

# ChromDB.org HELP

## I. Search Functions

Genes, proteins, organisms and information in ChromDB can be accessed in the following ways:

### A. Quick Search text box options:

#### 1. Gene name or synonym (alias):

Use the Quick Search text box as a rapid way to access a particular gene of interest or to determine if a gene is included in the database. This search function allows you to type in an official gene symbol, a synonym (alias), or the ChromDB Identifier (ID) for a particular gene or protein. The same ChromDB ID is used for both the gene and protein; gene and protein are used interchangeably in this document.

#### Example:

The ChromDB ID for the *Arabidopsis thaliana* DDM1 gene or protein is CHR1; either name will display the DDM1/CHR1 gene record page.

**Note:** If you enter an official gene symbol that is used for multiple organisms, you will be presented with a list of ChromDB IDs, each of which has that name as the official name or an alias.

**Example:** AGO1 typed into the Quick Search text box will return:

ChromDB ID = AGO1 for the *Arabidopsis* AGO1 gene/protein  
ChromDB ID = AGO401 for the *Drosophila* AGO1 gene/protein  
ChromDB ID = AGO601 for the *Schizosaccharomyces* AGO1 gene/protein

#### 2. Locus:

*Arabidopsis thaliana*, *Oryza sativa ssp. japonica* (japonica rice), *Drosophila melanogaster*, *Saccharomyces cerevisiae*, and *Schizosaccharomyces pombe* genes can be accessed by entering the proper locus designation in the "Quick Search" text box. The text box will accept only one locus at a time. Viewing of multiple loci can be accomplished using the Loci link listed under the "Reports" section found in the style sheet section on the left side of ChromDB web pages.

#### 3. Organism:

Database users can type in an organism, either the formal or common name, and retrieve an output containing three columns, i.e. the formal name and the common name of the organism, as well as a list that shows the number of proteins in the database which are associated with that organism. The formal name is a link to the appropriate taxonomy page at NCBI, and the number is a link for displaying a list of genes in the database.

This search function is flexible in that a query will bring up an exact match to a whole name, as well as any organism name that includes the string of letters used as a query. This search function is a way to retrieve the formal name of the organism, which is necessary for some of the information retrieval tools.

**Note:** When using the protein number link to form a list of proteins, please be patient as the retrieval process time depends on the number of proteins to be displayed.

## B. Advanced Search for Information Retrieval:

The Advance Search feature is a useful way to search for and retrieve information regarding ChromDB genes. This feature can be accessed from the left side style sheet or from the link next to the Quick Search Text Box. There are three different options, or entry points, associated with the Advanced Search feature.

### 1. Starting the process

#### a. Option 1: Search by a list of gene/protein names.

The first sentence on the Advanced Search page, "Select both the taxon and protein classification or [Enter gene names](#)", contains the Option 1 link that brings up a new page with a text box and three types of information retrieval: Protein Description, Synonyms, and Protein Domain Viewer.

The text box accepts multiple entries as ChromDB IDs or loci (use the Synonyms link under the Reports section on the left side style sheet to see which organisms have loci associated with genes in the database).

**Note:** Because a ChromDB ID has the protein group and organism encoded in the ID (see Nomenclature Link for more information) and a locus is specific for an organism and gene/protein; this option does not require selections for the organism or a protein group (Steps 2 and 3, respectively, which appear on the opening page of the Advanced Search Utility).

**Note:** When using Option1, the "[Make other selections](#)" link at the top of the new page returns you to the original menu.

#### b. Option 2: Customize your organism selection from a list

Use the opening page selection to access a single organism or to customize a combination of organisms (hold the control key while making multiple selections). It is possible to jump within the text box to a selection by typing in a name.

**Note:** This search option requires the use of formal names for organisms. If you do not know the formal names, use the Quick Search text box to enter a common name in order to retrieve the formal name.

#### c. Option 3: Search by taxon criteria using pre-set groups of organisms

An alternative to selecting a group of organisms from the list is to "[Select Taxon](#)" link in Step one. This method provides the means to carry out a comparative analysis of a group or groups of organisms. The taxon groups are pre-set and cannot be altered. Each taxon name is a link to view the organisms within that group. The rest of the page continues the menu selection for Step 2 which selects a protein group (see Nomenclature for link to protein group reference table) and Step 3 for data selection; Protein Description, Synonyms, and Protein Domain.

