

CisOrtho



Users' Guide



The purpose of this document is to instruct the reader on the use of the CisOrtho web based utility. This document is divided into three sections:

- Methods (Pages 2-4, similar to online version)
- Running CisOrtho (Pages 5-6, skip to here to jump right in)
- CisOrtho Output (Pages 7-10)

Methods

We implement a simple 5-step pipeline for the detection of transcription factor binding sites conserved between *C. elegans* and *C. briggsae*. Here are the details of the five steps:

1. Build a Position Weight Matrix (PWM). The user submits a binding site alignment file (below, right). This file is input to the **hmmbuild** command from the [HMMER](#) software suite, to build the position weight matrix. Only the emission parameters in the file are used in the search. These are the four numbers in lines starting with 1,2,3 etc. The sixth column in these lines is the position in the multiple alignment, which these parameters represent. Given the following weight matrix (left) and alignment used for its generation, we can see that there are 19 columns in the alignment but only 17 columns (displayed

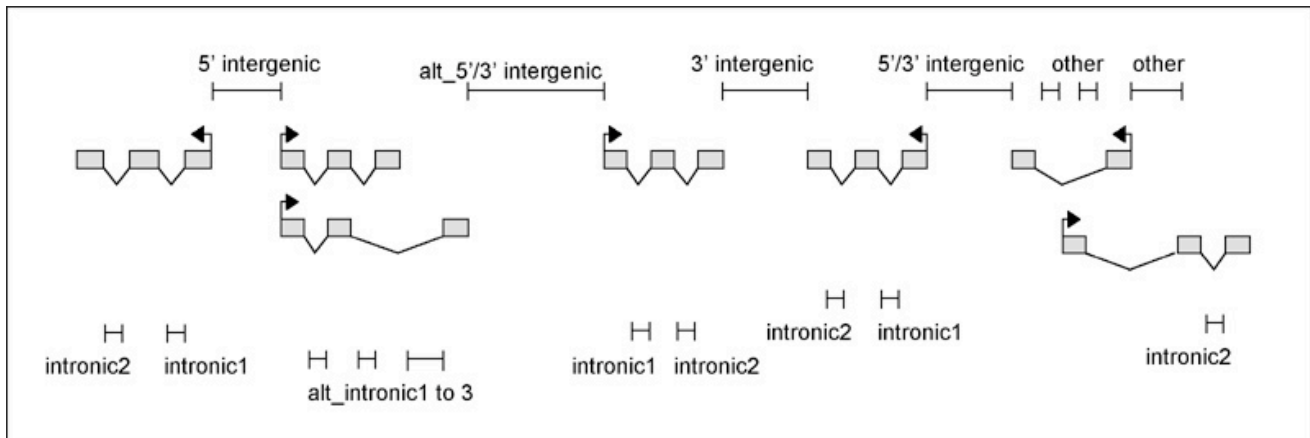
PWM_column	A	C	G	T	alignment_column
1	-231	-514	1472	-1392	1
2	568	158	834	-4321	2
3	-4321	273	2016	-4321	3
4	-1106	2016	-304	-4321	4
5	1357	-4323	-370	-4321	5
6	-93	560	1305	-4321	6
7	-4321	1915	690	-2407	7
8	-300	1801	-1414	-2189	9
9	-1229	2152	-1046	-4321	10
10	-4321	2486	-4323	-4321	11
11	1555	-4323	-4323	-4321	12
12	-126	-4323	-551	802	13
13	-1859	-4323	2181	-1399	14
14	-4321	2486	-4323	-4321	15
15	1555	-4323	-4323	-4321	16
16	-1612	1168	-4323	602	18
17	-2636	1020	83	348	19

```
GAGCAAGTCCCATGCAATT
TAGCAGGTCCCATGCACTT
AAGCAAGTCCCATGCAATG
GGGCAAGTCCCATGCAATA
AAGCAGTCCCATGCAGAC
AAGCAGTCCCATGCAGAC
AAGCACCTCCCATGCAGAC
GAGAAGCGCCCAAGCATTG
GCGCAACCACCAAGTCATTT
GCGCAACCACCAAGTCATTT
GGGCAGCGAACAGCATCT
CAGCAGCGCCCATGCAACC
GCGCAACCACCAAGTCAGCT
GGTGAGCCCAAGCATCT
ACGCAACCCATACACTC
AGCAGCGCCCATGCAATG
CAGCAGGATACATGCAACC
TGCGGCCGCCCATGCACTC
GACCACCAAGCAAGCACT
```

inverted as rows) in the position weight matrix. Columns **8** and **17** HMMER judged too close to background frequency to extract a log-odds scoring column. If this position weight matrix were used in an actual search, the nucleotides in positions **8** and **17** would be ignored and merely used as placeholders.

