA shuffling. *Nature* 370:389–391.
- Sternberg N, Hamilton D. 1981. Bacteriophage P1 site-specific recombination I. Recombination between loxP sites. *J Mol Biol* 150:467–486.



## BIBLIOGRAPHY

- Strecker J, Ladha A, Gardner Z, Schmid-Burgk JL, Makarova KS, Koonin EV, Zhang F. 2019. RNA-guided DNA insertion with CRISPR-associated transposases. *Science* 365:48–53.
- Suzuki E, Nakayama M. 2011. VCre/VloxP and SCre/SloxP: new site-specific recombination systems for genome engineering. *Nucleic Acids Res* 39:e49–e49.
- Takao T, Yamada D, Takarada T. 2022. Mouse Model for Optogenetic Genome Engineering. *Acta Med Okayama* 76:1–5.
- Thomson JG, Rucker EB, Piedrahita JA. 2003. Mutational analysis of loxP sites for efficient Cre-mediated insertion into genomic DNA. *Genesis* 36:162–167.
- Thyagarajan B, Guimarães MJ, Groth AC, Calos MP. 2000. Mammalian genomes contain active recombinase recognition sites. *Gene* 244:47–54.
- Tomoiaga D, Bubnell J, Herndon L, Feinstein P. 2022. High rates of plasmid cotransformation in *E. coli* overturn the clonality myth and reveal colony development. *Sci Rep-uk* 12:11515.
- Turan S, Bode J. 2011. Site-specific recombinases: from tag-and-target- to tag-and-exchange-based genomic modifications. *Faseb J* 25:4088–4107.
- Urnov FD, Rebar EJ, Holmes MC, Zhang HS, Gregory PD. 2010. Genome editing with engineered zinc finger nucleases. *Nat Rev Genet* 11:636–646.
- Vaser R, Sović I, Nagarajan N, Šikić M. 2017. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res* 27:737–746.
- Velappan N, Sblattero D, Chasteen L, Pavlik P, Bradbury ARM. 2007. Plasmid incompatibility: more compatible than previously thought? *Protein Eng Des Sel* 20:309–313.
- Voziyanova E, Li F, Shah R, Voziyanov Y. 2020. Genome targeting by hybrid Flp-TAL recombinases. *Sci Rep-uk* 10:17479.
- Wang H, He L, Li Y, Pu W, Zhang S, Han X, Lui KO, Zhou B. 2022. Dual Cre and Dre recombinases mediate synchronized lineage tracing and cell subset ablation *in vivo*. *J Biol Chem* 101965.
- Wang Z, Xiong G, Lutz F. 1995. Site-specific integration of the phage  $\Phi$ CTX genome into the *Pseudomonas aeruginosa* chromosome: characterization of the functional integrase gene located close to and upstream of attP. *Mol Gen Genetics Mgg* 246:72–79.
- Weinberg BH, Pham NTH, Caraballo LD, Lozanoski T, Engel A, Bhatia S, Wong WW. 2017. Large-scale design of robust genetic circuits with multiple inputs and outputs for mammalian cells. *Nat Biotechnol* 35:453–462.
- Weng W, Liu X, Lui KO, Zhou B. 2021. Harnessing orthogonal recombinases to decipher cell fate with enhanced precision. *Trends Cell Biol* 32:324–337.
- Wong TS, Roccatano D, Zacharias M, Schwaneberg U. 2006. A Statistical Analysis of Random Mutagenesis Methods Used for Directed Protein Evolution. *J Mol Biol* 355:858–871.

## BIBLIOGRAPHY

Xu X, Chemparathy A, Zeng L, Kempton HR, Shang S, Nakamura M, Qi LS. 2021. Engineered miniature CRISPR-Cas system for mammalian genome regulation and editing. *Mol Cell* 81:4333-4345.e4.

Xu X, Wagner K-U, Larson D, Weaver Z, Li C, Ried T, Hennighausen L, Wynshaw-Boris A, Deng C-X. 1999. Conditional mutation of *Brca1* in mammary epithelial cells results in blunted ductal morphogenesis and tumour formation. *Nat Genet* 22:37–43.

Yang SH, Jayaram M. 1994. Generality of the shared active site among yeast family site-specific recombinases. The R site-specific recombinase follows the FIp paradigm [corrected]. *J Biological Chem* 269:12789–96.

Yarnall MTN, Ioannidi EI, Schmitt-Ulms C, Krajeski RN, Lim J, Villiger L, Zhou W, Jiang K, Garushyants SK, Roberts N, Zhang L, Vakulskas CA, et al. 2022. Drag-and-drop genome insertion of large sequences without double-strand DNA cleavage using CRISPR-directed integrases. *Nat Biotechnol* 1–13.