**Note:** The speed of the data retrieval is decreased with increased numbers of selections. Please be patient.

### 2. Step 2: Selection of protein classifications.

Over 90 different protein groups exist in the database. The options listed below provide access to a functional-type of classification or proteins group and an alphabetical listing of the groups.

### **a. Option 1: Selection using functional groupings**

The menu selection in Step 2 presents a higher order grouping of the ChromDB protein groups. To understand the groupings, open the plus signs see the individual groups; the final enlargement of each group has the name and a short description of each ChromDB protein group.

### **b. Option 2: Alphabetical listing of protein groups**

“[Select a protein group](#)” in Step 2 is a link that brings up a new page with an alphabetical listing of all protein groups in the database. Each protein group includes a three to five letter abbreviation followed by a short description of that group. To make multiple selections, please hold down the control key while clicking on groups. Within the text box it is possible to jump to an abbreviation by typing in the name.

**Note:** Multiple protein groups can be made with either option, but increasing numbers of groups will slow the retrieval process. Please be patient.

## **3. Step 3: Select the type of reports or displays**

### **Protein Description:**

This selection displays all proteins classified into the selected group(s). The output includes the ChromDB IDs and a short description of each protein.

### **Synonyms:**

This selection displays formal names and aliases for the organisms and protein groups selected.

### **Pfam and/or SMART Domain View:**

The last two selections are for viewing either or both Pfam or SMART domains for the selected organisms and protein groups. An additional link for the protein domain viewer is included on the side tool bar.

**Note:** The speed of viewing a final report decreases with increasing the number of selections; please be patient.

## **C. BLAST ChromDB:**

This link brings up local BLAST tools for searching the ChromDB database. BLAST provides a means of determining if your gene or protein is included in the database or to find similar sequences.

Comprehensive help on BLAST is available at NCBI (<http://www.ncbi.nlm.nih.gov/>) and is not duplicated here.

# **II. Report Utilities**

## **A. ChromDB Contents:**

This utility provides the means to generate a list of genes/proteins in the database.

### **1. Option 1: View all proteins**

The number in parentheses after “All Proteins” indicates the total number available for viewing and is a link to generate a list of database contents. The “All” selection can take minutes to load. The output includes the ChromDB IDs and formal names and provides a means for filtering the output by groups of organisms.

**Note:** The speed of the retrieval process can be hastened by selecting subgroups or individual protein groups using Option 2.

## 2. Option 2: View proteins of functionally related proteins

If there is a plus sign to the left of "All Proteins", click on the sign to open the functional groups. To understand the groupings, click on the plus signs to see the contents of each group; the final enlargement of a group contains the name and a short description of a ChromDB protein group. The number at the end of each subgroup shows the number of genes/proteins in this group and is a link to generate an output. Please click on the plus signs to view individual protein groups comprising each general classification.

## 3. Organism filter:

The top of the output page provides the means to filter the results by groups of organism. The organism comprising each group can be viewed by clicking on the taxon group.

## B. Sequence Tools: Form FASTA Files

This utility generates DNA or protein sequences in a FASTA file format. Two functions or options are provided, one called "Sequence" and one called "Domain". The sequence function generates files containing the entire sequence (DNA or protein); the domain function limits the amino acids to a protein domain.

### 1. FASTA file format for protein, transcript, or genomic sequences.

The opening menu selection is the same format as the Advanced Search function.

#### 1. Starting the process

##### a. Option 1: Search by a list of gene/protein names.

The first sentence on the Advanced Search page, "Select both the taxon and protein classification or [Enter gene names](#)", contains the Option 1 link that brings up a new page with a text box and three types of information retrieval: Protein Description, Synonyms, and Protein Domain Viewer.

The text box accepts multiple entries as ChromDB IDs or loci (use the Synonyms link under the Reports section on the left side style sheet to see which organisms have loci associated with genes in the database).

**Note:** Because a ChromDB ID has the protein group and organism encoded in the ID (see Nomenclature for more information) and a locus is specific for an organism and gene/protein; this option does not require selections for the organism or a protein group (Steps 2 and 3, respectively, which appear on the opening page of the Advanced Search Utility).

**Note:** When using Option1, the "[Make other selections](#)" link at the top of the new page returns you to the original menu.

##### b. Option 2: Customize your organism selection from a list

Use the opening page selection to access a single organism or to customize a combination of organisms (hold the control key while making multiple selections). It is possible to jump within the text box to a selection by typing in a name.

