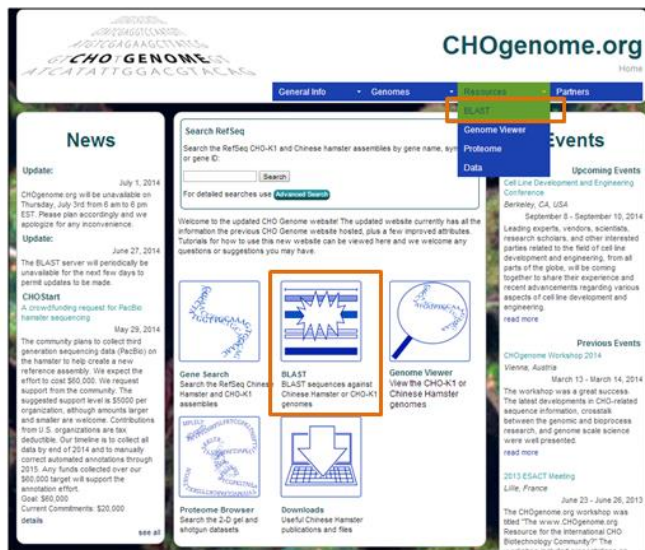


Tutorial 4 – BLAST Searching the CHO Genome

Accessing the CHO Genome BLAST Tool

The CHO BLAST server can be accessed by clicking on the BLAST button on the home page or by selecting “BLAST” from the menu bar under the Resources tab. This tab is available on all web pages within the CHO genome project. An additional link to the CHO genome BLAST web server is also provided on the CHO genome search pages, as well as a link to the NCBI BLAST web server.

Select the BLAST icon on the home page or from the resources tab



Click on the BLAST server link on the CHO genome search pages



Using the CHO Genome BLAST Tool

1) The CHO BLAST page allows for BLAST searches against the CHO and Chinese hamster (CH) genome databases.

CHOblast Search

BLAST Search - Required parameters help

Enter query sequences here in Fasta format

Or upload fasta file: No file chosen

Algorithm: blastn - Nucleotide Database

Database(s):

Genome (Scaffolds)

- 1) CHO-K1[ATCC]_RefSeq_2014
- 2) CH_RefSeq_2014
- 3) CHO-K1[ATCC]_GenBank_2011
- 4) CH_GenBank_2013
- 5) CH-17A/GY_Chr_GenBank_2013

Transcripts (RNA)

- 6) CHO-K1[ATCC]_RefSeq_2014
- 7) CHO-K1[ATCC]_RefSeq_2012
- 8) CH_RefSeq_2014

Currently available nucleotide and protein databases details

Nucleotide Databases:

Genome (Scaffolds)

- 1) CHO-K1[ATCC]_RefSeq_2014
- 2) CH_RefSeq_2014
- 3) CHO-K1[ATCC]_GenBank_2011
- 4) CH_GenBank_2013
- 5) CH-17A/GY_Chr_GenBank_2013

Transcripts (RNA)

- 6) CHO-K1[ATCC]_RefSeq_2014
- 7) CHO-K1[ATCC]_RefSeq_2012
- 8) CH_RefSeq_2014

Amino Acid Databases:

Proteins

- 1) CHO-K1[ATCC]_RefSeq_2014
- 2) CHO-K1[ATCC]_RefSeq_2012
- 3) CH_RefSeq_2014
- 4) CHO-K1[ATCC]_GenBank_2011
- 5) CH-17A/GY_Chr_GenBank_2013

Assembly Color Key:

RefSeq Assembly
GenBank Assembly

Assembly ID Key:

CHO-K1 RefSeq (GCF_000223135.1)
CH RefSeq (GCF_000419365.1)
CHO-K1 GenBank (GCA_000223135.1)
CH GenBank (GCA_000419365.1)
CH-17A/GY GenBank (GCA_000448345.1)

Database Naming Convention:

CHO Chinese hamster ovary cell line
CH Chinese hamster cell
CH(O)-xxxx Strain definition
[xxxx] Source of cells
genbank GenBank assembly
refseq RefSeq assembly
chr Chromosomal identification

BLAST Search - Other parameters help

Expect threshold:

Word size: 11

Max target sequences: 50

Match/Mismatch scores: 2-3

Gap costs: Existence: 5, Extension: 2

Filter: Low complexity regions

Mask: Mask for lookup table only
 Mask for lower case letters

Alignment: Perform ungapped alignment

Alignment output format: pairwise

Other parameters:

Hosted by Delaware Biotechnology Institute / CBCB at the University of Delaware
BLAST tool adapted from ViroBLAST v2.2 © 2005-2010 University of Washington. All rights reserved. (Terms of Service)

The nucleotide and amino acid databases hosted on the Chinese hamster genome database are listed to the right of the Basic Search panel. The nucleotide databases are divided into Genome (scaffold) and Transcript (RNA) databases, while the amino acid databases consist only of protein databases. The organism or cell line of origin is listed first, followed by the type of assembly (RefSeq or GenBank), and finally the year of release. The keys for the abbreviations and naming conventions are listed below these database lists.

