Tutorial 3 – Searching the Chinese Hamster Genome Database

Multiple Chinese hamster (CH) and CHO-K1 genome search pages are hosted on the CHO genome website. A comprehensive search function for the RefSeq CH and CHO-K1 genomes is located on the homepage. Additionally, an independent search page with additional selection details for the RefSeq genomes is available under the **Genomes** menu. Separate search pages for both the CH mitochondrial and GenBank CHO-K1 genomes are also available under the **Genomes** tab. The comprehensive RefSeq search function on the homepage can be accessed from any CHO genome webpage by selecting the CHO GENOME logo located in the top left corner.

Searching the RefSeq Chinese hamster (CH) or CHO-K1 Genomes

As of August 2014, there are 3 CH and CHO-K1 RefSeq genome databases. The initial CHO RefSeq genome that was hosted on CHOgenome.org is the CHO-K1 RefSeq (2012) genome database. The genomic information in this genome database has not been altered since its release in 2012, so it is not aligned with the most recent CHO-K1 genome database. The CHO-K1 RefSeq (2012) genome database is the database to search if you have previously used the CHO-K1 RefSeq (2012) database and are interested in obtaining identical results. In May 2014, NCBI released their first complete CH genome database and an updated CHO-K1 genome database. Currently, these two databases are the most recent CH and CHO genome databases to search against.

1) The comprehensive search function for the RefSeq CH and CHO-K1 genomes is located on the CHO genome homepage.
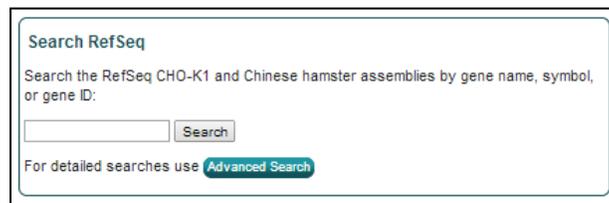
The gene name, gene symbol, or NCBI gene ID can be searched against all of the RefSeq CH and CHO genomes. After entering the search term and selecting the "Search" button, all relevant results are displayed.

2) To specify the search terms or RefSeq genomes to be searched, select the **CH & CHO RefSeq** link from the **Genomes** menu, the Advanced Search button on the homepage, or the Gene Search image button located on the homepage.

Select the type of keyword and genome(s) you would like to search, type the keyword into the search box, and press the Search button or hit enter.

The RefSeq genome databases can be searched by:

Gene symbol (*i.e.* **cfl1**)

The database can be searched using the official gene symbols assigned to gene products during the NCBI genome annotation projects. Unofficial gene symbols are not searchable and if the gene was not assigned a gene symbol, one was created. The created gene symbols consist of 'LOC' followed by the 9-digit NCBI gene ID (*i.e.* LOC100######).

Search Results for cfl1

1 Result Found

#	Assembly	Type	Symbol	GeneID	Product Descrition
1	CHO-K1 (2014)	mRNA	Cfl1	100760435	cofilin 1 (non-muscle)

Gene name (*i.e.* **cofilin 1**)

The database can be searched using the protein names assigned to the genes during the genome annotation projects.

Search Results for cofilin 1

1 Result Found

#	Assembly	Type	Symbol	GeneID	Product Descrition
1	CHO-K1 (2014)	mRNA	Cfl1	100760435	cofilin 1 (non-muscle)

NCBI Gene ID (*i.e.* **100760435**)

The database can be searched using the NCBI gene IDs. For the CHO-K1 and CH genomes, these gene IDs currently range between 100682525 and 103163833.

All CHO and Chinese hamster NCBI gene IDs can be found in the NCBI protein database by using the search term "*Cricetulus griseus.*"

Search Results for 100760435

1 Result Found

#	Assembly	Type	Symbol	GeneID	Product Descrition
1	CHO-K1 (2014)	mRNA	Cfl1	100760435	cofilin 1 (non-muscle)

Matches for the genome database searches are displayed in a tabular format. Searching for **cofilin** in the CHO-K1 (2014), Chinese Hamster (2014), and CHO-K1 (2012) assemblies returns 6 entries.

