Published online 11 November 2014 *Nucleic Acids Research, 2014, Vol. 42, No. 22* ***13839–13852***
doi: 10.1093/nar/gku1096

Massively parallel determination and modeling of endonuclease substrate specificity

Summer B. Thyme[1,*], Yifan Song[2], T. J. Brunette[2], Mindy D. Szeto[2], Lara Kusak[2], Philip Bradley[3] and David Baker[2,4,*]

[1]Harvard University, 16 Divinity Avenue, Cambridge, MA 02138, USA, [2]Department of Biochemistry, University of Washington, Seattle, WA 98195, USA, [3]Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue, Seattle, WA 98109, USA and [4]Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195, USA

Received April 24, 2014; Revised October 20, 2014; Accepted October 21, 2014

ABSTRACT

We describe the identification and characterization of novel homing endonucleases using genome database mining to identify putative target sites, followed by high throughput activity screening in a bacterial selection system. We characterized the substrate specificity and kinetics of these endonucleases by monitoring DNA cleavage events with deep sequencing. The endonuclease specificities revealed by these experiments can be partially recapitulated using 3D structure-based computational models. Analysis of these models together with genome sequence data provide insights into how alternative endonuclease specificities were generated during natural evolution.

INTRODUCTION

Homing endonucleases (also termed 'Meganucleases') are a family of enzymes that generate double-stranded DNA breaks (1) and are found in the genomes of a wide variety of organisms, such as fungi, algae, bacteria and archaea. They are encoded by mobile elements that typically correspond to an intron or intein that contains their own coding sequence (1). Of the many known types of homing endonucleases, the LAGLIDADG family has been used by several groups for genome engineering. There has been some level of success using both computational design and directed evolution to alter the specificity of these enzymes (2–8), but no approach has proven reliable enough to engineer an endonuclease for any target DNA sequence of interest. A possible strategy to increase the potential of these enzymes for gene targeting is to identify and characterize as many novel members of the LAGLIDADG family, along with their DNA target sites, as possible. That process, however, has represented a very labor-intensive investment of time and resources for each endonuclease being studied.

Putative native target sites of these enzymes can often be identified by analysis of the nucleotide sequences that flank the mobile element containing the endonuclease gene (9–12). However, the substrate specificity of these enzymes and how their protein sequences confer this specificity is not clear. For example, homing endonuclease target preferences that are not dependent upon direct protein–DNA interactions have been reported at certain positions in their target sites; these preferences are thought to arise from DNA bending required for catalysis. However, the drivers of this indirect readout are not well understood (13,14).

Previously, we carried out standard DNA cleavage assays to collect kinetic data on each single base-pair substitution in the target site of the I-AniI homing endonuclease and found that distinct interface domains function in ground-state and transition-state formation during the reaction (2). The approach required extensive experimental effort, and data on this single enzyme did not uncover the biophysical basis behind this segregation of target-site regions. Developing a more complete understanding of how interface residues participate in the cleavage reaction is an important step in increasing the success rate of engineering.

Deep sequencing has revolutionized genomics and human disease research, and has also recently begun to transform the study of how proteins evolve and interact with each other and with other biomolecules (15–18). Such high-throughput methods are well established for profiling DNA binding specificities (19–23), but substrate binding and catalysis are not always tightly correlated with one another (2). Approaches have recently been published for using deep sequencing to profile DNA cleavage specificity

*To whom correspondence should be addressed. Tel: +1 951 204 4067; Email: sthyme@gmail.com
Correspondence may also be addressed to David Baker. Tel: +1 206 543 1295; Email: dabaker@u.washington.edu
Present addresses:
Summer B. Thyme, Harvard University, 16 Divinity Avenue, BIOL 1020, Cambridge, MA 02138, USA.
David Baker, University of Washington, Molecular Engineering Building, 4th Floor, 4000 15th Ave. NE, Seattle, WA 98195, USA.

(24), but they have so far only been tested on a small scale. High-throughput methods are necessary for assaying the large numbers of native endonucleases or engineered variants needed to assess and guide improvements to computational methods for predicting specificity.

Here we integrate genomic database mining, high-throughput screening and computational modeling to identify and characterize new homing endonucleases, and develop a deep-sequencing approach for high-throughput profiling of endonuclease–substrate interactions. Using homology models of the newly characterized endonucleases, corroborated by experimental data and binding energy calculations, we relate interface interactions to target-site preferences. The method presented here enables assessment of the specificity and kinetic properties of many DNA-cleaving enzymes with minimal effort, which should greatly facilitate understanding of these endonucleases and improvement of computational models.

MATERIALS AND METHODS

Identifying endonucleases and predicting target sites

A program was developed to generate a database of homing endonuclease genes and DNA sequences predicted to contain the endonuclease cleavage site. The database and source code are available in a public github repository: https://github.com/tjbrunette/endonuclease.

Prospective homing endonucleases were identified (Figure 1) using two rounds of Position-Specific Iterative Basic Local Alignment Search Tool (PSI-BLAST) (25) starting with 1263 proteins labeled as LAGLIDADG endonucleases in the Genbank (26) and Refseq (27) databases and the previously crystallized homing endonucleases I-Vdi141I (28), I-SceI (29), I-OnuI (4), I-MsoI (30), I-LtrI (4), I-DmoI (31), I-CreI (30), I-CeuI (32) and I-AniI (33,34). This initial search resulted in 813,747 prospective endonucleases, many of which were likely not endonucleases due to the permissive nature of two rounds of Basic Local Alignment Search Tool (BLAST) with an e-value of 1e-5. These prospective endonucleases were filtered using HHsearch (35) to those that have 50% probability of being homologous to an endonuclease with known structure; duplicate sequences were removed at this point. HHsearch uses predicted secondary structure and sequence similarity to match distant homologs, making it more accurate than BLAST. Out of the prospective endonucleases, 8255 were recognized as unique homing endonucleases.

For the identified endonucleases, flanking DNA and intron–exon boundary annotations were then extracted from the Genbank ftp://ftp.ncbi.nih.gov/genbank and Refseq ftp://ftp.ncbi.nlm.nih.gov/refseq/release/complete/ databases. The versions of these two databases that were used contained only a subset of the sequences found from the BLAST search, resulting in identification of the flanking region (that contain potential target sites) for 2059 of the 8255 endonucleases.

The putative target site region (30 base pairs on each side of the intron containing the endonuclease gene) could be unambiguously identified for 384 endonucleases, based on complete annotations of the exon–intron boundaries. Target sites for the remaining endonucleases were predicted

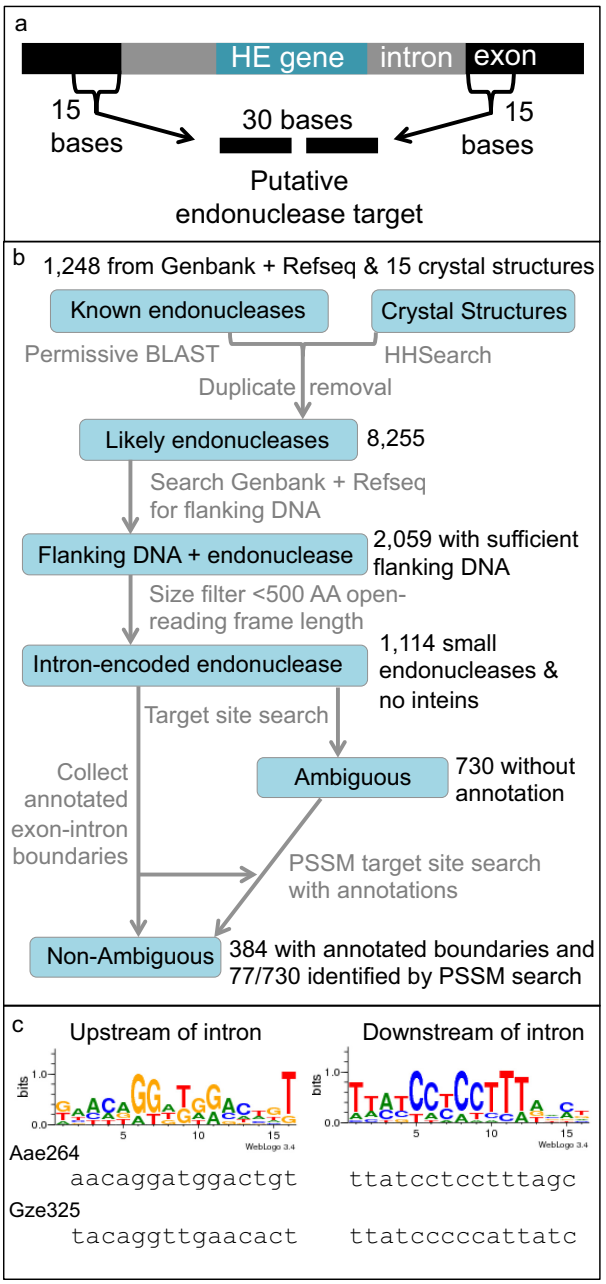

Figure 1. Mining genomic databases for endonuclease target sites. (**a**) Schematic of intron-encoded homing endonuclease (HE) genes and associated putative target-site regions. The target of an intron-encoded LAGLIDADG endonuclease is typically 20 base pairs in length and is likely contained within the 30 base-pair region assembled from the 15 base pairs on each side of the intron. (**b**) Protocol used to collect HE genes and putative target sequences. Many endonucleases reside in introns with clearly annotated boundaries thought to contain the putative targets. Using this information, the PSSM search program identified additional endonuclease-target pairs in the ambiguous classification (without clearly annotated boundaries) by searching the DNA sequence surrounding the endonuclease for similar target sequences. (**c**) Endonucleases were clustered by protein sequence identity and the clusters were found to contain similar predicted target sites. The site for the low-surviving Gze325 endonuclease (Supplementary Table S1) was not identified by boundary annotations and was considered ambiguous. Using the PSSM search, a putative site was determined for Gze325 and this protein was also matched with 12 similar endonucleases, including the highly active Aae264, by protein sequence clustering.

by searching the DNA sequence flanking the endonuclease gene with the putative target site of the most similar endonuclease from the group that was unambiguously determined. A previously described position-specific scoring matrix (PSSM) program was used to predict the site (3,14). Cut-sites were identified for an additional 77 endonucleases, based on criteria of the cut-sites matching ≥13 nucleotides out of 15 in one of the two exonic boundaries and a BLAST e-value of $10e^{-40}$ to the non-ambiguous homolog. Of the remaining endonucleases, 653 did not clearly cluster with a non-ambiguous endonuclease with these stringent criteria, and 945 were either inteins or endonucleases larger than 500 residues.

