Chapter 11

Quantifying the Information Content of Homing Endonuclease Target Sites by Single Base Pair Profiling

Joshua I. Friedman, Hui Li, and Raymond J. Monnat Jr.

Abstract

Homing endonucleases (HEs) are native proteins that recognize long DNA sequences with high site specificity in vitro and in vivo. The target site specificity of HEs is high, though not absolute. For example, members of the well-characterized LAGLIDADG family of homing endonucleases (the LHEs) recognize target sites of ~20 base pairs, and can tolerate some target site base pair changes without losing site binding or cleavage activity. This modest degree of target site degeneracy is practically useful once defined and can facilitate the engineering of LHE variants with new DNA recognition specificities. In this chapter, we outline general protocols for systematically profiling HE target site base pair positions in order to define their functional importance in vitro and in vivo, and show how information theory can be used to make sense of the resulting data.

Key words Position-specific-scoring/weight matrix (PSSM/PWM), Information theory, Information content, DNA target sequence specificity, Target sequence specificity profiling

1 Introduction

The DNA binding surfaces of homing endonucleases (HE) must be structurally and chemically complimentary to their cognate DNA target sites to specifically recognize and cleave DNA [1, 2]. These surfaces form stabilizing intermolecular interactions with target site DNA that facilitate target site recognition, and stabilize the high-energy transition state leading to catalytic cleavage of the DNA phosphodiester backbone. Efforts to redesign HEs to recognize novel DNA target sites requires knowledge of both the starting specificity as encoded by contacts in the DNA–protein interface, and how these contacts can be modified to alter HE target site recognition specificity [3].

Structural analyses of HEs bound to their target sites have provided many useful insights into HE structure–function relationships [4]. These data have guided efforts to design HEs with altered target sequence specificities, but cannot directly identify

David R. Edgell (ed.), *Homing Endonucleases: Methods and Protocols*, Methods in Molecular Biology, vol. 1123, DOI 10.1007/978-1-62703-968-0_11,

the changes needed in an existing HE interface to generate a new recognition specificity. The explanation for this is that design alterations to the binding interface induce unanticipated structural rearrangements as residues repack to accommodate a new structural and chemical environment. These structural rearrangements in turn can alter or destroy existing or newly designed contacts to suppress high affinity binding or cleavage of the new target site [5]. Despite these challenges, some positions within HE DNA–protein interfaces have been found to readily accommodate design changes. Thus the systematic identification of base pair positions in the HE-DNA interface that are most conducive to redesign can guide the engineering of HEs with novel DNA recognition specificities.

Information theory is widely used to quantify the contribution of specific positions and base pairs to target site recognition and catalysis [6–9]. By systematically measuring the catalytic activities of HEs against a large set of DNA targets, information theoretic approaches can be used to identify the sequence features that are most important to site recognition and catalysis. These statistical models can accelerate protein engineering and design by screening out prohibitively difficult targets early in the design process, identifying permissive positions and directing subsequent efforts to the regions of the DNA–protein interface that represent design challenges. In this chapter, we provide a brief review of information theory and how it can be used to understand HE target site specificity. We then provide experimental protocols for the generation of HE site specificity profiling data, and suggest several useful ways to interpret and visualize the resulting data to aid HE design and engineering.

1.1 Position-Specific Scoring Matrices

Information theoretic approaches to HE target site modeling are based on large datasets that describe the relative preference of an HE for each DNA base (P_A, P_T, P_G, P_C) at each target site base pair position. These preferences can be concisely represented in the form of a Position-Specific Scoring Matrix (PSSM; also referred to as a Position Weight Matrix, or PWM), in which HE activity (binding and/or catalysis) at a given base or base pair (in rows) is recorded at each of the "N" positions in the target sequence (in columns). These scores are normalized such that each of the columns will sum to 1.

$$PSSM = \begin{pmatrix} P_{1,A} & \cdots & \cdots & P_{N,A} \\ P_{1,T} & \ddots & & P_{N,T} \\ P_{1,G} & & \ddots & P_{N,G} \\ P_{1,C} & \cdots & \cdots & P_{N,C} \end{pmatrix}$$

In the case of HEs, probability terms can be derived by simply measuring the relative efficiency with which an HE cleaves a target

site that contains a specific base pair substitution in the native DNA target site under single turnover conditions. Fully populating the PSSM/PWM matrix for a given HE requires determining HE activity against all possible "one-off" target sites that contain an A, C, G, or T at each target site base pair position. The number of target site sequences that need to be assayed in this way to fully populate a PSSM matrix is $3N+1$, where N is the target site length in base pairs (*see* Protocols below).

1.2 Definition of Information

Information can be usefully thought of as a reduction in uncertainty about outcomes, or in the parlance of information theory a reduction in "information entropy." For example, when an HE exclusively cleaves target sequences containing only one base at a given position, e.g., only an A (adenine) at position x, position x can be described as having an information entropy of zero: there is no uncertainty as to the identity of a DNA substrate at that position that will be cleaved by the cognate HE.