2

2) Additional details regarding the multiple BLAST programs and databases are available. Clicking on the [Algorithm](#) link provides a brief description of the BLAST programs.

Programs available for CHOblast	
blastn	compares a nucleotide query sequence against a nucleotide sequence database
blastp	compares an amino acid query sequence against a protein sequence database
blastx	compares a nucleotide query sequence translated in all reading frames against a protein sequence database
tblastn	compares a protein query sequence against a nucleotide sequence database dynamically translated in all reading frames
tblastx	compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database

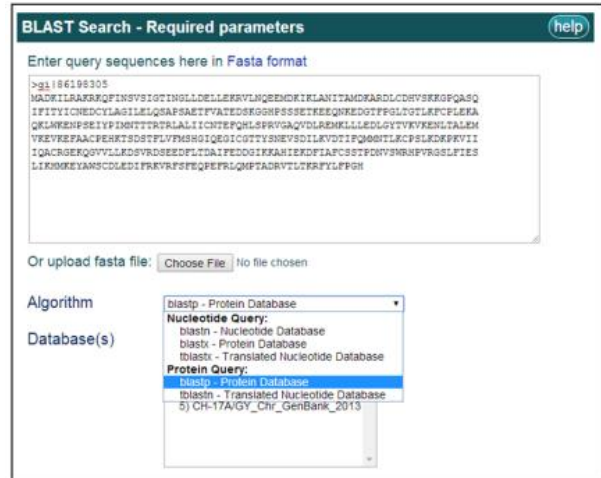
Clicking on the [Database\(s\)](#) link or the [details](#) button in the “Currently available nucleotide and protein databases” section title bar brings up a webpage with a brief description of the databases currently available for BLAST searching, including the name, version, date, and a link to the original publication article.

