Chapter 16

Identification and Analysis of Genomic Homing Endonuclease Target Sites

Stefan Pellenz and Raymond J. Monnat Jr.

Abstract

Homing endonucleases (HEs) are highly site-specific enzymes that enable genome engineering by introducing DNA double-strand breaks (DSB) in genomic target sites. DSB repair from an HE-induced DSB can promote target site gene deletion, mutation, or gene addition, depending on the experimental protocol. In this chapter we outline how to identify potential genomic target sites for HEs with known target site specificities and the different experimental strategies that can be used to assess site cleavage in living cells. As an example of this approach, we identify potential human genomic target sites for the LAGLIDADG HE I-CreI that, by nine different selection criteria, may be new "safe harbor" sites for gene insertion.

Key words Homing endonuclease (HE), DNA double-strand break (DSB), Position weight matrix (PWM), Position-specific scoring matrix (PSSM), Genomic target site, Safe harbor site (SHS)

1 Introduction

Homing endonucleases (HEs) are valuable reagents for genome engineering in many organisms because of their exceptionally long DNA target sites and high specificity of site cleavage [1–3]. HEs can be used to cleave specific genomic target sites to promote gene disruption, modification, or addition at the cleavage site. This engineering capability may be broadly generalizable to many genes and genomic regions, as HEs with different target specificities continue to be identified, and it has become easier to engineer new target site specificities for existing HEs [4]. Existing HEs can also be used to facilitate the most common gene therapy goal, which is therapeutic gene insertion. This chapter provides protocols for the identification and analysis of potential target sites for well-characterized HEs in sequenced genomes to facilitate basic science and enable therapeutic gene insertion.

The starting point for the identification of potential genomic HE target sites is a detailed knowledge of HE cleavage specificity. Homing endonucleases exhibit some flexibility in their DNA recognition

David R. Edgell (ed.), *Homing Endonucleases: Methods and Protocols*, Methods in Molecular Biology, vol. 1123, DOI 10.1007/978-1-62703-968-0_16,

sequence, i.e., they are able to tolerate some base pair changes within their target site without losing site-specific activity. This target site degeneracy can be quantified in the form of binding [5] or cleavage [6, 7] profiling of a target site: HE pair or the interrogation of complex target site libraries (ref. Chapter 11). The resulting binding or cleavage profiles can be integrated to generate an HE-specific target site position weight matrix (PWM) or position-specific scoring matrix (PSSM) to enable subsequent target site searches (ref. Chapter 11). Within a PWM, a numerical value reflecting binding or cleavage efficiency is assigned to each nucleotide at every target site position. PWM values are typically referenced to the native base at that position which is assigned to an activity of 100 %. PWM can also be generated that reflects the informational content of target site base pair positions (ref. Chapter 11).

Two useful, web-accessible tools can be used to convert HE-specific PWMs into lists of genomic target sites. The LAGLIDADG homing endonuclease database and engineering server (LAHEDES) [8] includes a growing list of LAGLIDADG HE target site PWM data and can be used directly to search for the best potential target sites in short DNA sequences, e.g., an individual gene. The NCBI's BLAST server [9] can be used with LAHEDES output to identify the best target sites for a given HE in genomic sequences. The use of these two search options in sequence is fast, as illustrated below, and typically identifies dozens or a few hundred potential genomic target sites in the human genome depending on how stringent the initial LAHEDES PWM search is and the quality of the genomic sequence being searched.

The quality of the starting genomic sequence, both in terms of accuracy and completeness, can strongly determine search output. The presence and nature of genomic variation, ranging from single nucleotide polymorphism (SNP) variants through short insertion–deletion variants (indels) to large-scale structural and copy number variants (CNVs) [10], can strongly influence the likelihood of experimental success. For example, global error rate estimates for the current, extensively analyzed and well-documented hg19 human genome build range from $1 \times 10e\text{-}6$ to $1 \times 10e\text{-}4$. This translates into thousands to hundreds of thousands of potential sequence differences between the genome of a person or human cell line and the corresponding genome sequence. Thus, it is essential that potential genomic HE target sites identified as described below be experimentally verified before embarking on any HE-enabled genome engineering protocol.

A simple way to verify potential HE target sites is to amplify and sequence the target site(s) from genomic DNA and use the same amplified fragment(s) as a substrate for HE digestion in vitro. This approach verifies the site is present, documents sequence differences between the genome sequence and genomic target, and provides a direct measure of the functional consequences of sequence differences between the native HE and genomic target sites. The cleavage

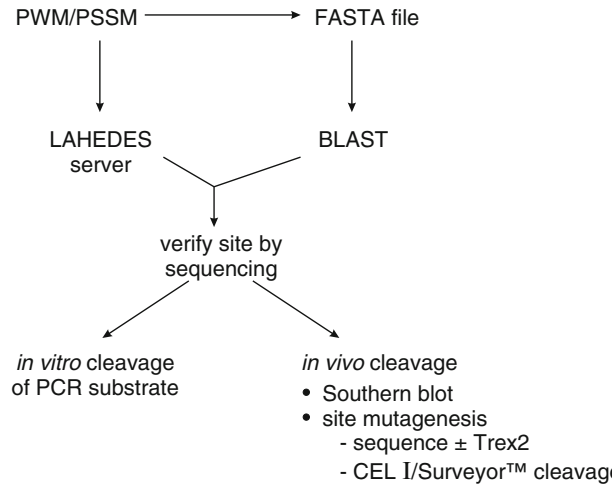


Fig. 1 Identification and analysis of genomic homing endonuclease target sites. Target site searches are driven by the use of HE-specific PWMs (position weight matrices, *upper left*) to drive target site searches in short sequences or in genomic DNA. Site searching and the generation of site libraries for FASTA conversion to facilitate BLAST searching (*top right*) of sequenced genomes or large blocks of sequence are facilitated by the LAHEDES (LAGLIDADG Homing Endonuclease Design and Engineering Server). Sites once identified by searches need to be verified, then can be further analyzed by in vitro (*lower left*) or in vivo (*lower right*) cleavage or sequence-based assays

sensitivity of a genomic HE target site in vivo can be assessed by direct measures of cleavage activity such as Southern blot analysis. Indirect measures of site cleavage in vivo are also very useful and can be more easily multiplexed than Southern blotting. For example, the efficiency of site cleavage in vivo can be estimated by determining how frequently a target site is mutated after HE expression. This approach takes advantage of the error-prone nature of both canonical and alternative nonhomologous DNA end-joining (NHEJ) pathways [11, 12] and can be rapidly implemented across many potential target sites to provide a minimum estimate of site cleavage efficiency. Additional data on the molecular nature of misrepair products can be obtained by DNA sequencing of small numbers of target sites. The sensitivity of all of these assays can be further enhanced by co-expressing an HE with the exonuclease TREX2 to promote error-prone rejoining up to ~25-fold [13] or by the use of highly accurate duplex sequencing protocols that can reliably identify target site mutations at frequencies as low as the background mutation frequency ($\leq 1 \times 10e\text{-}6$; [14]). The interrelationship of site searches and experimental site validation and analysis are shown schematically in Fig. 1. Protocols for these site analysis approaches are outlined below.