This relationship between information entropy and a statistical outcome can be further formalized by the Shannon entropy relation, Eq. 1 below, where H_x is the information entropy (a measure of uncertainty) associated with the DNA base at position x, and $P_{x,i}$ is the xth column and ith row of the PSSM matrix [10].

$$H_x = -\sum_{i=\mathrm{A,T,G,C}} P_{x,i} \cdot \log_2 \left| P_{x,i} \right| \tag{1}$$

If each of the four bases found in DNA is equally likely to occur at a given position, (i.e., if $P_\mathrm{A} = P_\mathrm{c} = P_\mathrm{G} = P_\mathrm{T} = 0.25$), then by evaluation of Eq. 1, the informational entropy of that base position would be 2 bits. The explanation for this value is that with four possible DNA bases at a position, two binary digits or bits are needed to uniquely specify the four possible bases at that position (e.g., in one possible encoding scheme A = (0 0), T = (0 1), G = (1 0), C = (1 1)). Site-specific DNA proteins by definition do not recognize all possible DNA bases with equal probability (*thus* $P_\mathrm{A} \neq P_\mathrm{T} \neq P_\mathrm{G} \neq P_\mathrm{C}$), and thus the informational entropy of specifically recognized DNA positions (following Eq. 1) will always be ≤ 2 bits of uncertainty. By extension, the information content of a single base, commonly written as R_info, is given by $R_\mathrm{info} = 2 - H_x$, where 2 is the information entropy of a randomly selected base and H_x is the information entropy of all the possible cognate bases at that position.

1.3 Information Content of a HE DNA Target Site

One way to calculate the information content of HE DNA target site would be to simply sum the R_info values across all target site base pair positions. This approach assumes that base pair recognition is independent of sequence context, but this is known not to be the case: specific base pair recognition often involves additional binding avidity contributions from adjacent base pairs.

This influence of additional positions X on the informational entropy of recognition at sequence position Υ is specified in Eq. 2 below, a conditional entropy equation.

$$H(X \mid \Upsilon) = -\sum_{i=\mathrm{A,T,G,C}} \sum_{j=\mathrm{A,T,G,C}} P(X_i \mid \Upsilon_j) \cdot log_2 \left| \frac{P(X_i \mid \Upsilon_j)}{P(\Upsilon_j)} \right| \quad (2)$$

Fully accounting for all the interdependencies in a target sequence using Eq. 2 would require evaluating an exponentially increasing number of terms $P(X_i|\Upsilon_{1j}, \Upsilon_{2j}, \ldots, \Upsilon_{nj})$ for each additional base in the target site. Experimentally evaluating these interdependent probabilities for long HE target sites is prohibitively difficult, even using new high-throughput approaches. Thus cases more complicated than the simplest assumption of Eq. 1 are rarely, if ever, considered.

A work-around that is still highly informative and practically useful is to experimentally determine "one-off" dependences, in which informational entropy is measured within a fixed sequence context where Eq. 1 remains phenomenologically valid. These experimentally derived measures of information content are target site sequence-dependent, and as a result capture a portion of the information contained in and contributed by adjacent or nearby base pairs. In the protocols below we describe how to perform target site single base pair scans, and show how information theory and visualization can be used to make sense of the resulting data.

2 Materials

2.1 Cloning HE Target Sites into Plasmid DNA

1. The pDR-GFP-universal plasmid harbors the target sites and is used for combined in vitro/in vivo cleavage analyses (*see* map and details at: http://depts.washington.edu/monnatws/plasmids/pDR-GFP%20univ.pdf).
2. Oligonucleotides need to be synthesized for all single base pair target site variants on a 25 nmol scale. The complementary pairs should be designed so that when annealed they generate a dsDNA target site insert with XhoI/SacI sticky ends to facilitate directional cloning into the pDR-GFP-universal vector. The number of oligonucleotides that need to be synthesized for a systematic scan of a target site that is N base pairs long is $2 \times (3N + 1)$.
3. DR-GFP target site sequencing primer, 5′-GGGGAGGGC CTTCGTGCGTCGC-3′. Primers should be synthesized on a 25 nmol scale and resuspended in oligo storage buffer (10 mM Tris-Cl, pH 8.5) to generate a 100 μM stock. Suspended oligos can be stored at −20 °C and thawed as needed.

4. Luria Broth: 10 g tryptone, 5 g yeast extract, 10 g NaCl in 1 L of water. Autoclave and store at room temperature or 4 °C.
5. *E. coli* DH5α chemically competent host cells.
6. 10× T4 polynucleotide kinase (PNK) reaction buffer: 0.7 M Tris–HCl pH 7.6 (at 25 °C), 0.1 M $MgCl_2$, 50 mM dithiothreitol, and 10 mM rATP.
7. T4 polynucleotide kinase (PNK): 10,000 U/mL.
8. DNA annealing buffer: 50 mM Tris–HCl pH 7.6, 0.5 M NaCl.
9. T4 DNA ligase: 400,000 U/mL.
10. PCR Cleanup Kit.

