

# Chapter 2

## Bioinformatic Identification of Homing Endonucleases and Their Target Sites

Eyal Privman

### Abstract

Homing endonuclease genes (HEGs) are a large, phylogenetically diverse superfamily of enzymes with high specificity for especially long target sites. The public genomic sequence databases contain thousands of HEGs. This is a large and diverse arsenal of potential genome editing tools. To make use of this natural resource, one needs to identify candidate HEGs. Due to their special relationship with a host gene, it is also possible to predict their cognate target sequences. Here I describe the HomeBase algorithm that was developed to this end. A detailed description of the computational pipeline is provided with emphasis on technical and methodological caveats of the approach.

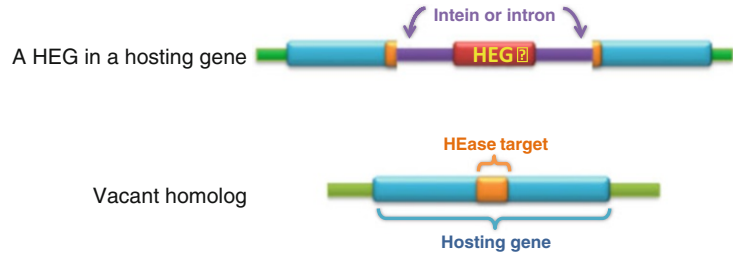
**Key words** Homing endonucleases, Homology search, Target-site prediction, HomeBase

---

### 1 Introduction

In this chapter, I describe a computational approach that identifies novel Homing endonuclease genes (HEGs) in nucleotide sequence databases, infers their native target sequence, and then generalizes this single target to a predicted range of possible targets. This approach was first used to construct the HomeBase collection [1], which is searchable using the HomeBase web server at <http://homebase-search.tau.ac.il/>. The first part of Subsection 3 describes the usage of the web server to search for HEGs in the existing HomeBase collection that are predicted to have targets within a given query DNA sequence. The second part describes running the HomeBase pipeline on a local computer in order to search for novel HEGs in a nucleotide sequence database.

The HomeBase pipeline involves first the inference of the native target sequence. This stage relies on the observation that the two halves of the target sequence are found in the exons/exteins flanking the intron/intein in which the HEG resides (Fig. 1). In the second stage HomeBase infers the range of nonnative targets that may be cleaved by the nuclease. This stage relies on the



**Fig. 1** A HEG residing in an intron/intein of a hosting gene compared to a homolog of the hosting gene that does not contain the intron/intein

observation that the long target sequences allow considerable plasticity: some nucleotides can be substituted while cleavage efficiency is retained [2]. For HEGs that reside in introns or inteins of a protein-coding host gene, the target sequence is typically coding for a conserved amino acid translation, yet synonymous (silent) mutations are still frequent. Therefore, HEGs evolved tolerance to variation in these synonymous sites [3, 4]. This observation is useful for the prediction of the range of targets that a nuclease can recognize beyond the native target sequence found in its host. In HomeBase, the range is defined by the translated amino acid sequence of the target site.

## 2 Materials

The computational protocol described below requires the following:

1. A query set consisting of the protein sequences of known HEGs. For example, HEG-containing intron sequences may be obtained from the Group I Intron Sequence and Structure Database [5] (<http://www.rna.whu.edu.cn/gissd>) and HEG-containing inteins from the intein database INBASE [6] (<http://www.neb.com/neb/inteins.html>). The manual curation of these databases ensures that these protein sequences do not include exonic or exteinic parts. This is essential for our purpose because otherwise the BLAST search will retrieve many homologs of the hosting gene instead of homologs of the HEGs. The query sequences will include only the HEG sequence for the case of introns. For the case of inteins we will use the full intein sequence, which includes both the nuclease domain and the protein-splicing domains.
2. The nucleotide database that will be searched for HEGs needs to be available in plain FASTA format, and also formatted as a BLAST-searchable database.
3. A local installation of a BLAST implementation such as the NCBI BLAST package (<ftp://ftp.ncbi.nlm.nih.gov/blast/>)

[executables/blast+/LATEST/](#)). Special-purpose software to parse and analyze the BLAST results, such as the Perl scripts described below, which were used in the original production of the HomeBase database. These scripts are available for download from <http://homebase-search.tau.ac.il/>. A BLAST parser implementation such as the BioPerl Bio::SearchIO module is also recommended [7] (<http://www.bioperl.org/>).