There can be poorly conserved columns anywhere in the alignment, but for a successful search, the alignment must contain at least 6 well conserved columns. **hmmbuild** then uses a tree-based sequence weighting scheme to calculate the weighted counts of nucleotides at each position in the multiple alignment. Background frequency priors are added to each count at each position. These are the frequencies of nucleotides in the non-exonic regions of *C. elegans* and *C. briggsae* genomes to be searched. (Using current GFF files to determine these regions, the frequencies are **34.23%** G/C and **65.77%** A/T content.) Then the ratio with the background nucleotide frequency of the sequence being searched is taken, and the \log_2 is taken of this ratio, giving the 'bits'. The resulting PWM is the $4 \times n$ matrix in which a cell i,j is the bits score of nucleotide i at sequence position j . If the bits are too low in flanking positions (resulting from a position in the alignment having close to background frequency), **hmmbuild** automatically excludes these positions from the PWM. Because of this, if the user is unsure where the region of conservation ends, it is best to err on the side of including unconserved flanking regions, as this won't affect the results. For details, see the HMMER [UserGuide](#).

2. Isolate and Classify non-exonic sequence Using [GFF](#) (General Feature Format) exon annotation files, the program **snip.plx** isolates all non-exonic sequence from either *C. elegans* or *C. briggsae* genomic sequence.



Sequences are classified as:

3' intergenic between the 3' ends of two genes on opposite strands

5' intergenic between the 5' ends of two genes on opposite strands

3'/5' intergenic between two genes on the same strand

intronic# between exons of the same gene (though not necessarily the same splice variant)

Note: if the gene has only one splice form, these fragments correspond to actual numbered introns

BEGIN or **END** at the beginning or ending of a chromosome (*C. elegans*) or sequence read (*C. briggsae*)

other all segments not in any above category (such as those bounded by exons from two different genes).

In addition, any segment will be prefixed by **alt_** if either of its bounding exons belongs to a gene with more than one splice variant. Note that no attempt is made to correctly number introns for each splice variant.

3. Search non-exonic regions using the PWM The program SimpleSearch is used as a scanning window to find up to N (user-defined) top-scoring hits in each genome. Technically, the program chooses the score cutoff to produce as many top-scoring hits as possible less than N. It estimates this cutoff by sampling 1% of the windows and sorting the resulting list of scores. Therefore, the number of hits retrieved will be slightly less than N in most cases. If the PWM has very little information (low bits in most positions) then there will be many windows achieving the same score, and the number of hits retrieved could be substantially less than N. This can arise if the binding sites are incorrectly aligned or poorly conserved.

4. Filter out 'orphan' hits We then use a 1-1 mapping of *C. elegans* to *C. briggsae* orthologs provided in the file [orthologs-2.00](#) to find only those hits in one species for which there is a hit in the matching ortholog of the other species. The filter traverses all ortholog pairs, and retrieves the top-scoring hit for each ortholog, called a 'hit-pair'. The percentage of top-scoring hits filtered out in each species during this step will depend on the number of hits (N) the user chooses to retrieve. A permissive (i.e. 20000) value for N will include many more hit-pairs during this step, but since the hit-pairs passing the filter are then sorted by either average or maximum or minimum score (see next step), often the value of N chosen won't affect which hits appear at the top of the list. The only case where this can affect the top-ranked hit-pairs is the results sorted by mismatch first, then by some score criterion (minimum, average, or maximum).

5. Output sorted Hit-Pairs in HTML tables Finally, the resulting hit-pairs are sorted by either minimum, average, or maximum score in hit-pair, and by number of mismatches between the hit sequences. Three possible combinations of primary/secondary sortings are provided. Ranking by average score first gives the most statistically correct measure of rank order in the context of our protocol. The other sortings are provided for more specialized analysis of the results.

For the web based CisOrtho utility each of these steps is executed automatically for the user once the binding site alignment is submitted.

Running CisOrtho

Starting with an alignment of known transcription factor binding sites/*cis*-regulatory sequences, CisOrtho can be used to predict additional targets of a transcription factor or additional genes controlled by a given *cis*-regulatory motif. The following details how to do this using the CisOrtho Prediction user interface page, shown below.

