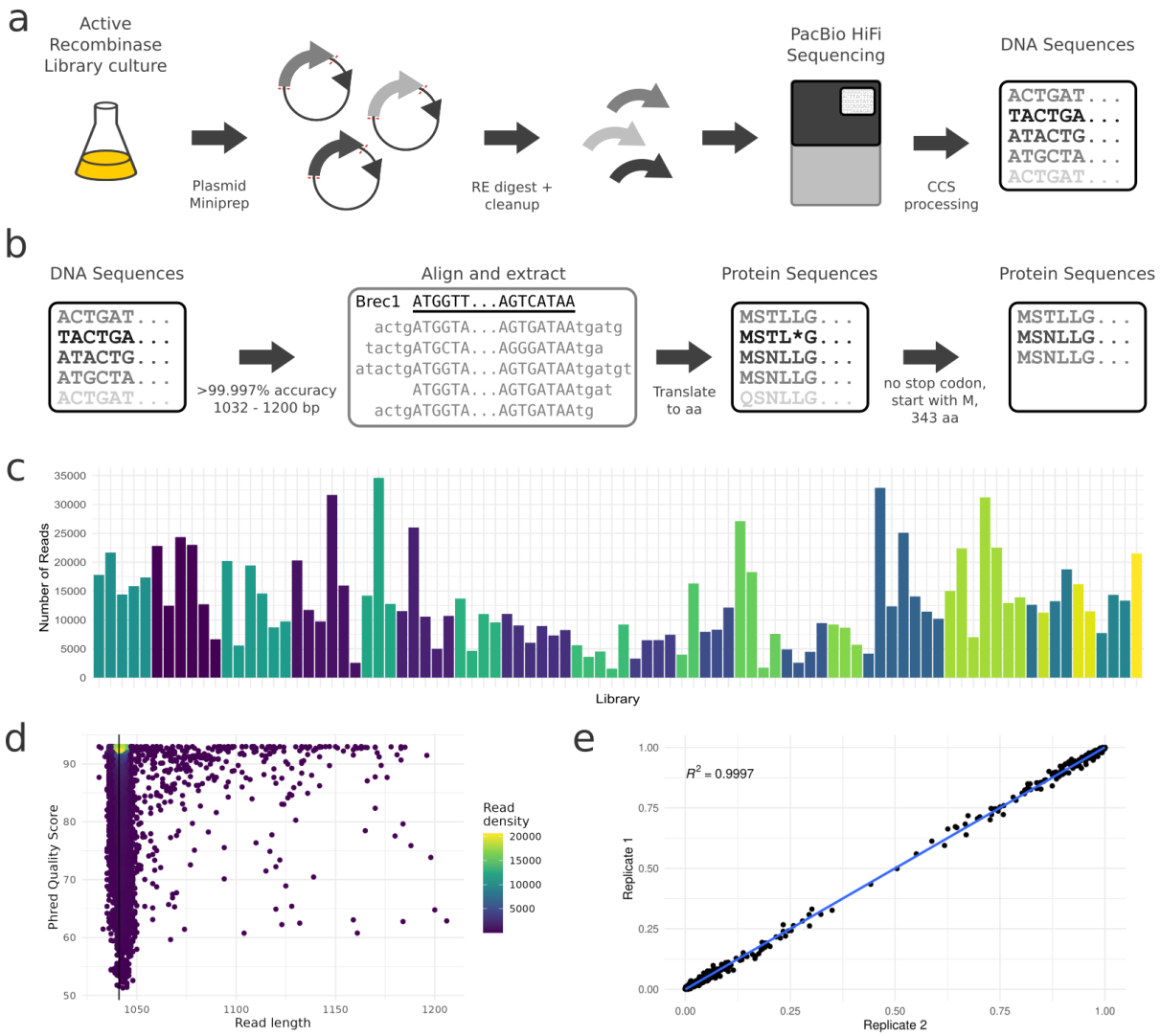
Supplementary Information

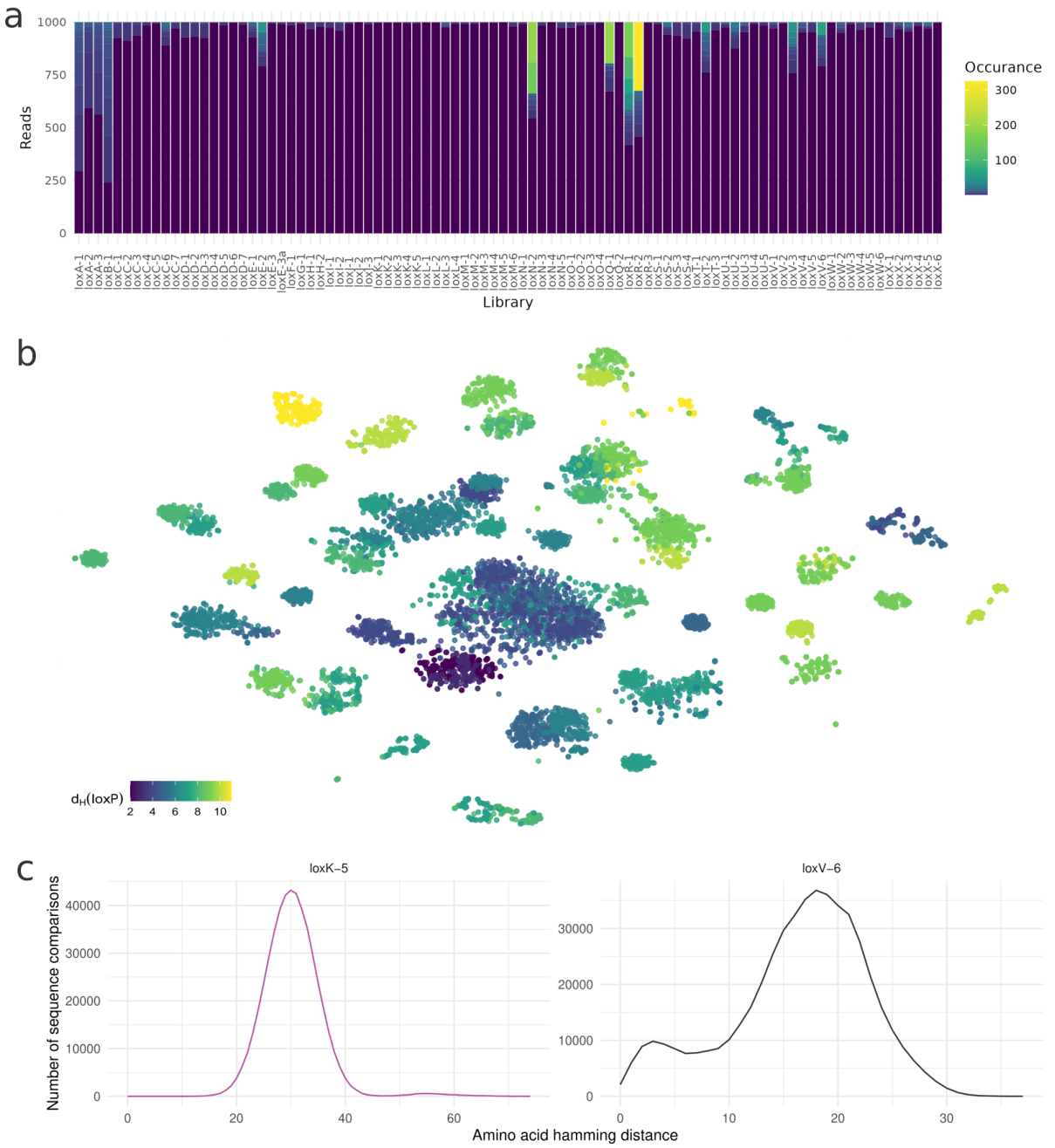
Prediction of designer-recombinases for DNA editing with generative deep learning

Authors: Lukas Theo Schmitt [1], Maciej Paszkowski-Rogacz [1], Florian Jug [2,3,4], Frank Buchholz [1]*