**Note:** This search option requires the use of formal names for organisms. If you do not know the formal names, use the Quick Search text box to enter a common name in order to retrieve the formal name.

### **c. Option 3: Search by taxon criteria using pre-set groups of organisms**

An alternative to selecting a group of organisms from the list is to “[Select Taxon](#)” link in Step one. This method provides the means to carry out a comparative analysis of a group or groups of organisms. The taxon groups are pre-set and cannot be altered. Each taxon name is a link to view the organisms within that group. The rest of the page continues the menu selection for Step 2 which selects a protein group (explained below) and Step 3 for data selection; Protein Description, Synonyms, and Protein Domain Viewer.

**Note:** The speed of the data retrieval is decreased with increased numbers of selections. Please be patient.

## **2. Step 2: Selection of protein classifications.**

Over 90 different protein groups exist in the database. The options listed below provide access to a functional-type of classification or proteins group and an alphabetical listing of the groups.

### **a. Option 1: Selection using functional groupings**

The menu selection in Step 2 presents a higher order grouping of the ChromDB protein groups. To understand the groupings, open the plus signs see the individual groups; the final enlargement of each group has the name and a short description of each ChromDB protein group.

### **b. Option 2: Alphabetical listing of protein groups**

“[Select a protein group](#)” in Step 2 is a link that brings up a new page with an alphabetical listing of all protein groups in the database. Each protein group includes a three to five letter abbreviation followed by a short description of that group. To make multiple selections, please hold down the control key while clicking on groups. Within the text box it is possible to jump to an abbreviation by typing in the name.

**Note:** Multiple protein groups can be made with either option, but increasing numbers of groups will slow the retrieval process. Please be patient.

## **3. Select the sequence type**

Select the type of sequence, either protein, transcript or genomic, and click on the “Form FASTA file” to complete the process.

## **2. FASTA file format for protein domains.**

This utility is the same as the “Sequence” function except Step 2 consists of two drop-down menus to select the name of a protein domain and the database (Pfam or SMART). Users can select only one domain at a time with this function. It is possible to jump to a selection by beginning to type the name. The “Optional” box allows users to include additional amino acids around a domain. Finally, click on the “Form FASTA file” to complete the process.

## **C. Plant Genomes**

Plant proteins are the major focus ChromDB. In order to highlight sequenced plant genomes (including those in progress), we have placed links on the left side style sheet to provide easy access to these organisms. Each link provides botanical information, genome sequencing information, links to genome and organism databases, and other useful links.

## D. Genomic-based Reports

### 1. Genomic-based versus transcript-based organisms

The database contains two types of sequences, genomic- and transcript-based. Genomic-based sequences are limited to plant genomes or to genomes which are annotated by the ChromDB staff for the DOE Joint Genome Institute (<http://www.jgi.doe.gov/>). Although genomic sequences are available for other genomes, *i.e.* *Homo sapiens* and *Drosophila melanogaster*, these important model organisms are available as transcript-based only at ChromDB. We depend on sequences from the NCBI Reference Sequence (RefSeq) collection (<http://www.ncbi.nlm.nih.gov/RefSeq/>). ChromDB users are cautioned that these RefSeq entries have varying levels of curation.

ChromDB does not display whole chromosome sequences for genomic-based organisms. Genomic sequences are limited to a span of nucleotides which is sufficient to contain the predicted transcript splice model and 5' and 3' untranslated regions (UTR).

For ChromDB purposes, GBrowse is a gene browser not a genome browser. The GBrowse thumbnail view on each Gene Record Page is a link to bring up a full GBrowse view. Transcript-based organisms have a GBrowse view that is limited to the transcript and protein domains.

### 2. Genome-based tools

This report generator uses some of the same features used for the Advanced Search utility.

#### 1. Starting the process; choose the organism(s)

##### a. Option 1: Manual entry of gene names

“[Enter gene names](#)” is a link that brings up a new page with a text box for manual entry of ChromDB IDs or loci (if available).

**Note:** Because a ChromDB ID has the protein group and organism encoded in the ID (see Nomenclature for more information); this selection does not use Step 2 which appears on the opening page of “Genomic-based Reports”.

##### b. Option 2: Selection of organisms from a list

All current genomic-based organisms are listed in the box in Step 1 of the opening page of “Genomic-based Reports”. Use this selection to access a single organism or to select multiple organisms (hold the control key while making multiple selections). Within the text box it is possible to jump to a selection by beginning to type in a name.