- Depending on the size of the database to be searched, considerable computing resources may be necessary. As an example, the original HomeBase collection was constructed in 2009 by searching a total of 36 Gbp genomic and metagenomic sequences, and the run-time of the pipeline was roughly 5 weeks on a 40-core cluster. A key factor is the number of queries in the BLAST-2 stage (see below). In that case we had about 10,000 queries, which were eventually narrowed down to less than 1,000 in the final output.

### 3 Methods

#### 3.1 Using the HomeBase Web Server to Search the Existing HomeBase Collection of Putative HEGs

The results of the original HomeBase pipeline that was run in 2009 are online and searchable using the HomeBase web server at the address: <http://homebase-search.tau.ac.il/>. The web server allows searching for HEGs in the existing HomeBase collection that are predicted to cut within a query DNA sequence provided by the user.

- Enter the query sequence by either copy-paste into the text box or by choosing to upload a file. You may enter the sequence either as a plain DNA sequence or as a FASTA formatted file.
- Click “submit” and wait for the search to complete, which normally takes up to a few minutes for an input sequence of a few thousands of bases. A results page will be shown with a table in the following format:

HEN name	Cleavage site position	Blast pairwise alignment	Match score
ref XM_001554561.1 _TRUNC_1_4210_INS_774_3208	24058	EKASGFEESM EKASGFEESM EKASGFEESM	9/10
gb AY836254.1 _TRUNC_1_1788_INS_177_1744	24058	EKASGFEESM EK+SGFEESM EKSSGFEESM	8/9
gb AACY023780064.1 _TRUNC_1_1533_INS_431_1389	20035	AGLPAQPYSMS AG P QP+S+S AGFPCQPFSSIS	6/14
gb AY863213.1 _TRUNC_14408_21647_INS_742_1726	12808	LHQGKTNNPLTV LHQ +NNPL + LHQNGSNNPLGI	7/12

*HEN name:* The ID for the nuclease in the HomeBase collection, which includes the NCBI Nucleotide ID in which this gene was identified by the HomeBase pipeline. Clicking the link gives the coding sequence for this gene, as defined by the HomeBase pipeline.

*Cleavage site position:* The position of the predicted target site in the query sequence.

*Blast pairwise alignment:* An alignment of translated sequences of the native target sequence (from the HomeBase database) and the predicted target sequence (in the query sequence).

*Match score:* The number of identical amino acid residues out of the total alignment length.

### 3.2 Searching for Novel HEGs in a Sequence Database Using a Local Implementation of the HomeBase Pipeline

Two consecutive rounds of BLAST searches will be run: The first identifies candidate novel HEGs (BLAST-1). These sequences are used as queries in the second search (BLAST-2) to identify *vacant homologs*: for a HEG-hosting gene that contains a HEG inside one of its introns/inteins, a vacant homolog is a homolog of the hosting gene that does not contain the intron/intein (Fig. 1). Subsequently, the alignments to the vacant homologs are used to infer the target sequences.

#### 3.2.1 Search for HEGs (BLAST-1)

1. The set of known HEGs is used as queries in a translated BLAST (tblastn) search against the nucleotide database to find novel HEGs. Using the NCBI BLAST package:
  - (a) Create a BLAST formatted database from a FASTA formatted file of the DNA sequences that will be searched for novel HEGs:
 

```
makeblastdb -in database.fasta -dbtype nucl
```
  - (b) Run translated BLAST (tblastn) of the known HEG protein sequences against the database:
 

```
blastn -query known_hegs.fa -db database.fasta -out blast1.out
```
  - (c) Retain all hits of  $E\text{-value} < 10$ , which is the default in NCBI BLAST (*see Note 1*).
2. For long hit sequences it is necessary to divide the hit sequence to disjoint loci by clustering the HSPs from all queries on the same hit sequence (*see Note 2*). Overlapping HSPs are clustered, as well as neighboring HSPs less than 2,000 bases apart. Additional 1,000 bp flanking sequences are included on either side. An implementation of this procedure is available in the script getBlastSeqs.pl (*see lines 81–96*).