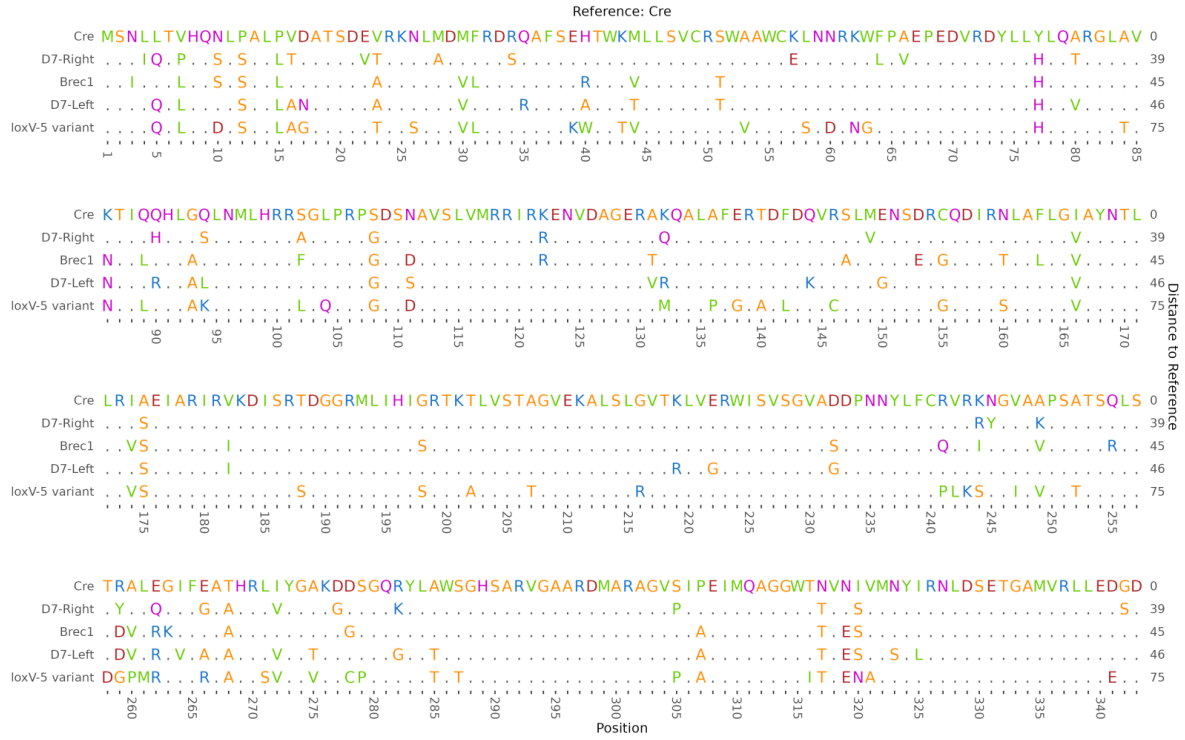
[1] Medical Systems Biology, Medical Faculty, Technical University Dresden, 01307 Dresden, Germany
[2] Fondazione Human Technopole, Milano, Italy
[3] Center for Systems Biology Dresden, Dresden, Germany
[4] Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany
* Corresponding Author: frank.buchholz@tu-dresden.de



