**HEfinder: Locating Homing Endonucleases within Genome Sequences**

**Background**

Homing endonucleases (HEs) are selfish genetic elements found in every domain of life, often within type I introns or inteins [1]. The protein product of the homing endonuclease gene (HEG) makes a double stranded cut at the HEG- allele on the homologous chromosome. DNA repair copies over the HEG, and surrounding intein or intron, into the previously vacant allele. In this way, HEs confer horizontal mobility to the inteins or introns that house them [1].

HEs also show promise as biotechnology agents, since they make a site specific double stranded cut, and are amenable to manipulations of their target site [2-5]. Reprogrammable homing endonucleases may eventually act as therapeutic agents, alongside CRISPR based approaches [3].

Given their importance, there is significant value in being able to identify possible active HEs within genome sequences. Under this motivation, I wrote a tunable program, HEfinder, that takes a genome as input, and outputs those loci that match homing endonucleases to a user-supplied specificity. I evaluated the script to ensure it worked as intended. I then applied it to test the canonical view that active homing endonucleases are not found within the genome sequences of multicellular Eukaryotes [4,5]. This view was recently challenged by the discovery of an HE in *Arabidopsis thaliana* [6]. I found that the script worked effectively, and that there are several examples of close homing endonuclease homologs in eukaryotic genomes, further challenging this view.

**Methods**

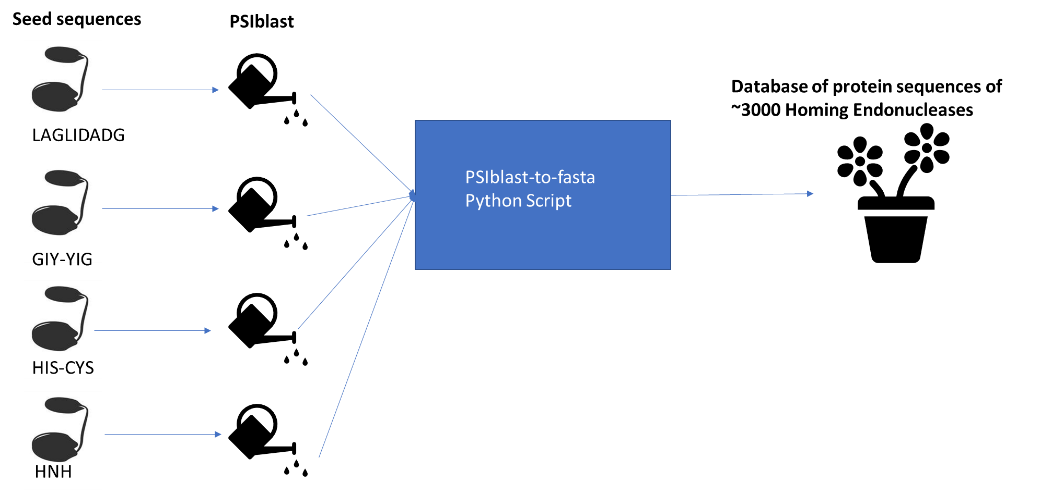
To create HEfinder, I first assembled a representative database of HEs (fig 1). The major HE families (LAGLIDADG, GIY-YIG, HIS-CIS, and HNH) share little sequence similarity [5], so I started with representative seed sequences (with experimentally verified activity, where possible) from each family. I ran a 5 iteration Psi-Blast against NCBI’s NR database, with an E-value inclusion cutoff of 1e-5. I wrote two short Python scripts called select\_HEs, and Psiblast\_to\_fasta to parse the Psi-Blast output. Given the last round of a Psi-Blast, select\_HEs culls all hits not annotated as homing endonucleases. Psiblast\_to\_fasta takes the select\_HEs output, and outputs a multiple sequence file of the protein sequences of all unique homing endonucleases in the input file. I then wrote the HEfinder python script. The script can be called from the command line and takes the parameters listed in table 1.

Figure : graphical overview of HE database creation.