The general characteristics displayed for the search results include the parent assembly, gene feature type, gene symbol, NCBI gene ID, and gene name/product description.

To access more details about a single entry, click on the NCBI gene ID [*i.e.* **100757641**].

3) Searching and selecting the NCBI gene ID **100762756** from the CHO-K1 (2014) genome assembly will open the 'Gene Details' page for **Pfkfb3, Variant X2**.

The **General Information** section provides the gene name, gene symbol, NCBI gene ID, feature type, and the genome assembly of origin.

The **Genomic Location** section provides the scaffold on which the gene is located, the coordinates of the gene's coding region, the NCBI transcript ID, a link to the NCBI graphics page, a link to the FASTA nucleotide or amino acid sequences, and links to the NCBI protein homologs in human, mouse, and rat.

Gene Details: Pfkfb3

Database: CHO_RefSeq_052014

General Information

Name:	6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 3, transcript variant X2
Symbol:	Pfkfb3 [Propose a new symbol]
Gene ID:	100762756
Feature Type:	mRNA
Genome Assembly:	CHO-K1 (2014) - GCF_000223135.1

Genomic Information

Scaffold:	NW_003613580.1
Graphic (Scaffold):	NCBI View
Range (Graphic):	3304483 - 3390245
Transcript:	XM_007645552.1
Nucleotide Sequence:	Download sequence
Protein Homologs:	Human: 4758900 | Mouse: 295293217 | Rat: 16923988

Related Entries

Type	NCBI Link	Sequence	Graphics	Transcript #	
gene			Graphic		
mRNA	XM_007645552.1	Download	Graphic	X2	
- Protein	XP_007643742.1	Download	Graphic	X2	16 CDS results found
mRNA	XM_003494936.2	Download	Graphic	X1	
- Protein	XP_003494984.1	Download	Graphic	X1	16 CDS results found
mRNA	XM_003494930.2	Download	Graphic	X4	
- Protein	XP_003494978.1	Download	Graphic	X4	14 CDS results found
mRNA	XM_003494934.2	Download	Graphic	X3	
- Protein	XP_003494982.1	Download	Graphic	X3	15 CDS results found

The **Related Entries** section provides information and links to all of the features associated with each gene, including all of the relevant transcripts and proteins, in one table. Each row represents one feature associated with the gene and the highlighted row identifies the feature thats content is displayed on the current page. The table columns contain the following information:

- **Type** lists the row's feature type.
- **NCBI Link** displays the NCBI transcript/protein IDs.
- The **Sequence Download** links to the FASTA nucleotide or protein sequence.
- The **Graphics Graphic** links to the NCBI Map Viewer graphic associated with the row's gene feature type.
- **Transcript #** contains the NCBI transcript # if multiple transcripts exist for the gene of interest.

4) All hyperlinked text serves as a link to a corresponding webpage. Any green button with white text located on the 'Gene Details' webpage enables direct communication with the editors of CHO genome. These webpages are opened in a new tab, conserving the user's current 'Gene Details' webpage.

The **General Information** section contains one button and one hyperlink.

- The Propose a new symbol button allows the user to propose a new/different gene symbol with an accompanying reason for the proposed change. These submissions are cataloged and periodically evaluated.
- The **Genome Assembly** link [*i.e.* **GCF_000223135.1**], opens the webpage of the NCBI genome assembly associated with the selected gene.

The **Genomic Location** section provides links to multiple pages related to the genomic location information for the selected gene.

- The **Scaffold** link [*i.e.* **NW_003613580.1**] links to the NCBI scaffold information page of the scaffold containing the selected gene.
- The **Graphic (Scaffold)** link [**NCBI View**] links to the full NCBI Genome Viewer image of the scaffold containing the selected gene.
- The **Range (Graphic)** link [*i.e.* **3304483-3390245**] links to the NCBI Genome Viewer image zoomed in to the selected gene.
- The **Transcript** link [*i.e.* **XM_007645552.1**] links to the NCBI mRNA information page for the selected gene.
- The **Nucleotide Sequence** link [**Download sequence**] opens a page with the selected gene's mRNA FASTA file.
- The **Protein Homologs** link [*i.e.* **Human: 4758900**] lists the NCBI protein homolog IDs of the human, mouse, and rat homologs and link to the selected gene's NCBI protein webpage.

