

ChromDB: The Chromatin Database

Karla Gendler, Tara Paulsen and Carolyn Napoli*

BIO5 Institute, University of Arizona, Tucson, AZ 85719, USA

Received August 16, 2007; Revised September 4, 2007; Accepted September 11, 2007

ABSTRACT

The ChromDB website (<http://www.chromdb.org>) displays chromatin-associated proteins, including RNAi-associated proteins, for a broad range of organisms. Our primary focus is to display sets of highly curated plant genes predicted to encode proteins associated with chromatin remodeling. Our intent is to make this intensively curated sequence information available to the research and teaching communities in support of comparative analyses toward understanding the chromatin proteome in plants, especially in important crop species such as corn and rice. Model animal and fungal proteins are included in the database to facilitate a complete, comparative analysis of the chromatin proteome and to make the database applicable to all chromatin researchers and educators. Chromatin biology and chromatin remodeling are complex processes involving a multitude of proteins that regulate the dynamic changes in chromatin structure which either repress or activate transcription. We strive to organize ChromDB data in a straightforward and comparative manner to help users understand the complement of proteins involved in packaging DNA into chromatin.

INTRODUCTION

Eukaryotic nuclear DNA, along with a variety of proteins, is organized and packaged into a complex structure called chromatin. The basic, repeating unit of chromatin is the nucleosome which consists of the DNA duplex wound approximately twice around an octamer consisting of two copies each of the four core histones H2A, H2B, H3 and H4. Nucleosomes undergo further compaction to form the distinct structures seen as chromosomes (1). The function of chromatin is to maintain a restrictive ground state wherein the DNA is inaccessible to the transcriptional machinery but is accessible to protein complexes capable of remodeling chromatin locally to allow transcription initiation. Additionally, chromatin has a pivotal role in a number

of DNA-associated processes, e.g. replication (2), repair (3,4), kinetochore and centromere formation (5). The covalent modification of histones is extremely important in nucleosome and chromatin dynamics. Histones are ideally suited for this role, as both the histone tails and globular domains are subjected to a variety of posttranslational modifications such as acetylation, methylation, phosphorylation, ubiquitination, sumoylation, ADP ribosylation, deimination and proline isomerization (6). These covalent modifications either repress or activate transcription. Nucleosomes are remodeled by the action of at least five classes of ATP-dependent chromatin remodelers (7). There are many other facets to chromatin biology and chromatin remodeling which involve a variety of proteins associated with nucleosome assembly and disassembly (8,9), DNA methylation (10,11) and more, e.g. histone variants (12) and the involvement of RNAi components in heterochromatic gene silencing (13). Chromatin biology is complicated and multifaceted and a comprehensive review of this subject is beyond the scope of this article. In addition to the limited number of reviews cited above, readers are referred to a special issue of *Cell*, volume 128 (2007), 'Epigenetics and Chromatin Organization', which provides a series of excellent reviews on different aspects of chromatin biology and epigenetics.

The ChromDB public database (<http://www.chromdb.org>) serves as a repository for chromatin-related proteins. ChromDB was initiated as a National Science Foundation Plant Genome project database and focused on *Arabidopsis thaliana* and *Zea mays* (maize). The database was populated using *Saccharomyces cerevisiae* and animal chromatin-associated proteins as BLAST queries to search the nearly completed *Arabidopsis* genome and the maize EST collection to identify corresponding plant homologs. ChromDB has grown from several hundred plant proteins to over 7000 proteins representing over 30 organisms (7474 proteins total: 3328 plants, 1779 animals, 2143 fungi, 167 stramenopiles, 57 protists). Currently, the database focuses on chromatin-related proteins that are conserved widely across eukaryotic species. Our main research interest lies in the analysis of the evolution of the plant chromatin proteome. To facilitate this study and to make the database applicable to all researchers and educators

*To whom correspondence should be addressed. Tel: +1 520 626 3824; Fax: +1 520 626 4824; Email: cnapoli@email.arizona.edu

interested in chromatin biology, we have moved beyond plants to include fungal and animal proteins. We strive to organize ChromDB data in a straightforward and comparative manner and provide users with a variety of tools to visualize sequence information and to extract data by way of user-generated customized reports. Our goal is to make information on chromatin-associated proteins readily accessible despite the relative complexity of the processes.