Yoshimura Y, Ida-Tanaka M, Hiramaki T, Goto M, Kamisako T, Eto T, Yagoto M, Kawai K, Takahashi T, Nakayama M, Ito M. 2018. Novel reporter and deleter mouse strains generated using VCre/VloxP and SCre/SloxP systems, and their system specificity in mice. *Transgenic Res* 27:193–201.

Zhang M, Yang C, Tasan I, Zhao H. 2021. Expanding the Potential of Mammalian Genome Engineering via Targeted DNA Integration. *Acs Synth Biol*.

Zurek PJ, Knyphausen P, Neufeld K, Pushpanath A, Hollfelder F. 2020. UMI-linked consensus sequencing enables phylogenetic analysis of directed evolution. *Nat Commun* 11:6023.

## BIBLIOGRAPHY

## ACKNOWLEDGMENTS

### ACKNOWLEDGMENTS

This PhD was an exciting, sometimes rocky journey full of ups and downs. It shaped me as a researcher and as a person. 'It takes a village' they say, and I couldn't have done it without all of the support of the people around me. Thank you all!

First and foremost, I would like to express my deepest gratitude to my advisor, **Frank Buchholz**, for his unwavering support, guidance, and mentorship throughout my PhD journey. Without his invaluable advice, encouragement, and patience, this thesis would not have been possible. His understanding of my potential even in times when I was doubting myself really kept me motivated all the time.

I am also grateful to the members of my thesis advisory committee, **Francis Stewart**, **Michael Schlierf**, and **Michal Sarov**, for their insightful feedback, constructive criticism, and commitment to my academic and personal growth over these four years. Their expertise and encouragement have been instrumental in shaping my research and helping me achieve my goals.

Furthermore, I would like to extend my appreciation to my fellow lab members and colleagues in the Buchholz lab for their camaraderie, collaboration, and the intellectually stimulating environment they have created. A very special thanks to **Duran Sürün**, who adopted me under his wing and gave me continuous support and friendship. The countless hours spent discussing our research and life in general, sharing ideas and dreaming about our institute really made my days in the lab, and shaped me as an independent researcher. Also, thank you for keeping your service fees at 5 euros all this time, despite the huge inflation. Otherwise, I would have to say goodbye to two of my postdoc salaries instead of this one 😊. I further extend my gratitude to **Manavi** and **Jonas** for all the laughs, compliments, shamelessly stolen racks, good music and hugs! It is a pleasure working side-by-side with you! I would also like to thank **Angelika Walder**, who accepted me as her supervisor during her master thesis work. I am really grateful that you allowed me to practice my teaching skills on you, and I hope you enjoyed our time together as much as I did! I am very proud with how much you accomplished these months! Big thanks to **Nadja Schubert**, for all the travel tips, but also amazing western blot skills she shared with me and Angie when we needed it! A unique thanks goes to **Jovan Mircetic**, for being a small, but highly significant, piece of home to me in the lab and in Dresden in general. I appreciate you so much! I further extend my gratitude to **Lukas Schmitt** for being my bioinformatic tandem in the lab from the beginning! I am very grateful that Frank's visionary thinking brought us to

## ACKNOWLEDGMENTS

work together, this whole experience wouldn't be same without you, our long discussions, jokes, your ideas and critical thinking. I extend my gratitude also to **Maciej**, who helped me analyze a lot of my data! Big thanks to all the former and current 'Recombinators', especially: **Liliya** – for being a good friend, always eager to share a few words in the hallways; **Teresa** – for all the fun facts I learned from you and for being the best Mejillones sous-chef; **Felix** – for being so exaggerated and always helpful with insightful advice and finally, **Jenna** – for helping me improve my illustrator skills, for talking so much when we run, for being my favorite cartoon character and being one of my closest friends in and out of the lab. I extend my gratitude to **Moustafa** my night shift buddy, and also **Shady** for his warm and welcoming personality. Last but not least I would like to thank **Mandy Erlitz**, **Ira Ilgen**, **Martina Augsburg** and **Sebastian Rose** for keeping the lab running and for making my time at the lab a pleasant and safe experience.

My heartfelt thanks go to my family, especially my parents, **Nenad** and **Gabrijela**, for their love, understanding, and unwavering belief in my abilities. Thank you for teaching me how to be independent, resourceful and persistent. Your support, and encouragement have been my source of strength throughout this journey. I am also grateful to my sister, **Joca** for her constant love, trust and support. Additionally, I want to acknowledge the huge role my uncle (**Pera**) and my aunt (**Rada**) had in my life as a second pair of parents 😊. Your emotional support, parental love and advice are very valuable to me and something that I will always cherish. I extend my love to my cousins, **Lela** and **Nena** for being like sisters to me. Even if not officially family yet, I want to thank **Goca** and **Rale**, for making me feel like part of their family since the beginning, with all their love and understanding!