2.2 In Vitro "Barcode" Cleavage Assay

1. Reaction buffer: 10 mM $MgCl_2$, 20 mM Tris–HCl pH 8.0 (*see* **Note 1**).
2. Stop buffer (3×): 300 mM EDTA, 0.3 % SDS (w/v), 3.9 % Ficoll 400 (w/v).
3. 1× TBE buffer: mix 10.8 g Tris base, 5.5 g boric Acid, and 20 mL of 0.5 M EDTA and add water to 1 L.
4. Taq thermophilic DNA polymerase: 5,000 U/mL.
5. Taq PCR buffer (10×): 500 mM KCl, 15 mM $MgCl_2$, 100 mM Tris–HCl pH 8.3.
6. dNTP stock: equimolar mix of 10 mM dATP, 10 mM dTTP, 10 mM dGTP and 10 mM dCTP.
7. Betaine: 4 or 5 M stock solution in H_2O.
8. PCR Cleanup Kit.
9. Purified homing endonuclease protein.
10. Primer pairs for amplification of target sites from pDR-GFP-universal: for 1.3 kb substrate fragments forward primer 5′-GG GGAGGGCCTTCGTGC GTCGC-3′ and reverse primer 5′-GTGGTATGGCTGATTATGATCTAGA GTCGC-3′; for 1.6 kb substrate fragments forward primer for 1.6 kb fragment 5′-TTTATGGTAATCGTGCGAGAGGGCGCAGGG-3′ and reverse primer 5′-TTGTGATGCTATTGCTTTATTTGTAAC CATTATAAGCTGC-3′; for 1.9 kb substrate fragments forward primer 5′-GCGGGG CTCGCCGTGCC-3′ and reverse primer 5′-CCTCTGTTCCACATACACTT CATTCTCAGT ATTGTTTTGCC-3′; and for 2.2 kb substrate fragments forward primer 5′-GGGCTGCGAGGGGAACAAAGGCTGCGT GCGGGG-3′ and reverse primer 5′-CCAAATTAAGGGCCA GCTCATTCCTCCCAC TCATG-3′. Primers are synthesized on a 25 nmol scale, and resuspended in nuclease-free water to generate 100 μM stocks that are store at –20 °C until use.
11. Electrophoresis gel image quantification software (e.g., ImageQuant and Image J).

2.3 In Vivo Cleavage Assay in Human Cells

All of the following reagents should be purchased or prepared sterile in order to ensure the success of the in vivo cleavage profiling protocol outlined below.

1. Complete growth medium: Dulbecco-modified Eagle's medium supplemented with 10 % (v/v) fetal bovine serum and 1 % penicillin/streptomycin.
2. Human HEK 293 T cells.
3. Sterile 0.25 mM $CaCl_2$.
4. Sterile 2× BBS buffer.
5. 1× phosphate-buffered saline.
6. 0.25 % trypsin–EDTA.
7. Flow cytometry analysis software (e.g., FlowJo).
8. Plasmids: pEGFP-C1 plasmid (Clontech), a transfection efficiency and flow cytometry positive control plasmid; expression plasmids for the homing endonuclease being profiled; pDR-GFP-universal reporter target site plasmids with single base pair variant target sites cloned into the XhoI/SacI cloning site.

3 Methods

An overview of the experimental protocols outlined below for in vitro and in vivo cleavage profiling assays is shown in Fig. 1.

3.1 Cloning HE Target Sites into Plasmid DNA

This protocol details how to generate a library of all possible single base pair variant HE target sites in a common plasmid vector backbone that can then be used for both in vitro and in vivo cleavage profiling assays.

1. The synthetically prepared top and bottom strands of each target site should be designed to form XhoI and SacI sticky-ends when annealed to facilitate directional cloning into the pDR-GFP-universal plasmid. Resuspend individual top and bottom strand oligonucleotides at 100 μM in nuclease-free sterile water.
2. Target site oligonucleotides need to be phosphorylated prior to annealing to facilitate cloning. In separate PCR tubes combine 5 μL of each DNA oligonucleotide with 1 μL of 10× T4 polynucleotide kinase (PNK) reaction buffer and 4 μL of nuclease-free H_2O. Mix and then add 1 μL of 10,000 U/mL T4 PNK, mix again gently by pipetting up and down, then incubate at 37 °C for 1 h.
3. Mix pairs of complementary, phosphorylated oligonucleotides in a single PCR tube to anneal the top and bottom strands of each test target site. Dilute to a final concentration of ~50 nM dsDNA by adding 180 μL of DNA annealing buffer. Heat to 95 °C for 5 min followed by a slow cooling to 25 °C (at –1 °C/min).