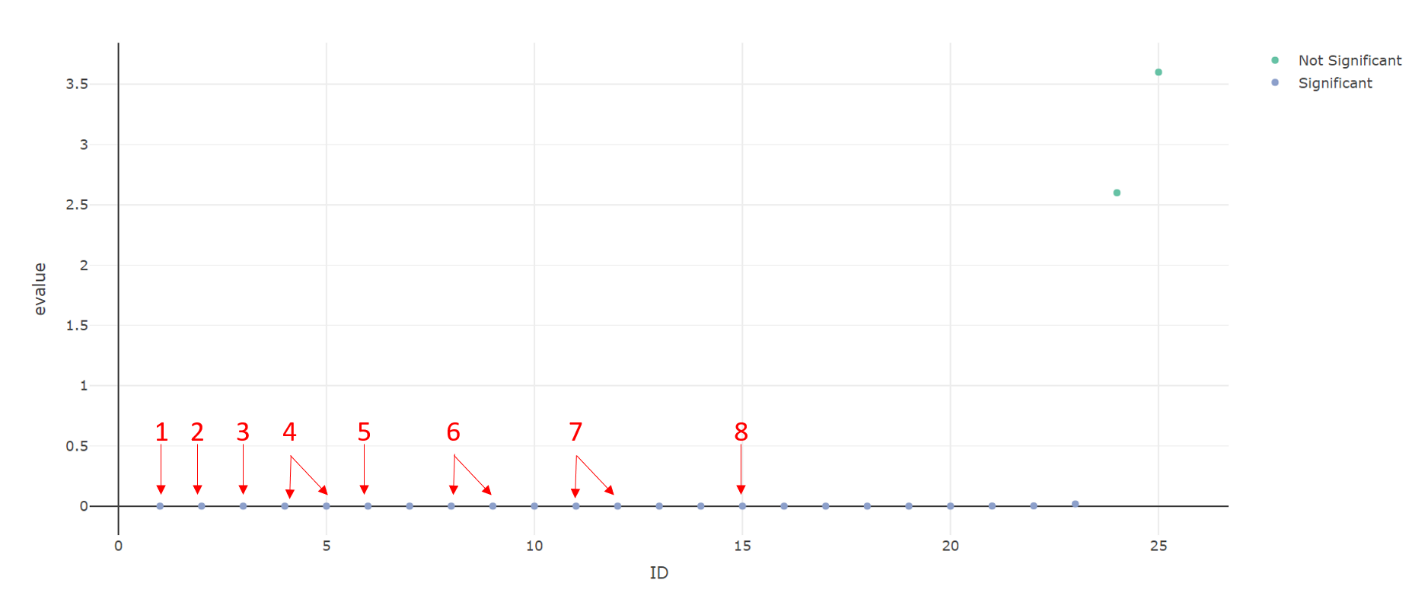
|  |  |  |
| --- | --- | --- |
| **Flag** | **Description** | **Default** |
| **Required Parameter** | | |
| Not applicable | Genome | Not applicable |
| **Optional Parameters** | | |
| -use\_bitscore | use bit score rather than e value for our significance cutoff? | False |
| -name | Name of output file prefix | Blast |
| -c | Significance cutoff | If E-value is used: 1  If bit score: 50 |
| -o | Should we write out significant results to a multiple sequence file | True |

Table 1: parameters taken by HEfinder

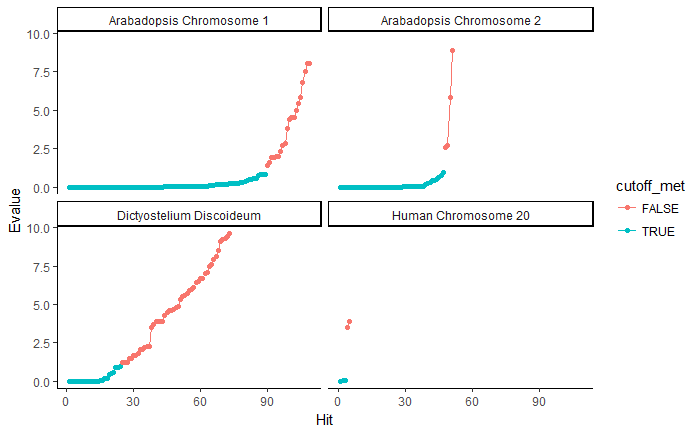
The script first calls a bash script (hefinder.sh) to run a blastx of the genome input query against the HE database, with a culling limit imposed to prevent many HEs mapping to the same loci. It outputs the blast results to a table called name\_full (name supplied by user). Hefinder.sh then calls an Rscript called Blastparse.R that parses and visualizes the data, and outputs 4 graphs and a table of those hits that meet the significance cutoff called name\_filtered. Finally, unless the -o flag is used, HEfinder uses Emboss extractseq to write out significant sections of the genome to a file, called name\_significant\_hit\_sequences.seq.