The **Related Entries** table provides links to the Chinese hamster genome pages, gene sequences, NCBI database information, and NCBI graphics pages.

The gene & RNA entries contain links to relevant webpages associated with the nucleotide entries.

- The **NCBI Link** [*i.e.* **XM_007645552.1**] links to the NCBI RNA-specific information page corresponding to the selected gene.
- The **Sequence** link [**Download**] provides the nucleotide FASTA file for the corresponding gene transcript.
- The **Graphics** link [**Graphic**] links to the NCBI Genome Viewer image of the selected gene's scaffold (for a **gene** row entry) or of the zoomed-in, transcript-specific portion of the scaffold (for a **RNA** row entry).

The protein row entries contain links to relevant webpages associated with the amino acid entries.
Note: The protein entry following each RNA entry is the corresponding protein, as often indicated by the transcript #.

- The **Type** link [**Protein**] links to the CHO genome results page associated with that protein entry.
- The **NCBI Link** [*i.e.* **XP_007643742.1**] links to the NCBI information page for the selected protein.
- The **Sequence** link [**Download**] provides the amino acid FASTA file for the selected protein.
- The **Graphics** link [**Graphic**] links to the zoomed NCBI Genome Viewer image of the selected protein.

Searching the CHO-K1 GenBank (2011) Genome

1) Select the **CHO-K1 GenBank (2011)** genome from the **Genomes** menu.
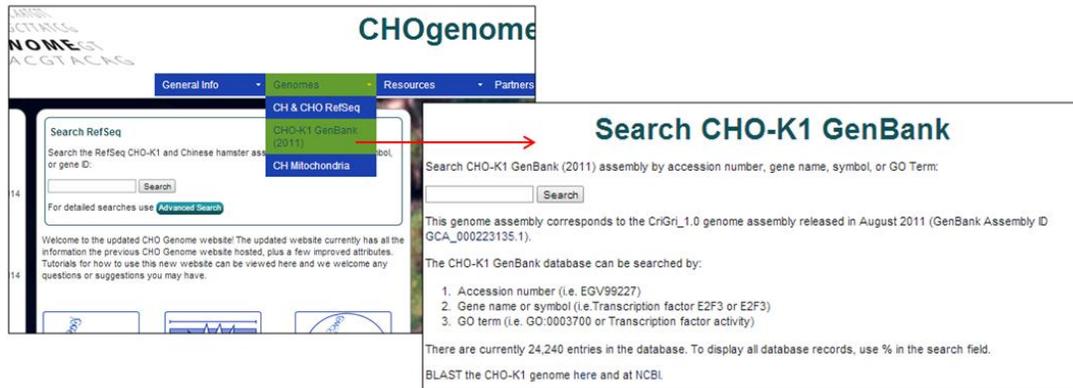
The CHO-K1 GenBank (2011) genome database is the initial genome database that was hosted on CHOgenome.org. The genome information does not align with the most recent CHO-K1 genome information, as the content has not been altered since its release in 2011. Selection of the CHO-K1 GenBank (2011) genome database is advantageous if the objective is the replication of previous results from the CHO-K1 GenBank (2011) genome database.

2) To search this database, type a keyword into the search box at the top of the page and submit your query. Typing **%** in the search box will list all 24,240 entries currently in the database.

The CHO-K1 genome database can be searched by:
Accession number (*i.e.* **EGV99227**)

> The database can be searched using the GenBank WGS protein accession IDs. For the CHO-K1 WGS project, these accession numbers are EGV##### or EGW#####, where # is any digit 0-9. These accession numbers can also be found in the NCBI protein database by restricting the search to "*Cricetulus griseus*" and the locus tag to "I79_######."

Gene name or **gene symbol** (*i.e.* Transcription factor E2F3 or E2F3)

> The database can be searched using gene names or gene symbols. The gene names were assigned to gene products during the annotation of the WGS project, while the gene symbols were assigned based on the annotation of homologous proteins.

GenBank Results for Transcription factor E2F3

Found 1 search result for "Transcription factor E2F3" in the GenBank assembly.