2 Materials

1. Oligonucleotide primers: these are designed to allow specific amplification of genomic target sites for target site confirmation by sequencing and for mutation analyses. Genomic primers of approximately 20 bases with melting temperatures around 55 °C work well for the amplification of human genomic target sites and can be designed using Primer3, Primer3Plus, Primer-BLAST, or other widely available PCR primer design tools [15, 16].
2. Genomic DNA purification kit.
3. PCR cleanup kit.
4. Spectrophotometer to measure DNA concentration.
5. HE cleavage buffer: optimized for the specific HE(s) to be used.
6. HE digestion stop solution: again, optimized for a specific HE(s) to be used.
7. Image analysis software: e.g., ImageJ (http://rsb.info.nih.gov/ij/) or equivalent.
8. Nylon hybridization membrane for Southern blot analysis.
9. Chemiluminescent kit.
10. Surveyor™ Mutation Detection Kit (Transgenomic, Omaha).
11. TA cloning vector with a high fidelity PCR polymerase that leaves 3′ A-tails.
12. Luria broth bacterial agar plates with ampicillin (50 μg/ml): protocols for this and other standard molecular biology and microbiology protocols can be found in several widely available protocols manuals.
13. IPTG, 1.2 g in 50 ml of H_2O, filter sterilized and stored at 4 °C.
14. X-gal, 100 mg in 2 ml *N,N′*-dimethylformamide, stored away from light at −20 °C.

3 Methods

Two types of target site searches are useful, either alone or in sequence, depending on whether you are looking for potential HE target sites in a specific gene or small number of genes or for sites located in a sequenced genome. The first, more limited search strategy can be efficiently implemented by making use of the HE-specific search matrices and search function contained in the LAHEDES homing endonuclease web server (http://homingendonuclease.net/). The second search protocol is to

identify potential HE-specific target sites in sequenced genomes. This search protocol makes use of the NCBI's BLAST search engine (BLAST: http://blast.ncbi.nlm.nih.gov/) together with a list of high-quality HE-specific target site sequences generated from the LAHEDES HE web server. The protocols for each search type are given below.

3.1 LAHEDES Server HE Target Site Searches

The LAHEDES web server facilitates HE target site searches in single genes or small number of genes. These searches make use of the previously defined HE-specific PWMs contained in the LAHEDES server that can be found by following "Browse>PWM Browser." Custom PWMs can also be defined by following "Entry'>'Custom PWM Entry." Weights or values for cleavage or binding activity at each target site position across all nucleotide combinations should sum to 1.0 to ensure proper handling of the new matrix in searches. An example of a LAHEDES search using a predefined search matrix is given below.

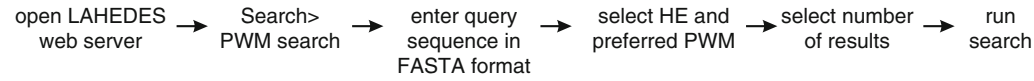
1. Open the LAHEDES web server in a browser window (*see* **Note 1**).
2. Go to "Search'>'PWM search."
3. Enter the query sequence in FASTA format into the input box. The current version of the server can accommodate searches in typical human genes (~100 kb of contiguous sequence), though lacks the capacity to do genome-scale searches.
4. Select the HE you wish to search against your target sequence and the corresponding HE-specific PWM you would like to use.
5. Select the number of search results you want returned.
6. Run the search.

Figure 2 shows this sequence of steps in outline and provides an example of the output from a search of a 5,020 bp long query sequence using two different PWMs for mCreI, the monomerized version of the canonical LAGLIDADG homing endonuclease I-CreI [6]. These PWMs are based on I-CreI/mCreI single base pair profiling and degeneracy data or represent the output of a straight identity search. Search output in each case is in the form of a tabular list of target sites in the query sequence, their location and orientation, and the location of base pair differences between the input DNA target site sequence and the mCreI target site. Quantitative assessment of the target site matches is given by a target quality score and number of mismatches compared to the wild-type sequence.

3.2 BLAST Server Genomic HE Target Site Searches

The NCBI's BLAST server [9] can be used to search query sequences against large target sequences, e.g., entire genomes. BLAST searches can be set up using HE-specific PWM data once it has been converted into a FASTA file format. The example

a



b

LAHEDES - PWM (PSSM) search

gene	position	-11 -10 -9 -8 -7 -6 -5 -4 -3 -2 -1 1 2 3 4 5 6 7 8 9 10 11	score	mismatches
WT sequence	N/A	G A A A A C G T C G T G A G A C A G T T T C	7.97976e-8	0
>231	f2500/5020	t A A t c t G T C a T a t G A C A G a T T a	5.82473e-7	8
>231	r2500/5020	t A A A c t G T C a T a t G A C A G T T T a	1.43114e-07	8
>231	f3112/5010	t A A t A t t T t G T G t a A t A a T T T t	6.32896e-11	10
>231	f1243/5020	t A A g A t t c C a c a g a A C A c T T T g	3.33558e-11	12
>231	r263/5020	G A A A A C a T C G T G A G A C g G c c T t	2.77944e-11	9

c

LAHEDES - identity search

gene	position	-11 -10 -9 -8 -7 -6 -5 -4 -3 -2 -1 1 2 3 4 5 6 7 8 9 10 11	score	mismatches
WT sequence	N/A	G A A A A C G T C G T G A G A C A G T T T C	0.00311348	0
>231	r2500/5020	t A A t c t G T C a T a t G A C A G T T T a	3.11348e-11	8
>231	f2500/5020	t A A A c t G T C a T a t G A C A G a T T a	3.11348e-11	8
>231	r60/5010	G g A A A a G T C c T G c a A C A a T a T t	3.11348e-11	8
>231	f4055/5020	G A A t A C G g C c T G g c A C A t T g c C	3.11348e-11	8
>231	f3047/5020	G A t A A g c T g G c a t a A t A G T T T C	3.11348e-12	9