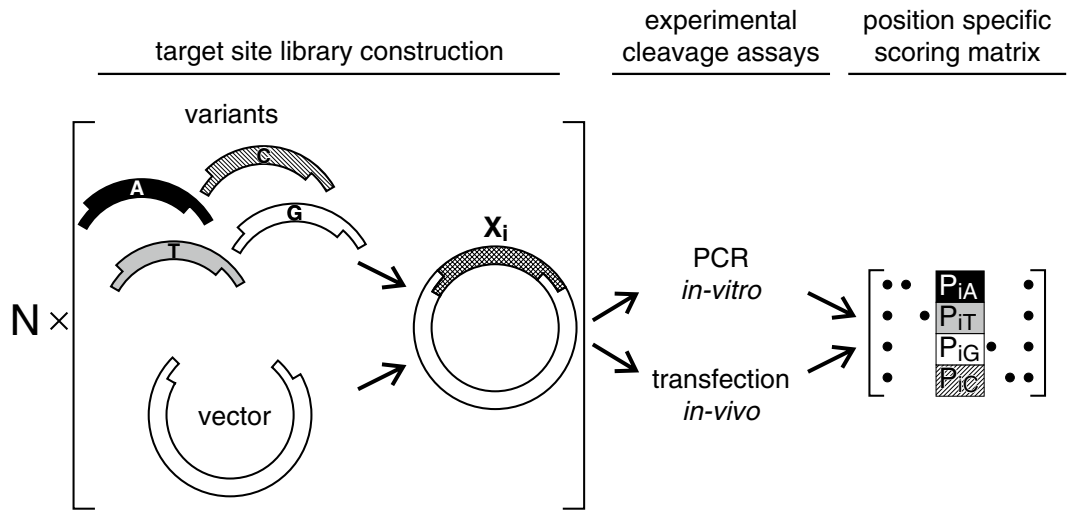

Fig. 1 General outline of profiling protocols. A library of plasmids harboring all target site single base pair variants of a target site *N* bases long is first assembled. HE specificity is assessed by in vitro cleavage of substrate DNAs PCRed from the plasmid site library, or by the in vivo, cleavage-dependent generation of GFP+ cells. The cleavage activity in both assays can be used to generate HE-specific target site cleavage matrices that form the basis for assessing the information content of an HE target site. These data can also be used as a statistical description of HE target site specificity and activity

4. Prepare pDR-GFP universal target site plasmid vector by double-digest 50 ng of plasmid (or ~3 μg for a library of 61 sites) with the restriction enzymes XhoI and SacI using manufacturer's recommended double digest conditions. Stop the reaction by heating at 80 °C for 20 min, then purify cleaved plasmid DNA using a PCR cleanup kit. Adjust the concentration of cleaved plasmid stock with nuclease-free water to a final concentration of ~25 ng/μL (or A_{260} ~0.5). A small aliquot can be run on an agarose gel to check the completeness of digestion if desired.
5. Ligate target sites into plasmid DNA by combining 1 μL of cleaved pDR-GFP-universal plasmid DNA from **step 4** in a fresh tube with 1 μL of phosphorylated, annealed target site insert from **step 3**. Add 6 μL nuclease-free water and 1 μL 10× T4 DNA ligase buffer, then mix gently. Add 1 μL (10 U) of T4 DNA ligase and mix gently, then incubate at room temperature for ≥1 h. To facilitate subsequent steps it is desirable to do all of the ligations needed to generate a library in parallel in a 96-well PCR plate (*see* **Note 2**).
6. Following ligation, chill samples on ice for 15 min, then add ~30 μL of chemically competent DH5α *E. coli* host cells/well. Heat shock by placing plates in a thermocycler pre-equilibrated at 42 °C. After a 45 s return the plate to an ice bucket for 1 min. Add 500 μL of sterile LB media to each well, then transfer well contents into individual deep-well 96-well plate

wells and cover with a gas-permeable top prior to shaking gently to recover at 37 °C for 1 h.

7. Plate 200 μL of each transformation onto LB-agar plates containing 100 μg/mL ampicillin. Grow plates overnight at 37 °C, or until well-delineated individual colonies are visible.
8. Use sterile toothpicks to transfer well isolated bacterial colonies into a 96-well PCR plate-containing 20 μL of ultrapure water/well. Mix gently to break up colony cell clumps.
9. Transfer 10 μL of each resuspended colony into a fresh, deep-well, 96-well plate containing 500 μL of LB or TB media/well supplemented with 50 μg/mL ampicillin. Grow overnight in a 37 °C plate shaker.
10. Isolate target site plasmids from the overnight cultures using your preferred plasmid DNA isolation method.
11. Combine 10 μL of purified plasmid DNA with 5 μL of a 1:100 dilution of sequencing primer and submit for DNA sequencing to verify all the targeted sequences (*see* **Note 3**).

3.2 In Vitro "Barcode" Cleavage Profiling of HE Target Sites

This protocol uses pooled sets of oligonucleotide substrates generated by PCR amplification of target sites cloned into pDR-GFP-universal in competitive cleavage reactions. In each reaction the cleavage sensitivity of all four base pair variants at a target site base pair position are directly compared in the same single tube digest. Full "one off" target site libraries can be easily profiled using this "barcode" cleavage protocol, and the resulting cleavage reactions displayed on a single agarose gel, as shown in outline in Fig. 2a.

1. Four forward and four reverse primer pairs need to be designed and synthesized, to amplify individual target site variants cloned into pDR-GRP-universal in Subheading 3.1 above. Primer sets should be designed to generate DNA fragments that are different enough in size to be easily resolved on a 1.0–1.2 % agarose gel, and in which the DNA fragment size is coded to be directly informative of the base pair variant present at base pair positions in that fragment. For example, our primer sets generate PCR products of approximately 1.3, 1.6, 1.9 and 2.2 kb, with the HE cleavage site located at the center of the fragment and in which all 2.2 kb fragments contain A's (adenines), 1.9 kb fragments C's, 1.6 kb fragments G's and 1.3 kb fragments T's as the variable base pair in amplified target sites across all base pair positions. This design allows four PCR fragments containing all four base pairs at each target site base pair position to be combined, digested and displayed in a single tube-1 lane agarose gel assay to generate target site position- and base pair-specific cleavage "barcodes" (*see* **Note 4**).
2. Adjust the volume of each site primer to a final concentration of 10 μM, and each target site plasmid to ~50 ng/μL ($A_{260} = 1.0$).