Genes

No.	Assembly	Gene	Accession	Parent Accession	Locus Tag	Product Qualifier
1	CHO-K1 GenBank	E2f3	EGV99227	JH000056	I79_002249	Transcription factor E2F3

Gene ontology (GO) term (*i.e.* GO:0003700 or Transcription factor activity)

> The database can be searched using a GO accession ID or term. Searching by GO accession ID or term will return all database entries annotated with that GO term.

GenBank Results for GO:0003700

Found 472 search results for "GO:0003700" in the GenBank assembly.

Genes

No.	Assembly	Gene	Accession	Parent Accession	Locus Tag	Product Qualifier
1	CHO-K1 GenBank	Etv3,Mets,Pe1	EGW07031	JH000344	I79_009579	ETS translocation variant 3
2	CHO-K1 GenBank	Etv3l	EGW07032	JH000344	I79_009580	ETS translocation variant 3-like protein
3	CHO-K1 GenBank	Mef2d	EGW07051	JH000344	I79_009600	Myocyte-specific enhancer factor 2D

3) The results from the GenBank searches are listed with 6 identifying characteristics per row.

GenBank Results for caspase

Found 13 search results for "caspase" in the GenBank assembly.

Genes

No.	Assembly	Gene	Accession	Parent Accession	Locus Tag	Product Qualifier
1	CHO-K1 GenBank	Casp8	EGV99437	JH000741	I79_014611	Caspase-8
2	CHO-K1 GenBank	Pacap	EGV97626	JH000776	I79_014961	Proapoptotic caspase adapter protein
3	CHO-K1 GenBank	Card11	EGV95787	JH000039	I79_001680	Caspase recruitment domain-containing protein 11

The returned entries' characteristics include the assembly of origin (CHO-K1 GenBank), gene symbol, protein accession ID, scaffold (parent accession), locus tag, and gene name/description (product qualifier).

4) To find more details about a single entry, select the gene symbol or accession ID [*i.e.* **Casp8** or **EGV99437**].

The **General Information** section provides the gene name, symbols, synonyms, locus tag, descriptions, and assembly of origin.

The **Genomic Information** section provides the WGS scaffold on which the gene is located, the coordinates of the gene coding region, and links to graphics of the scaffold and the gene.

The **Sequence Information** section provides links to download the nucleotide and protein sequence in FASTA format.

The **External Links** section provides links to the corresponding CHO protein entries in the NCBI and EMBL databases.

GenBank Details for Caspase-8

General Information

Name	Caspase-8
Symbol	Casp8
Synonyms	(not available)
Locus tag	I79_014611
Description	(temporary unavailable)
Genome Assembly	CHO-K1 GenBank (GenBank Assembly ID GCA_000223135.1)

Genomic Information

WGS Scaffold	JH000741
CDS	complement(join(JH000741:197973..198108,JH000741:199527..200025,JH000741:203332..203473,JH000741:205312..205376,JH000741:206895..206945,JH000741:207026..207164,JH000741:209567..209675,JH000741:214348..214652))
Graphics	Scaffold Gene

Sequence Information

Nucleotide sequence	Download sequence
Protein sequence	Download sequence

External Links

NCBI (protein)	EGV99437
NCBI (protein graphics)	EGV99437
EMBL-EBI (protein)	EGV99437

Homologs

	Mouse	Rat	Human
UniProtKB/Swiss-Prot	O89110		
UniProtKB/TrEMBL	B2CMK5		
NCBI Entrez Gene	12370	64044	841

Annotation

GO Terms	GO_process: GO:0042981 - regulation of apoptosis [Evidence IEA]
	GO_process: GO:0006915 - apoptosis [Evidence IEA]
	GO_process: GO:0006508 - proteolysis [Evidence IEA]
	GO_function: GO:0008234 - cysteine-type peptidase activity [Evidence IEA]
	GO_function: GO:0005515 - protein binding [Evidence IEA]
	GO_function: GO:0004197 - cysteine-type endopeptidase activity [Evidence IEA]

The **Homologs** section provides links to homologous proteins in other species in the UniProt databases and to homologous proteins in mouse, rat, and human in the NCBI Entrez databases.