Fig. 2 LAHEDES search steps, options, and outputs. (**a**) Outline of steps for a target site search in a short sequence using one of the predefined HE PWMs contained in the LAHEDES server. (**b**) Search output of potential I-CreI/mCreI target sites in a 5,020 bp query sequence from human genomic region chr4 58974113 to 58979132 in assembly GRCh37.p10 using a PWM/PSSM of the homing endonuclease mCreI that incorporates single base pair cleavage degeneracy information [7]. The search results display the HE native site sequence at top ("WT sequence") followed by a list of target sites in the query sequence (here designated ">231"). The position and strand on which the potential sites are located together with target site coordinates are listed next, followed by site sequences referenced to the 22 bp I-CreI/mCreI target site. Lower case, *red letters* indicate base pair positions where there is a difference between the target and the mCreI target site sequence. Nucleotide positions that match central four nucleotides of the native site are in *upper case* and *blue*. "Score" and "mismatch" columns give a quantitative assessment of site quality and the number of base differences, respectively. (**c**) An equivalent search using an "identity" PWM that identifies the closest matches between the native I-CreI/mCreI target site and the target or query sequence. This emphasizes the value of using matrices that incorporate site degeneracy data for searches

below illustrates how to search for I-CreI/mCreI sites in the human genome.

1. Develop a list of all HE-specific target sites you wish to BLAST search from PWM data. Cleavage degeneracy matrices are often the most useful for this step, as they are typically the best populated with data and reflect the most common goal of genomic target site identification which is to cleave and/or

modify these genomic target sites in vivo. The more stringent your site selection is at this step (e.g., only for sites that have a high likelihood of being cleaved with high efficiency), the fewer the sites your BLAST search is likely to return (*see* **Note 2**).

2. Generate a text file in which each line consists of a candidate target site sequence. This list should include all possible combinations of nucleotide positions and base pairs that exceed a defined functional threshold as outlined in **step 1**.
3. Convert this list of potential target sites sequences to FASTA format by preceding each sequence with a ">" and a unique site identifier.
4. Open the BLAST web server in a browser window.
5. From the list of BLAST Assembled RefSeq Genomes, choose "Human."
6. Upload your file of FASTA-compatible candidate target sites as the "Query Sequence."
7. Run BLAST with the following parameters (*see* **Note 3**):

   Database: Genome (reference only).

   Optimize for: Somewhat similar sequences (blastn).

   Max target seqs: 50.

   Short queries: Adjust for short sequences.

   Expect threshold: 1.

   Word size: 7.

   Match/mismatch: 4, −5.

   Gap cost: Existence: 12/Extension: 8.
8. In the results page that opens, chose each of the query sequences from the "Results for" drop-down menu.
9. Check the "Alignments" for query sequences that align perfectly to the target sequence (Fig. 3).
10. Follow the Sequence ID hyperlink to the NCBI reference sequence.
11. In the window that opens, expand the box "Change region shown"; chose "Selected region"; enter the coordinates for the hit in the BLAST results window and verify that the displayed sequence and the sequence for the BLAST hit are identical (Fig. 3).
12. Expand the region shown by changing the "selected region" to 2,500 bp upstream and downstream of the candidate target site.
13. Save this sequence by selecting "Send'>'Complete Record'>'File'>'Format:FASTA'>'Create File" to capture your putative target sites.

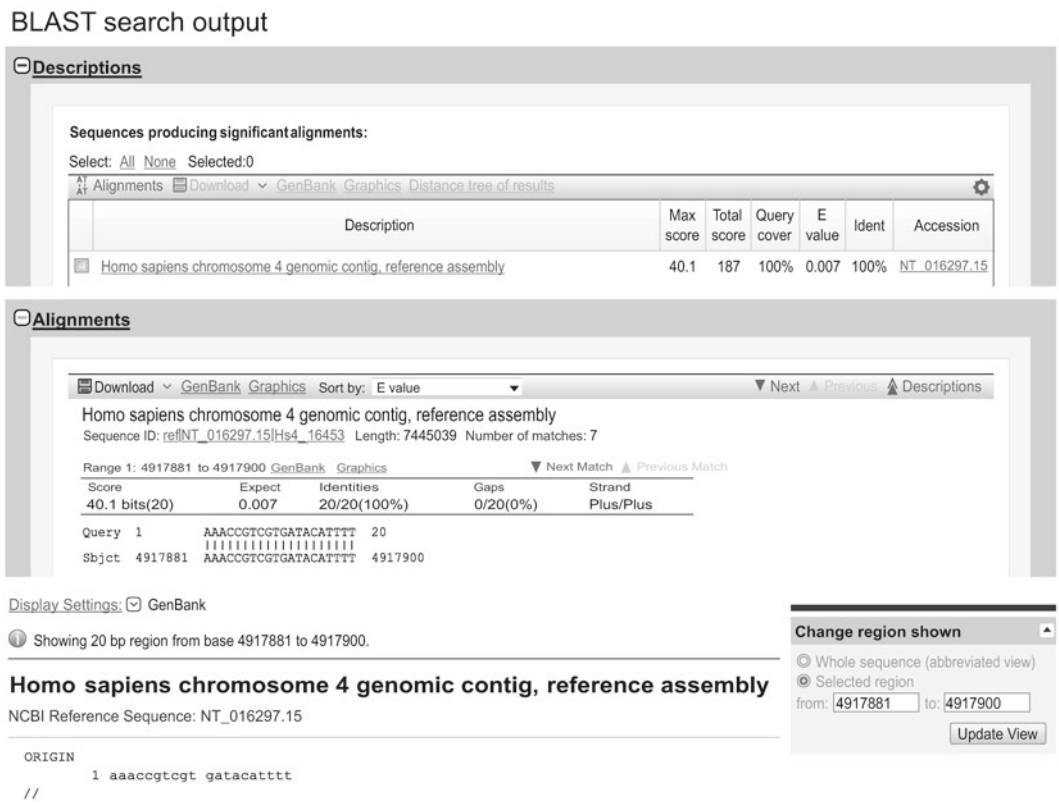


Fig. 3 BLAST search results for a potential I-CreI/mCreI human genomic target site. The *upper* and *center panel* shows a BLAST search hit. The alignment reveals a perfect match for one of the candidate target sequences on chromosome 4. The "Accession" hyperlink opens the identified target site match in a new window (*lower panel*). By changing the region shown in the *gray box* at right, it is possible to recover the flanking sequences of the candidate target site. This example was obtained using Build 36.3 of the "reference only" database

3.3 Sequence Verification of Genomic HE Target Sites

Search results are *potential* target sites: their existence and sequence need to be confirmed in your cells or host organism of interest before proceeding. This is most easily done by using the flanking genomic sequence captured in your BLAST search above to design oligonucleotide primers that can be used to amplify the putative site region from genomic DNA as a PCR product of >500 bp. Ideally, the target site is in the middle of the PCR product. This way, successful cleavage of the target site results in replacement of one substrate band by a second, smaller product band doublet.