I am forever indebted to all of my friends, with whom I shared most cherished moments of my life, and without whose emotional support and moments of laughter I wouldn't be who I am today. Thank you: **Kale** (for being my best friend and biggest comfort), **Tišma** (for admitting that I am the funniest person you know, and for being my motivation through enormous peer pressure), **Boco** (for being so proud of me) , **Jano** (for being my biggest fan since childhood), **Džoni** (for being Moj Debil), **Saro** (for never giving up on me), **Nedeljka**, **Đ**, **Pajo**, **Uroše** – for friendship that lasts and all skiing, vacations and life ups and downs we've been through together, **Elisa** (for being my greatest presence, and the best wifey), **Vasanth** (my kick-ass training partner), **Karen & Diego** (for all the talks, sushi, games and wall kisses shared together), **Irina** (for the best ever first year in Dresden), **Caro** (for all the support, Sunday brunches and naps), **Laura**, **Ivan**, **Roberto**, **Luka**, **Marina**, **Fede** – for all the funny and caring moments we shared during our time in Dresden.

## ACKNOWLEDGMENTS

Finally, an enormous thank you to my **Jankić**, my soulmate, the most exciting person in every room, now and forever. My appreciation and love for you transcend the limits of this universe. Working theory is that is why the space is expanding, 5/5 scientists confirmed so. Thank you for all your love, encouragement and support that gave me strength to overcome all doubts, fears and difficult moments in professional and personal life. Thank you for making our team awesome, for nurturing my inner child, making me laugh and inspiring me to be the best version of myself.

In closing, I dedicate this thesis to everyone who has played a role in my academic journey, both directly and indirectly. Your support and encouragement have been instrumental in helping me reach this milestone, and I will always be grateful for your contributions.

# Anlage 1

Technische Universität Dresden

Medizinische Fakultät Carl Gustav Carus

Promotionsordnung vom 24. Juli 2011

## Erklärungen zur Eröffnung des Promotionsverfahrens

1. Hiermit versichere ich, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe; die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht.

2. Bei der Auswahl und Auswertung des Materials sowie bei der Herstellung des Manuskripts habe ich Unterstützungsleistungen von folgenden Personen erhalten:

3. Weitere Personen waren an der geistigen Herstellung der vorliegenden Arbeit nicht beteiligt. Insbesondere habe ich nicht die Hilfe eines kommerziellen Promotionsberaters in Anspruch genommen. Dritte haben von mir weder unmittelbar noch mittelbar geldwerte Leistungen für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen.

4. Die Arbeit wurde bisher weder im Inland noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt.

5. Die Inhalte dieser Dissertation wurden in folgender Form veröffentlicht:

*Jelicic M, Schmitt LT, Paszkowski-Rogacz M, Walder A, et al. 2023. Discovery and characterization of novel Cre-type tyrosine site-specific recombinases for advanced genome engineering. Nucleic Acid Res. Future DOI: 10.1093/nar/gkad366 – accepted to be published*

*Jelicic M, Buchholz F, Schmitt LT (2023). Novel recombinase enzymes for site-specific DNA-recombination. Patent pending; Application No. 23168351.7*

6. Ich bestätige, dass es keine zurückliegenden erfolglosen Promotionsverfahren gab.

7. Ich bestätige, dass ich die Promotionsordnung der Medizinischen Fakultät der Technischen Universität Dresden anerkenne.

8. Ich habe die Zitierrichtlinien für Dissertationen an der Medizinischen Fakultät der Technischen Universität Dresden zur Kenntnis genommen und befolgt.

Dresden, Datum

Milica Jelicic

## Anlage 2

Hiermit bestätige ich die Einhaltung der folgenden aktuellen gesetzlichen Vorgaben im Rahmen meiner Dissertation

- das zustimmende Votum der Ethikkommission bei Klinischen Studien, epidemiologischen Untersuchungen mit Personenbezug oder Sachverhalten, die das Medizinproduktegesetz betreffen: **entfällt**
- die Einhaltung der Bestimmungen des Tierschutzgesetzes  
*Aktenzeichen der Genehmigungsbehörde zum Vorhaben/zur Mitwirkung: entfällt*
- die Einhaltung des Gentechnikgesetzes *Projektnummer:*  
AZ 54-8451/197 (S1)  
AZ 54-8452/89 (S2)
- die Einhaltung von Datenschutzbestimmungen der Medizinischen Fakultät und des Universitätsklinikums Carl Gustav Carus.

Dresden, den

Milica Jelacic