CisOrtho

[Introduction](#) | [Methods](#) | [Download](#) | [Prediction](#)

Transcription Factor Binding Site Prediction using Position Weight Matrices and *C. elegans/C. briggsae* Ortholog-based Filtering (Phylogenetic Footprinting)

Output name (optional):

500 Maximum number of retrieved hit-pairs desired
10000 Maximum number of hits retrieved per species before phylogenetic filtering
3 Maximum number of next-highest hits reported for each gene

Filename:

or Paste it here:

Please submit a TF Binding Site Alignment Formatted as in this [Example](#)
Alignment must have lines all the same length (minimum 6 characters), containing only [acgtnACGTN] Results will appear on a randomly named webpage which will be deleted after two hours

Note: Binding Site Alignment must have at least 6 conserved columns (as judged by HMMER) or the search cannot be run. Please read the [CisOrtho manual](#) for further details.

1. Enter the ‘Output name’ with which you would like to have your output files prefixed. If no name is entered, ‘Results’ will be used as a default prefix (see CisOrtho Output section).
2. Use the pull down tab to select the ‘Maximum number of hits retrieved per species before phylogenetic filtering.’ This is the approximate number of N-highest hits that will be retrieved from searching the non-exonic sequences of each genome. CisOrtho actually returns all hits in non-exonic sequences that are above a certain cutoff score. The cutoff score that is used is determined by CisOrtho to approximate the number of initial hits that are desired by the user (see Methods, Sections 3,4). The number chosen here ultimately will determine the upper limit on the number of hit pairs that can be achieved. 10,000 is the default value.
3. Use the pull down tab to select the ‘Maximum number of retrieved hit pairs desired.’ This sets the upper limit on the number of hit pairs (hits above the cutoff, that are in the regulatory regions of one or more *C. elegans/C. briggsae* orthologous gene pairs) that are returned by

CisOrtho. This number does not affect the number of hit pairs generated by CisOrtho, which is determined by the number of initial hits (#2 above) that are retrieved. This parameter allows the user to determine how many of these hit pairs, ranked in order by different criteria (see 'CisOrtho Output,' C-E), will be returned to the user in Html tables. If this number is higher than the number of hit pairs found by CisOrtho, all hit pairs will be displayed.

4. **Use the pull down tab to select the 'Maximum number of next highest hits reported for each gene.'** CisOrtho sorts hit pairs based on the highest scoring hit in the regulatory regions of each gene. However, since some genes will have multiple high-scoring hits in their regulatory regions, CisOrtho has the option to have additional hits above the initial cutoff (#2 above) displayed as well. This number is the maximum number of additional hits, ranked by score, that will be displayed by CisOrtho.
5. **Enter your binding site alignment.** The user can enter an alignment of binding sites by a) saving the alignment as a text file and using the 'Browse...' tab to specify the filename of the alignment or b) typing or pasting the alignment into the box below 'or Paste it here:.' In either case, the alignment must be given in the form shown in the example in 'Methods,' Section 1 and 'CisOrtho Output,' B. This sample alignment can also be accessed online by Clicking on the 'Example' link on this page. Capital or lower case letters can be used for each base, but each binding site in the alignment must be the same length. Note that CisOrtho does not align binding sites. This must be done by the user. See 'Methods,' Section 1 and 'CisOrtho Output,' B, for a more detailed explanation of how Hmmer generates PWMs from alignments. This will aid the users in determining what sequence should be included in their alignments.
6. **Click 'Find Hits' tab to retrieve results.**

CisOrtho Output

Upon clicking the 'Find Hits' button the following screen is displayed:

The system is working on your request. This page will update every minute until finished. (Approx. 5 minutes)

The search has already started using the [Position Weight Matrix](#) generated from your input. There is an example Position Weight Matrix file and explanation in the Methods section of this site as well.

At this point, the user can click on the 'Position Weight Matrix' link to display, in a new window, the PWM that has been generated (see below for explanation of file). The page will then refresh every minute until the results page is displayed (below, red and green annotations are for this manual and are not displayed). The browser refresh button can also be used to check the search status (A).

The System is Working on your request. This page will update every minute until done

Searching and Filtering...Done.
Zipping results...Done.
GZipping Results...Done. } A

Please find all results files here.