**Note:** The speed of the report generator is decreased with increased numbers of selections. Please be patient.

#### 2. Selection of protein classes (Step 2).

##### a. Option 1: Functional classification of protein groups

To understand the groupings, open the plus signs see the individual groups; the final enlargement of each group has the name and a short description of each ChromDB protein group.

##### b. Option 2: Alphabetical listing of protein groups in the database

Each protein group includes a three to five letter abbreviation followed by a short description of the group. To make multiple selections, please hold down the control key while clicking on groups. Within the text box it is possible to jump to an abbreviation by beginning to type in the name.

### 3. Select the type of reports or displays.

#### a. Protein Description:

This selection displays all proteins classified into that group, as ChromDB IDs and a short description of the protein group.

#### b. Synonyms:

This selection displays formal names and aliases for the organisms and protein groups selected.

#### c. Locus:

This tool prepares a list of ChromDB IDs and loci for each gene, if available.

#### d and e. Protein Domain View for Pfam and SMART domains:

These two separate selections display Pfam or SMART domains views for the selected organism(s) and protein group(s).

#### f. Number of exons.

This genomic-based specific tool will generate the number of exons comprising a transcript splice model.

#### g. Exon Viewer.

This genomic-based specific tool will generate a graphical view of alternating exons and introns comprising a transcript splice model. When used in conjunction with the “Number of Exons” selection, a number plus the graphic will appear for each gene selected.

#### h. Splice Model Status.

This genomic-based specific tool shows the extent of biological data for a transcript splice model. There are three classes of support:

- i. Transcript splice model confirmed by cDNAs
- ii. Transcript splice model partially confirmed by cDNAs
- iii. No cDNAs available

## E. Protein Information

There are three different options, or entry points, associated with this feature.

### 1. Starting the process

#### a. Option 1: Search by a list of gene/protein names.

The first sentence on the Advanced Search page, “Select both the taxon and protein classification or [Enter gene names](#)”, contains the Option 1 link that brings up a new page with a text box and three types of information retrieval: Protein Description, Synonyms, and Protein Domain Viewer.

The text box accepts multiple entries as ChromDB IDs or loci (use the Synonyms link under the Reports section on the left side style sheet to see which organisms have loci associated with genes in the database).

**Note:** Because a ChromDB ID has the protein group and organism encoded in the ID (see Nomenclature for more information) and a locus is specific for an organism and gene/protein; this option does not require selections for the organism or a protein group

(Steps 2 and 3, respectively, which appear on the opening page of the Advanced Search Utility).

**Note:** When using Option1, the “[Make other selections](#)” link at the top of the new page returns you to the original menu.

#### **b. Option 2: Customize your organism selection from a list**

Use the opening page selection to access a single organism or to customize a combination of organisms (hold the control key while making multiple selections). It is possible to jump within the text box to a selection by typing in a name.

**Note:** This search option requires the use of formal names for organisms. If you do not know the formal names, use the Quick Search text box to enter a common name in order to retrieve the formal name.

#### **c. Option 3: Search by taxon criteria using pre-set groups of organisms**

An alternative to selecting a group of organisms from the list is to “[Select Taxon](#)” link in Step one. This method provides the means to carry out a comparative analysis of a group or groups of organisms. The taxon groups are pre-set and cannot be altered. Each taxon name is a link to view the organisms within that group. The rest of the page continues the menu selection for Step 2 which selects a protein group (explained below) and Step 3 for data selection; Protein Description, Synonyms, and Protein Domain Viewer (these subjects will be explained below; click on a subject to jump to any of the links).

### **2. Step 2: Selection of protein classifications.**

Over 90 different protein groups exist in the database.

#### **a. Option 1: Selection using functional groupings**

The menu selection in Step 2 presents a higher order grouping of the ChromDB protein groups. To understand the groupings, open the plus signs see the individual groups; the final enlargement of each group has the name and a short description of each ChromDB protein group.

**Note:** Multiple protein groups can be made, but increasing numbers of groups will slow the retrieval process. Please be patient.

**Note:** The “[Select a protein group](#)” link in Step 2 brings up an alternative protein group selection described below in number 5.

#### **b. Option 2: Alphabetical listing of protein groups**

“[Select a protein group](#)” in Step 2 is a link that brings up a new page with an alphabetical listing of all protein groups in the database. Each protein group includes a three to five letter abbreviation followed by a short description of that group. To make multiple selections, please hold down the control key while clicking on groups. Within the text box it is possible to jump to an abbreviation by typing in the name.