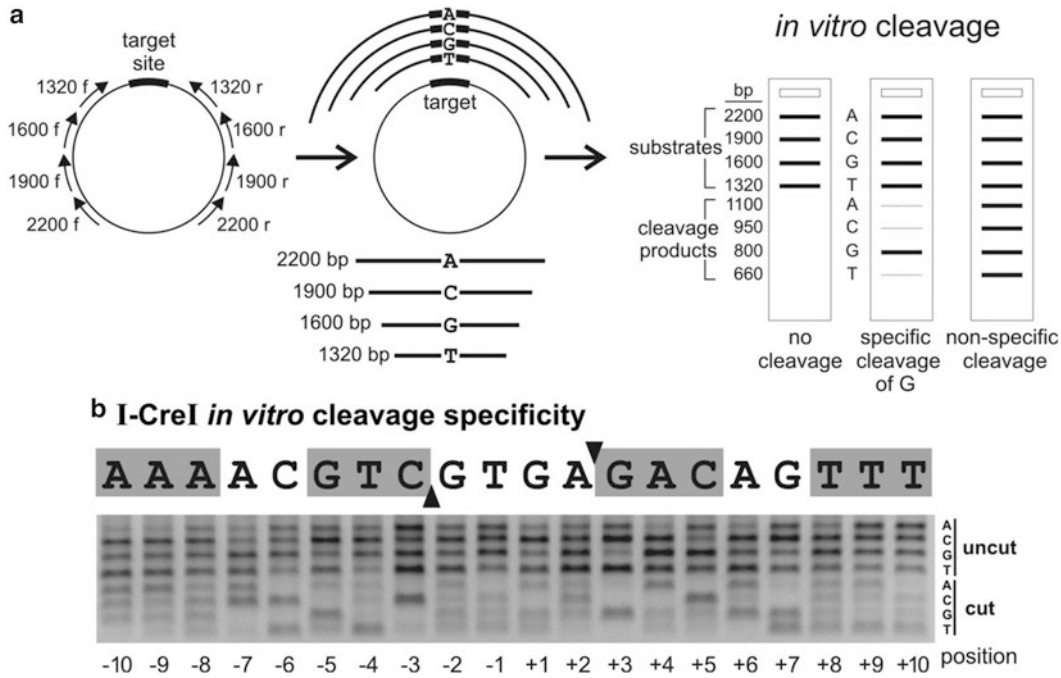

Fig. 2 In vitro "barcode" cleavage profiling of HE target sites. (**a**) Substrate DNA fragments are generated by PCR amplifying individual HE target site variants from the pDR-GFP-universal plasmid backbone using primer pairs that generate different sized substrate molecules in which fragment size encodes the variant base pair identity. Pools of the four PCR fragments covering each target site position and nucleotide possibility are then used in a 1-tube competitive cleavage assay. This approach allows all target site single base pair variants and their cleavage products to be assayed and quantified in a single experiment on an agarose gel. (**b**) Example of barcode cleavage profiling of the I-CreI HE cleavage site using the monomerized version of I-CreI (i.e., mCreI; [ref]) as the cognate HE. (Panels (**a**) and (**b**) are taken from Li H, Ulge UY, Hovde BT, Doyle LA, Monnat RJ Jr. (2012) Comprehensive homing endonuclease target site specificity profiling reveals evolutionary constraints and enables genome engineering applications. *Nucleic Acids Res.* 40(6):2587–2598. Epub 2011 Nov 25. PMID:22121229)

3. Set up PCR amplification reactions for each target site variant. For each target site plasmid mix 2 μL target site plasmid DNA, 4 μL base pair-specific forward primer, 4 μL base pair-specific reverse primer, 1 μL dNTP mix, 5 μL 10× Taq Buffer, 12.5 μL 4 M betaine, 2 μL formamide, 24 μL nuclease-free water, and 1 μL of Taq DNA polymerase (*see* **Note 4**).
4. PCR amplify target sites using 30 cycles of 95 °C for 30 s, 60 °C for 30 s, and 68 °C for 120 s followed by a single final incubation at 68 °C for 5 min.
5. Clean up PCR reactions using a PCR cleanup kit and elute DNA in ~30 μL nuclease-free water. Calculate the amount of DNA and its concentration based on A_{260} using the equation $[C] = A_{260} \times (50/\lambda \times 650)$, where [C] is concentration in μM and λ is the length of the PCR product in kb.

6. For each target site base pair position, mix in a single tube an equimolar ratio of the four PCR products (10 nM each) corresponding to each of the base pair variants (A, G, C, and T) at that position in a final volume of 10 μL (*see* **Note 5**).
7. Add homing endonuclease protein to each pooled template tube to a final concentration of 40 nM (this corresponds to 1:1 HE molecule/substrate DNA molecule in the example here). Adjust volume to 20 μL using Reaction Buffer and mix gently. Incubate the reaction mixture at 37 °C for 15 min, then stop digestions with 4 μL/tube of 6× Stop Buffer (*see* **Note 6**).
8. Load digests into single lanes of a 1.0–1.2 % agarose/1× TBE gel and run at 90 V for 2 h (*see* Fig. 2b for an example). Stain the gel in 1 μg/mL ethidium bromide gel buffer for 40 min with gentle shaking, then destain in water for 10 min prior to visualizing bands under 302 nm UV illumination (wear proper eye protection!). Take care not to oversaturate the detector in any part of the gel (*see* **Note 7**).
9. Integrate the intensities of each band on the gel using image quantification software. The signal intensity of each cleaved band (the two cleaved products in the substrates designed as described above run as a double-intensity, unresolved doublet; *see* Fig. 2b and **Note 4**) should be divided by the signal intensity of its corresponding un-cleaved substrate band plus the cleaved product band(s) to find the fractional cleavage of each substrate. The relative cleavage efficiency of target sites with single base pair changes is calculated by dividing the cleavage efficiency of target sites with single base pair changes by the cleavage efficiency of native target site base in the same lane. These results can be used directly to populate different displays of the results (*see* Subheading 4 below, **Note 8**).