Endonucleases with accurately identified cut-sites have been clustered and target-site logos generated with WebLogo (36,37). Clustering was done using k-means clustering in NumPy (38) with distances measured by Clustal W (39). For experimentally tested endonucleases that did not automatically cluster, due to the inability of our program to parse all types of boundary annotation in the Genbank and Refseq, putative target-site sequences were compared to other highly homologous endonucleases and were clustered manually if both target site and protein sequences were similar.

Characterizing activity of putative endonucleases

All reagents, methods and vectors—the bacterial selection plasmids pENDO-HE and pCcdB and the His-tagged protein expression vector pET15-HE—are previously described (14). The LAGLIDADG endonuclease genes, sequences available in the supplement, were assembled from oligonucleotides and codon-optimized for expression in *Escherichia coli* (40). A previously described bacterial screen (Figure 2a) (14,41,42) was used to characterize the activity of these endonucleases. In brief, the pCcdB plasmid contains arrays of predicted target sites for the putative endonucleases, and encodes a toxin that is expressed (resulting in cell death) if not cut by a corresponding active nuclease. DH12S *E. coli* (Invitrogen) containing this plasmid were transformed with endonuclease expression constructs (pENDO-HE). Two bacterial lines were used, each containing approximately half the putative target sites on pCcdB (Supplementary Table S1). A single round of selection was completed, followed by collection of the plasmids and re-transformation to obtain more accurate values for bacteria survival.

Endonuclease genes were transferred from the pENDO-HE plasmid into the pET15-HE plasmid for protein expression. To facilitate expression, maltose-binding protein (MBP) with an N-terminal His-tag was fused upstream of each endonuclease. The fusion sequence and all additional sequence modifications are detailed in the supplemental information. Proteins were expressed in BL21 Star cells (Invitrogen) using a half-liter of media and autoinduction (43) and purified with nickel affinity chromatography. Proteins were stored in 20 mM Tris, pH 7.5, 500 mM NaCl, 50% (v/v) glycerol; purity was assessed with sodiumdodecyl sulphate-polyacrylamide gel electrophoresis, and the concentration was determined by absorbance at 280 nm collected with a NanoDrop.

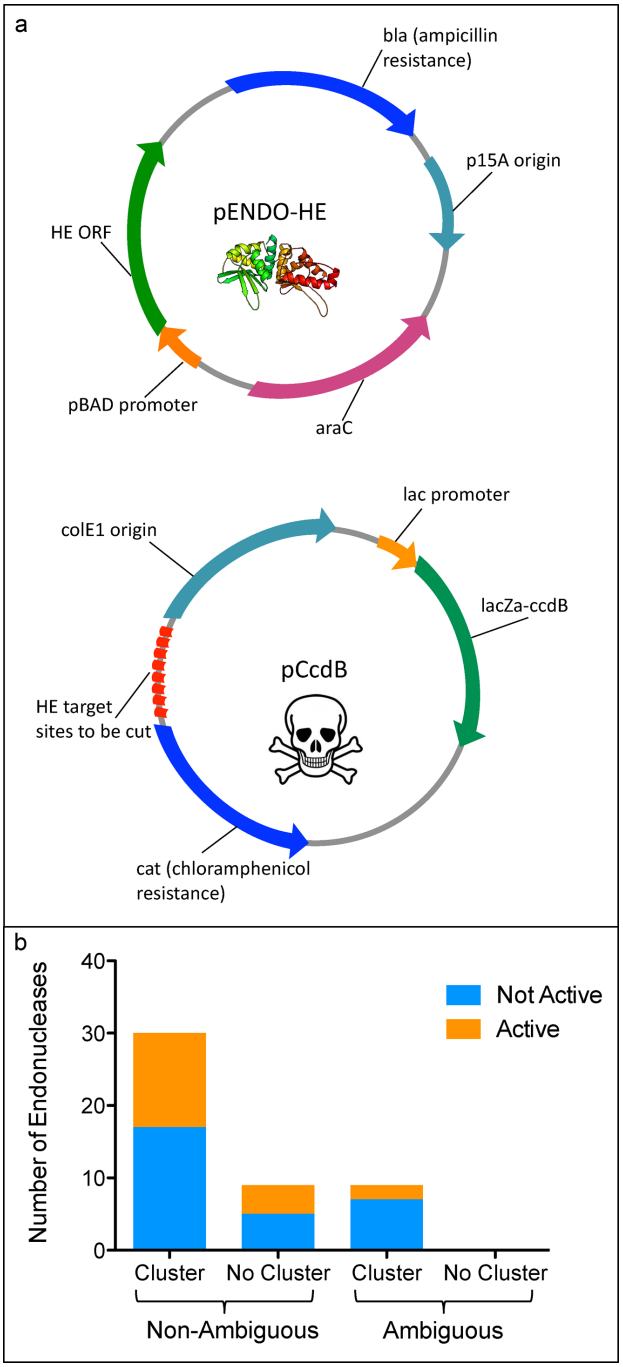

Figure 2. Activity of new homing endonucleases against predicted target sites. (**a**) Bacterial selection system used to screen endonucleases for activity by linking bacterial survival to target-site cleavage (14,42). A high-throughput adaptation of the original selection (42) was used (14), where many putative target sites were placed in tandem on pCcdB. (**b**) Of the 48 experimentally tested endonucleases, active enzymes were found in all categories of target-site identification, suggesting accuracy of site predictions. Active endonucleases included the 17 with high activity in the bacterial selection system and two additional endonucleases that were shown to have some activity in the Sanger sequencing experiments. Most of the tested endonucleases, 39 of 48, were classified as non-ambiguous and had clear exon–intron boundary annotations. All endonucleases with ambiguous targets or without enough DNA to find a target (considered as ambiguous in this figure) as well as the majority of the non-ambiguous endonucleases clustered with other homologs, sharing similar protein and predicted target-site sequences.

Activity of the expressed endonucleases was measured with *in vitro* cleavage assays (9,14), using the target-site arrays amplified from the pCcdB plasmids as substrates. The enzyme reaction buffer was 170 mM KCl, 10 mM $MgCl_2$, 20 mM Tris, pH 9.0 and 1 mM dithiothreitol (DTT). Reactions were completed for 30 min at 37°C and halted with approximately 17 nM EDTA, followed by 60°C incubation for 5–10 min. Cleavage products were separated on a 1.2% agarose tris-borate-EDTA (TBE) gel and stained with ethidium bromide. To identify the exact location of cleavage, the same reaction procedure was followed by polymerase chain reaction (PCR) cleanup (Qiagen) and Sanger sequencing instead of agarose gel separation.

Next-generation specificity and kinetic profiling

Methods used for single-turnover kinetic analyses of homing endonucleases were described in detail in previous work (2,44). In brief, enzyme concentrations and reaction times are varied, and the DNA concentration is significantly lower than the K_M of the enzyme.

The DNA substrate tested with each endonuclease was a library of all single nucleotide substitutions in the known or putative target site for that endonuclease, added to the end of a constant 1584 base-pair DNA sequence. This substrate was further amplified to incorporate phosphorothioate bonds on both 5′ ends to prevent exonuclease degradation of molecules not cleaved in the endonuclease reaction. In these kinetic experiments, the DNA substrate concentration was 2.5 nM in a 50 μl reaction with the same buffer conditions used to test enzyme activity, and samples were removed at eight time-points (15 s, 30 s, and 1, 2, 4, 8, 16 and 32 min). Six enzyme concentrations were tested (Supplementary Table S2). In previous kinetics experiments, the reaction was stopped with EDTA-containing buffer, but the new method required that the reactions be stopped in a different way that did not necessitate a cleanup step prior to enzymatic digestion of the cut portion of the substrate population. Therefore, the enzymatic reaction was halted by lowering the pH to approximately 4.5, because it is known that LAGLIDADG endonucleases cannot cleave DNA at low pH (45). The reaction buffer was the same as used for the enzyme activity assays described in the previous methods section. Samples of 5 μl were removed from the reactions and halted with 20 μl of a 15 mM Glycine-HCl solution with a pH of 3. To eliminate future endonuclease activity, the samples were then heated at 70°C in the low pH solution for at least 10 min.

To degrade the cut substrate, 5 μl of a mix of 0.5 μl lambda exonuclease, 0.5 μ exonuclease I, 1 μl of water and 3 μl of an equal mix of their respective buffers was added to each halted 25 μl sample. The two buffers neutralized the relatively low concentration of low pH glycine and the solution was returned to the optimal pH of approximately 9. All enzymes and buffers were obtained from New England Biolabs. The degradation step was completed at 37°C for a minimum of 1.5 h. The activity of lambda exonuclease and exonuclease I was halted by incubating the reaction at 80°C for 20 min.

To amplify each tested condition and incorporate a unique barcode for each condition, 20 μl PCR reactions were assembled including 1 μl of the reaction mix, 10 μl of 2X taq master mix (GoTaq green master mix, Promega), 7 μl of water, and 1 μl each of a 10 μM constant forward primer and reverse barcoding primer. Eight cycles of PCR amplification were completed to minimize the effect of the amplification. Following this barcoding PCR, all barcoded conditions for each individual enzyme were mixed together equally. This mix was column purified and the concentration was determined by NanoDrop. These clean samples of all conditions for each tested endonuclease were then combined with each other at equal concentrations to ensure sufficient sequencing coverage for all experiments. Duplicate reactions were included for each tested condition and the samples were sequenced twice to ensure reliability at the sequencing step.