- Each webpage contains 100 hits. View pages with suffix '1' first.
- ehits.txt and bhits.txt are the complete set of individual hits input into the phylogenetic filtering procedure.
- The .tgz (linux/mac osX) and .zip (windows/mac) files contain all files here

[PWMfile](#) } B
[Results ave min 1.html](#)
[Results ave min 2.html](#)
[Results ave min 3.html](#)
[Results ave min 4.html](#)
[Results ave min 5.html](#) } C
[Results mat min 1.html](#)
[Results mat min 2.html](#)
[Results mat min 3.html](#)
[Results mat min 4.html](#)
[Results mat min 5.html](#) } D
[Results ave mat 1.html](#)
[Results ave mat 2.html](#)
[Results ave mat 3.html](#)
[Results ave mat 4.html](#)
[Results ave mat 5.html](#) } E
[Results ehits.txt](#)
[Results bhits.txt](#) } F
[Results zip](#)
[Results tgz](#) } G

Please download promptly. This webpage will be deleted after two hours for your privacy.

The following files are then available for viewing/download:

B) Position Weight Matrix file (below). This is an example of a PWM generated by Hmmer using the

sample alignment shown in the methods section and below. (Color coding was added in this text as an aid in the explanation of the Hmmer output). As discussed in the Methods section, only the **emission parameters** (highlighted in green) are used by SimpleSearch. The numbers on the left correspond to the columns of the **PWM** (cyan) and the numbers on the right

correspond to the columns of the **alignment** (red). Note that positions **8** and **17** from the alignment are not represented in the PWM (see methods for explanation).

```
GAGCAAGTCCCATGCAATT
TAGCAGGTCCCATGCACTT
AAGCAAGTCCCATGCAATG
GGGCAAGTCCCATGCAATA
AAGCAGTTCCTATGCAGAC
AAGCAGTTCCTATGCAGAC
AAGCAGTTCCTATGCAGAC
GAGAAGCGCCCAAGCATTG
GCGCAACCACCATGCACTT
GCGCAACCACCATGCACTT
GGGCAGCGAACAAGCATCT
CAGCAGCGCCCATGCAACC
GCGCAACCACCATGCACTT
GGTGAGCCCAAGCATCTT
ACGCAACACCATATCACTC
AGCAGCGCCCATGCAATG
CAGCAGGATACATGCAACC
TGCGGCGCCCATGCACTC
GACCACCAAGCAAGCACTT
```

SimpleSearch then uses the matrix shown below as its search window to scan sequence.

SimpleSearch calculates the score of each search window as it scans through non-exonic sequences by summing the values for any given base in each position. For example, for a search window containing the sequence ‘GAGCAAG**C**CCTATG**C**AA**A**CT,’ the score would be the sum of the numbers in red type in the matrix below. Note that positions **8** and **17** are given zero values for all bases, serving as spacers, since Hmmer found the frequency of bases at these positions to be near background distribution.

```
HMMER2.0 [2.2g]
NAME align1_stockholm
LENG 17
ALPH Nucleic
RF no
CS no
MAP yes
COM hmmbuild --null ../per/eb.null --prior ../per/eb.pri -F align_ex1.hmm
align1_stockholm
NSEQ 19
DATE Thu Oct 30 20:32:29 2003
CKSUM 4418
XT -9967 -1 -1000 -1000 -9967 -1 -9967 -1
NULT -1 -9967
NULE 396 -547 -547 396
HMM
  A C G T
m->m m->i m->d i->m i->i d->m d->d b->m m->e
-70 * -4392
1 -231 -514 1472 -1392 0
- -396 547 547 -396
- -3 -9479 -10521 -894 -1115 -701 -1378 -70 *
2 568 158 834 -4321 2
- -396 547 547 -396
- -3 -9479 -10521 -894 -1115 -701 -1378 * *
3 -4321 273 2016 -1959 3
- -396 547 547 -396
- -3 -9479 -10521 -894 -1115 -701 -1378 * *
4 -1106 2016 -304 -4321 4
- -396 547 547 -396
- -3 -9479 -10521 -894 -1115 -701 -1378 * *
5 1357 -4323 -370 -4321 5
- -396 547 547 -396
- -3 -9479 -10521 -894 -1115 -701 -1378 * *
6 -93 560 1305 -4321 6
- -396 547 547 -396
- -3 -9479 -10521 -894 -1115 -701 -1378 * *
7 -4321 1915 690 -2407 7
- -573 519 685 -347
- -4643 -60 -10521 -10 -7179 -701 -1378 * *
8 -300 1801 -1414 -2189 8
- -396 547 547 -396
- -3 -9479 -10521 -894 -1115 -701 -1378 * *
9 -1229 2152 -1046 -4321 9
- -396 547 547 -396
- -3 -9479 -10521 -894 -1115 -701 -1378 * *
10 -4321 2486 -4323 -4321 10
- -396 547 547 -396
- -3 -9479 -10521 -894 -1115 -701 -1378 * *
11 1555 -4323 -4323 -4321 11
- -396 547 547 -396
- -3 -9479 -10521 -894 -1115 -701 -1378 * *
12 -126 -4323 -551 802 12
- -396 547 547 -396
- -3 -9479 -10521 -894 -1115 -701 -1378 * *
13 -1859 -4323 2181 -1399 13
- -396 547 547 -396
- -3 -9479 -10521 -894 -1115 -701 -1378 * *
14 -4321 2486 -4323 -4321 14
- -396 547 547 -396
- -3 -9479 -10521 -894 -1115 -701 -1378 * *
15 1555 -4323 -4323 -4321 15
- -482 643 -12 -6
- -4643 -60 -10521 -10 -7179 -701 -1378 * *
16 -1612 1168 -4323 602 16
- -396 547 547 -396
- -3 -9479 -10521 -894 -1115 -701 -1378 * *
17 -2636 1020 83 348 17
- * * * * * * * * * *
//
```