DATABASE DESCRIPTION

The ChromDB database was initiated in 2000 as a project database for a National Science Foundation Plant Genome Research Program grant (DBI-9975930; R. Jorgensen, PI) which focused on the identification and functional analysis of *Z. mays* (maize) and *A. thaliana* genes that contribute to chromatin-based control of gene expression. One aspect of the aforementioned project was to produce and analyze RNAi lines for a set of Arabidopsis and maize chromatin-associated genes and display the results at a public database. ChromDB has evolved from a project database to a community database with renewed funding from the Plant Genome Research Program (NSF DBI-0421679; PI Napoli). ChromDB consists of a web interface, Perl modules and a relational database. The web interface is built in HTML::Mason, taking advantage of Mason's ability to embed Perl within HTML. Perl modules have been developed to access data based on user queries and pass this information back to the web pages for Mason to interpret and display the results. The database is a MySQL relational database running in a UNIX background with the schema developed to account for the type of data used and generated at the website.

Database contents

ChromDB sequences fall into two categories: genomic-based and transcript-based. Genomic-based sequences are limited to plant genomes [*A. thaliana*, *Oryza sativa* (japonica cultivar- and indica cultivar-groups), *Medicago truncatula*, *Populus trichocarpa*, *Physcomitrella patens* (moss) and *Z. mays*] and algal and diatom genomes (*Chlamydomonas reinhardtii*, *Ostreococcus lucimarinus* and *Phaeodactylum tricornerutum*). The plant genomes are highlighted on the side toolbar on the ChromDB homepage (shown on the left side in Figure 1). Other important plant species are included in the database as transcript-based sequences which are derived from EST contigs or singlets. The use of EST contigs results in partial sequences especially for larger proteins. For example, ~200 transcript-based sequences are included in the database for *Hordeum vulgare* (barley) but only 36% of these transcripts represent the entire, predicted coding sequence. Partial protein sequences, usually protein domains or the C-termini, are used as BLAST queries when identifying EST contigs. The use of a limited span of protein, rather than the entire sequence, limits redundancy that could result from the inclusion of multiple, non-overlapping contigs representing different

regions of the same transcript. Transcript-based plant sequences are converted to genome-based as sequencing projects produce sufficient data to make a conversion worthwhile. For example, maize is being converted from transcript-based to genomic-based due to the rapid accumulation of sequence data from large-scale maize genome sequencing projects.

ChromDB does not display whole chromosomes; thus for genomic-based organisms, the genomic sequence is limited to a span of nucleotides containing the predicted transcript splice model and 5' and 3' untranslated regions. Plant sequences are obtained from a variety of sources, e.g. NCBI databases (<http://www.ncbi.nlm.nih.gov/>), the Department of Energy Joint Genome Institute (<http://www.jgi.doe.gov/>), The Arabidopsis Information Resource (<http://www.arabidopsis.org/>); the J. Craig Venter Institute [<http://www.tigr.org/>, formally The Institute for Genomic Research (TIGR)] and PlantGDB (<http://www.plantgdb.org/>). All plant sequences are curated by ChromDB staff members to provide the best transcript models. In many cases, we have derived our own transcript models from genomic sequences using FGNEISH or FGENESH+ licensed from Softberry (<http://www.softberry.com>). We make use of multiple sequence alignments to analyze plant proteins and correct models where biological support of the model (cDNA sequences) is not available.

Important animal and fungal model organisms, such as *Homo sapiens*, *Drosophila melanogaster* and *S. cerevisiae*, are available as transcript-based sequences and are obtained from the NCBI Reference Sequence (RefSeq) collection (<http://www.ncbi.nlm.nih.gov/RefSeq/>). We focus on sequenced genomes and do not derive EST contigs for non-plant organisms. These transcripts are rarely curated by ChromDB staff, except for predicted transcript models that need substantial improvement as indicated by multiple sequence analysis and only when RefSeq accessions affect the quality of a phylogenetic tree.

All database sequences are assigned a ChromDB ID (identifier) which denotes both the transcript and the protein. These identifiers, as well as formal gene names, loci and aliases are included in the database and can be used to search for gene records. An explanation of the ChromDB identifiers is provided in the help manual under section IV entitled ChromDB Identifiers. ChromDB identifiers are not gene names and are not intended to take the place of recognized formal names; these are database identifiers and serve as one level of database organization.