**Results**

To test HEfinder, I ran the script on defaults against the chloroplast sequence of *Haematococcus Lacustris*. I chose this alga as a target because it has eight unique annotated HEs. The results (fig 2) show that while the script was able to identify the 8 annotated HEs, there were several false positives (significant matches that did not match one of the 8).

*Figure 2: results of testing HEfinder on H. lacustris chloroplast sequence. Annotated homing endonuclease hits marked with red arrows*

However, when I took a multiple sequence alignment of the proteins sequences of the false positives and blasted them against the NR database, each match a homing endonuclease within other algae with an E value of less than 1\*10^-17. These are almost certainly not false positives, but actual homing endonucleases within other proteins that are not explicitly annotated as such, bringing the total unique homing endonuclease count, within this 1.35 MB sequence, to 15.

Given the high saturation of homing endonucleases in every test sequence I had tried, I decided to test the theory that homing endonucleases were not found within the genome sequences of multicellular. I tested the script on *Dictyostelium Discoideum,* Arabadopsis chromosome 1 and 2, and human chromosome 20 (fig 3)

*Figure 3: Testing HEfinder on a variety of multicellular eukaryotic genomes. Hits colored in blue indicate sections of genome that match a homing endonuclease in the HEdatabase with an E-value less than 1. See Eukaryotic\_genomes\_results\_and\_figure\_3\_code for raw data and code for graphs*

**Discussion**

The HEfinder *H. Lacustris* demonstrate that HEfinder works as intended, and that algae chloroplasts are replete with homing endonucleases, many of which are unannotated.

HEfinder found at least a few very significant hits within the genomes of every multicellular eukaryote I tried. It is important to note that HEfinder cannot prove that these HEs are active. It can, however, provide loci of putative homing endonucleases, which can then be experimentally evaluated for activity. However given the volume of HEs observed and their close homology to active HEs, these results certainly cast doubt on the conclusion that HEs are not found within eukaryotic genomes, and call for further inquiry.

One strength of the current script is its ability to identify all homing endonucleases within a genome, including those in pseudogenes. However, this strength also comes with a performance tradeoff- blasting the six frame translation is computationally intensive for large queries. On chromosome 20, the program was run for several hours. If I were to continue attempting to demonstrate that HEs are present in eukaryotes, I would likely modify the script to instead take in a proteome and do a blastp against the HE database. The closest HE homologs are likely to be in protein coding sequences anyway, and the speed of the script would be improved.

**References**

1. Gogarten, J. P., & Hilario, E. (2006). Inteins, introns, and homing endonucleases: Recent revelations about the life cycle of parasitic genetic elements. *BMC Evolutionary Biology*. https://doi.org/10.1186/1471-2148-6-94
2. Antunes, M. S., Smith, J. J., Jantz, D., & Medford, J. I. (2012). Targeted DNA excision in Arabidopsis by a re-engineered homing endonuclease. *BMC Biotechnology*, *12*. <https://doi.org/10.1186/1472-6750-12-86>
3. Barzel, A., Privman, E., Peeri, M., Naor, A., Shachar, E., Burstein, D., … Kupiec, M. (2011). Native homing endonucleases can target conserved genes in humans and in animal models. *Nucleic Acids Research*, *39*(15), 6646–6659. <https://doi.org/10.1093/nar/gkr242>
4. Belfort, M., & Bonocora, R. P. (2014). Homing endonucleases: From genetic anomalies to programmable genomic clippers. *Methods in Molecular Biology*, *1123*, 1–26. <https://doi.org/10.1007/978-1-62703-968-0_1>
5. Stoddard, B. L. (2011). Homing endonucleases: From microbial genetic invaders to reagents for targeted DNA modification. *Structure*. <https://doi.org/10.1016/j.str.2010.12.003>
6. Cho, S. K., Ryu, M. Y., Poulsen, C., Kim, J. H., Oh, T. R., Choi, S. W., … Yang, S. W. (2017). HIGLE is a bifunctional homing endonuclease that directly interacts with HYL1 and SERRATE in Arabidopsis thaliana. *FEBS Letters*. <https://doi.org/10.1002/1873-3468.12628>