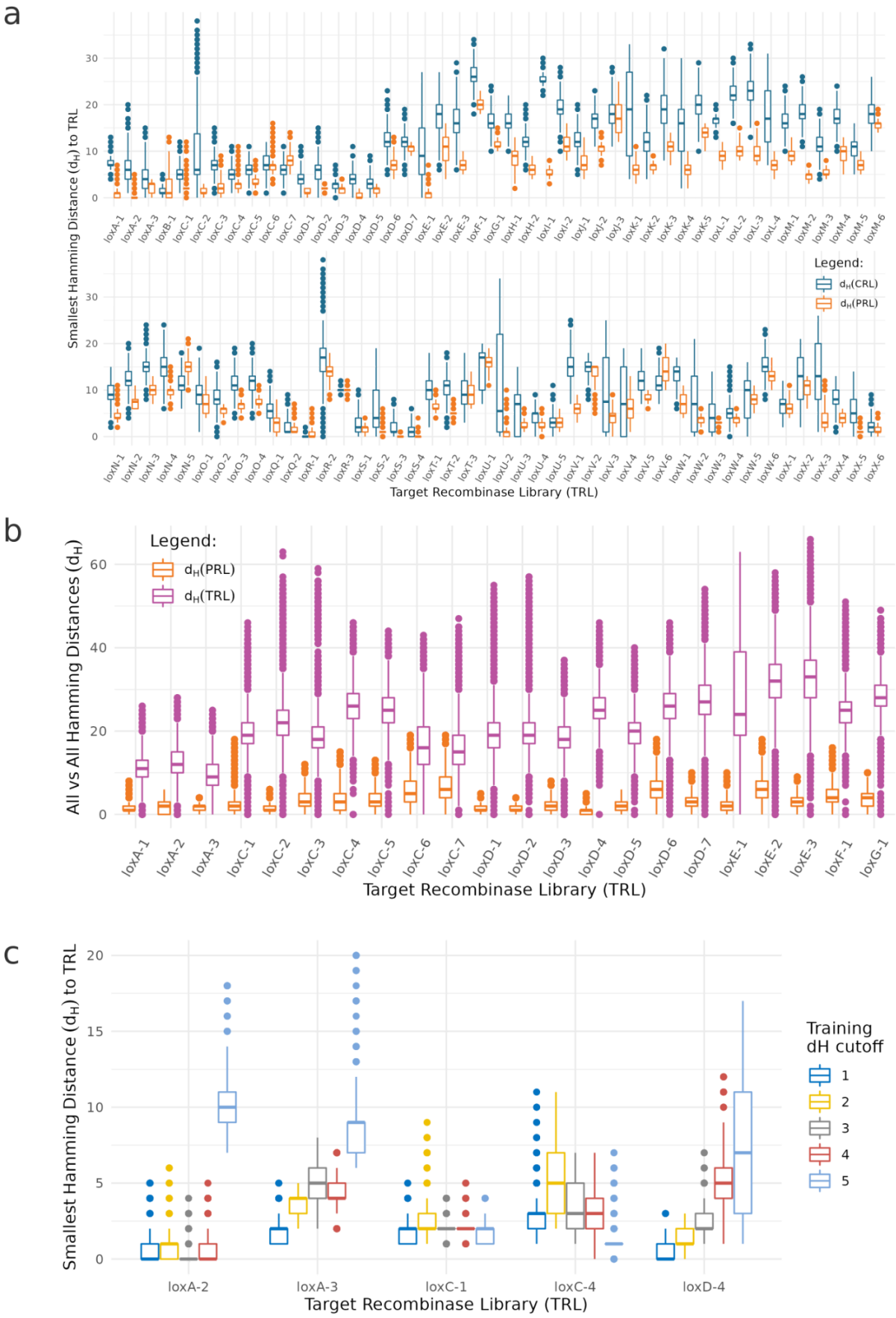
Supplementary Figure 1. Sequencing pipeline scheme and sequence data properties. **a** Recombinase library sequencing preparations. Plasmids encoding active recombinases are extracted from *E. coli* culture. Recombinase genes are then cut by restriction enzyme digestion and enriched by SPRI-bead cleanup. Sequencing of the genes is then performed on a PacBio Sequel with the HiFi sequencing approach. The acquired subreads were then processed with ccs to generate high quality consensus sequences. **b** Processing of sequencing data. DNA sequences are filtered for accuracy and length. Gene sequences are then isolated by alignment to Brec1 [1] and converted to amino acid. Finally, resulting protein sequences are filtered for correct length, sequences that start with methionine, and sequences that do not contain a stop codon in the middle of the sequence. **c** Total number of recombinase sequences acquired per Library, colored by evolution project. **d** Example analysis of loxE-3 DNA read length and read quality acquired by PacBio HiFi sequencing. Calculated sequencing quality is beyond the required 99.997 % accuracy (Phred Score of ~25). **e** Correlation of repeated sequencing of loxE-1. Replicate 1 and Replicate 2 were prepared from independent plasmid extractions from separate cultures and sequenced on different sequencing runs. Each point represents the observed frequencies of the amino acids on each residue position of the proteins. Replicate sequences are highly correlated, R^2 was calculated as 0.9997 using R 4.1.1. Source data are provided as a Source Data file.



Supplementary Figure 2. Sequence data composition. **a** Number of sequence reads occurring multiple times per library. The color code depicts sequence occurrence. Note the high diversity of clones within most libraries. **b** t-SNE dimensionality reduction of 100 recombinase sequences from all sequenced libraries. Color indicates the base pair hamming distance (dH(loxP)) of the associated target sites to the loxP sequence. **c** All vs. all amino acid sequence hamming distances of loxK-5 and loxV-6 libraries. Comparisons were made with 1000 sequences per library. Source data are provided as a Source Data file.