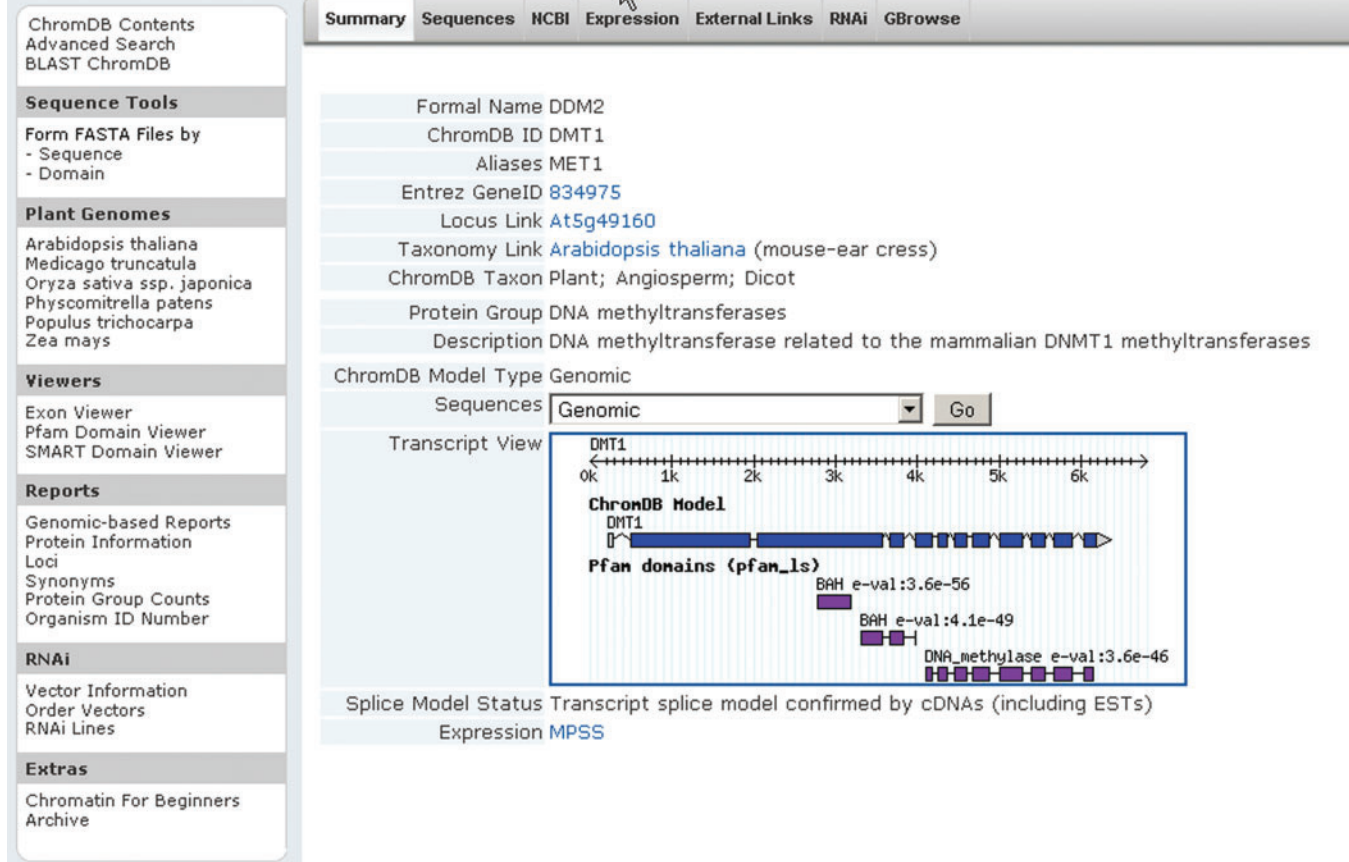
Database protein groups

One of the more difficult aspects of the database relates to providing a straightforward hierarchical organization of the range of ChromDB protein groups associated with chromatin remodeling. This difficulty relates to the complexity of the full range of proteins associated with chromatin remodeling. There are over 90 protein groups displayed at ChromDB; however, they can be grouped into parent categories reflecting different functional

ChromDB: Chromatin Database

Quick Search  [advanced search](#)

Go



ChromDB Contents
Advanced Search
BLAST ChromDB

Sequence Tools

Form FASTA Files by
- Sequence
- Domain

Plant Genomes

Arabidopsis thaliana
Medicago truncatula
Oryza sativa ssp. japonica
Physcomitrella patens
Populus trichocarpa
Zea mays