Alignment and quality filtering of the sequencing data from raw Illumina reads was completed by the sequencing facility (htSEQ, University of Washington). Reads were assigned to the correct pool on the basis of a unique eight base-pair barcode identifier (Supplementary Information). The number of reads for each included substrate was counted and compared between each reaction condition and an uncleaved control reaction with the same substrate mix. Dividing each reaction condition by the uncleaved control sample produced a substrate ratio, equivalent to the endonuclease specificity for each position in its putative target site. Substitutions that abrogated endonuclease cleavage increased most in the substrate pool, while substitutions in the region flanking the endonuclease target decreased the most. Endonuclease kinetic properties were determined by comparing substrate ratios for different concentrations of endonuclease. Positions with that are influenced by enzyme concentration, K_M positions, will decrease in the substrate population with higher enzyme concentration. Processing of deep-sequencing data to generate profiles for specificity and kinetics is described in further detail in the Supplementary Material. Scripts for analysis of deep-sequencing data and generation of graphs are available by request. Sequencing data is also available upon request.

Computational modeling

Structure models of endonucleases were generated using the recently developed RosettaCM protocol (46), which samples protein conformations based on all homolog structures. Recent CASP10 experiments showed that this protocol generates more accurate atomic models for homology modeling compared to other widely used methods. For the protein structure modeling, templates and alignments are first identified by HHsearch (47), SPARKS-X (48) and RaptorX (49). A total of 41 endonuclease structures were used to generate these models and are listed at the end of this section. For DNA modeling, the backbone of the DNA in a crystal structure of I-OnuI in 3QQY (4) was used as the template. Base pairs are placed based on the given target sequence using DNA substitution methods described earlier (3,41). Both the forward and reverse orientation of the DNA sequence and 17 DNA threading possibilities with different registration in each direction were considered.

The top 10 templates from each alignment method were selected for each endonuclease being modeled and were su-

perimposed to 3QQY using TMalign (50). RosettaCM was then applied to recombine and refine the structure based on the input templates with threaded DNA in place (46). For each protein, 1000 models were generated for each DNA target site being considered. Lowest total energy models were selected as input for ΔΔG calculations. First, 20% (200) of the lowest energy models were selected. To reduce the noise resulting from uneven sampling of high-energy structures, average total energy was calculated among the selected models. Those models with total energies higher than one standard deviation above the average energy were filtered. For a given model, ΔΔG of the protein–DNA interface was calculated as the difference between the energy of protein–DNA complex and the energy of isolated protein and DNA. The average and standard deviation of the five lowest ΔΔG values are reported here. The total time for all calculations for each target was approximately 5000 CPU hours. All models are available upon request and the best model for each characterized endonuclease is included with this submission.

PDB codes for RosettaCM protein modeling templates:

1dq3, 1j27, 1kaf, 1lwt, 1m5x, 1mow, 1n3e, 1p8k, 1t9i, 1t9j, 1u0c, 2ab5, 2cw8, 2dch, 2ex5, 2fld, 2i3p, 2i3q, 2qoj, 2qrr, 2vbj, 2vbl, 2vbn, 2vs7, 2xe0, 3c0w, 3cxj, 3e54, 3eh8, 3fd2, 3hyi, 3mip, 3mis, 3mx9, 3mxa, 3mxb, 3qqy, 3r7p, 3uvf, 4aae, 4efj.

RESULTS AND DISCUSSION

Endonuclease and target-site identification

Candidate LAGLIDADG homing endonucleases (1) were identified by BLAST (25) searches against the non-redundant NCBI database with the sequences of previously characterized members of the family. Initially, putative target sites were assembled by manually examining the exon boundaries of the intron that contained the endonuclease gene (Figure 1a). The binding sites of these enzymes are

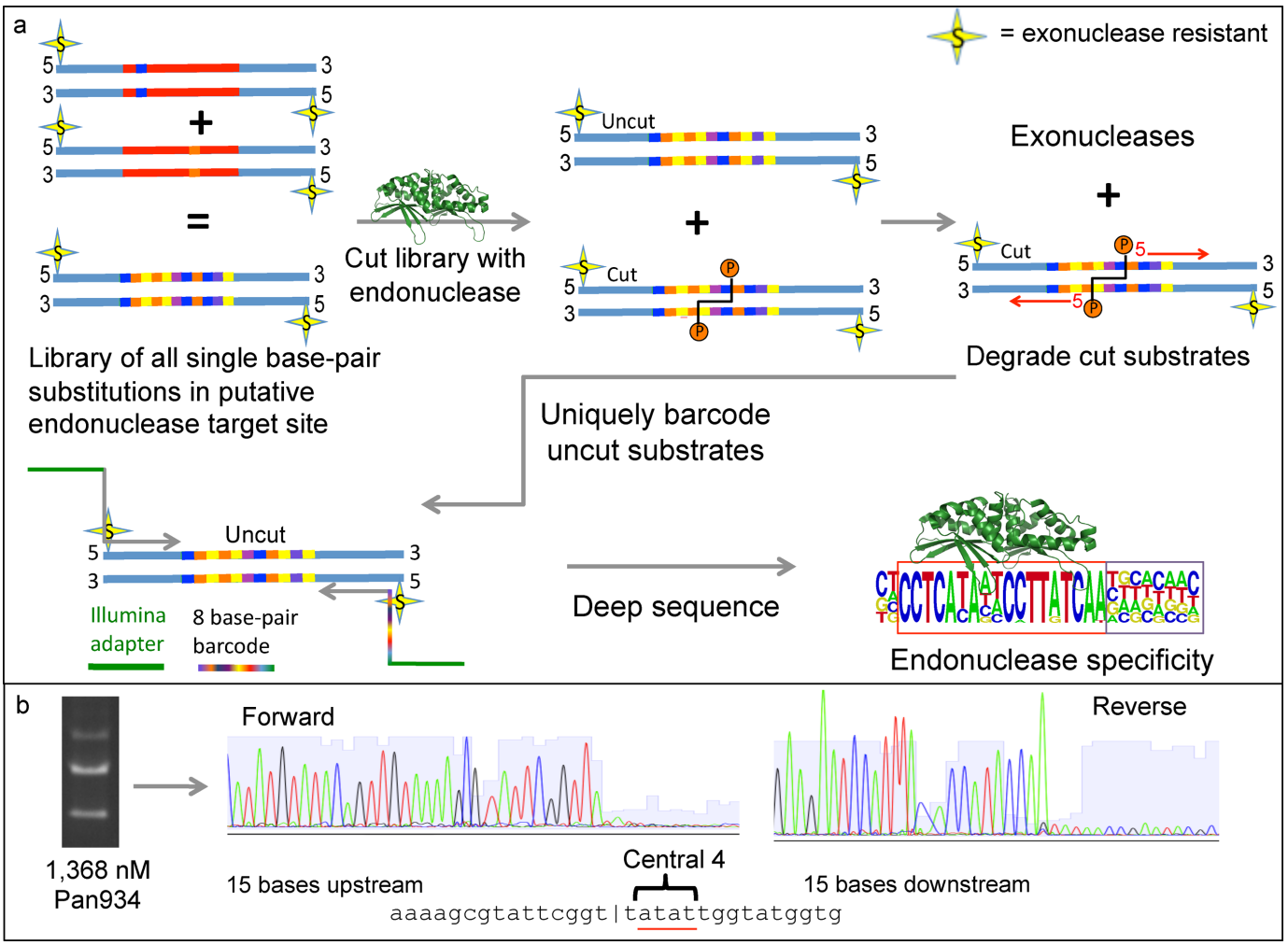

Figure 3. Identification of endonuclease cleavage specificity and location. (**a**) A new method of profiling DNA cleavage specificity using deep sequencing was developed. A substrate library was constructed for each tested endonuclease, consisting of all single base-pair substitutions in its putative target site. This library was reacted with endonuclease under multiple conditions, the cut substrate was degraded by exonucleases, and the uncut portion of the library for each reaction condition was marked with a unique barcode and deep sequenced. (**b**) Sanger sequencing of cut plasmid substrate revealed the cleavage location and central four bases for each expressed endonuclease (Supplementary Figure S3), corroborating the specificity profiles generated with deep sequencing. An additional adenine nucleotide is often found added at the end of the trace from the sequencing reaction.