Databases available for CHOblast		
Nucleotide sequence databases (blastn, tblastn, tblastx):		
Genome (Scaffolds)		
Name	Release Date/ Assembly Version	Reference
1) CHO-K1 ATCC RefSeq_2014	08 May 2014 Assembly v1.0	Xu X, Nagarajan H, Lewis NE et al. The genomic sequence of the Chinese hamster ovary (CHO)-K1 cell line. <i>Nature Biotechnology</i> , 29(8), 735-741 (2011). [LINK]
2) CH_RefSeq_2014	08 May 2014 Assembly v1.0	Lewis NE, Liu X, Li Y et al. Genomic landscapes of Chinese hamster ovary cell lines as revealed by the <i>Cricetulus griseus</i> draft genome. <i>Nature Biotechnology</i> , 31(8), 759-765 (2013). [LINK]
3) CHO-K1 ATCC GenBank_2011	26 August 2011 Assembly v1.0	Xu X, Nagarajan H, Lewis NE et al. The genomic sequence of the Chinese hamster ovary (CHO)-K1 cell line. <i>Nature Biotechnology</i> , 29(8), 735-741 (2011). [LINK]
4) CH_GenBank_2013	12 July 2013 Assembly v1.0	Lewis NE, Liu X, Li Y et al. Genomic landscapes of Chinese hamster ovary cell lines as revealed by the <i>Cricetulus griseus</i> draft genome. <i>Nature Biotechnology</i> , 31(8), 759-765 (2013). [LINK]
5) CH-17A/GY_Chr_GenBank_2013	29 August 2013 Assembly v1.0	Brinkhoff K, Rupp O, Laux H et al. Chinese hamster genome sequenced from sorted chromosomes. <i>Nature Biotechnology</i> , 31(8), 694-695 (2013). [LINK]
Transcripts (RNA)		
Name	Release Date/ Assembly Version/ Annotation Version	Reference
6) CHO-K1 ATCC RefSeq_2014	08 May 2014 Assembly v1.0 Annotation v101	Xu X, Nagarajan H, Lewis NE et al. The genomic sequence of the Chinese hamster ovary (CHO)-K1 cell line. <i>Nature Biotechnology</i> , 29(8), 735-741 (2011). [LINK]
7) CHO-K1 ATCC RefSeq_2012	15 March 2012 Assembly v1.0 Annotation v1	Xu X, Nagarajan H, Lewis NE et al. The genomic sequence of the Chinese hamster ovary (CHO)-K1 cell line. <i>Nature Biotechnology</i> , 29(8), 735-741 (2011). [LINK]
8) CH_RefSeq_2014	08 May 2014 Assembly v1.0 Annotation v101	Lewis NE, Liu X, Li Y et al. Genomic landscapes of Chinese hamster ovary cell lines as revealed by the <i>Cricetulus griseus</i> draft genome. <i>Nature Biotechnology</i> , 31(8), 759-765 (2013). [LINK]
Amino Acid Sequence Databases (blastp, blastx):		
Proteins		
Name	Release Date/ Assembly Version/ Annotation Version	Reference
1) CHO-K1 ATCC RefSeq_2014	08 May 2014 Assembly v1.0 Annotation v101	Xu X, Nagarajan H, Lewis NE et al. The genomic sequence of the Chinese hamster ovary (CHO)-K1 cell line. <i>Nature Biotechnology</i> , 29(8), 735-741 (2011). [LINK]
2) CHO-K1 ATCC RefSeq_2012	15 March 2012 Assembly v1.0 Annotation v1	Xu X, Nagarajan H, Lewis NE et al. The genomic sequence of the Chinese hamster ovary (CHO)-K1 cell line. <i>Nature Biotechnology</i> , 29(8), 735-741 (2011). [LINK]
3) CH_RefSeq_2014	08 May 2014 Assembly v1.0 Annotation v101	Lewis NE, Liu X, Li Y et al. Genomic landscapes of Chinese hamster ovary cell lines as revealed by the <i>Cricetulus griseus</i> draft genome. <i>Nature Biotechnology</i> , 31(8), 759-765 (2013). [LINK]
4) CHO-K1 ATCC GenBank_2011	26 August 2011 Assembly v1.0	Xu X, Nagarajan H, Lewis NE et al. The genomic sequence of the Chinese hamster ovary (CHO)-K1 cell line. <i>Nature Biotechnology</i> , 29(8), 735-741 (2011). [LINK]
5) CH-17A/GY_Chr_GenBank_2013	29 August 2013 Assembly v1.0	Brinkhoff K, Rupp O, Laux H et al. Chinese hamster genome sequenced from sorted chromosomes. <i>Nature Biotechnology</i> , 31(8), 694-695 (2013). [LINK]
Database Naming Convention:		Assembly Color Key:
CHO Chinese hamster ovary cell line		RefSeq Assembly
CH Chinese hamster cell		GenBank Assembly
CH(O)xxxxx Strain definition		Assembly ID Key:
xxxx Source of cells		CHO-K1 RefSeq (GCF_000223135.1)
GenBank GenBank assembly		CH RefSeq (GCF_000419365.1)
RefSeq RefSeq assembly		CHO-K1 GenBank (GCA_000223135.1)
Chr Chromosomal identification		CH GenBank (GCA_000419365.1)
		CH-17A/GY GenBank (GCA_000448345.1)

3) Query sequences in FASTA format can be pasted into the search box at the top of the page or uploaded as a FASTA file. Multiple query sequences may be entered for each search.

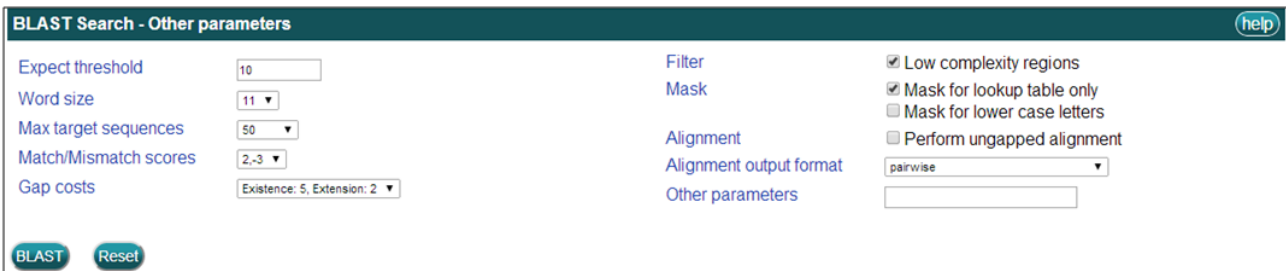
The BLAST program and database are then selected from the currently available options.