Viewers

Exon Viewer
Pfam Domain Viewer
SMART Domain Viewer

Reports

Genomic-based Reports
Protein Information
Loci
Synonyms
Protein Group Counts
Organism ID Number

RNAi

Vector Information
Order Vectors
RNAi Lines

Extras

Chromatin For Beginners
Archive

Summary Sequences HNCBI Expression External Links RNAi GBrowse

Formal Name DDM2
ChromDB ID DMT1
Aliases MET1
Entrez GeneID 834975
Locus Link At5g49160
Taxonomy Link Arabidopsis thaliana (mouse-ear cress)
ChromDB Taxon Plant; Angiosperm; Dicot
Protein Group DNA methyltransferases
Description DNA methyltransferase related to the mammalian DNMT1 methyltransferases

ChromDB Model Type Genomic

Sequences Genomic Go

Transcript View

DMT1
0k 1k 2k 3k 4k 5k 6k

ChromDB Model

DMT1

Pfam domains (pfam_ls)

BAH e-val:3.6e-56
BAH e-val:4.1e-49
DNA_methylase e-val:3.6e-46

Splice Model Status Transcript splice model confirmed by cDNAs (including ESTs)

Expression MPSS

Figure 1. A screen shot of a gene record page along with the side tool bar for accessing database searching, tools, reports and viewers.

aspects of chromatin biology, as shown in Supplementary Table 1. The next, lower level of organization is the individual protein groups, i.e. the three- to five-letter designations. The more complex groups such as the CHR proteins (SWI/SNF chromatin remodeling ATPase super family) can be broken down further into distinct phylogenetic groups, e.g. SNF2, CHD1, RAD16 (14,15). This protein group classification scheme forms the basis for advanced searching and generating reports, two of our important database tools. A full description of the hierarchy is provided in the Supplemental Data Table 1. However, we warn readers that these groups may change slightly by the time this manuscript is published, as we continue to find new ways to organize data in a straightforward manner that allows for ease of navigation through the website. The major divisions of protein groups are as follows: Histones and Histone Linker Proteins, Nucleosome Organization (includes assembly and displacement), Histone Modifications, Histone Modification Binding-Proteins, Modified-Histone-Binding Proteins, DNA Modifying Proteins, Non-Histone DNA-Binding Proteins, RNAi Components and Chromosome Dynamics.

Database access and interface

ChromDB is a public database available at: <http://www.chromdb.org>. The contents of the database can be searched, compared, and visualized using a variety of search functions, viewers and report tools. A user manual is available through a 'Help' link at the top of each web page (Figure 1). Two search options are provided, a limited 'Quick Search' text box and a menu-driven 'Advanced Search' that provides the means for comparative searching. The 'Quick Search' text box is located at the top of every webpage and accepts single entries for gene names, either the ChromDB ID, the formal gene name, an alternative alias or a locus. In those cases where the same formal gene name is used for multiple organisms, a list is generated showing each gene name and organism, as well as the ChromDB ID. For example, the Argonaute gene designated as AGO1 is used for homologous genes in *Arabidopsis*, *D. melanogaster*, and *Schizocaccharomyces pombe* and a quick search using AGO1 will bring up all three entries. Additionally, this text box accepts an organism name (either the scientific or common name) or an NCBI accession.

An 'Advanced Search' is available from the link on the side menu shown in Figure 1. This link brings up a menu-driven format that allows the user to customize a search in a variety of ways using three different criteria: organisms, protein groups and the type of report. The first two criteria have alternative options. For the organisms, the default is an alphabetical list of scientific names, and a link is provided to switch to a taxon classification (e.g. plants, animals, fungi). For the Protein Group selection, the default is the functional groups shown in Supplementary Table 1, and a link is provided to display an alphabetical list of all protein groups. Alternatively, a link is provided that displays a text box for entering a list of gene names as well as the report selection.

The gene record page is the central navigation portal for accessing information relating to each database gene. Regular users of the database will notice a new look to the ChromDB website and the gene record pages. New additions are NCBI Entrez Gene link (if available), a ChromDB taxon description, a drop-down menu of FASTA formatted sequences, the organism/sequence classification (transcript- or genomic-based), and a thumbnail view of the GBrowse display. Figure 1 shows a screen shot of the summary or default page of a Gene Record Page. The tabular format at the top of the gene record page provides access for additional information such as 'decorated' sequences, NCBI accessions and the GBrowse display. Two more tabs, titled Expression and RNAi, will appear for some maize and Arabidopsis genes, as we have retained the RNA gels and RNAi information from the previous grant. More information regarding each tab can be found on the Help page at the website (<http://www.chromdb.org/help/genePage.html>).