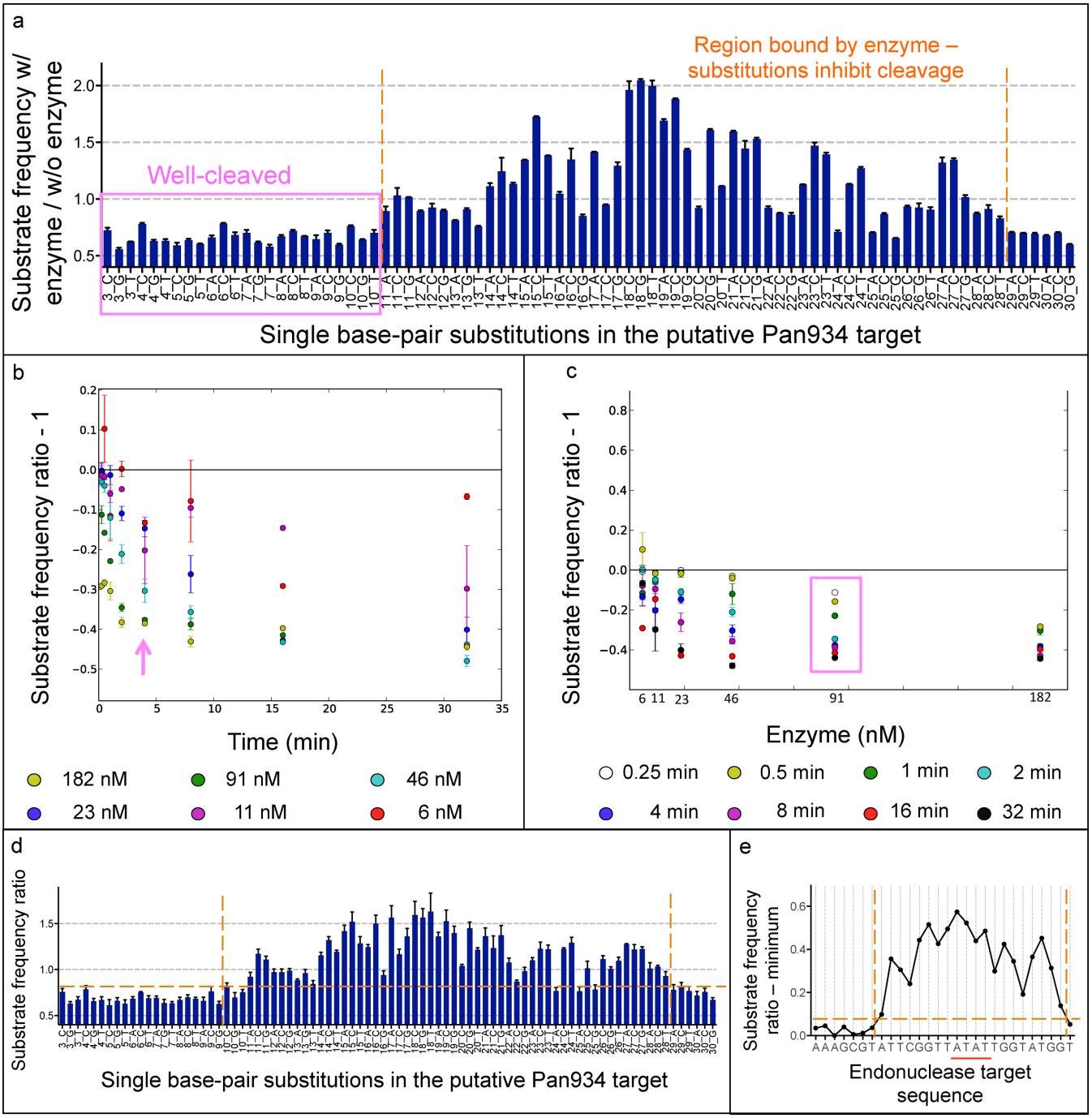

Figure 4. Processing of deep-sequencing data to generate specificity profiles. (**a**) Example of deep-sequencing data collected for the highest tested concentration of the Pan934 enzyme (182 nM) and the longest reaction time (32 min), averaged across two independent experiments (standard deviation shown). A substrate frequency ratio was determined by dividing the frequency of each single base substitution in the reaction condition with enzyme by the control condition with no enzyme added. The best-cleaved substitutions decreased the most in the substrate library and are located in the region adjacent to the approximately 20 base-pair endonuclease target site. (**b**) The best-cleaved substitutions in the substrate library are adjacent to the endonuclease target site, and their changes across reaction conditions need to be eliminated from the specificity profile. Therefore, the second step in the data processing was to identify the reaction time when the substrate ratio no longer decreased for well-cleaved substrates (pink arrow). The substrate 3_G in the Pan934 site is shown as an example and the ratios were compared to 0 instead of 1 to more easily identify changes relative to the starting sample. This data are an average of the two separate runs completed for each reaction condition and the standard deviation error bars are shown (except for the 16-min time, as one of these two barcodes was used for the uncleaved substrates for this highest enzyme concentration). (**c**) Similarly to the data-processing step shown in (b), the approximate enzyme concentration with which the well-cleaved substrates no longer changed significantly was identified (pink box). (**d**) The final specificity plots were generated from an average of three timepoints (the time before and after the time identified in (b)) to more comprehensively represent the sequencing results and reduce experimental noise. The standard deviation is shown for the average of the data from these three (each already averaged across duplicates) different reaction times with the enzyme concentration identified in (c). (**e**) Summarized specificity plots were generated by averaging the values shown in the full specificity graph in (d) for the three substitutions at each position. To facilitate comparisons between profiles, the substrate frequencies for the average of three best-cleaved substitutions are set equal to zero and all other values are correspondingly adjusted.

typically 20 bases in length, so we hypothesized that 30 bases, 15 from each surrounding exon, were likely sufficient to contain the site if the boundaries were accurately annotated. We then automated this process to collect the endonuclease genes and corresponding target site containing sequences from Genbank and Refseq databases (Figure 1b). A previously reported approach for identifying endonuclease targets compared alleles with and without introns and inteins (10). Our algorithm instead finds sites of intron-encoded LAGLIDADG endonucleases by comparing published intron–exon boundaries to each other and to longer sequence regions in less well annotated genomes. Similar target sites were identified for endonucleases with high protein sequence similarity, supporting these target-site predictions (Figure 1c). While the majority of endonuclease genes reside in the fungi mitochondrial sequences (4,51), several were tested from organisms in other kingdoms, such as cucumber (52) and coral (53).

Enzyme activity

Active endonuclease–substrate pairs were identified using a selection system that couples survival of bacteria to cleavage of a plasmid containing the substrate (Figure 2a) (14,42). We tested 48 enzymes in the selection system and found 17 that were highly active, targeting 12 unique sites (Supplementary Table S1). The low survival of the remaining endonucleases is either due to activity that is not high enough for the stringent bacterial selection (42,54), poor stability or incorrect target-site prediction. The inactive endonucleases clustered with many other endonucleases predicted to cleave the same target and even with high-surviving endonucleases, suggesting that their sites were accurately determined (Figure 2b, Supplementary Table S1). Homing endonuclease protein sequences can degrade and lose nuclease activity following the homing process, as they no longer need to cleave within their host genome (34,54,55), and such degradation probably occurred in some of the 32 enzymes that did not display high activity. For 11 of the inactive endonucleases we retrospectively identified potentially deleterious mutations (sequences in Supplementary Information), such as the conversion of a catalytic aspartate to asparagine in the inactive Pak761 endonuclease. A computational approach to more reliably detect this degradation by comparison to a consensus enzyme sequence (56,57) would likely increase success rate in future work. Additionally, we found that activity could be recovered for one enzyme by swapping in protein regions from a related high-activity endonuclease (Supplementary Figure S1).

For further *in vitro* characterization, we initially explored *in vitro* translation and compared cleavage activity to survival in the bacterial system (Supplementary Figure S2). However, only a subset of the active enzymes could be made with this method, so the proteins were instead expressed and purified from *E. coli* as His-tagged MBP fusions. All but two of the highly active enzymes (Glu729 and Pan933) expressed well and displayed high activity (Supplementary Figure S3), while almost all low-activity enzymes either did not express or showed no activity. Scu342 and Ade066 were exceptions, showing some activity in plasmid cleavage experiments (Supplementary Figure S3), giving a total of 19 endonucleases with some activity.

Substrate specificity

To further profile the target-site preferences of the expressed and active endonucleases described above, we developed a high-throughput protocol using deep sequencing (Figure 3a, Supplementary Figure S4). In brief, a DNA substrate library was generated containing all single base-pair substitutions in the putative endonuclease target and exposed to endonuclease under varying conditions. This method directly indicates the identity of base-pair substitutions that inhibit cleavage: cleaved substrates are degraded while the uncut portion of the library is preserved and passed on to next-generation sequencing. The uncut substrates remaining in each reaction condition are identifiable by a unique sequence tag (barcode) that is added after the degradation step. To complement the new method and further clarify the precise point of cleavage and central four bases, the same plasmids used in the bacterial selection were digested with homing endonucleases and Sanger sequenced (Figure 3b). From the deep-sequencing data we generated specificity profiles by taking the ratio of the frequency of each DNA substrate in samples exposed to endonuclease to the corresponding substrate frequency in a no-enzyme control, and averaging these ratios across several reaction times (Figure 4a–e). Target sites with substitutions that are not tolerated by these newly characterized endonucleases increased in the substrate pool while cleaved target sites decreased, thus identifying their approximately 20 base-pair binding regions within the 30 base pairs surrounding the intron containing the endonuclease gene. Similar specificity profiles were obtained with protein produced via *in vitro* translation and with the MBP fusions (Supplementary Figure S5).

Substrate specificity profiles were also generated using this method for previously characterized enzymes with published profiles (2,4,58), enzymes with published target sites but uncharacterized specificity profiles (4,42), and for nine of the newly identified high-activity enzymes (Figure 5, Supplementary Figures S6 and S7). The binding sequences are not always centered on the intron break point, but each endonuclease has a similar length target site and high level of specificity (Figure 5). For three enzymes with published specificity profiles, the general trends matched well with previous results (Figure 5, Supplementary Figure S6), although the dynamic range of the profiles derived from deep-sequencing data was lower. The cause of the reduced dynamic range and difference between actual and theoretical deep-sequencing results (Supplementary Figure S8) is not clear; it is not due to the reliability of the deep-sequencing data (Supplementary Figure S9), degradation of cut substrate (Supplementary Figure S10) or the presence of competing substrates (Supplementary Figure S11). We also used the method to identify new and unexpected substrate preferences for previously engineered endonuclease variants (Supplementary Figure S12) (14).