#### 3.2.2 Defining Target Sites Based on Vacant Homologs (BLAST-2)

1. Each hit sequence from the BLAST-1 results (as defined by the above clustering procedure) is used as a query in the second BLAST search (BLAST-2) to identify vacant homologs (*see Note 3*). A translated BLAST (tblastx) search is performed

to allow any of the three possible translations of the strand that was aligned to the known HEG in BLAST-1. The BLAST database should be the same as for BLAST-1:

```
tblastx -query blast1_hits.fa -db database.
fasta -out blast2.out -strand plus
```

2. Filter the BLAST-2 hits using two Perl scripts: findVacantAllele.pl that parses the BLAST-2 output and prints summary information for candidate hits; and findTargetSite.pl that reads this information, applies the more complicated filters and predicts the target sequence of the putative HEG:
  - (a) The alignment to the hit sequence must fit the expected homology between a full and a vacant homolog, as shown in Fig. 1: the homology is in the two exonic sequences, which flank the intron/intein that contains the HEG in the query sequence, and are adjacent in the hit sequence (the vacant homolog). Each of these two homology regions will result in a separate HSP. The two HSPs should be approximately adjacent in the hit sequence and separated by a large insertion in the query sequence. We also require that they be on the same strand and the same reading frame in the hit (findVacantHomolog.pl lines 73–114).
  - (b) The insertion is classified as an intein if both BLAST-2 HSPs are in the same reading frame. Otherwise, the insertion is classified as an intron.
  - (c) Similarly, the BLAST-1 HSP inside the insertion (which represents region coding for the homing endonuclease) must be in the same frame of the BLAST-2 HSPs to be classified as an intein (*see Note 4*). Where the insertion contains several BLAST-1 HSPs in different frames a majority vote is taken (findTargetSite.pl lines 464–501).
  - (d) If all three reading frames are consistent, we translate the insertion sequence in this frame and check for stop codons, which are not expected in an intein. If stop codons are found the insertion is classified as an intron (findTargetSite.pl lines 503–516).
  - (e) Discard insertions classified as introns if they are <450 bases long, and insertions classified as inteins if they are <930 bases long (*see Note 5*) (findVacantHomolog.pl lines 103–114, findTargetSite.pl lines 913–920).
  - (f) Check that some BLAST-1 HSP (an alignment to a known HEG) lies inside the insertion in the query sequence, which represents the intron/intein. Do not allow the BLAST-1 HSP(s) to overlap with more than 15 bases of either of the BLAST-2 HSPs (*see Note 6*) (findVacantHomolog.pl lines 116–148, findTargetSite.pl lines 899–911).