3.3 In Vivo Cleavage Profiling of HE Target Sites

The same pDR-GFP-universal target site library used in Subheading 3.2 above can be used directly to profile the cleavage sensitivity of all single base pair variant HE target sites in human cells. This is done by co-transfecting each target site plasmid together with an HE coding plasmid into cells, then using flow cytometry to detect and quantify cleavage events that promote recombination and the generation of GFP+ cells. An example of this assay is shown in Fig. 3.

This in vivo cleavage assay takes advantage of having target sites cloned into the 5′ copy of the GFP (green fluorescence protein) genes contained in the pDR-GFP-universal plasmid. Target site insertion inactivates the 5′ GFP gene, which can be repaired after cleavage off the downstream, inactive 3′ GFP copy to restore the GFP open reading frame (Fig. 3). In this system the efficiency of in vivo cleavage of pDR-GFP-universal target site plasmids can be estimated from the frequency of GFP+ cells.

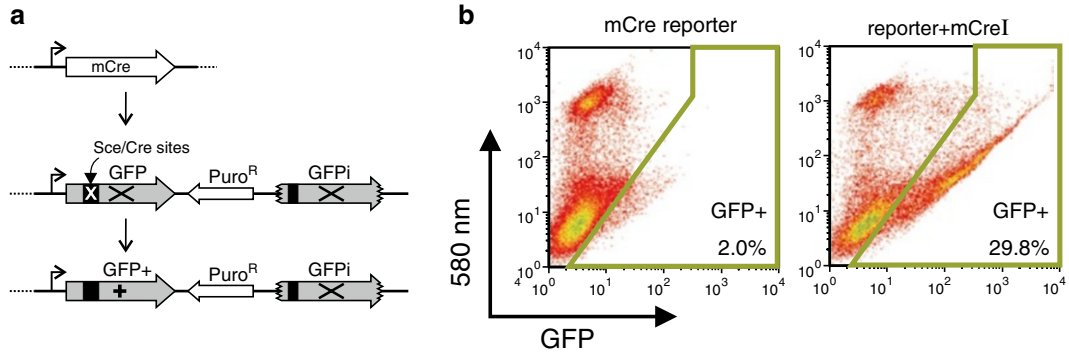

Fig. 3 In vivo cleavage profiling of HE target sites in human cells. (**a**) The pDR-GFP-universal target site plasmid facilitates in vivo cleavage profiling of HE target site variants that are cloned into the 5′ copy of GFP (green fluorescence protein gene) contained in pDR-GFP-universal. The HE target site insertion together with stop codons inactivate the 5′ GFP gene, whereas the downstream duplicated 3′ GFP gene is inactivated by open reading frame truncations. (**b**) Cleavage of the 5′ HE target site in cells by a co-transfected and co-expressed HE stimulates homology-dependent repair of the 5′ GFP copy off the downstream, inactive 3′ GFP copy to restore the GFP open reading frame. GFP+ cells can then be detected and quantified by flow cytometry. GFP+ cell generation closely parallels in vivo cleavage of the pDR-GFP substrate, and thus can be used to profile HE cleavage activity in vivo on target site variants. (Figure from Li H, Ulge UY, Hovde BT, Doyle LA, Monnat RJ Jr. (2012) Comprehensive homing endonuclease target site specificity profiling reveals evolutionary constraints and enables genome engineering applications. *Nucleic Acids Res.* 40(6):2587–2598. Epub 2011 Nov 25. PMID:22121229)

1. Seed 1.5×10^5 HEK 293 T cells in 500 µL of complete growth medium into each well of a 24-well plate, then incubate in a 37 °C humidified, 5 % CO_2 incubator for 24 h. Set up at least duplicate wells for each target site sample you plan to assay.
2. Prepare a transfection mix for each target site plasmid in a 1.5 mL microfuge tube that consists of 1.5 µg of plasmid DNAs in 10 µL H_2O to which 40 µL 0.25 M $CaCl_2$ and 40 µL 2× BBS buffer are added. The plasmid DNA amount for 24-well format transfections should be a total of 1.5 µg/well with a 3:1 M ratio mix of HE expression plasmid to pDR-GFP-universal target site plasmid. For control transfections, pUC19 plasmid DNA can be used to make up the difference in plasmid DNA amount to ensure a consistent 1.5 µg for transfections.
3. Mix transfection reactions by finger flicking tubes, incubate for 15 min at room temperature (~22 °C), then add drop by drop into plate wells. Place cells in a humidified, 37 °C, 3 % CO_2 incubator for 24 h to allow precipitate to form and transfection to occur (*see* **Notes 9** and **10**).
4. Refeed transfections by gently aspirating medium from transfected cells, and replacing it with 500 µL fresh complete growth medium before moving the plates to a humidified, 37 °C 5 % CO_2 incubator for an additional 24 h.
5. Harvest cells for flow cytometry: 48 h after transfection, aspirate the growth medium from each well and gently wash cells

once with pre-warmed PBS. Add 100 μL of trypsin–EDTA per well, and incubate the plate in 37 °C incubator for a couple of minutes to allow cells to detach. Add 400 μL fresh medium per well and pipet cells up and down gently to generate a single cell suspension. Transfer cells in media to a flow cytometry tube (a sterile 5 or 7 mL snap cap polystyrene or polypropylene tube), then transport on ice to the cytometer.