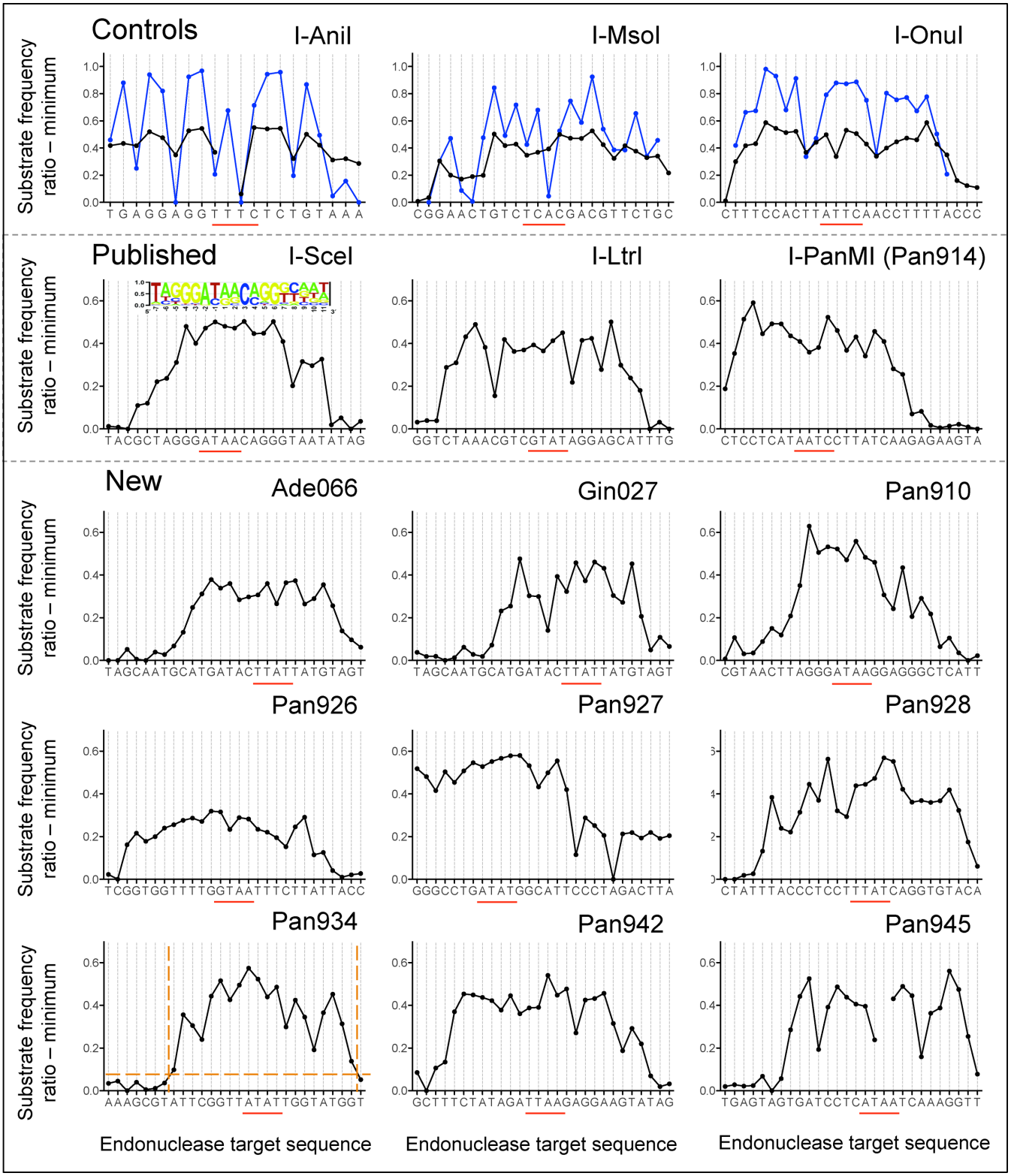

Figure 5. Endonuclease cleavage specificity profiles. Condensed cleavage profiles, generated by averaging the substrate ratio for the three possible base substitutions at each target-site position and setting the best-cleaved substitutions equal to zero. Full specificity profiles are available in Supplementary Figure S7. The central four bases, identified by comparisons of Sanger sequencing and deep-sequencing data, are underlined in red. The specificities of the control enzymes I-AniI, I-MsoI and I-OnuI were previously published (blue) (2,4,58) and are compared to the cleavage profile obtained from deep sequencing (black). A specificity profile (sequence logo) has also been published for I-SceI (42) and closely matches the deep-sequencing profile, but the necessary data for a quantitative comparison wre not available. Target sequences for both I-LtrI and I-PanMI were previously published without specificity profiles (4).

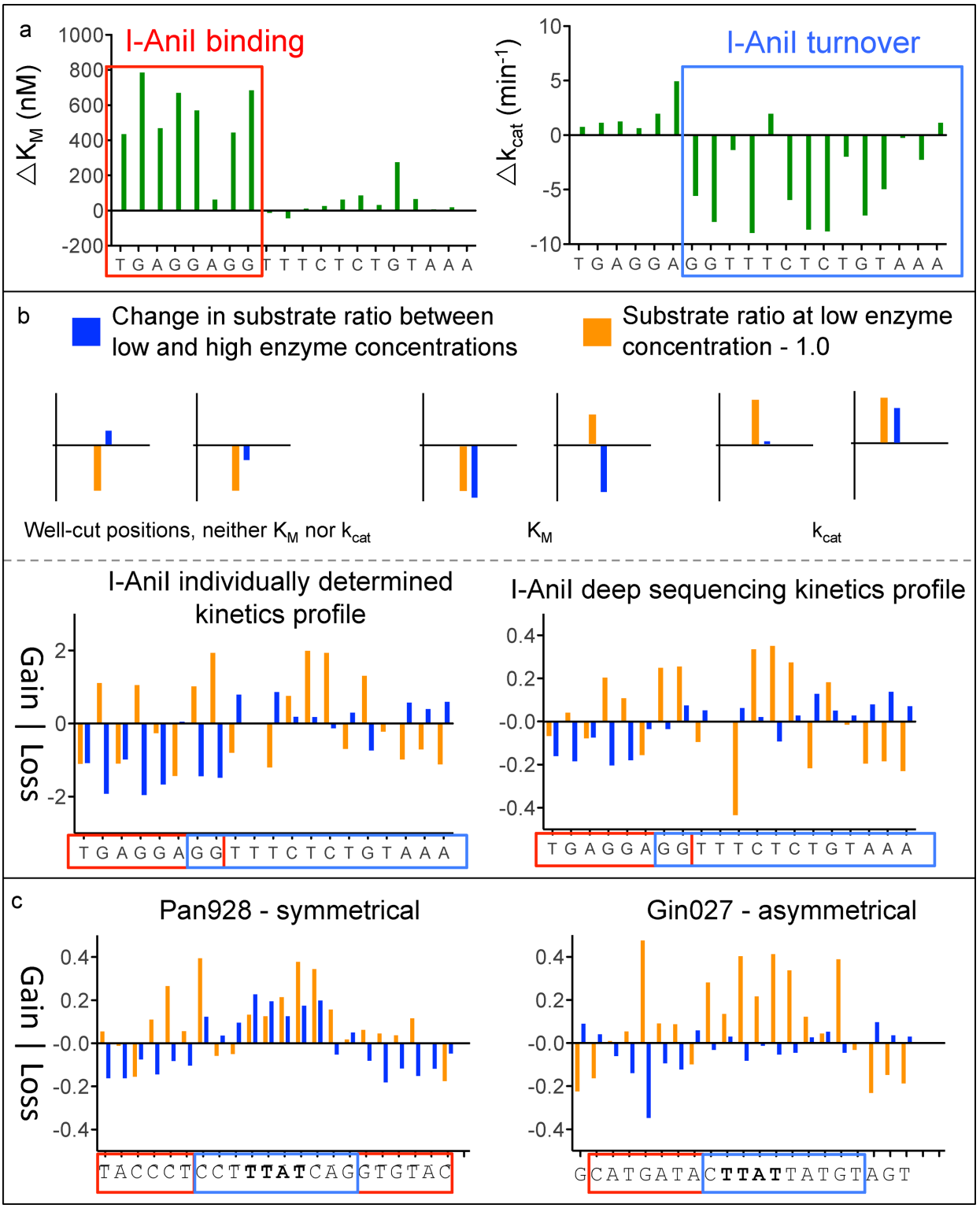

Figure 6. Determination of cleavage kinetics using deep sequencing. (**a**) Previously published kinetic data for the I-AniI endonuclease (2) revealed regions of the interface involved in ground-state formation (binding), where target-site substitutions resulted in increased K_M, and those involved in transition-state formation (turnover), where target-site substitutions resulted in decreased k_{cat}. The kinetic data for all three possible single base-pair substitutions were averaged to generate single values for each position in the I-AniI target. (**b**) Comparison between the kinetic profiles generated for I-AniI using the deep-sequencing method and using the traditional kinetics approach. The profiles are based on the response of each substrate from the mix of reacted substrates to changing enzyme concentration. Targets with substitutions in the region of the I-AniI interface involved in ground-state formation displayed a loss in the substrate pool with increased enzyme concentration. In contrast, positions in the turnover region of the interface displayed an increase in concentration at short reaction times and were either unaffected or showed a gain in response to increased enzyme. (**c**) Kinetic profile for Pan928 and Gin027, with regions of the interface that show similar characteristics to I-AniI regions boxed in the same color as in panels (a) and (b).