### **3. Step 3: Select the type of reports or displays**

**Protein Status:** Transcript-based proteins can be partial, if sufficient cDNA information is lacking to construct a putative, full-length protein. In some cases gaps in genomic sequences will result in incomplete predicted transcripts. This utility generates a list on the completeness of predicted proteins

**Protein Length:** This utility presents information on the number of amino acids comprising the selected proteins.

**Pfam or SMART domains:** This utility generates a list (not a view) showing the Pfam and/or SMART domains within the selected proteins.

**Pfam or SMART e-values:** This utility generates a list showing the e-values of Pfam and/or SMART domains within database proteins.

## F. Gene Counts

This utility brings up an alphabetical listing of all organisms by formal name and provides the total number of genes/proteins in the database.

## G. Loci

This utility displays loci for genes in the database. The selection box in Step 1 shows the organisms for which this tool is applicable. Step 2 provides the means for selecting all proteins or a functional group of proteins. Alternatively, the [Enter gene names](#) link provides a text box to enter gene names as ChromDB IDs to retrieve a list of associated loci. Note: as the ChromDB ID includes the protein group in the name, no selection of protein group is necessary for this option.

## F. Synonyms

This utility displays synonyms for database genes. There are three different options, or entry points, associated with this feature.

### 1. Starting the process

#### a. Option 1: Search by a list of gene/protein names.

The first sentence on the Advanced Search page, "Select both the taxon and protein classification or [Enter gene names](#)", contains the Option 1 link that brings up a new page with a text box and three types of information retrieval: Protein Description, Synonyms, and Protein Domain Viewer.

The text box accepts multiple entries as ChromDB IDs or loci (use the Synonyms link under the Reports section on the left side style sheet to see which organisms have loci associated with genes in the database).

**Note:** Because a ChromDB ID has the protein group and organism encoded in the ID (see Nomenclature for more information) and a locus is specific for an organism and gene/protein; this option does not require selections for the organism or a protein group (Steps 2 and 3, respectively, which appear on the opening page of the Advanced Search Utility).

**Note:** When using Option1, the "[Make other selections](#)" link at the top of the new page returns you to the original menu.

#### b. Option 2: Customize your organism selection from a list

Use the opening page selection to access a single organism or to customize a combination of organisms (hold the control key while making multiple selections). It is possible to jump within the text box to a selection by typing in a name.

**Note:** This search option requires the use of formal names for organisms. If you do not know the formal names, use the Quick Search text box to enter a common name in order to retrieve the formal name.

#### c. Option 3: Search by taxon criteria using pre-set groups of organisms

An alternative to selecting a group of organisms from the list is to "[Select Taxon](#)" link in Step one. This method provides the means to carry out a comparative analysis of a group

or groups of organisms. The taxon groups are pre-set and cannot be altered. Each taxon name is a link to view the organisms within that group. The rest of the page continues the menu selection for Step 2 which selects a protein group (explained below) and Step 3 for data selection; Protein Description, Synonyms, and Protein Domain Viewer.

## 2. Step 2: Selection of protein classifications.

Over 90 different protein groups exist in the database. The options listed below provide access to a functional-type of classification or proteins group and an alphabetical listing of the groups.

### a. Option 1: Selection using functional groupings

The menu selection in Step 2 presents a higher order grouping of the ChromDB protein groups. To understand the groupings, open the plus signs see the individual groups; the final enlargement of each group has the name and a short description of each ChromDB protein group.

### b. Option 2: Alphabetical listing of protein groups

“[Select a protein group](#)” in Step 2 is a link that brings up a new page with an alphabetical listing of all protein groups in the database. Each protein group includes a three to five letter abbreviation followed by a short description of that group. To make multiple selections, please hold down the control key while clicking on groups. Within the text box it is possible to jump to an abbreviation by typing in the name.

**Note:** Multiple protein groups can be made with either option, but increasing numbers of groups will slow the retrieval process. Please be patient.

Press “[go](#)” to display the results.

## III. Viewers

Two viewers are provided that show the transcript splice model for genomic sequences (Exon Viewer) and the position of protein domains for both genomic-based and transcript-based sequences (Pfam Domain Viewer and SMART Domain Viewer).

### A. Exon Viewer:

This viewer is applicable for genomic-based organisms only.