1. Design PCR primers flanking your putative genomic HE target site using search output from Subheading 3.2 and the primer design tools listed in Subheading 2. Design a third sequencing primer that is located ~100 bp upstream or downstream of the putative HE target site.

2. Prepare genomic DNA using a suitable molecular biology kit.
3. Use the PCR primer pair from **step 1** above to amplify the region of interest from your genomic DNA sample, and run an aliquot on an agarose check gel with flanking size standards to determine whether the predicted size product has been generated and how many other potentially contaminating PCR products are present.
4. Gel-purify your target site band of interest, and use this as a template together with your site-specific sequencing primer to determine the DNA sequence of the putative target site and flanking genomic DNA.
5. Compare the sequence of the target site region of your genomic PCR product with both the reference sequence and your predicted target site sequence to confirm that the site exists and has the expected sequence. If unexpected sequence differences are present between your sequenced genomic site and the reference genome you are starting, PWN can be used to assess their potential functional consequences. You may be able to assess whether a sequence difference is a known human genomic sequence variant by using the dBSNP database (http://www.ncbi.nlm.nih.gov/projects/SNP/).

3.4 In Vitro Cleavage Analysis of Genomic HE Target Sites

Sequence confirmation of genomic HE target sites is reassuring, but often does not directly predict the cleavage sensitivity of the site, especially if there are multiple base pair changes between the genomic and native HE target site sequences. The PCR product from Subheading 3.3 above can be used as substrate in a cleavage reaction to confirm target site cleavage sensitivity. Digesting the PCR product with different concentrations of HE and including a native target site as a control in addition will reveal how cleavage sensitive a genomic target site sequence is relative to the native site. This protocol is shown in outline in Fig. 4.

1. Clean up the PCR product using a suitable purification protocol or kit.
2. Determine the concentration of the PCR product using a spectrophotometer. Calculate the molar concentration of the PCR product.
3. Prepare control and experimental sample reactions for each substrate using ~100 ng of PCR substrate in a final reaction volume of 15 μl. Each sample reaction should contain the same amount of substrate to simplify interpretation. A good starting point for the molar ratio of enzyme to substrate is equimolar (1:1), followed by a second reaction with ten times more enzyme than substrate.

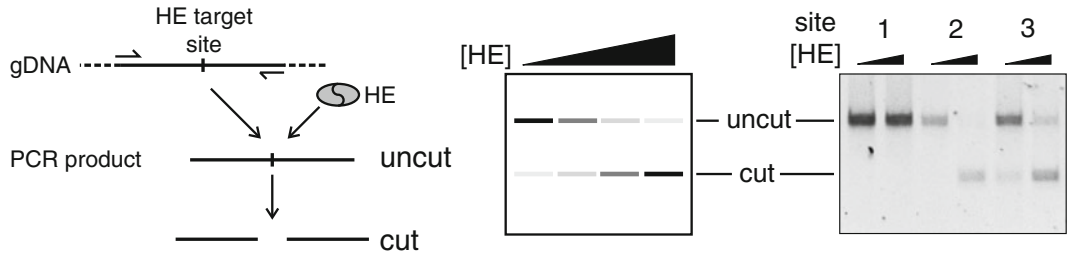


Fig. 4 In vitro cleavage verification of a potential HE genomic target site. (**a**) Schematic outline of Protocol in Subheading 3.4, in which a putative genomic target site is PCR-amplified from genomic DNA, then digested with a cognate HE. (**b**) Schematic and gel photo of panel (**a**) in which target site PCR substrate DNA is digested with an increasing amount of HE to verify site cleavage sensitivity. The gel examples show a cleavage-resistant target site (*left*) and partially (*center*) and largely (*right*) cleavage-sensitive sites. For each reaction 15 fmol of the substrate PCR product was digested with equal amounts of enzyme (*left lane* for each target) or ten times more enzyme (*right lane*)

Sample reaction:

Substrate: 15 fmol.

Reaction buffer 10×: 1.5 µl.

Homing endonuclease: 15 fmol.

H_2O: add to 15 µl.

4. Incubate the reaction mix for 1 h at 37 °C.
5. Depending on the homing endonuclease, add 1/10 volume stop buffer to the reaction (*see* **Note 3**).
6. Separate the digestion products on an agarose gel.
7. Determine the intensity of the bands corresponding to the digested and undigested PCR product bands with reference to the native site control using ImageJ or other image analysis software.

3.5 Southern Blot Analysis of In Vivo Target Site Cleavage

Southern blot analysis of target site cleavage in vivo is still a "gold standard" assay for HE activity in living cells. Cleavage time course profiles can provide a good sense of steady-state cleavage levels, and integration of these data over time can be used to estimate cleavage efficiency and investigate other aspects of HE-induced DSB repair such as repair kinetics and the genetic or functional requirements for DSB repair (see, e.g., [17]). Southern blot analysis can detect low frequencies of target site-specific cleavage (~0.5 % of potential target sites) that are difficult or impossible to detect by other strategies. The following is a general protocol that provides an overview of major steps in Southern blot analysis. For additional technical detail and more explicit protocols, see [18, 19] or one of the widely available molecular biology methods manuals.