6. Flow cytometry analysis: use positive- and negative-control cell/plasmid combinations to determine gating to reliably identify GFP+ cells, then count 50,000 events for each transfected sample. Quantify the fraction of events that are GFP+ over total events.
7. Calculate target site cleavage efficiency: divide the number of GFP+ positive cells in co-transfected samples by the number of GFP+ positive cells observed in reporter-only transfections. The relative cleavage efficiency of a given target site variant can be determined by dividing the GFP+ frequency of that target site by the GFP+ frequency of the native target site determined in the same assay. This transient transfection assay is simple to multiplex and can detect even low levels of target site-specific cleavage despite a relatively high background that results from reporter plasmid DNA breakage upon transfection.

3.4 Data Analysis and Visualization

The protocols above yield information on the relative preferences of HEs for different target sequences. Thus a first step in visualizing these data is to appropriately normalize these ratios. There are two ways these relative data can be normalized: (1) by assigning the native DNA target site a relative activity value of 1.00, or (2) alternatively, normalizing the relative activities across all four target site base pair possibilities at a position such that each column of the PSSM sums to 1.00. This second normalization is required to calculate information entropy via Eq. 1. As this chapter is focused on the information content in DNA target sites, we typically use normalization 2, where the probability terms can be thought of as representing the fraction of HEs that would bind and cleave a specific substrate when in the presence of the other three competing target sites. Once appropriately normalized, PSSMs can be displayed as a matrix (Fig. 4 top) or transformed into a graphical representation of the data.

Sequence logos are a common way to visualize the data in PSSMs: the height of each base letter in a sequence logo is proportional to its corresponding value in the PSSM. To generate a sequence logo from a PSSM, first translate it into TRANSFAC motif format [11]. Briefly, TRANSFAC format is an ASCII text file that begins with a header line "P0 A C G T," and lists the contents of the PSSM as "xx $P_{xx,A}$, $P_{xx,C}$ $P_{xx,G}$ $P_{xx,T}$" in separate lines where xx

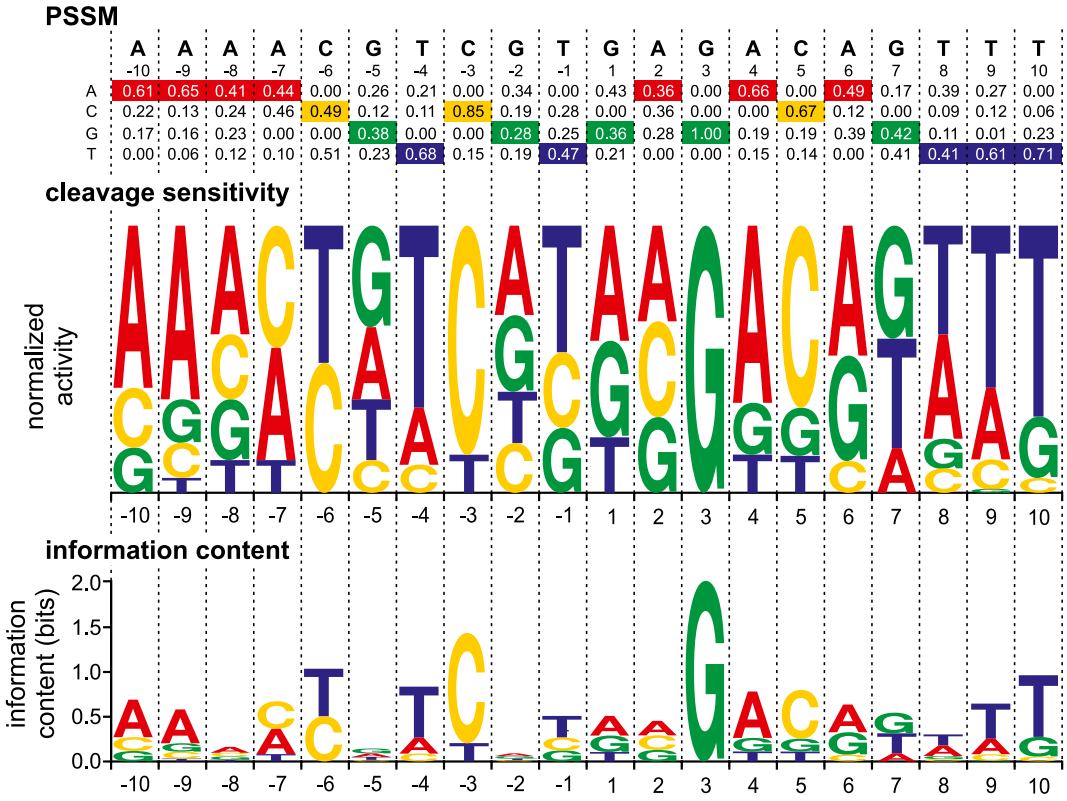

Fig. 4 Common methods for the visualization of target site specificity information. (**a**) Position Specific Scoring Matrix (PSSM) for the HE I-CreI where the native base is colored at each position; (**b**) the PSSM depicted as a sequence iconograph where the height of each letter is proportional to the preference of the HE for target sites containing that base; and (**c**) a sequence iconograph where the height of each letter is proportional to the information in bits that base provides to the HE I-CreI upon binding

is the target sequence position number beginning with "01" and the $P_{xx,i}$ terms are the appropriate entries from the PSSM. This file can be fed directly into WebLogo [12] (*see*: http://weblogo.berkeley.edu), or into many other freely available logo generators that will produce plots representing sequence preference or information content.