For example, to BLAST the most recent CH genome protein database, select the **blastp** program and the **CH_RefSeq_2014** database from the “Proteins” list.



To perform a basic BLAST search, click the **BLAST** button after all the above information is entered and selected. If you wish to perform a more advanced search, do not hit basic search yet and proceed to instruction #4.

4) In the BLAST Search – Other parameters section, the default BLAST parameters can be varied to perform an altered, more advanced BLAST search.



Clicking on the highlighted blue terms (such as **Expect threshold**, etc.) will provide a brief description of each advanced search parameter that can be varied.

To perform an advanced BLAST search, click the **BLAST** button once all the required information is entered and the advanced parameters are altered.

5) The results of the BLAST alignment are summarized in a table with the query sequence name, the subject sequence name, the bit score, the identity length, the identity percentage, and the *E*-value.

Query	Subject	Bit_Score	Identity (Query_Len)	Similarity	E-Value
gj 86198305	gi 625278770 ref XP_007631029.1 PREDICTED: caspase-1 isoform X2 [Cricetulus griseus]	598	284/363 (402)	78	0.0
gj 86198305	gi 625278768 ref XP_007631028.1 PREDICTED: caspase-1 isoform X1 [Cricetulus griseus]	664	318/402 (402)	79	0.0
gj 86198305	gi 625278766 ref XP_007631027.1 PREDICTED: caspase-12-like isoform X2 [Cricetulus griseus]	280	172/437 (402)	39	3e-89
gj 86198305	gi 625278791 ref XP_007631040.1 PREDICTED: caspase-1-like [Cricetulus griseus]	135	65/89 (402)	73	7e-37
gj 86198305	gi 625260408 ref XP_007621635.1 PREDICTED: caspase-14 [Cricetulus griseus]	72.8	54/195 (402)	28	2e-14
gj 86198305	gi 625241515 ref XP_007611988.1 PREDICTED: caspase-6 isoform X2 [Cricetulus griseus]	68.9	67/230 (402)	29	6e-13
gj 86198305	gi 625241513 ref XP_007611987.1 PREDICTED: caspase-6 isoform X1 [Cricetulus griseus]	68.9	65/229 (402)	28	9e-13
gj 86198305	gi 625249028 ref XP_007615799.1 PREDICTED: caspase-7 isoform X3 [Cricetulus griseus]	67.4	65/246 (402)	26	4e-12
gj 86198305	gi 625249259 ref XP_007615919.1 PREDICTED: caspase-9-like isoform X2, partial [Cricetulus griseus]	62.4	46/153 (402)	30	3e-11
gj 86198305	gi 625258876 ref XP_007619804.1 PREDICTED: putative caspase-16 [Cricetulus griseus]	48.5	47/164 (402)	29	6e-06
gj 86198305	gi 625249026 ref XP_007615798.1 PREDICTED: caspase-7 isoform X1 [Cricetulus griseus]	43.0	50/236 (402)	25	1e-04

The results can be filtered by score (showing only the top 1, 5, or 10 alignments), by Similarity Cutoff Percentage, or by BLAST Bit Score.

After entering the filter parameter, click either the “Filter” or the “Parse again” buttons to refresh the results table.

To view the RefSeq/GenBank entry for each subject sequence, click on the sequence name in the Subject column (i.e. gi|625278770|ref|XP_007631029.1|).

To view the pair-wise alignment for a specific alignment, click on the value in the Score column for any alignment (i.e. 598).

Query	Subject	Bit_Score	Identity (Query_Len)	Similarity	E-Value
gj 86198305	gi 625278770 ref XP_007631029.1 PREDICTED: caspase-1 isoform X2 [Cricetulus griseus]	598	284/363 (402)	78	0.0
gj 86198305	gi 625278768 ref XP_007631028.1 PREDICTED: caspase-1 isoform X1 [Cricetulus griseus]	664	318/402 (402)	79	0.0
gj 86198305	gi 625278766 ref XP_007631027.1 PREDICTED: caspase-12-like isoform X2 [Cricetulus griseus]	280	172/437 (402)	39	3e-89
gj 86198305	gi 625278791 ref XP_007631040.1 PREDICTED: caspase-1-like [Cricetulus griseus]	135	65/89 (402)	73	7e-37
gj 86198305	gi 625260408 ref XP_007621635.1 PREDICTED: caspase-14 [Cricetulus griseus]	72.8	54/195 (402)	28	2e-14