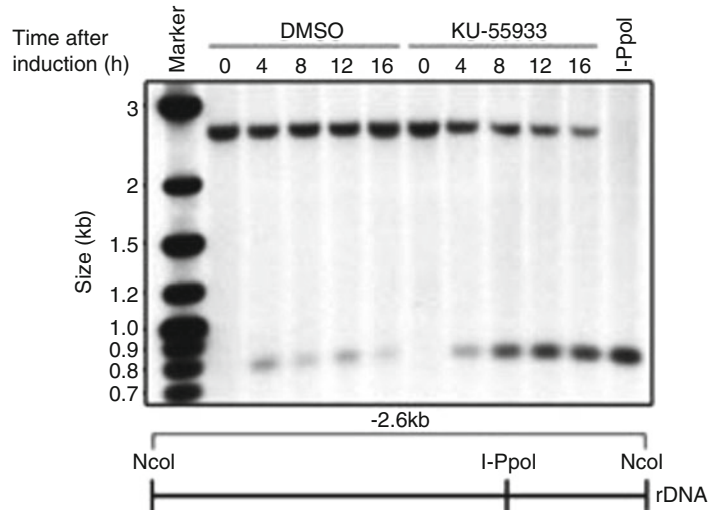


Fig. 5 Southern blot analysis of in vivo HE target site cleavage. Southern blot analysis of genomic DNA from cells expressing the I-PpoI HE in the presence or absence of the ATM inhibitor KU-55933 for the indicated times. The I-PpoI target is shown in a representation of the NcoI digestion product. The probe used for the Southern blot anneals to the smaller, 0.8 kb long cleavage fragment. This blot panel was previously published as fig. 16.4e in [18]

An example of a Southern-based analysis of cleavage of a genomic target site for the I-PpoI HE is shown in Fig. 5.

1. Express your HE in host cells by transfecting or infecting cells with an expression vector. Prepare a mock-infected control sample and any desired time point samples. Short time courses (24–36 h post-transfection) work well with most HEs unless your aim is to drive target site mutagenesis, when longer time points up to 72 or more hours may be advantageous (*see* **Note 4**).
2. Prepare genomic DNA from transfected cells using a genomic DNA purification kit.
3. Cut the genomic DNA with a restriction enzyme(s) to generate a target site fragment for blot analysis. The best starting fragments are ~5 kb long and have the HE cleavage site located asymmetrically in the resulting restriction fragment (*see* **Note 5**).
4. Determine the concentration of the digestion products using a spectrophotometer.
5. For each time point load 5 μg of digested genomic DNA/lane of a 1 % agarose gel. Separate the digestion products by electrophoresis in 1× TBE buffer for 16 h at 20 V on a 10-cm long agarose gel. Include as controls a genomic DNA sample that you have cleaved in vitro with your HE or a restriction endonuclease that is close to or in the HE target site. Include size standards.

6. Blot the separated DNA fragments from your agarose gel onto a nylon hybridization membrane.
7. Prepare labeled probe(s) to detect and quantify site-specific cleavage events. The PCR product used for target site sequence verification above can be used as a probe if it is free of repeats and known to have little or no cross-hybridization issues. This probe fragment can be captured by cloning if desired for future use. Probes can be designed to anneal to one or both of the HE/restriction endonuclease digestion products and can be labeled with radioactivity or a nonradioactive detection system (e.g., biotinylation) depending on your experience and radiation licensing status.
8. Detect probe hybridized to your membrane-bound digestion product. Radioactive probes can be detected directly by imaging on film or phosphorimager screens. Biotinylated probes can be detected using the chemiluminescent kit.

3.6 Analysis of In Vivo Target Site Cleavage by Site Amplification and Cleavage

A simpler, though less sensitive, approach to assess in vivo site cleavage is to amplify the target site from cells after HE expression to determine their cleavage sensitivity. This approach takes advantage of the fact that HE target sites cleaved in vivo may undergo error-prone repair [11, 12]. The mutagenic "footprints" of error-prone DSB repair can be detected by HE cleavage, restriction endonuclease cleavage, or mismatch nuclease cleavage of target site DNA fragments PCR-amplified from HE-expressing cells. All of these methods can be further enhanced by co-expressing an HE with the TREX2 3′ repair exonuclease in vivo: TREX2 degrades free DNA ends, antagonizes the error-free religation of cleaved target sites, and thus promotes the generation of mutant target site repair products. TREX2 co-expression is discussed first below.

3.6.1 Co-expression of HEs and TREX2 in Human Cells

Several different types of expression systems can be used to co-express an HE and the exonuclease TREX2 [13]. The open reading frames can be cloned together in one expression plasmid, separated either by an internal ribosome entry site (IRES) or a 2A ribosome skipping sequence [20, 21] to ensure co-expression of the two gene products. It is advantageous to integrate a fluorescent protein (e.g., mCherry) into the same expression plasmid downstream of the two ORFs to allow for easy screening (and, if desired, sorting) of cells that co-express an HE and TREX2. Alternatively, the HE and TREX2 proteins can be expressed from two different plasmids that are co-transfected at the same time.

The presence and frequency of HE target site mutations can be assayed by digesting genomic DNA target site sequences with the cognate HE or with a restriction enzyme that cleaves within the HE target site as outlined above in Subheading 3.4. Target site PCR products from HE-expressing and control cells can also be annealed to generate mismatches between mutant and control target sites

that can be detected with the mismatch-cleaving nuclease CEL I (available commercially as the Surveyor™ nuclease cleavage assay). The CEL I/Surveyor™ endonuclease [22] is a member of the plant-derived CEL nuclease family [23] that cuts DNA at nucleotide mismatches.

3.6.2 CEL I/Surveyor Cleavage of Target Site PCR Products

1. Prepare two cell cultures: one will be transfected to express HE (and TREX2, if desired), and the other will be mock-transfected to serve as a control.
2. Prepare genomic DNA from both cultures using the genomic DNA purification kit.
3. PCR amplify the putative homing endonuclease target site region from both samples.
4. Clean up the PCR products using a suitable purification protocol or kit.
5. Verify the quality of the PCR products on a 1 % agarose gel.
6. Determine the concentration of the PCR products using a Nanodrop spectrophotometer.
7. Mix, denature, and anneal equimolar amounts of the two template populations (the HE-expressing experimental and mock-transfected control) then digest with CEL I nuclease following the protocol included in the Surveyor™ Mutation Detection Kit manual.
8. Separate the digestion products on a 1 % agarose gel.
9. Determine the intensity of the bands corresponding to the digested heteroduplex and undigested homoduplex DNA molecules using ImageJ or other image analysis software.
10. Calculate the percentage of heteroduplex, mutant-containing DNA molecules by dividing the intensity of the digested band by the total of the digested and undigested bands.