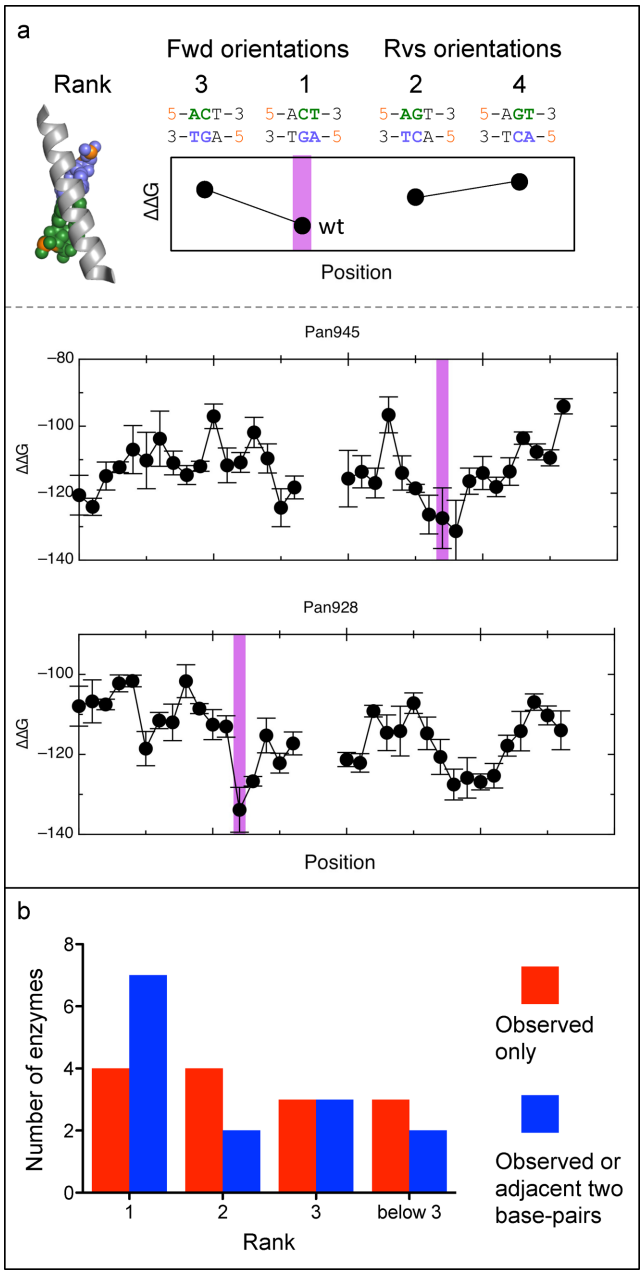

Figure 7. Predicting endonuclease target-site preferences using homology modeling and binding energy calculations. (**a**) Endonuclease–DNA complexes were modeled and the interface binding energy (ΔΔG) was calculated for each protein with 34 possible target-site orientations. The putative target site for these endonucleases, identified through experimental characterization (Figure 5, Supplementary Figure S3), is highlighted with a magenta bar. A simplified scheme of how the target-site orientations are presented for the endonucleases is shown in the upper part of the panel. Binding energy plots are shown for Pan945 (49% identity to template, I-OnuI), for which the putative target site is ranked second by the computation, and Pan928 (42% identity to I-OnuI), for which the actual target site is ranked best. (**b**) Summarized prediction results for 14 newly characterized endonucleases (Supplementary Figure S15). If the experimentally identified target site were chosen as the best site by the computational prediction, then the result is ranked 1 (red). If either the experimental target or either of the two adjacent base pairs were predicted as the best site, then the result is ranked 1 (blue) in order to capture the energetic funnel seen for some endonucleases such as Pan945.

Kinetic profiling

Specificity data only provide a static view of the importance of each target-site interaction for DNA cleavage, rather than revealing the role of interface regions at different stages of catalysis, and do not distinguish the contributions of interface residues to substrate binding versus transition state stabilization. Previously we demonstrated, for the endonuclease I-AniI, that two distinct interface regions (the N-terminal and C-terminal domains) respectively dominate the enzyme's activity and specificity during substrate binding and turnover (Figure 6a) (2). To uncover sequence determinants of kinetics for the newly identified endonucleases, we carried out a similar analysis by sequencing cleavage reactions at multiple concentrations and times. To determine which substitutions influenced substrate binding (K_M) or turnover (k_{cat}) we calculated how the amounts of uncut substrate in the sequencing reaction changed in response to varying conditions (Figure 6b, Supplementary Figure S13). Substitutions that decrease substrate abundance in the population with increasing enzyme concentration influence K_M, while substitutions that increase substrate abundance even at high enzyme concentrations influence k_{cat} (2). Comparing the deep-sequencing derived kinetic profile with the previously published I-AniI kinetic profile (2), the regions involved in formation of the ground-state complex (where target-site substitutions impact initial substrate binding) and of the transition state (where substitutions impact turnover) were found to be very similar between the two profiles. Evaluation of newly generated profiles for other high-activity endonucleases revealed differences in their degree of catalytic asymmetry (Figure 6c, Supplementary Figure S14): the Pan928 profile is highly symmetric, with target-site substitutions in and surrounding the central four on both sides reducing turnover, while the Gin027 profile resembles that of I-AniI, with each target-site half having distinctive characteristics.

Computational modeling

Without a crystal structure or reliable model it is difficult to connect these substrate preferences to particular interface interactions or to use these endonucleases as starting points for further structure-based engineering. Even if the cleavage site is precisely defined, the orientation of the enzymes on their target sites is not clear without structural data. Since crystal structures are not available for the new enzymes, we chose to model these protein–DNA complexes using RosettaCM (46). The main challenge of this approach was building accurate models of the protein–DNA interface with the putative target-site bases substituted. The DNA backbone used in the computation was copied from the template crystal structure, with the DNA bases substituted with the sequences of putative target sites and rigid-body shifts allowed during the optimization process, altering the relative orientation of the protein and DNA molecules. Fourteen active endonucleases, nine with newly generated specificity profiles (Figure 5) and several with only Sanger sequencing data (Supplementary Figure S3), were modeled with 34 possible target sequences, 17 in each orientation centering around the original 30 base-pair sites. Sequence registries were then ranked by the calculated protein–DNA binding

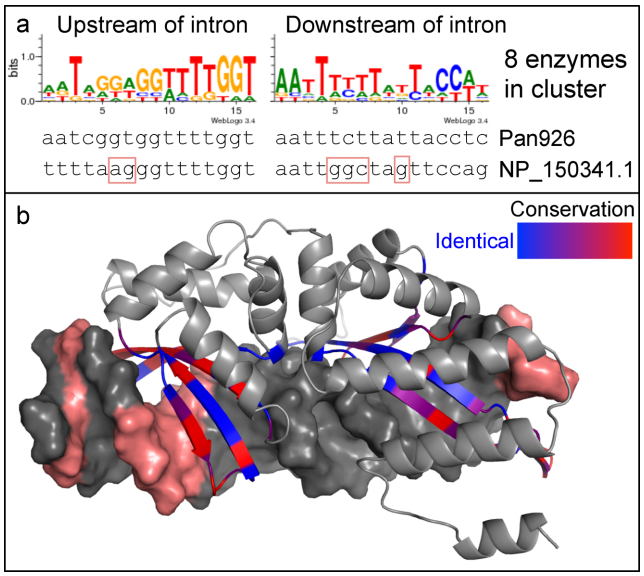

Figure 8. Shifts in target-site preference correlate with protein sequence changes in homologous endonucleases. (**a**) The sequence cluster containing the tested Pan926 endonuclease also included a homologous endonuclease predicted to cleave a target site containing several substitutions. These base substitutions were located with regions of the Pan926 target site that has high specificity (Figure 5), indicating that there must also be amino acid changes in this homolog to accommodate the new target sequence. (**b**) Comparing the residues in the protein–DNA interface of Pan926 and the homolog with the differing target site indicated that one-half of the interface was more conserved than the other (blue = identical, purple = similar, red = divergent). Pan926 was predicted to bind in a reverse complement orientation with binding-energy calculations (Supplementary Figure S15). Comparing the interfaces of Pan926 and the homologous endonuclease supports this binding model, as the region with more target-site changes is interacting with the more divergent protein sequence half in the reverse complement orientation.

energy (Figure 7a). In some cases there is a funnel-like energy landscape, where the registries near the correct position have low binding energies. This rigid-body shifting occurs because the long side chains characteristic of protein–DNA interfaces, such as arginine residues, can make multiple different energetically favorable interactions with the correct base. Half of the proteins have the best binding energy either at the experimentally supported or adjacent site, and 11 have a lower binding energy for one of the two possible orientations of the experimentally identified target site than for the majority of the competing sites (Figure 7b, Supplementary Figure S15).

Connecting substrate preferences and interface interactions

For those endonucleases where experimental data and binding-energy calculations corroborate each other, we can use the homology models to understand how changes to the protein sequence lead to new target-site specificities. The sequence cluster containing the Pan926 endonuclease includes an endonuclease with a predicted target site differing by almost half the nucleotides (Figure 8a). Some target-site bases are conserved and are near similarly conserved protein interface residues, while some are completely different and are contacted by correspondingly evolved protein residues (Figure 8b). Structural models can explain how specificity shifts are produced by evolution, an essential step toward being able to engineer similar shifts. Comparisons between models and experimentally derived target-site preferences can generate hypotheses for further investigation; for example, these comparisons suggest a possible role for aromatic residues in promoting endonuclease catalysis (Supplemental Discussion, Supplementary Figures S14 and S16).

Table 1. Endonucleases with activity against predicted target sites

Name	Accession #	Target Site (orientation)
Endonuclease with deep sequencing specificity profiles		
Pan910	NP_074910.1	acttaggg**ataa**ggagggct (R)
Pan914	NP_074914.1	ctcctcat**aatc**cttatcaa (F)
Pan926	NP_074926.1	tggttttg**gtaa**tttcttat (R)
Pan927	NP_074927.1	tgggcctg**atat**ggcattcc (R)
Pan928	NP_074928.1	accctcct**ttat**caggtgta (F)
Pan934	NP_074934.1	attcggtt**atat**tggtatgg (F)
Pan942	NP_074942.1	tctatagat**ttaa**gaggaagt (R) or tctatagat**ttaa**gaggaagt (F)
Pan945	NP_074945.1	tgatcctc**ataa**tcaaaggt (R)
Gin027	YP_002587027.1	catgatac**ttat**tatgtagt (F)
Ade066	XP_002620066.1	similar to Gin027
Endonucleases with targets identified by Sanger sequencing		
Scu342	YP_203342.1	taggtgtt**aatt**taacattt (F) or aggtgtta**atttaacatttt** (R)
Aae264	AAC72264.1	gatggact**gttt**atcctcct (F)
Bcu047	YP_003127047.1	similar to Aae264
Cre842	ABX82842.1	similar to Aae264
Pan940	NP_074940.1	similar to Gin027
Pma729	NP_943729.1	similar to Gin027
Endonucleases with high survival that could not be expressed		
Glu729	AAO13729.1	ggtaggaaataggattagataccccggtac
Pan933	NP_074933.1	aatggattcttcggtcatcccgaggtttat likely based on Pan934 similarity: cttcggtc**atcc**cgaggttt (F)
Soc368	NP_700368.1	similar to Aae264

We have characterized homing endonucleases cleaving 13 unique target sites. For 10 endonucleases, eight with unique and newly identified target sites (the site for Pan914 or PanMI was previously published (4)), a specificity profile was collected using our new deep-sequencing method (Figure 5, Supplementary Figure S7). Several additional endonucleases were shown to cleave their predicted targets by Sanger sequencing (Figure 3b, Supplementary Figure S3). The central four bases for each target are indicated with bold font. The orientation of these endonucleases binding to their putative target sites was predicted by homology modeling and binding-energy calculations (Figure 7, Supplementary Figure S14) and is indicated by (F), binding to the site in the orientation shown, or (R), binding to the site in the reverse complement orientation. In two cases, the exact orientation was not clear from the modeling and both possible target sites are shown. For three endonucleases that did not express, target sites could not be verified outside of the bacterial selection system. Two of these, Glu729 and Pan933, are highly active in the bacterial system and are predicted to cleave unique sites. The 30 base-pair region likely containing their targets is shown, as well as a putative site for Pan933 that is based on the conservation between DNA-interacting residues in its N-terminal domain and in that of Pan934 (see attached homology model of Pan934).