The **Annotation** section provides functional information and GO terms assigned to the gene product. The GO terms link to the AmiGO gene ontology browser.

5) Hyperlinked text also serves as a link to additional relevant webpages.

The **General Information** section provides a link associated with the WGS assembly.

- In the **Genome Assembly** field, clicking on the genome assembly ID [*i.e.* **GCA_000223135.1**] opens the corresponding NCBI assembly report the selected gene was annotated from.

The **Genomic Information** section provides links associated with the WGS scaffold on which the selected gene is located and graphics of the scaffold and gene.

- In the **WGS Scaffold** field, clicking on the scaffold accession ID [*i.e.* **JH000741**] opens the NCBI database entry for the WGS scaffold the selected gene is located on.
- In the **Graphics** field, clicking on **Scaffold** opens the NCBI Map Viewer to view the genome annotation of the entire scaffold, while clicking on **Gene** opens the NCBI Map Viewer to view the genome annotation of the selected gene model portion of the scaffold.
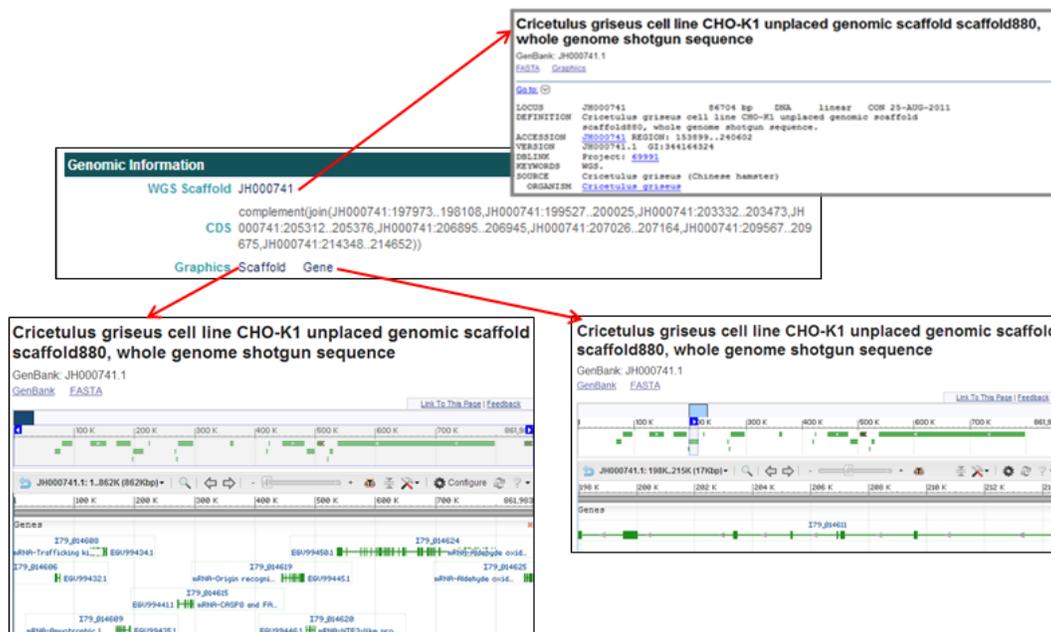
The **Sequence Information** section provides links to download the nucleotide and protein sequences in FASTA format.

- In the **Nucleotide sequence** field, clicking on **Download sequence** opens the nucleotide FASTA file.
- In the **Protein sequence** field, clicking on **Download sequence** opens the amino acid FASTA file.
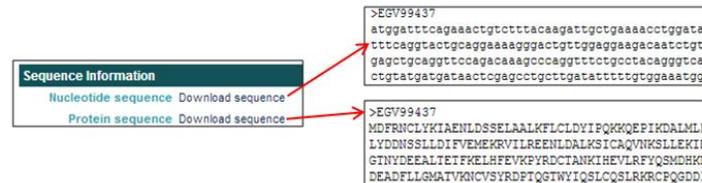
The **External Links** section provides links to the selected CHO protein pages in the NCBI and EMBL databases.