Sequence logos are also a useful way to visualize the information content of each position in a target site sequence, and can again be automatically generated by WebLogo using the same TRANSFAC input file described above. In contrast to direct representations of the PSSM, representations of information content emphasize DNA target site positions most important to recognition by a site-specific binding protein. The reader is referred to http://weblogo.berkeley.edu for detailed instructions on how to generate and customize these target site representations.

4 Notes

1. Optimal cleavage reaction conditions vary between HEs, and should be determined and optimized in advance.
2. The ligation can be extended at 16 °C overnight to improve ligation efficiency. In addition, inactivation of ligase by heating the ligation reaction at 70 °C for 20 min before transformation can also improve ligation efficiency.
3. Many commercial DNA sequencing services will now perform colony sequencing and PCR cleanup for an additional fee.
4. To optimize signal in barcode cleavage assays, the target sites in amplified fragments should be located in the center of the amplicon. This ensures the cleaved products will appear as a single band of double intensity on agarose gels. Added betaine is essential for efficient amplification of target sites from pDR-GFP-universal plasmid DNA.
5. Accurate quantification of enzymatic activity requires equal molar ratios of each DNA substrate in the substrate mixture.
6. The cleavage conditions used here were chosen to favor 50 % native target site cleavage to provide the best dynamic range for assessing target site cleavage sensitivities. Cleavage conditions and sampling of the cleavage time course need to be determined and optimized for each HE of interest.
7. Using 1× TBE buffer typically results in better separation of DNA fragments in agarose gels than the use of 1× TAE buffer. Electrophoresis conditions should again be optimized in advance in order to achieve the best separations to facilitate easy quantification of substrate and product bands.
8. As the PCR "barcode" substrate fragments are of different lengths, it is important to note the band intensity is not directly proportional to concentration. Each substrate and corresponding reaction product band should be self-normalized before comparing the different substrates. These complications can be avoided if the fragments are instead end-labeled.
9. We have had good luck using CMV promoter plasmids to drive HE expression in several different human cell types. A wide range of mammalian expression vectors can be used for this purpose once verified.
10. pEGFP C1 (Clontech) is transfected as positive control to monitor transfection efficiency. Sterile water or DNA buffer can be used as negative control. pDR-GFP-universal plasmid containing a native HE target site is also included.

Acknowledgements

This work was supported by US National Institutes of Health Training Grant award to H.L. (5RL9HL092555); by a US National Institutes of Health U54 Interdisciplinary Research Roadmap award (1RL1 CA133831) to R.J.M. Jr; and by a Bill and Melinda Gates Foundation/Foundation for the National Institutes of Health Grand Challenges in Global Health award to R.J.M. Jr.

References

1. Choo Y, Klug A (1997) Physical basis of a protein-DNA recognition code. Curr Opin Struct Biol 7(1):117–125
2. Garvie CW, Wolberger C (2001) Recognition of specific DNA sequences. Mol Cell 8(5): 937–946
3. Chevalier BS, Kortemme T, Chadsey MS, Baker D, Monnat RJ Jr, Stoddard BL (2002) Design, activity, and structure of a highly specific artificial endonuclease. Mol Cell 10(4): 895–905
4. Stoddard BL (2005) Homing endonuclease structure and function. Q Rev Biophys 38(1):49
5. Thyme SB, Baker D, Bradley P (2012) Improved modeling of side-Chain-Base interactions and plasticity in Protein-DNA interface design. J Mol Biol 419:255–274
6. Chevalier B, Turmel M, Lemieux C, Monnat RJ, Stoddard BL (2003) Flexible DNA target site recognition by divergent homing endonuclease isoschizomers I-CreI and I-MsoI. J Mol Biol 329(2):253–270
7. Rosen LE, Morrison HA, Masri S, Brown MJ, Springstubb B, Sussman D et al (2006) Homing endonuclease I-CreI derivatives with novel DNA target specificities. Nucleic Acids Res 34(17):4791–4800
8. Ulge UY, Baker DA, Monnat RJ (2011) Comprehensive computational design of mCreI homing endonuclease cleavage specificity for genome engineering. Nucleic Acids Res 39(10):4330–4339
9. Li H, Ulge UY, Hovde BT, Doyle LA, Monnat RJ (2012) Comprehensive homing endonuclease target site specificity profiling reveals evolutionary constraints and enables genome engineering applications. Nucleic Acids Res 40(6):2587–2598
10. Schneider TD, Stormo GD, Gold L, Ehrenfeucht A (1986) Information content of binding sites on nucleotide sequences. J Mol Biol 188(3):415–431
11. Wingender E (1988) Compilation of transcription regulating proteins. Nucleic Acids Res 16(5 Pt B):1879
12. Crooks GE, Hon G, Chandonia JM, Brenner SE (2004) WebLogo: a sequence logo generator. Genome Res 14(6):1188–1190