PREDICTED: caspase-1 isoform X2 [Cricetulus griseus]
 NCBI Reference Sequence: XP_007631029.1
[FASTA](#) [Graphics](#)

[Go to:](#)

LOCUS XP_007631029 363 aa linear ROD 30-APR-2014
 DEFINITION PREDICTED: caspase-1 isoform X2 [Cricetulus griseus].
 ACCESSION XP_007631029
 VERSION XP_007631029.1 gi:625278770
 DBLINK BioProject: PRJNA239316
 DBSOURCE REFSEQ: accession XP_007631029.1

```
> gj|86198305 on gi|625278770|ref|XP_007631029.1| PREDICTED: caspase-1 isoform X2 [Cricetulus griseus]
Length=363

Score = 598 bits (1542), Expect = 0.0, Method: Compositional matrix adjust.
Identities = 284/363 (78%), Positives = 317/363 (87%), Gaps = 0/363 (0%)

Query 40 MDKIKLANITAMDKARDLGDHVSKKGPGASQIFITYICNEDCVLAGLELQSAPSAETVF 99
M++IK N T MDKARDLCD V+KKGPG ASQI IITYIC EDCVLAG+LEL+S P AE +
Sbjct 1 MERIKCINATVMDKARDLCDVTKKGPLASQICITYICKEDCVLAGVLELESPPAENSM 60

Query 100 ATEDSKGGHPSSSETKEEQNKEDGTFPGLTGLTKFCPLKAQKLMKNFSEIYIPIMNTTI 159
T+D +GG+PSSSETKEEQ KE GT PG +G+LK C LE AQR+ KENFSEIYIPIM+T+T
Sbjct 61 RTDDFGQGYVPSSETKEEQKKEGGTCGPGSGSLKLCLETAQRKIRKENFSEIYIPIMDTST 120
```


BLAST searching the CHO-K1 Genome at NCBI

A link to the NCBI BLAST web server is also provided on the CHO-K1 genome search pages. To BLAST the CHO genome using the NCBI BLAST web server, enter the required BLAST information and select the “*Cricetulus griseus* WGS” database under the “Choose Search Set” menu.

Search Page

Search Term: Search

Search Term	Genome
Symbol	CHO-K1 (RefSeq Assembly GCF_000223151.1 2May2014 - Release 101)
Gene Name	Chinese Hamster (RefSeq Assembly GCF_000415085.1 2May2014 - Release 101)
Gene ID	CHO-K1 (RefSeq Assembly GCF_000223151.1 15Mar2012 - Release 1)

The CHO-K1 RefSeq database can be searched by:

- Gene name (i.e. Caspase 1)
- Gene symbol (i.e. Casp1)
- Gene ID (i.e. 100759171)

BLAST the CHO-K1 RefSeq and Chinese Hamster RefSeq genomes here and **NCBI**

Tips for using the database:

- Search by gene name, symbol, or ID to find individual gene pages.
- Multiple genomes may be selected at once, but the time required for the query may increase.
- Each gene, transcript, and protein has a unique, individual entry. To obtain the relevant protein information or download the protein sequence, select the gene or transcript entry of interest, scroll to the bottom of the "Gene Details" page, and select the protein entry associated with the relevant transcript in the "Gene Relations" table.
- Many pseudogenes do not have a gene name or symbol, but all have a gene ID and may be searched.

NCBI/BLAST/blastn suite **Standard Nucleotide BLAST**

blastn blastp blastx tblastn tblastx

BLASTN programs search nucleotide databases using a nucleotide query. [more...](#) [Reset page](#) [Bookmark](#)

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear Query subrange

From To

Or, upload file No file chosen

Job Title

Align two or more sequences

Choose Search Set

Database Human genomic + transcript Mouse genomic + transcript Others (nr etc.)

Nucleotide collection (nr/nt)

Organism Exclude

Exclude Models (XM/XP) Uncultured/environmental sample sequences

Limit to

Entrez Query [YouTube](#) [Create custom database](#)

Program Selection

Optimize for Highly similar sequences (megablast) More dissimilar sequences (discontiguous megablast) Somewhat similar sequences (blastn)

Choose a BLAST algorithm

BLAST Search database Nucleotide collection (nr/nt) using Megablast (Optimize for highly similar sequences) Show results in a new window