#### Step 1. Starting the process; choose the organism(s)

##### a. Option 1: Manual entry of gene names

“[Enter gene names](#)” is a link that brings up a new page with a text box for manual entry of ChromDB IDs or loci (if available).

**Note:** Because a ChromDB ID has the protein group and organism encoded in the ID (see Nomenclature Link for more information); this selection does not use Step 2 which appears on the opening page of “Genomic-based Reports”.

##### b. Option 2: Selection of organisms from a list

All current genomic-based organisms are listed in the box in Step 1 of the opening page of “Genomic-based Reports”. Use this selection to access a single organism or to select multiple organisms (hold the control key while making multiple selections). Within the text box it is possible to jump to a selection by beginning to type in a name.

**Note:** The speed of the report generator is decreased with increased numbers of selections. Please be patient.

## Step 2. Selection of protein classes

### a. Option 1: Functional classification of protein groups

To understand the groupings, open the plus signs see the individual groups; the final enlargement of each group has the name and a short description of each ChromDB protein group.

### b. Option 2: Alphabetical listing of protein groups

Each protein group includes a three to five letter abbreviation followed by a short description of the group. To make multiple selections, please hold down the control key while clicking on groups. Within the text box it is possible to jump to an abbreviation by beginning to type in the name.

Press “go” to view the selections

## B. Protein Domain Viewer:

This viewer will generate a view of the placement of either Pfam or SMART domains within a predicted protein sequence. There are three different options, or entry points, associated with this feature.

### 1. Starting the process

#### a. Option 1: Search by a list of gene/protein names.

The first sentence on the Advanced Search page, “Select both the taxon and protein classification or [Enter gene names](#)”, contains the Option 1 link that brings up a new page with a text box and three types of information retrieval: Protein Description, Synonyms, and Protein Domain Viewer (these choices will be explained below; click on a subject to jump forward).

The text box accepts multiple entries as ChromDB IDs or loci (use the Synonyms link under the Reports section on the left side style sheet to see which organisms have loci associated with genes in the database).

**Note:** Because a ChromDB ID has the protein group and organism encoded in the ID (see Nomenclature Link for more information) and a locus is specific for an organism and gene/protein; this option does not require selections for the organism or a protein group (Steps 2 and 3, respectively, which appear on the opening page of the Advanced Search Utility).

**Note:** When using Option1, the “[Make other selections](#)” link at the top of the new page returns you to the original menu.

#### b. Option 2: Customize your organism selection from a list

Use the opening page selection to access a single organism or to customize a combination of organisms (hold the control key while making multiple selections). It is possible to jump within the text box to a selection by typing in a name.

**Note:** This search option requires the use of formal names for organisms. If you do not know the formal names, use the Quick Search text box to enter a common name in order to retrieve the formal name.

### c. Option 3: Search by taxon criteria using pre-set groups of organisms

An alternative to selecting a group of organisms from the list is to “[Select Taxon](#)” link in Step one. This method provides the means to carry out a comparative analysis of a group or groups of organisms. The taxon groups are pre-set and cannot be altered. Each taxon name is a link to view the organisms within that group. The rest of the page continues the menu selection for Step 2 which selects a protein group (explained below) and Step 3 for data selection; Protein Description, Synonyms, and Protein Domain Viewer (these subjects will be explained below; click on a subject to jump to any of the links).

## 2. Step 2: Select a protein group or a domain

### a. Option 1: Select a protein group from a list

This option provides an alphabetical listing of protein groups. Multiple choices can be made by holding down the control key.

### b. Option 2: View by domain

[Select Domains](#) in Step 2 is a link to generate a report by domain, either Pfam or SMART, rather than by protein group.

## IV. ChromDB Nomenclature

All sequences entered into the database are given a ChromDB ID (identifier). The same ID is used to denote both a transcript and a protein. Formal names, as well as alias, are presented on individual gene record pages and can be entered into the “Quick Search” text box to access genes. ChromDB IDs are composed with of a three- to five-letter symbol followed by a number, e.g., BRD4 or BRD106.

If you do not know the ChromDB ID, you can use the “Quick Search” text box at the top right of each webpage as an “ID converter”.

### Letter symbols:

Protein groups unified by homologous features are represented by different letter.

### Numerals:

Because it is often not possible to determine orthologous relationships unambiguously, especially as IDs are assigned to chromatin genes in newly sequenced genomes, ChromDB IDs do NOT imply orthology. Each organism has a unique number series, Arabidopsis genes are indicated by the numerals 1-99, maize by 101-199, and japonica rice by 701-799, etc.