3.7 Analysis of In Vivo Target Site Cleavage by Site Sequencing

Target site sequencing from cells expressing an HE (and, if desired, TREX2) can provide additional information beyond the above protocols on the frequency and molecular nature of target site mis-repair and mutagenesis events. Sequencing is potentially the most revealing of the target site analysis methods beyond Southern blot analysis and can be performed on small numbers of cloned target sites or by high throughput DNA sequencing (HTS) with bar coding if desired. The protocol below is designed for the analysis of small numbers (tens to dozens) of mutant sites. The use of HTS to analyze target sites is covered in Chapter 12 (*see* **Note 6**).

1. Prepare genomic DNA from experimental (±HE/±TREX2) and control cell cultures using a genomic DNA preparation kit.
2. Design PCR primers to amplify the HE target site that anneal ~250 bp upstream and downstream of the target site. Design a

second set of sequencing primers that anneal within the predicted PCR product and are located ~100 bp from the target site region.

3. Use the flank primer pair to PCR amplify target sites from cellular DNA samples with a high fidelity PCR polymerase that leaves 3′ A-tails.
4. Clean up the PCR products using a suitable purification protocol or kit.
5. Clone the PCR products into protocol vector suitable for TA cloning.
6. Transform the ligation products into an *E. coli* strain that allows for blue/white selection, e.g., DH5α, and plate on LB plates with ampicillin/IPTG/X-Gal.
7. Sequence 96 white colonies using the DNA sequencing primer(s) designed in **step 2** above (*see* **Note 7**).
8. Compile and compare the sequencing results from experimental and control samples with the genomic target site sequence defined in Subheading 3.3 above.

3.8 Worked Example: Identification of Potential Human Genomic "Safe Harbor" Sites Cleaved by the LAGLIDADG HE I-CreI/mCreI

HEs are being used in a growing number of organisms to target the disruption (or "knockouts") or modification of specific genes. Another less common though practically important genome engineering goal is to use HE cleavage of a genomic "safe harbor" site to facilitate transgene insertion without disrupting adjacent gene structure or expression. The inserted transgene may have therapeutic value or may provide a convenient and consistent way to "tag" the same site in different cells with a molecular bar code or other easily selected or scored marker gene such as a fluorescent protein coding cassette.

This section provides an example of how the protocols described above can be used to identify potential genomic cleavage sites for a HE based on sequential PSSM/PWM and BLAST searching, and then determine whether these sites could serve as new genomic "safe harbor" sites (SHS) for a range of genome engineering applications [24, 25]. The outline of this series of experiments is shown in Fig. 6.

1. Identify potential genomic SHS by LADHEDES PWM analysis: We used the protocol outlined in Subheading 3.1 to identify 128 I-CreI/mCreI target sites predicted by PWM data to be highly cleavage sensitive. The design criteria used with PWM data to generate this site list required that individual base pair differences, when combined in all possible target site combinations, did not reduce the predicted cleavage sensitivity of any site below 90 %.
2. BLAST search high-quality target site variants against the human genome: The list of 128 potential target sites from **step 1** was converted into FASTA format as described in Subheading 3.2

a

best potential sites from degeneracy data		BLAST search output		best potential SHS's
128 sites	→	29 sites / 37 locations	→	3 sites

b

position in hg19	site	criteria match
chr4:58,976,613 - 58,976,632	AAACTGTCATA**t**GACAGATT	8/9
chr2:48,830,185 - 48,830,204	AAACTG**a**CATAAGACAGATT	5/9

Fig. 6 Search for potential I-CreI "safe harbor" sites (SHSs) in the human genome. (**a**) The human genome was searched for high-quality I-CreI target site variants by the sequential use of LADHEDES I-CreI PWM data and BLAST. This search yielded 128 possible sites, of which 29 were identified in the human genome at 37 different locations. Only three of these sites, predicted to be highly cleavage sensitive by the LADHEDES I-CreI degeneracy PWM, met ≥8 of the 9 SHS criteria detailed in Table 1. (**b**) Two examples of human genomic I-CreI target sites that have high potential (*upper row*, 8 of 9 SHS criteria met) or low potential (*lower row*, 5 of 9 SHS criteria met) to serve as new human genomic SHS that could be specifically targeted with I-CreI or mCreI

then used to BLAST search the human genome sequence. A total of 29 of the 128 sites on our starting list were found at a total of 37 locations in the human genome.

3. Verify predicted target sites by amplification, sequencing, and cleavage analysis: The protocols in Subheadings 3.3 and 3.4 were next used to verify the sequence and predicted cleavage sensitivity of 6 of the 29 different target site sequences identified in **step 2** (results not shown).
4. Determine suitability to serve as a safe harbor site (SHS): There are no generally accepted criteria for SHS identification, so we assembled a list of nine different, stringent SHS scoring criteria in order to rank order the 29 different sites identified in **step 2** above. These criteria included uniqueness, accessibility, and likely safety as assessed by site proximity and activity measures. Table 1 summarizes these criteria and the most useful data sources including UCSC Genome Browser tracks to facilitate additional SHS assessments. Three of the 29 potential I-CreI/mCreI SHS from **step 2** met 8 of these 9 criteria and were judged to be of high value as potential new human genomic safe harbor sites (Fig. 6).
5. Next experimental steps: The next step to verify the utility of all 29 and the three highest scoring SHS candidates is to assess their cleavage sensitivity in vivo using the protocols outlined in Subheadings 3.6 and 3.7 in cells expressing mCreI ± TREX2 protein. Target sites that appear the most cleavage sensitive in vivo from these data will be used to design donor cassettes that include flank homology arms to facilitate homology-dependent, site-specific recombination, together with two initial transgene constructs that express either a drug-resistance marker or a fluorescent protein marker (*see* **Note 8**).