CONCLUSIONS

The rapidly increasing availability of whole genome sequences from diverse organisms has enabled the discovery of large numbers of homing endonucleases (4,10). Here we present a new method for identifying these endonucleases and their corresponding target sites from these sequence data. The accuracy of our method was evaluated

with high-throughput experimental approaches for profiling DNA cleavage activity and specificity. We characterized 19 active enzymes, targeting 13 unique target sites, and generated full specificity profiles for 10 endonucleases (Table 1), nine of which were newly identified. The approaches tested here are readily applicable to studying other DNA cleavage enzymes, such as transcription activator-like effector nucleases (TALENs) (59,60) and Cas9 (61–63). Discovery and characterization of Cas9 nucleases with different specificities is a current challenge (64–66) that could employ the pipeline we have established for homing endonucleases.

The deep-sequencing method for profiling DNA cleavage specificity allows characterization of the role of specific base interactions in substrate binding and transition state formation by monitoring cleavage across a wide range of enzyme reaction conditions. This deep-sequencing approach is useful both for discovery and characterization of new enzymes and for providing feedback during protein engineering endeavors, identifying causes of low activity or specificity at multiple stages of the design process. The method could also be adapted for high-throughput study of single-strand nicking or RNA cleavage.

We find that the Rosetta homology-modeling platform can be used to model protein–DNA interfaces and probe the exact DNA target site for endonucleases. While the approach described here is not perfectly accurate, it has considerable potential; for half of the targets, it predicted the binding site no more than a single base off from the correct target. These models, combined with database mining and sequence clustering, can inform our understanding of how amino acid mutations in homologous endonucleases result in natural target-site specificity shifts and can be used as starting points for further interface engineering. As high-throughput assays provide more data for guiding model improvement, the accuracy of modeling should increase and be extendable to other enzyme-substrate classes where high-throughput experimental methods are unavailable.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors would like to thank the entire Rosetta Commons community for contributions to the Rosetta code base. Justin Ashworth provided the PSSM search protocol used in the program for finding homolog target sites, as well as helpful discussion. Eva-Maria Strauch provided helpful discussion on the sequencing aspects of the work. Next-generation sequencing support was provided mainly by the University of Washington htSEQ facility, with early experiments completed in the Shendure lab. The authors would particularly like to thank Audra K. Johnson, Daniel Bates and Morgan Diegel from the htSEQ facility, and Charlie Lee from the Shendure lab. We also thank anonymous reviewers, Michelle McCully and Ratika Krishnamurty for helpful paper edits. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author Contributions: S.B.T. and D.B. wrote the paper, with contributions from Y.S. and T.J.B. S.B.T. conceived of the project, designed the experiments and completed experimentation with assistance from Y.S. and M.D.S. Both Y.S. and M.D.S. helped with testing the deep-sequencing method, Y.S. helped with protein expression and M.D.S. participated in gene construction and the bacterial screen for endonuclease activity. Y.S. developed the computational protocol for homology modeling with DNA and target-site prediction using binding energy. T.J.B. developed the program to collect endonuclease genes and putative target sites, with input from S.B.T. L.K. used the deep-sequencing method to profile specificities of engineered I-AniI variants. P.B. provided helpful discussion and contributed methods for analyzing DNA geometry and modeling DNA flexibility.

FUNDING

National Science Foundation graduate research fellowship [to S.B.T.]; U.S. National Institutes of Health and the Foundation for the National Institutes of Health through the Gates Foundation Grand Challenges in Global Health Initiative [GM084433, RL1CA133832 to D.B.]; Howard Hughes Medical Institute. Funding for open access charge: Howard Hughes Medical Institute.

Conflict of interest statement. None declared.

REFERENCES

1. Stoddard,B.L. (2011) Homing endonucleases: from microbial genetic invaders to reagents for targeted DNA modification. *Structure*, **19**, 7–15.
2. Thyme,S.B., Jarjour,J., Takeuchi,R., Havranek,J.J., Ashworth,J., Scharenberg,A.M., Stoddard,B.L. and Baker,D. (2009) Exploitation of binding energy for catalysis and design. *Nature*, **461**, 1300–1304.
3. Ashworth,J., Taylor,G.K., Havranek,J.J., Quadri,S.A., Stoddard,B.L. and Baker,D. (2010) Computational reprogramming of homing endonuclease specificity at multiple adjacent base pairs. *Nucleic Acids Res.*, **38**, 5601–5608.
4. Takeuchi,R., Lambert,A.R., Mak,A.N.-S., Jacoby,K., Dickson,R.J., Gloor,G.B., Scharenberg,A.M., Edgell,D.R. and Stoddard,B.L. (2011) Tapping natural reservoirs of homing endonucleases for targeted gene modification. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, 13077–13082.
5. Redondo,P., Prieto,J., Muñoz,I.G., Alibés,A., Stricher,F., Serrano,L., Cabaniols,J.-P., Daboussi,F., Arnould,S., Perez,C. *et al.* (2008) Molecular basis of xeroderma pigmentosum group C DNA recognition by engineered meganucleases. *Nature*, **456**, 107–111.
6. Popplewell,L., Koo,T., Leclerc,X., Duclert,A., Mamchaoui,K., Gouble,A., Mouly,V., Voit,T., Pâques,F., Cédrone,F. *et al.* (2013) Gene correction of a duchenne muscular dystrophy mutation by meganuclease-enhanced exon knock-in. *Hum. Gene Ther.*, **24**, 692–701.
7. Djukanovic,V., Smith,J., Lowe,K., Yang,M., Gao,H., Jones,S., Nicholson,M.G., West,A., Lape,J., Bidney,D. *et al.* (2013) Male-sterile maize plants produced by targeted mutagenesis of the cytochrome P450-like gene (MS26) using a re-designed I-CreI homing endonuclease. *Plant J.*, **76**, 888–899.
8. Chan,Y.S., Takeuchi,R., Jarjour,J., Huen,D.S., Stoddard,B.L. and Russell,S. (2013) The design and in vivo evaluation of engineered i-onui-based enzymes for HEG gene drive. *PLoS One*, **8**, e74254.
9. Szeto,M.D., Boissel,S.J.S., Baker,D. and Thyme,S.B. (2011) Mining endonuclease cleavage determinants in genomic sequence data. *J. Biol. Chem.*, **286**, 32617–32627.
10. Barzel,A., Privman,E., Peeri,M., Naor,A., Shachar,E., Burstein,D., Lazary,R., Gophna,U., Pupko,T. and Kupiec,M. (2011) Native homing endonucleases can target conserved genes in humans and in animal models. *Nucleic Acids Res.*, **39**, 6646–6659.
11. Baxter,S., Lambert,A.R., Kuhar,R., Jarjour,J., Kulshina,N., Parmeggiani,F., Danaher,P., Gano,J., Baker,D., Stoddard,B.L. *et al.*