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
A	-231	568	-4321	-1106	1357	-93	-4321	0	-300	-1229	-4321	1555	-126	-1859	-4321	1555	0	-1621	-2636
C	-514	158	273	2016	-4323	560	1915	0	1801	2152	2486	-4323	-4323	-4323	2486	-4323	0	1168	1020
G	1472	834	2016	-304	-370	1305	690	0	-1414	-1046	-4323	-4323	-551	2181	-4323	-4323	0	-4323	83
T	-1392	-4321	-4321	-4321	-4321	-4321	-2407	0	-2189	-4321	-4321	-4321	802	-1399	-4321	-4321	0	602	348

C) Html tables of hits sorted primarily by average score of the hits from *C. elegans* and *C. briggsae*, then by the minimum of the two scores. The first link lists the top 100 hits, the second the next 100, etc. Navigation between these pages can be achieved by clicking on the links at the top of the page (see example below). A sample of the output obtained running the sample matrix is shown below.

1..100 101..200 201..300 301..400 401..500													
Rank	Score	Site	Mis	Type	Str1	Offset1	ID1	Common1	Str2	Offset2	ID2	Common2	Species
1	21340	GCGCAGCAGCGCATGCACCG	2	alt_5' intergenic	N	-13967	C18E3.9	-	P	-280	C43E11.6c	nab-1	<i>C.elegans</i>
	12503	GACCGGATGCCAAGCATTTG		alt_5' intergenic	N	-9859	C18E3.9		P	-4388	C43E11.6c	nab-1	
	10289	ACCAACCGACCATCCAATT		alt_5' intergenic	N	-1728	C18E3.9		P	-12519	C43E11.6c	nab-1	<i>C.briggsae</i>
	21637	GCGCACCACGCAAGCACCG		5' intergenic	N	-268	CBG12099	-	P	-1604	CBG12100	-	
	10889	TACGGGGCGCCCATGGAAATG		5' intergenic	N	-182	CBG12099		P	-1690	CBG12100		
	10856	ACCTAGCACCCAAGCCTCC		5' intergenic	N	-1755	CBG12099		P	-117	CBG12100		
2	15084	ACTGGACACCCATGCAATT	6	3'/5' intergenic	P	+4579	C09F12.1	clc-1	P	-234	C09F12.2	-	<i>C.elegans</i>
	24395	GGCCACCACCCATGCAATT		3'/5' intergenic	N	-264	CBG17510	-	N	+4371	CBG17511	-	<i>C.briggsae</i>
3	15537	GAGCAGCTGGCAAACAGCG	9	5' intergenic	N	-219	F32A6.5	sto-2	P	-17604	C26B9.5	-	<i>C.elegans</i>
	10994	AGGAAACCAGCAGTCAGCT		5' intergenic	N	-6364	F32A6.5	sto-2	P	-11459	C26B9.5		
	10067	GGTCAGGAAACAGCAGTC		5' intergenic	N	-6360	F32A6.5	sto-2	P	-11463	C26B9.5		<i>C.briggsae</i>
	22996	GGGCACCACCAAGCAAAT		5' intergenic	N	-5724	CBG14656	-	P	-8550	CBG14660	-	
	14253	GGCCAGCAACCCCTTCATT		5' intergenic	N	-7536	CBG14656		P	-6738	CBG14660		
	10209	GAGCAACACACACCCAATT		5' intergenic	N	-8501	CBG14656		P	-5773	CBG14660		