Table 1
Criteria for human genomic "safe harbor" sites (SHS)

	SHS criterion	Useful UCSC browser track	Refs.
Unique/ consistent accessible	Uniqueness (one copy in human genome)	None (BLAST search result)	–
	Not located in copy number variation (CNV)/segmental duplication region	*Variations and repeats/segmental dups*	[35, 36]
	Located in open chromatin	*Regulation/ENC DNase/FAIRE*	[37, 38]
Safety	Proximity to genes (>50 kb from the 5′ end of any gene)	*Genes and gene prediction tracks/ RefSeq genes*	[39]
	Proximity to miRNA/other functional small RNAs (>300 kb away from any miRNA)	*Genes and gene prediction tracks/ sno/miRNA*	[40–44]
	Proximity to cancer-related genes or mutations (>300 kb from any cancer-related gene)	*Phenotype and disease associations/ COSMIC*	[45, 46]
Functional silence	Low transcriptional activity	*mRNA and EST tracks/human mRNAs*	[47, 48]
	Located outside known replication origins (no origin within >50 kb)	*Regulation/UW Repli-seq/peaks*	[49, 50]
	Location outside ultraconserved elements (>50 kb from UCEs)	*Regulation/Vista Enhancers*	[51]

4 Notes

1. The LAHEDES server works well with most common Internet browsers, i.e., Internet Explorer, Mozilla Firefox, or Safari.
2. The stringency of the initial genomic sites search and correspondingly the number of potential target sites returned can be adjusted depending on the search aim. Our starting search in Subheading 3.8 focused on only those base substitutions that have near-native levels of activity (e.g., 90–95 % of the activity observed on the native target site base pair) in order to identify a small number of genomic target sites that had a high likelihood of being cleavage sensitive as DNA target sites and perhaps in chromatin as well.
3. The parameters for BLAST searches should again, at least initially, be kept restrictive to identify the most potentially useful genomic targets. Once these sites are defined, the expected threshold and/or the seed (word) length can be increased, and

penalties for mismatches and gaps reduced, to provide a more exhaustive site search. This type of secondary "relaxed" search can give a useful sense of potential genomic site numbers and their distribution (or "landscape") for a given HE and thus the potential for off-target or "collateral damage" by an HE with a defined specificity.

4. Product release may be a rate-limiting step for some HEs (e.g., I-CreI and derivatives). This requires the use of a stop buffer containing a denaturant such as SDS to unambiguously identify cleavage products.
5. A time course should be performed with sampling at least every 12 h over a 48 h interval to identify the time point with the highest fraction of cleaved molecules. Alternatively, addition of the ATM inhibitor KU-55933 [26] to 10 μM in the growth medium during HE expression interferes with DSB repair and increases the steady-state level of cleavage products and, by extension, mutant target sites.
6. Restriction fragments of ~5 kb run and transfer well in Southern blot analyses. Digest excess genomic DNA (10–20 μg) when possible to guarantee that enough sample is available to do equal lane loadings to detect cleavage products. The location of an asymmetric HE cleavage site allows both products to be visualized if the hybridization probe that is used covers both of the flanking DNA segments. An alternative, mentioned in Subheading 3.4, is to use a substrate band with the HE site placed in or near the center to double the intensity of the produce band. This placement is more difficult to achieve with restriction-generated as opposed to PCR-amplified substrates.
7. PCR suppression is an alternative technique to distinguish intact versus cleaved and native versus mutated target sites following homing endonuclease expression in vivo. While this approach can work, it is less sensitive than the methods described, may reveal only a minority of misrepair events, and is more prone to false positive and negative results [27–29].
8. There are many variations on this general protocol that can include, e.g., an enrichment step for mutant target sites if the goal of sequencing is to define a mutant repair spectrum [3, 30, 31].
9. Successful in vivo cleavage of the targeted SHS can be monitored by insertion of a selectable marker or a fluorescent protein into the generated DSB exploiting the cellular homology-directed repair (HDR) [25]. The reporter gene ORF can be preferentially inserted in the SHS by adding flanking homology arms of 400–800 bp adjacent to the intended homing endonuclease target site to facilitate homology-directed repair. Shorter homology arms of 50–100 bp length have been shown to work, but are less effective [32].

For standard laboratory and mammalian cell culture techniques, refer to Molecular Cloning—A Laboratory Manual [33], Current Protocols in Molecular Biology [34], and Protocols Online (http://www.protocol-online.org/).

References

1. Rouet P, Smih F, Jasin M (1994) Introduction of double-strand breaks into the genome of mouse cells by expression of a rare-cutting endonuclease. Mol Cell Biol 14:8096–106
2. Choulika A, Perrin A, Dujon B, Nicolas JF (1995) Induction of homologous recombination in mammalian chromosomes by using the I-SceI system of *Saccharomyces cerevisiae*. Mol Cell Biol 15:1968–73
3. Monnat RJ Jr, Hackmann AF, Cantrell MA (1999) Generation of highly site-specific DNA double-strand breaks in human cells by the homing endonucleases I-PpoI and I-CreI. Biochem Biophys Res Commun 255:88–93
4. Arnould S, Perez C, Cabaniols JP, Smith J, Gouble A, Grizot S, Epinat JC, Duclert A, Duchateau P, Paques F (2007) Engineered I-CreI derivatives cleaving sequences from the human XPC gene can induce highly efficient gene correction in mammalian cells. J Mol Biol 371:49–65
5. Zhao L, Pellenz S, Stoddard BL (2009) Activity and specificity of the bacterial PD-(D/E)XK homing endonuclease I-Ssp6803I. J Mol Biol 385:1498–1510
6. Li H, Pellenz S, Ulge U, Stoddard BL, Monnat RJ Jr (2009) Generation of single-chain LAGLIDADG homing endonucleases from native homodimeric precursor proteins. Nucleic Acids Res 37:1650–1662
7. 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. Nucl Acids Res 40:2587–2598
8. Taylor GK, Petrucci LH, Lambert AR, Baxter SK, Jarjour J, Stoddard BL (2012) LAHEDES: the LAGLIDADG homing endonuclease database and engineering server. Nucleic Acids Res 40:W110–W116
9. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402
10. 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061–1073
11. Shrivastav M, De Haro LP, Nickoloff JA (2008) Regulation of DNA double-strand break repair pathway choice. Cell Res 18: 134–147
12. Wang M, Wu W, Wu W, Rosidi B, Zhang L, Wang H, Iliakis G (2006) PARP-1 and Ku compete for repair of DNA double strand breaks by distinct NHEJ pathways. Nucleic Acids Res 34:6170–6182
13. Certo MT, Gwiazda KS, Kuhar R, Sather B, Curinga G, Mandt T, Brault M, Lambert AR, Baxter SK, Jacoby K et al (2012) Coupling endonucleases with DNA end-processing enzymes to drive gene disruption. Nat Methods 9:973–975
14. Schmitt MW, Kennedy SR, Salk JJ, Fox EJ, Hiatt JB, Loeb LA (2012) Detection of ultra-rare mutations by next-generation sequencing. Proc Natl Acad Sci U S A 109:14508–14513
15. Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol 132:365–386
16. Untergasser A, Nijveen H, Rao X, Bisseling T, Geurts R, Leunissen JA (2007) Primer3Plus, an enhanced web interface to Primer3. Nucleic Acids Res 35:W71–W74
17. Berkovich E, Monnat RJ Jr, Kastan MB (2007) Roles of ATM and NBS1 in chromatin structure modulation and DNA double-strand break repair. Nat Cell Biol 9:683–690
18. Berkovich E, Monnat RJ, Kastan MB (2008) Assessment of protein dynamics and DNA repair following generation of DNA double-strand breaks at defined genomic sites. Nat Protocols 3:915–922
19. Southern E (2006) Southern blotting. Nat Protoc 1:518–525
20. Donnelly ML, Gani D, Flint M, Monaghan S, Ryan MD (1997) The cleavage activities of aphthovirus and cardiovirus 2A proteins. J Gen Virol 78(Pt 1):13–21
21. Luke GA, de Felipe P, Lukashev A, Kallioinen SE, Bruno EA, Ryan MD (2008) Occurrence, function and evolutionary origins of "2A-like"