(2012) Engineering domain fusion chimeras from I-OnuI family LAGLIDADG homing endonucleases. *Nucleic Acids Res.*, **40**, 7985–8000.
12. Jacoby,K., Metzger,M., Shen,B.W., Certo,M.T., Jarjour,J., Stoddard,B.L. and Scharenberg,A.M. (2012) Expanding LAGLIDADG endonuclease scaffold diversity by rapidly surveying evolutionary sequence space. *Nucleic Acids Res.*, **40**, 4954–4964.
13. Molina,R., Redondo,P., Stella,S., Marenchino,M., D'Abramo,M., Gervasio,F.L., Epinat,J.C., Valton,J., Grizot,S., Duchateau,P. *et al.* (2012) Non-specific protein-DNA interactions control I-CreI target binding and cleavage. *Nucleic Acids Res.*, **40**, 6936–6945.
14. Thyme,S.B., Boissel,S.J.S., Arshiya Quadri,S., Nolan,T., Baker,D.A., Park,R.U., Kusak,L., Ashworth,J. and Baker,D. (2013) Reprogramming homing endonuclease specificity through computational design and directed evolution. *Nucleic Acids Res.*, **42**, 2564–2576.
15. Whitehead,T.A., Chevalier,A., Song,Y., Dreyfus,C., Fleishman,S.J., De Mattos,C., Myers,C.A., Kamisetty,H., Blair,P., Wilson,I.A. *et al.* (2012) Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing. *Nat. Biotechnol.*, **30**, 543–548.
16. Starita,L.M., Pruneda,J.N., Lo,R.S., Fowler,D.M., Kim,H.J., Hiatt,J.B., Shendure,J., Brzovic,P.S., Fields,S. and Klevit,R.E. (2013) Activity-enhancing mutations in an E3 ubiquitin ligase identified by high-throughput mutagenesis. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, E1263–E1272.
17. Fowler,D.M., Araya,C.L., Fleishman,S.J., Kellogg,E.H., Stephany,J.J., Baker,D. and Fields,S. (2010) High-resolution mapping of protein sequence-function relationships. *Nat. Methods*, **7**, 741–746.
18. Araya,C.L. and Fowler,D.M. (2011) Deep mutational scanning: assessing protein function on a massive scale. *Trends Biotechnol.*, **29**, 435–442.
19. Maerkl,S.J. and Quake,S.R. (2007) A systems approach to measuring the binding energy landscapes of transcription factors. *Science*, **315**, 233–237.
20. Kinney,J.B., Murugan,A., Callan,C.G. and Cox,E.C. (2010) Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 9158–9163.
21. Jolma,A., Kivioja,T., Toivonen,J., Cheng,L., Wei,G., Enge,M., Taipale,M., Vaquerizas,J.M., Yan,J., Sillanpää,M.J. *et al.* (2010) Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res.*, **20**, 861–873.
22. Wang,J., Lu,J., Gu,G. and Liu,Y. (2011) In vitro DNA-binding profile of transcription factors: methods and new insights. *J. Endocrinol.*, **210**, 15–27.
23. Geertz,M. and Maerkl,S.J. (2010) Experimental strategies for studying transcription factor-DNA binding specificities. *Brief. Funct. Genomics*, **9**, 362–373.
24. Pattanayak,V., Lin,S., Guilinger,J.P., Ma,E., Doudna,J.A. and Liu,D.R. (2013) High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. *Nat. Biotechnol.*, **31**, 839–843.
25. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
26. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2005) GenBank. *Nucleic Acids Res.*, **33**, D34–D38.
27. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–D65.
28. Nomura,N., Nomura,Y., Sussman,D., Klein,D. and Stoddard,B.L. (2008) Recognition of a common rDNA target site in archaea and eukarya by analogous LAGLIDADG and His-Cys box homing endonucleases. *Nucleic Acids Res.*, **36**, 6988–6998.
29. Moure,C.M., Gimble,F.S. and Quiocho,F.A. (2003) The crystal structure of the gene targeting homing endonuclease I-SceI reveals the origins of its target site specificity. *J. Mol. Biol.*, **334**, 685–695.
30. Chevalier,B., Turmel,M., Lemieux,C., Monnat,R.J. and Stoddard,B.L. (2003) Flexible DNA target site recognition by divergent homing endonuclease isoschizomers I-CreI and I-MsoI. *J. Mol. Biol.*, **329**, 253–269.
31. Marcaida,M.J., Prieto,J., Redondo,P., Nadra,A.D., Alibés,A., Serrano,L., Grizot,S., Duchateau,P., Pâques,F., Blanco,F.J. *et al.* (2008) Crystal structure of I-DmoI in complex with its target DNA provides new insights into meganuclease engineering. *Proc. Natl. Acad. Sci. U.S.A.*, **105**, 16888–16893.
32. Spiegel,P.C., Chevalier,B., Sussman,D., Turmel,M., Lemieux,C. and Stoddard,B.L. (2006) The structure of I-CeuI homing endonuclease: evolving asymmetric DNA recognition from a symmetric protein scaffold. *Structure*, **14**, 869–880.
33. Scalley-Kim,M., McConnell-Smith,A. and Stoddard,B.L. (2007) Coevolution of a homing endonuclease and its host target sequence. *J. Mol. Biol.*, **372**, 1305–1319.
34. Bolduc,J.M., Spiegel,P.C., Chatterjee,P., Brady,K.L., Downing,M.E., Caprara,M.G., Waring,R.B. and Stoddard,B.L. (2003) Structural and biochemical analyses of DNA and RNA binding by a bifunctional homing endonuclease and group I intron splicing factor. *Genes Dev.*, **17**, 2875–2888.
35. Söding,J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics*, **21**, 951–960.
36. Crooks,G.E., Hon,G., Chandonia,J.-M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.
37. Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.
38. Van Der Walt,S., Colbert,S.C. and Varoquaux,G. (2011) The NumPy array: a structure for efficient numerical computation. *Comput. Sci. Eng.*, **13**, 22–30.
39. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
40. Stemmer,W.P., Crameri,A., Ha,K.D., Brennan,T.M. and Heyneker,H.L. (1995) Single-step assembly of a gene and entire plasmid from large numbers of oligodeoxyribonucleotides. *Gene*, **164**, 49–53.
41. Thyme,S.B., Baker,D. and Bradley,P. (2012) Improved modeling of side-chain–base interactions and plasticity in protein–DNA interface design. *J. Mol. Biol.*, **419**, 255–274.
42. Doyon,J.B., Pattanayak,V., Meyer,C.B. and Liu,D.R. (2006) Directed evolution and substrate specificity profile of homing endonuclease I-SceI. *J. Am. Chem. Soc.*, **128**, 2477–2484.
43. Studier,F.W. (2005) Protein production by auto-induction in high density shaking cultures. *Protein Expr. Purif.*, **41**, 207–234.
44. Halford,S.E., Johnson,N.P. and Grinsted,J. (1980) The EcoRI restriction endonuclease with bacteriophage lambda DNA. Kinetic studies. *Biochem. J.*, **191**, 581–592.
45. Geese,W.J., Kwon,Y.K., Wen,X. and Waring,R.B. (2003) In vitro analysis of the relationship between endonuclease and maturase activities in the bi-functional group I intron-encoded protein, I-AniI. *Eur. J. Biochem.*, **270**, 1543–1554.
46. Song,Y., DiMaio,F., Wang,R.Y.-R., Kim,D., Miles,C., Brunette,T., Thompson,J. and Baker,D. (2013) High-resolution comparative modeling with RosettaCM. *Structure*, **21**, 1735–1742.
47. Remmert,M., Biegert,A., Hauser,A. and Söding,J. (2012) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods*, **9**, 173–175.
48. Yang,Y., Faraggi,E., Zhao,H. and Zhou,Y. (2011) Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates. *Bioinformatics*, **27**, 2076–2082.
49. Peng,J. and Xu,J. (2009) Boosting protein threading accuracy. *Res. Comput. Mol. Biol.*, **5541**, 31–45.
50. Zhang,Y. and Skolnick,J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.*, **33**, 2302–2309.
51. Thiéry,O., Börstler,B., Ineichen,K. and Redecker,D. (2010) Evolutionary dynamics of introns and homing endonuclease ORFs in a region of the large subunit of the mitochondrial rRNA in Glomus species (arbuscular mycorrhizal fungi, Glomeromycota). *Mol. Phylogenet. Evol.*, **55**, 599–610.
52. Cho,Y., Qiu,Y.L., Kuhlman,P. and Palmer,J.D. (1998) Explosive invasion of plant mitochondria by a group I intron. *Proc. Natl. Acad. Sci. U.S.A.*, **95**, 14244–14249.

53. Fukami,H., Chen,C.A., Chiou,C.-Y. and Knowlton,N. (2007) Novel group I introns encoding a putative homing endonuclease in the mitochondrial cox1 gene of Scleractinian corals. *J. Mol. Evol.*, **64**, 591–600.
54. Takeuchi,R., Certo,M., Caprara,M.G., Scharenberg,A.M. and Stoddard,B.L. (2009) Optimization of in vivo activity of a bifunctional homing endonuclease and maturase reverses evolutionary degradation. *Nucleic Acids Res.*, **37**, 877–890.
55. Longo,A., Leonard,C.W., Bassi,G.S., Berndt,D., Krahn,J.M., Hall,T.M.T. and Weeks,K.M. (2005) Evolution from DNA to RNA recognition by the bI3 LAGLIDADG maturase. *Nat. Struct. Mol. Biol.*, **12**, 779–787.
56. Lehmann,M., Loch,C., Middendorf,A., Studer,D., Lassen,S.F., Pasamontes,L., van Loon,A.P.G.M. and Wyss,M. (2002) The consensus concept for thermostability engineering of proteins: further proof of concept. *Protein Eng.*, **15**, 403–411.
57. Bershtein,S., Goldin,K. and Tawfik,D.S. (2008) Intense neutral drifts yield robust and evolvable consensus proteins. *J. Mol. Biol.*, **379**, 1029–1044.
58. Li,H., Ulge,U.Y., Hovde,B.T., Doyle,L.A. and Monnat,R.J. (2012) Comprehensive homing endonuclease target site specificity profiling reveals evolutionary constraints and enables genome engineering applications. *Nucleic Acids Res.*, **40**, 2587–2598.
59. Miller,J.C., Tan,S., Qiao,G., Barlow,K.A., Wang,J., Xia,D.F., Meng,X., Paschon,D.E., Leung,E., Hinkley,S.J. *et al.* (2011) A TALE nuclease architecture for efficient genome editing. *Nat. Biotechnol.*, **29**, 143–148.
60. Li,T., Huang,S., Zhao,X., Wright,D.A., Carpenter,S., Spalding,M.H., Weeks,D.P. and Yang,B. (2011) Modularly assembled designer TAL effector nucleases for targeted gene knockout and gene replacement in eukaryotes. *Nucleic Acids Res.*, **39**, 6315–6325.
61. Mali,P., Yang,L., Esvelt,K.M., Aach,J., Guell,M., DiCarlo,J.E., Norville,J.E. and Church,G.M. (2013) RNA-guided human genome engineering via Cas9. *Science*, **339**, 823–826.
62. Cradick,T.J., Fine,E.J., Antico,C.J. and Bao,G. (2013) CRISPR/Cas9 systems targeting β-globin and CCR5 genes have substantial off-target activity. *Nucleic Acids Res.*, **41**, 9584–9592.
63. Hsu,P.D., Scott,D.A., Weinstein,J.A., Ran,F.A., Konermann,S., Agarwala,V., Li,Y., Fine,E.J., Wu,X., Shalem,O. *et al.* (2013) DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol.*, **31**, 827–832.
64. Esvelt,K.M., Mali,P., Braff,J.L., Moosburner,M., Yaung,S.J. and Church,G.M. (2013) Orthogonal Cas9 proteins for RNA-guided gene regulation and editing. *Nat. Methods*, **10**, 1116–1121.
65. Jinek,M., Jiang,F., Taylor,D.W., Sternberg,S.H., Kaya,E., Ma,E., Anders,C., Hauer,M., Zhou,K., Lin,S. *et al.* (2014) Structures of Cas9 Endonucleases Reveal RNA-Mediated Conformational Activation. *Science*, **343**, 1247997.
66. Nishimasu,H., Ran,F.A., Hsu,P.D., Konermann,S., Shehata,S.I., Dohmae,N., Ishitani,R., Zhang,F. and Nureki,O. (2014) Crystal structure of Cas9 in complex with Guide RNA and target DNA. *Cell*, **156**, 935–949.