### Prefixes:

Prefixes, such At for *Arabidopsis thaliana* and Zm for *Zea mays*, would be redundant in this ID system, because species are indicated by the numeric series. However, we may institute an additional name in the future that implies orthology.

### Synonyms:

Previously published gene designations and formal names are presented in ChromDB as synonyms or alias. Note that the designation with priority in literature should be used in publications. ChromDB IDs may be used for genes not previously named in the literature.

All sequences entered into the database are given a ChromDB ID (identifier). The same ID is used to denote both a transcript and a protein. Formal names, as well as alias, are presented on individual gene record pages and can be entered into the “Quick Search” text box to access genes. ChromDB IDs are composed with of a three- to five-letter symbol followed by a number, e.g., BRD4 or BRD106.

## V. Gene Record Page

The Gene Record Page is the central navigation portal for accessing information relating to each database gene.

### A. Database Sequences:

The database contains two types of sequences, genomic- and transcript-based. Genomic-based sequences are limited to plant genomes or to genomes which are annotated by the ChromDB staff for the DOE Joint Genome Institute (<http://www.jgi.doe.gov/>). Although genomic sequences are available for other genomes, *i.e.* *Homo sapiens* and *Drosophila melanogaster*, these are important model organisms are available as transcript-based only at ChromDB. ChromDB staff members do not curate non-plant sequences; we depend on sequences from the NCBI Reference Sequence (RefSeq) collection (<http://www.ncbi.nlm.nih.gov/RefSeq/>). ChromDB users are cautioned that these RefSeq entries have varying levels of curation.

ChromDB does not display whole chromosome sequences for genomic-based organisms, *i.e.* *Arabidopsis thaliana*. Genomic sequences are limited to the span of nucleotides which is sufficient to contain the predicted transcript splice model and 5' and 3' untranslated sequences as defined by cDNAs.

### B. Gene Record Page Contents:

#### Formal Name and ChromDB ID:

All sequences entered into the database are given a ChromDB ID (identifier). The same ID is used to denote both a transcript and a protein. An explanation of our nomenclature system for IDs can be viewed in the help section for ChromDB Nomenclature. In cases where ChromDB genes have a pre-assigned or formal gene symbol, these are listed in as Formal Names. The Formal Name precedes the ChromDB ID on the gene record page.

#### Aliases:

Additional assigned gene names, other than the formal or published gene symbol, are listed as Aliases on the Gene Record Page.

#### Entrez GeneID:

The NCBI Entrez Gene ID is displayed on each Gene Record Page, if it is available. The ID is a link that takes users out of the ChromDB database to the NCBI Entrez Gene page. Please visit this resource as ChromDB does not try to duplicate data and information when users can be directed to another database.

#### Taxonomy Link:

This entry shows the formal name of the organism; the name is a link to the appropriate NCBI taxonomy page. Common names for organisms are shown in parentheses.

#### ChromDB Taxon:

Organisms are grouped into convenient taxon groups to facilitate comparative analyses among the different organisms in the database. Database users can see the overall classification scheme by using the "Advanced Search" feature. The first page of this

search shows the groups and subgroups, each of which is a link to display the organisms within each group.

**Protein Group:**

ChromDB proteins are placed into protein groups which are unified by similarity and/or homology. Each protein group is designated by a three to five letter abbreviation. This protein classification scheme can be viewed with this link: [protein classification](#)

**Description:**

A short description of the protein group is provided here.

**ChromDB Model Type:**

ChromDB genes are either genomic-based or transcript-based; this entry shows which category the organism fits into. More information on this subject is included in the Database Sequences help section.

**Transcript View:**

A thumbnail view of the transcript is shown here as a GBrowse view. For genomic-based organisms, this thumbnail shows the transcript splice model and the placement of Pfam domains with reference to the translation of each exon. The thumbnail of a transcript-based sequence shows a transcript with introns removed.

**Splice Model Status:**

This entry pertains to genomic-based sequences only and shows the amount of biological proof (cDNA sequences) that are available to support a predicted transcript splice model. The categories are:

- Confirmed by cDNAs (including ESTs)
- Partially confirmed by cDNAs (including ESTs)
- No biological support available

**Expression:**

*Arabidopsis thaliana* and *japonica* rice have links to the MPSS website.