- (g) Correct overlap or gap between the two HSPs (*see Note 7*): To correct an overlap some bases must be removed from either or both HSPs (findTargetSite.pl lines 395–430). To correct for a gap some bases must be added (findTargetSite.pl lines 647–887). To decide on the correct number of bases that should be removed/added to each HSP, iterate over all possibilities for the position of the insertion of the intron/intein in the hit sequence, and choose the position that gives the highest similarity between the query and the hit in the extended/shortened HSPs. The rationale behind this is that true homology is expected to produce higher similarity than chance similarity. Formally, let  $L$  be the overlap length and  $S$  be the splice site position relative to the 5' start of the overlap region, and so  $0 \leq S \leq L$ . We seek  $S$  that maximizes the sum of similarity in the two HSPs after resolving the overlap region. Similarity is measured between the translated amino acid sequences using the same BLOSUM62 matrix that is used by translated BLAST. For introns, the insertion site should be allowed to reside in the middle of a codon. That is, one nucleotide is added/removed to one HSP and two nucleotides to the other HSP. This should not be allowed for sequences identified as inteins.
- (h) At least three out of the five amino acid positions at the end of each HSP after overlap/gap correction (the end of the exon) are required to be identical or similar between the query and hit sequences. Similarity is defined as in BLAST: a positive score in the BLOSUM62 matrix. This criterion ensures that the homology is sufficient to resolve the splice sites reliably (findTargetSite.pl lines 432–462).
- (i) For introns, search the insertion for the largest open reading frame (ORF) that overlaps one or more BLAST-1 hit in the same frame. This ORF is the putative coding sequence of the HEG. We require it to be at least 85 codons long, to accommodate the shortest known HEGs (findTargetSite.pl lines 518–645).
- (j) Filter noncoding host genes (*see Note 8*): Translate 50 codons (where available) on either side of the intron in the reading frame of the BLAST-2 HSPs. If a stop codon was found the host gene is classified as a noncoding RNA. For classification of partial sequences as a protein-coding host, we require that 50 codons will be available for at least one of the exons (findTargetSite.pl lines 889–897).
- (k) The final output of this procedure is the target sequence flanking the inferred splice sites in the query (the HEG-containing sequence), and the coding sequence of the HEG (in introns) or the HEG-containing intein. As the predicted

target sequence we report the first seven codons after the splice sites of each exon, giving a total of 14 codons or 42 bases (*see* **Note 9**) (*findTargetSite.pl* lines 968–990).

- (1) For many BLAST-2 hit sequences, several overlapping pairs of HSPs satisfy the above requirements. Therefore, it is necessary to choose the splice site positions that have the highest support, from more HSP pairs (*findTargetSite.pl* lines 1017–1051).
3. One BLAST-2 query sequence may contain several HEGs. After identifying the first HEG-containing intron/intein in a given BLAST-2 query, repeat the above procedure using only BLAST-1 HSPs that do not overlap this intron/intein. Only BLAST-2 HSP pairs that contain these BLAST-1 HSPs are considered for additional introns/inteins. We repeat this process until no BLAST-1 HSPs remain for the given BLAST-2 query (Loop starting in *findTargetSite.pl* line 155).

### 3.2.3 Quality Assurance and Control

The above protocol inevitably outputs some proportion of false-positive results. Therefore, it is necessary to estimate this proportion and adjust the protocol if it is higher than acceptable. False positives can be classified into two classes: (1) sequences that do not contain a HEG and (2) sequences that contain a HEG but whose predicted target sequence is wrong. To assess the accuracy of prediction putative HEGs can be sampled randomly and manually inspected for all stages of the automatic pipeline, including: confidence in the BLAST-1 homology and conservation of known HEG motifs; confidence and accuracy of the BLAST-2 alignment to the vacant homology, especially of the exon/extein boundaries after correction of gaps/overlaps. Where possible the prediction can be compared to existing annotation of intron position in the host genes. One can also check for the intron/intein splice sites consensus sequences (e.g., inteins sequences typically start with a C and end with HN). If too many results appear to be false positives, then some of the criteria and thresholds in the relevant stages of the protocol may be adjusted.

### 3.2.4 Searching for Possible Target Sites in a Genome of Interest

Predicted HEGs and their target sequences are potentially useful to target specific sites in a genome of interest, as a tool for genetic manipulation. Potential targets can be identified using a translated BLAST (*tblastn*) search to find a match between the translations of the predicted target sequences and all possible six-frame translations of the genome. This approach relies on the observation that target specificity is defined mainly by the non-synonymous sites of the target sequence [1]. Since the actual target sequence is a subsequence of the 14 codons of the predicted target, then candidate hits should be preferred to have no mismatches in the central region of the target sequence.