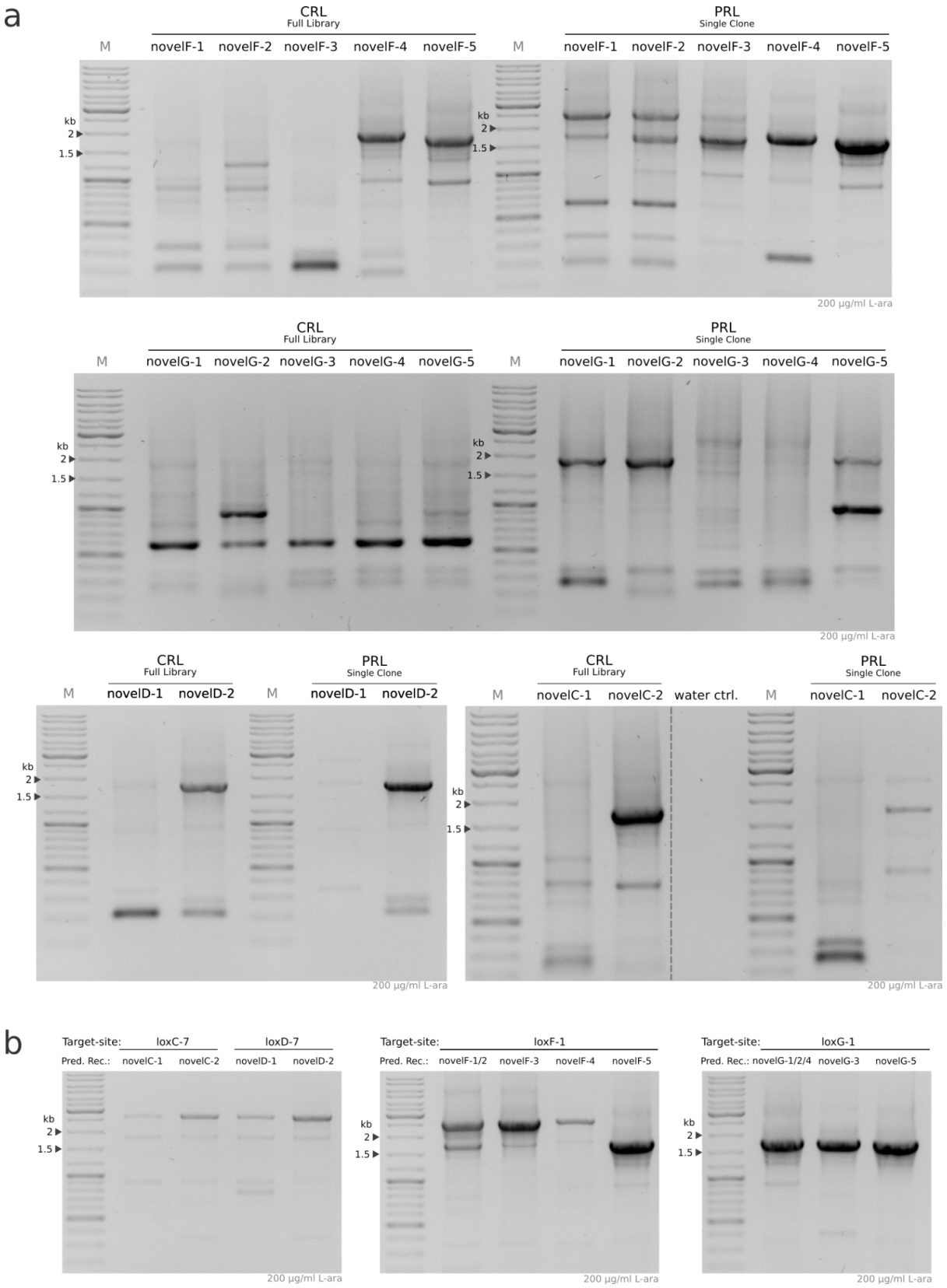


Supplementary Figure 3. Alignment of recombinase sequence with highest observed difference to Cre. D7-Left, D7-Right [2] and Brec1 [1] are included for comparison.



Supplementary Figure 4. a Leave-one-out cross-validation (LOOCV) results for all target recombinase libraries (TRL). Boxplots of the smallest hamming distances of each LOOCV

predicted sequence to all TRL sequences are shown. Additionally, the closest recombinase library (CRL) with the smallest hamming distance to TRL is included for comparison. n = 1000 per condition. **b** Sequence variance of selected TRL and LOOCV predicted recombinase library (PRL). Variance of the sequences was determined by all vs. all sequences comparison. Distance is measured in hamming of the amino acid sequences. TRL has a much higher variance than PRL in all libraries. n = 1000000 per condition. **c** LOOCV prediction of selected TRL with increasing number of neighboring libraries removed from training data. Training dH cutoff defines the minimum base pair distance to the TRL target site still allowed in the training data. Increasing the cutoff value causes the PRL to be less similar to the TRL in some samples, while others are less affected. n = 1000 per condition. All boxplots in a, b, and c are according to standard definition: median for the center line, upper and lower quartiles for the box limits, 1.5x interquartile range for the whiskers, and the points show outliers. Source data are provided as a Source Data file.