## VI. RNAi

### A. General Information

ChromDB was established through funding from the Plant Genome Research Program of the National Science Foundation (NSF award #9975930; Functional Genomics of Chromatin: Global Control of Plant Gene Expression; PI: R.A. Jorgensen). A broad goal of this project was to identify and functionally characterize the complement of genes in maize and *Arabidopsis* that contribute to chromatin-based control of gene expression. A more narrow focus of the project was to produce RNAi lines for a set of *Arabidopsis* and maize chromatin-associated genes.

With renewed funding from the Plant Genome Research Program (NSF award #DBI-0421679; ChromDB: Integrating Information on Plant Chromatin Proteins and Complexes; PI: C.A. Napoli). The grant does not fund additional RNAi line work; although the Maize Chromatin Consortium is continuing this work for maize.

As a service to the community and a goal of the previous funding, ChromDB will continue to display RNAi information generated from the previous grant. Please note that ChromDB staff members do not distribute the *Arabidopsis* RNAi lines. These are distributed through ABRC/TAIR. The maize lines are distributed through the Maize Stock Center and information on the lines,

including ordering information, is available at MaizeGDB (<http://www.maizegdb.org/>). Please use the Arabidopsis and maize links shown below to access order information.

## B. Vector Information

This document describes a set of binary vectors for producing dominant negative RNAi mutants using a target sequence cloning strategy that is based on the inclusion of two restriction enzyme cleavage sites in each of two primers used to amplify gene-specific fragments from cDNA. This design minimizes the number of PCR primers and results in the placement of unique restriction enzyme recognition sites to allow for flexibility in future manipulations of the plasmid, *e.g.*, moving the inverted repeat target sequence to a different vector. ChromDB's RNAi vectors are based on pCAMBIA binary vectors, a set of plasmids developed by the Center for Application of Molecular Biology to International Agriculture (CAMBIA). Importantly, CAMBIA vectors contain two origins of replication, a wide-host-range origin for plasmid replication in *Agrobacterium tumefaciens* and the pBR322 origin for replication in *Escherichia coli*, following the design of the Maliga lab (Plant Mol Biol 1994; 25(6):989-94). Detailed descriptions of pCAMBIA vectors, including conditions for use and distribution, can be obtained at the [CAMBIA web site](#).

Please do not request plasmids from ChromDB. The RNAi vectors are distributed by ABRC/TAIR. [Table 1](#) (and "Order Vectors" link) provides a list of available vectors and includes the ABRC stock number (direct link to the vector page at the TAIR database), brief information about each vector and links to bring up plasmid maps and FASTA sequence files. For those researchers using [Gene Construction Kit](#) software, a GCK file for each sequence is included as a link

## C. Order Vectors.

The RNAi vectors have been deposited at ABRC; they are not available through the ChromDB database. Please use the Order Vectors link for instructions on how to submit an order to ABRC.

## D. RNAi Lines

As a service to the community and a goal of the previous funding, ChromDB will continue to display RNAi information generated from the previous grant. But, please note that ChromDB staff members do not distribute the Arabidopsis RNAi lines. These are distributed through ABRC/TAIR. The maize lines are distributed through the Maize Stock Center and information on the lines, including ordering information, is available at MaizeGDB (<http://www.maizegdb.org/>). Please use the Arabidopsis and maize links shown below to access order information.

The RNAi Lines link brings up two further links, one to Arabidopsis lines and one to Maize line.

## VII. ChromDB Mission

The broad mission of ChromDB is display, annotate, and curate sequences of two broad functional classes of biologically important proteins: chromatin-associated proteins (CAPs) and RNA interference-associated proteins. Plant proteins are the major focus of the work support by The Plant Genome Research Program (PGRP) of the National Science Foundation. Our intent is to produce intensively curated sequence information and make it available to the research and teaching community in support of comparative analyses toward understanding the chromatin proteome in plants, especially in important crop species. In order to take do a comparative analysis, it is necessary to include non-plant proteins in the database. Non-plant genes are not curated to the degree carried out for plants and to automate the process of data import, our non-plant genes are from the RefSeq database of NCBI. We reason that the inclusion of non-plant, model organisms will broaden the relevance and usefulness of ChromDB to the entire chromatin community and will provide a more complete data set for phylogenetic analyses in support of the evolution of the plant chromatin proteome.

## **VIII. Acknowledgements**

ChromDB is funded by a grant from the National Science Foundation Plant Genome Research Project (#DBI-0421679). The ChromDB server is hosted by the Biotechnology Computing Facility at the University of Arizona.