11	18107	GCCAGCCGCCAAGCAAAT	8	alt_intronic1	P	+1565	M03A1.1b	vab-1					<i>C.elegans</i>
	12624	CCGGAGGATACAGGCATTC		alt_3'/5' intergenic	P	+4708	M03A1.7	dao-2	P	-3931	M03A1.1b	vab-1	
	11144	AAGCTGCACCCATGCGCAT		alt_intronic5	P	+11779	M03A1.1b	vab-1					<i>C.briggsae</i>
	18171	GGCGCCCCCAAGCCCT		intronic4	P	+12156	CBG13447	-					
	10848	CCGAGGGAACCACTACTG		intronic4	p	+9459	CBG13447						
.

The information listed in the Html tables is the following:

Rank: rank order of hit pair on list for any given sorting criteria. Hit pairs are sorted and listed with blank lines between each hit pair.

Score: score of site in the weight matrix.

Site: sequence of potential binding site.

Mis: Number of base mismatches between the highest scoring hits in *C. elegans* and *C. briggsae*.

Type: Type of non-exonic region (see Section 2 of Methods).

Each of the next four columns lists information for the gene(s) that flank the conserved site in each species. Note that only one gene will be listed for segment types 'intronic,' 'BEGIN,' and 'END.' Additionally, depending on synteny between *C. elegans* and *C. briggsae*, one or both of the flanking genes may be orthologous.

Str1/2: negative (N) or positive (P), strand on which the first/second of the gene(s) that flank the identified target site are located.

Offset1/2: distance of the target site to the flanking gene(s), given in relation to the start codon if the target site is 5' or located in an intron and given in relation to the stop codon if the site is 3' to the gene; in the latter two cases, the number has a positive value.

ID1/2: cosmid name or CBG name of the gene(s) flanking the site.

Common1/2: 3-letter gene names of flanking gene(s) (if available). This column has links to the WormBase gene model page for each gene, which contains further information about the gene.

Species: species in which hit was found, *C. elegans* hits are listed first, then *C. briggsae* hits for each hit pair.

Color-coding: Orthologous *C. elegans*/*C. briggsae* genes (“hit-pairs”) are color coded in blue (C09F12.1 and CBG17511 are orthologs, second ranked hit pair above) and green (C09F12.2 and CBG17510 are orthologs, second ranked hit pair above). Red color-coding is used if a flanking gene is not orthologous to another listed gene.

If the option to report the next highest scoring hits for each ortholog is used, the top-scoring *C. elegans* or *C. briggsae* hit in each hit-pair will be in black type, and the next *n-1* hits will be gray.

D) Html tables of hits sorted primarily by the number of mismatches between the sites found in *C. elegans* and *C. briggsae*, then by the minimum of the two scores. The first link lists the top 100 hits, the second the next 100, *etc.* The output tables are in the same format as in **C**.

E) Html tables of hits sorted primarily by average score of the hits from *C. elegans* and *C. briggsae*, then by the number of mismatches between the two sites. The first link lists the top 100 hits, the second the next 100, *etc.* The output tables are in the same format as in **C**.

F) Text files of all *C. elegans* ([Results_ehits.txt](#)) and *C. briggsae* ([Results_bhits.txt](#)) sites ranked in order by score before phylogenetic filtering. The columns in these files, from left to right, are **Score**, **Site**, **Type**, **Str1**, **ID1**, **Offset1**, **Str2**, **ID2**, and **Offset2**.

G) Zipped files containing all of the files (**B-E**) mentioned above.

Files are named as shown on the sample results page above. The default names of the output files are ‘Results_*.‘ (green box). This naming scheme is used if no ‘Output name’ is entered by the user. Otherwise, the ‘Output name’ supplied by the user replaces ‘Results’ in each of these file names.

Note that the web address of the results page contains a randomly generated number to help insure the security of your results. Because these results are deleted after two hours, the user is encouraged to download the output promptly.