Supplementary Figure 5. Untrimmed gel images of the predicted recombinase PCR-assays (as shown in Fig. 4B). **a** Activity of closest recombinase library (CRL) and predicted recombinase library (PRL) single clones on the indicated target sites are shown. “M” indicates the DNA marker. **b** Activity of predicted recombinases on the target site of the CRL

target site. The expected fragment size is 1.7 kb. A correct recombination product could be confirmed by sequencing for novelF-5, novelG-1/2/4, novelG-3, and novelG-5. Source data are provided as a Source Data file.

Supplementary Methods

Depletion of plasmid backbone with SPRI-beads

We followed dx.doi.org/10.17504/protocols.io.n7hdhj6 to make custom SPRI-beads with Tween. However, instead of using 0.4% Sera-Mag SpeedBeads Carboxyl Magnetic Beads (GE Healthcare) we used 2%. Additionally, we also made a concentrated SPRI-beads mix. We took 50ul Sera-Mag beads and washed them 3 times with nuclease free water, followed by complete removal of the liquid. To this we added a version of the custom SPRI-bead mix without the Sera-Mag beads and nuclease free water so that a mixture of beadless custom SPRI-bead mix and water with Sera-Mag beads equals 0.75x to 1 in a total volume of 50 µl.

To deplete the pEVO plasmid backbone, a fragment of ~5 kb size, we mix 0.75x custom SPRI-bead mix with 1x of the XbaI + BsrgI digested pEVO. After 15 minutes of incubation we placed the mixture on a magnet and transferred the supernatant to a new microcentrifuge tube. To this new tube we then mixed in 2 µl of the concentrated beads and incubated for 3 minutes. The supernatant is then transferred again and another 2 µl of concentrated beads is added. The liquid at this stage almost exclusively contained the recombinase fragments. We transferred the liquid to another tube and mixed in 0.35x Ampure XP (Beckman Coulter) relative to the volume of our supernatant. After incubating for 15 min we performed a normal Ampure XP bead cleanup with 2x 75% EtOH wash and an elution with 15 µl nuclease free water.

Conditional Variational Autoencoders

The aim of VAE is the inference of the characteristics of *z* (the latent space) based on *x* in other words *P(z|X)*. To get *P(z|X)* we can use Bayes theorem (Eq. 1).

$$P(z|X) = \frac{P(X|z)P(z)}{P(X)} \quad (1)$$

However, computing *P(x)* is intractable (Eq. 2).

$$P(X) = \int P(X|z)P(z)dz \quad (2)$$

Instead *P(z|X)* is estimated via variational inference. The parameters of *Q(z|X)* (the encoder) can be defined so that it is similar to *P(z|X)*. It can be used to perform approximate inference

of the intractable distribution. To make sure that *Q(z|X)* is similar to *P(X|z)* Kullback–Leibler divergence of the two distributions is minimized (Eq. 3).

$$logP(X) - D_{KL}[Q(z|X)||P(z|X)] = E[logP(X|z)] - D_{KL}[Q(z|X)||P(z)] \quad (3)$$

To adapt the objective to a Conditional Variational Autoencoder we introduce the condition *c* to the encoder: *Q(z|X,c)* and the decoder: *P(X|z,c)* (Eq. 4).

$$logP(X|c) - D_{KL}[Q(z|X,c)||P(z|X,c)] = E[logP(X|z,c)] - D_{KL}[Q(z|X,c)||P(z|c)] \quad (4)$$

Additionally to this objective a binary cross entropy loss function is used to penalize reconstruction error (Eq. 5).

$$Loss = -\frac{1}{n}\sum_{i=1}^{n} y_i \cdot log\,\hat{y}_i + (1 - y_i) \cdot log(1 - \hat{y}_i) \quad (5)$$

Where $\hat{y}_i$ is the output value at position *i,* y_i is the desired output and *n* is the size of the output.

References

1. Karpinski, J. *et al.* Directed evolution of a recombinase that excises the provirus of most HIV-1 primary isolates with high specificity. *Nature Biotechnology* **34**, 401–409 (2016).
2. Lansing, F. *et al.* Correction of a Factor VIII genomic inversion with designer-recombinases. *Nature Communications* **13**, (2022).