- In the **NCBI (protein)** field, clicking on the protein accession ID [*i.e.* **EGV99437**] opens the NCBI webpage of the selected protein entry.
- In the **NCBI (protein graphics)** field, clicking on the protein accession ID [*i.e.* **EGV99437**] opens the NCBI Map Viewer and visually shows the selected protein, region, and site features.
- In the **EMBL-EBI (protein)** field, clicking on the protein accession ID [*i.e.* **EGV99437**] opens the EMBL-EBI webpage containing both a visual and informational overview of the selected protein.
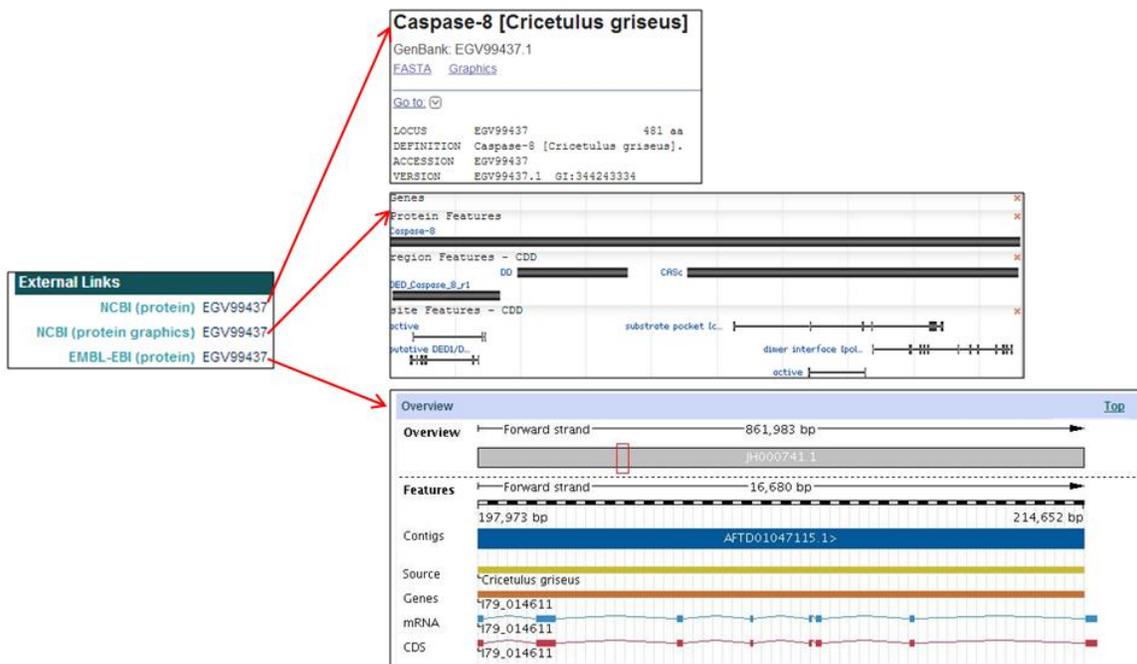
The **Homologs** section provides links to homologs of the selected CHO gene in the NCBI and UniProt databases.

- In the **UniProtKB/Swiss-Prot** field, clicking on the protein accession ID [*i.e.* **OB9110**] opens the UniProtKB/Swiss-Prot webpage displaying the information for a homolog to the CHO protein.
- In the **UniProtKB/TrEMBL** field, clicking on the protein accession ID [*i.e.* **B2CMK5**] opens the UniProtKB/TrEMBL webpage displaying the information for a homolog to the CHO protein.
- In the **NCBI Entrez Gene** field, clicking on the **Mouse** protein accession ID [*i.e.* **12370**] opens the NCBI gene webpage for the mouse homolog to the CHO protein.
- In the **NCBI Entrez Gene** field, clicking on the **Rat** protein accession ID [*i.e.* **64044**] opens a NCBI gene webpage for the rat homolog to the CHO protein.
- In the **NCBI Entrez Gene** field, clicking on the **Human** protein accession ID [*i.e.* **841**] opens the NCBI gene webpage for the human homolog to the CHO protein.

The **Annotation** section provides links to the GO Terms in the AmiGO databases.