---

## 4 Notes

1. Each BLAST-1 hit sequence often has several high scoring pairs (HSPs): pairwise alignments of a segment of the query to a segment of the hit. Furthermore, several queries often match the same hit and yield overlapping HSPs. Especially for hits in whole chromosome sequences, which may contain several different HEGs, the number of HSPs can be large.
2. These subsequences of the hit sequences are extracted to serve as the query sequences for the second BLAST search (BLAST-2). This procedure may result in a cluster of several HEG-containing intron/inteins in the same host gene.
3. In this search we are looking for a homolog of the hosting gene in which the HEG is embedded, so we will be interested in alignments involving the sequences flanking the BLAST-1 hit. These coding sequences may be in a different reading frame than the HEG, but have to be on the same strand.
4. By chance alone we expect one out of nine introns to satisfy the requirement for the three reading frames to be consistent (the two exons and the HEG). These introns will be mistaken for inteins according to this criterion.
5. We require that insertions classified as introns are at least 450 bases long, because this is the length of the shortest known HEG-containing introns. For inteins we require 930 bases, which is the length of the shortest HEG-containing inteins. This allows filtering out hits to homologs that contain mini-inteins (inteins lacking a HEG). Such homologs appear to have a large deletion compared to the homolog with the full (HEG-containing) intein. However, these deletions would be shorter than a deletion of the whole intein. Thus, they will not pass the 930 bases cutoff.
6. A large overlap is not expected because BLAST-1 queries do not include any exonic sequences, and so the BLAST-2 HSPs (which are supposed to be the exons) should not bear homology to the known HEG. This requirement is necessary to filter out hits to repetitive sequences as well as homologs containing a mini-intein that result in pairs of HSPs with a similar structure resembling a vacant homolog.
7. Ideally, the two HSPs (the exon sequences) should be exactly adjacent in the hit sequence (the vacant homolog that is missing the insertion of the intron/intein). However, in practice, they often overlap because BLAST extends the alignment from the true homology of the exons, to chance similarities in the intron/intein. In other cases, the HSPs may have a small gap between them (in the hit sequence) if the homology is distant enough so that some substitutions have obscured

the similarity in the positions adjacent to the splice sites. To correct an overlap some bases must be removed from either or both HSPs. To correct for a gap some bases must be added.

8. Many HEG-containing introns reside in noncoding RNA genes, especially rRNA. To identify protein-coding host genes we search for stop codons in the exons surrounding the intron. For noncoding sequences of 150 bases, we expect <10 % probability of not containing stop codons, and thus being erroneously classified as coding. Note that this step is required for the use of the translated target sequence as the range of putative targets for the novel HEG. However, other applications that are not based on this approach may not require this filter. For example, if the goal is only to identify novel HEGs (prediction of the target sequence is not required) then this step can be removed.
9. This is enough to encompass the target sequence of any known HEG, but for many the actual target would be a shorter subsequence.

## References

1. Barzel A, Privman E, Peeri M, Naor A, Shachar E, Burstein D, Lazary R, Gophna U, Pupko T, Kupiec M (2011) Native homing endonucleases can target conserved genes in humans and in animal models. *Nucleic Acids Res* 39: 6646–6659
2. Gimble FS, Wang J (1996) Substrate recognition and induced DNA distortion by the PI-SceI endonuclease, an enzyme generated by protein splicing. *J Mol Biol* 263:163–180
3. Kurokawa S, Bessho Y, Higashijima K, Shirouzu M, Yokoyama S, Watanabe KI, Ohama T (2005) Adaptation of intronic homing endonuclease for successful horizontal transmission. *FEBS J* 272:2487–2496
4. 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
5. Zhou Y, Lu C, Wu QJ, Wang Y, Sun ZT, Deng JC, Zhang Y (2008) GISSD: group I intron sequence and structure database. *Nucleic Acids Res* 36:D31–D37
6. Perler FB (2002) InBase: the Intein Database. *Nucleic Acids Res* 30:383–384
7. Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JG, Korf I, Lapp H et al (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res* 12:1611–1618

Homing Endonucleases

Methods and Protocols

Edgell, D. (Ed.)

2014, X, 284 p. 60 illus., 39 illus. in color., Hardcover

ISBN: 978-1-62703-967-3

A product of Humana Press