sequences in virus genomes. J Gen Virol 89: 1036–1042

22. Qiu P, Shandilya H, D'Alessio JM, O'Connor K, Durocher J, Gerard GF (2004) Mutation detection using Surveyor nuclease. Biotechniques 36:702–707
23. Oleykowski CA, Bronson Mullins CR, Godwin AK, Yeung AT (1998) Mutation detection using a novel plant endonuclease. Nucleic Acids Res 26:4597–4602
24. Papapetrou EP, Lee G, Malani N, Setty M, Riviere I, Tirunagari LMS, Kadota K, Roth SL, Giardina P, Viale A et al (2011) Genomic safe harbors permit high b-globin transgene expression in thalassemia induced pluripotent stem cells. Nat Biotech 29:73–78
25. Silva G, Poirot L, Galetto R, Smith J, Montoya G, Duchateau P, Paques F (2011) Meganucleases and other tools for targeted genome engineering: perspectives and challenges for gene therapy. Curr Gene Ther 11:11–27
26. Hickson I, Zhao Y, Richardson CJ, Green SJ, Martin NM, Orr AI, Reaper PM, Jackson SP, Curtin NJ, Smith GC (2004) Identification and characterization of a novel and specific inhibitor of the ataxia-telangiectasia mutated kinase ATM. Cancer Res 64:9152–9159
27. Seyama T, Ito T, Hayashi T, Mizuno T, Nakamura N, Akiyama M (1992) A novel blocker-PCR method for detection of rare mutant alleles in the presence of an excess amount of normal DNA. Nucleic Acids Res 20:2493–2496
28. Orum H, Nielsen PE, Egholm M, Berg RH, Buchardt O, Stanley C (1993) Single base pair mutation analysis by PNA directed PCR clamping. Nucleic Acids Res 21:5332–5336
29. Rand KN, Ho T, Qu W, Mitchell SM, White R, Clark SJ, Molloy PL (2005) Headloop suppression PCR and its application to selective amplification of methylated DNA sequences. Nucl Acids Res 33:e127
30. Argast GM, Stephens KM, Emond MJ, Monnat RJ Jr (1998) I-PpoI and I-CreI homing site sequence degeneracy determined by random mutagenesis and sequential in vitro enrichment. J Mol Biol 280:345–353
31. Scalley-Kim M, McConnell-Smith A, Stoddard BL (2007) Coevolution of a homing endonuclease and its host target sequence. J Mol Biol 372:1305–1319
32. Orlando SJ, Santiago Y, DeKelver RC, Freyvert Y, Boydston EA, Moehle EA, Choi VM, Gopalan SM, Lou JF, Li J et al (2010) Zinc-finger nuclease-driven targeted integration into mammalian genomes using donors with limited chromosomal homology. Nucleic Acids Res 38:e152
33. Sambrook J, Russell DW (2001) Molecular cloning – a laboratory manual. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY
34. Ausubel FM, Brent R, Kingston RE, Moore DD, Seidman JG, Smith JA, Struhl K (2013) Current protocols in molecular biology, John Wiley & Sons, Hoboken, NJ
35. Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE (2002) Recent segmental duplications in the human genome. Science 297:1003–1007
36. Bailey JA, Yavor AM, Massa HF, Trask BJ, Eichler EE (2001) Segmental duplications: organization and impact within the current human genome project assembly. Genome Res 11:1005–1017
37. Ho L, Crabtree GR (2010) Chromatin remodelling during development. Nature 463:474–484
38. Geiman TM, Robertson KD (2002) Chromatin remodeling, histone modifications, and DNA methylation—how does it all fit together? J Cell Biochem 87:117–125
39. Pruitt KD, Tatusova T, Maglott DR (2005) NCBI reference sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 33:D501–D504
40. Griffiths-Jones S (2004) The microRNA registry. Nucleic Acids Res 32:D109–D111
41. Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ (2006) miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res 34:D140–D144
42. Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ (2008) miRBase: tools for microRNA genomics. Nucleic Acids Res 36:D154–D158
43. Lestrade L, Weber MJ (2006) snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. Nucleic Acids Res 34:D158–D162
44. Weber MJ (2005) New human and mouse microRNA genes found by homology search. FEBS J 272:59–73
45. Forbes SA, Bhamra G, Bamford S, Dawson E, Kok C, Clements J, Menzies A, Teague JW, Futreal PA, Stratton MR (2008) The catalogue of somatic mutations in cancer (COSMIC). Curr Protoc Hum Genet Chapter 10, Unit
46. Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, Beare D, Jia M, Shepherd R, Leung K, Menzies A et al (2011) COSMIC: mining

complete cancer genomes in the catalogue of somatic mutations in cancer. Nucleic Acids Res 39:D945–D950

47. Benson DA, Karsch-Mizrachi I, Clark K, Lipman DJ, Ostell J, Sayers EW (2012) GenBank. Nucleic Acids Res 40:D48–D53
48. Kent WJ (2002) BLAT–the BLAST-like alignment tool. Genome Res 12:656–664
49. Hansen RS, Thomas S, Sandstrom R, Canfield TK, Thurman RE, Weaver M, Dorschner MO, Gartler SM, Stamatoyannopoulos JA (2010) Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. Proc Natl Acad Sci U S A 107: 139–144
50. Thurman RE, Day N, Noble WS, Stamatoyannopoulos JA (2007) Identification of higher-order functional domains in the human ENCODE regions. Genome Res 17:917–927
51. Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, Shoukry M, Minovitsky S, Dubchak I, Holt A, Lewis KD et al (2006) In vivo enhancer analysis of human conserved non-coding sequences. Nature 444:499–502