- In the **GO Terms** field, clicking on the GO Term ID [i.e. **GO_Process: GO:0042981**] opens the AmiGO webpage displaying the description and characteristics of the GO Term.

Searching the Chinese hamster mitochondrial genome

1) Select the Chinese hamster mitochondrial genome from the **Genomes** menu [*i.e.* **CH Mitochondria**].

2) Type a keyword into the search box at the top of the page.
   Typing **%** in the search box will list all 13 entries currently in the database.
   The Chinese hamster mitochondrial genome database can only be searched by **Gene symbol** [*i.e.* **COX**].

Searching for "**COX**" will return 3 entries from the database displayed in a tabular format.

General information for the returned entries include the gene symbol, NCBI Entrez Gene ID, gene name, and genomic coordinates (start and end positions).

To find more details about a single entry, click on a gene symbol.

3) Clicking on **COX1** will open the **Gene Details** page.

The **General Information** section
contains the gene name, symbol, ID,
and description (if available).

The **Genomic Information** section
provides the genomic coordinates in
the mitochondrial genome and Entrez
protein ID.

The **Sequence Information** section
provides links to download the
nucleotide and protein sequence in
FASTA format.

The **Homologs Information** section
provides links to homologous proteins
in other species in the NCBI and
Ensembl databases.

The **External Links** section provides
links to CHO gene and protein entries
in the NCBI databases.

Chinese Hamster Mitochondria Details

General Information

Gene Symbol	COX1
Name	cytochrome c oxidase subunit I [Cricetulus griseus]
Gene ID	3979185
Description	(temporary blank)

Genomic Location

Genomic Accession (start position)	5309
Genomic Accession (end position)	6853
Map	YP_537121.1

Sequence Information

Nucleotide Sequence	Download sequence
Protein Sequence	Download sequence

Homologs Information

	Human	Mouse	Rat
Ensembl Gene	ENSG00000198804	ENSMUSG00000064351	ENSRNOG00000034234
Entrez Gene	4512	17708	26195

External Links

Entrez Gene	3979185
Entrez Protein	YP_537121.1

Annotation

	GO:0016020 membrane
	GO:0070469 respiratory chain
	GO:0005743 mitochondrial inner membrane
	GO:0016021 integral to membrane
	GO:0005739 mitochondrion
	GO:0046872 metal ion binding
	GO:0004129 cytochrome-c oxidase activity
GO Terms	GO:0020037 heme binding
	GO:0009055 electron carrier activity
	GO:0005506 iron ion binding
	GO:0016491 oxidoreductase activity

The **Annotation** section reports GO terms and provides links to the AmiGO gene ontology browser.

4) Hyperlinked text also serves as a link to relevant informational webpages.

The **Genomic Location** section **Map** entry [*i.e.* **YP_537121.1**] links to the NCBI Map Viewer to view the selected gene's annotation within the mitochondrial genome.
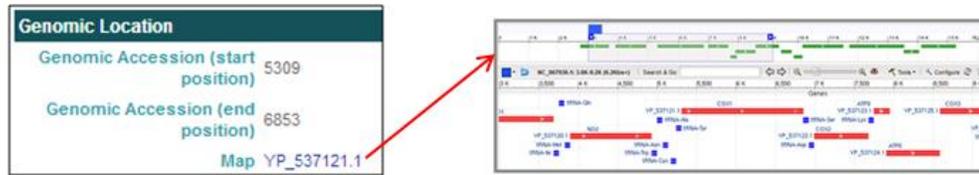
The **Sequence Information** section provides **Nucleotide Sequence** and **Protein Sequence** links [*i.e.* **Download sequence**] to download the nucleotide and protein sequences in FASTA format.

The **Homologs Information** and **External Links** sections provide links to CHO proteins and homologs.

- The **Homologs Information** section provides **Ensembl Gene** and **Entrez Gene** links [*i.e.* **ENSRNOG000000034234**] to the Ensembl and NCBI Entrez databases for CHO proteins and homologs in other species, including human, mouse, and rat.
- The **External Links** section provides **Entrez Gene** and **Entrez Protein** links [*i.e.* **3979185**] to CHO gene and protein pages in the NCBI Entrez database.