ChromDB uses the GMOD tool, GBrowse (16) as an individual gene-based visualization tool and not as a genome wide or chromosome visualization tool. For genomic-based organisms, the GBrowse view is based on the genomic sequence and the display includes the transcript splice model, protein domains (aligned against the transcript model) and NCBI accessions. The inclusion of the protein domains aligned to exons is useful in discerning the effect of alternate transcript splicing on protein domain structure. Each individual plant can have specialized tracks, for example Arabidopsis displays have a track for *Agrobacterium* T-DNA insertion events (17; <http://signal.salk.edu/>) For transcript-based plant organisms, the GBrowse display is based on the transcript and tracks include NCBI accessions and protein domains. For non-plant organisms, the GBrowse display is limited to the transcript, protein domains and the RefSeq accession.

ChromDB also provides a local BLAST (18) server. In addition to similarity searching, this tool is useful in determining if a gene is present in the database. Users can select the standard BLAST programs as well as preset databases such as plants, animals, or fungi. On the results page, each match is linked back to that gene's Gene Record Page where more information can be obtained about that gene. External links are provided to users in several places. There is a link on the homepage on the left tool bar and within the web pages. For example, each plant genome page has a list of appropriate links,

e.g. TAIR (The Arabidopsis Information Resource), among others, for *A. thaliana*, The Craig Venter Institute (TIGR) for a number of organisms.

Comparative tools

Part of our mission is to provide the community with the means to make comparative analyses of chromatin-associated proteins among a diverse group of organisms. The links for these comparative tools and viewers are provided on the side tool bar on each web page. Most of these features use the same menu-driven interface discussed above for the 'Advanced Search' feature. Examples of these features are: the ability to form FASTA files (entire sequence or a protein domain) and viewers for Pfam and SMART protein domains and exon structure (see Supplementary Data for examples of the protein and exon viewers). A full description of all ChromDB tools will not be listed here, but the contents can be seen in the side tool bar in Figure 1. We encourage readers to explore the website homepage to discover all database features. Information about the selection menus and the tools can be found on the general Help page (<http://www.chromdb.org/help/help.html>).

FUTURE DIRECTION

We have not included data on the number of genes for each organism in this article, as those numbers will change by the publication date. We add new protein groups to reflect new discoveries in the literature, and we continue to devise new tools to enable users to extract information from MySQL tables. New datasets (organisms and protein groups) and tools are prepared at our development site, and weekly updates are run to sync production and development to reflect data releases. A link is provided on the homepage so users can access a list of updated contents. A phylogenetic classification scheme will be introduced in the future to further subdivide protein groups. We will post multiple sequence alignments, as well as the phylogenetic trees, in support of the classifications. The ChromDB website changes continually, both in content and appearance, as we strive to present users with a comprehensive database of chromatin-associated proteins for an ever-increasing number of organisms, and as we endeavor to find new ways to display the data in a straightforward and comparative manner.

SUPPLEMENTARY DATA

Supplementary data are available at NAR Online.

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 and we acknowledge the support and help provided by Gavin Nelson and Nirav Merchant. Funding to pay the Open Access publication charges for this article was provided by NSF PGRP DBI-0421679.

Conflict of interest statement. None declared.

REFERENCES

1. Tremethick,D.J. (2007) Higher-order structures of chromatin: the elusive 30 nm fiber. *Cell*, **128**, 651–654.
2. Groth,A., Rocha,W., Verreault,A. and Almouzni,G. (2007) Chromatin challenges during DNA replication and repair. *Cell*, **128**, 721–733.
3. Costelloe,T., FitzGerald,J., Murphy,A.F. and Lowndes,N.F. (2006) Chromatin modulation and the DNA damage response. *Exp. Cell Res.*, **12**, 2677–2686.
4. Osley,M.A., Tsukuda,T. and Nickoloff,J.A. (2007) ATP-dependent chromatin remodeling factors and DNA damage repair. *Mutat. Res.*, **618**, 65–80.
5. Bloom,K. (2007) Centromere dynamics. *Curr. Opin. Genet. Dev.*, **17**, 151–156.
6. Kouzarides,T. (2007) Chromatin modifications and their function. *Cell*, **128**, 693–705.
7. Jerzmanowski,A. (2007) SWI/SNF chromatin remodeling and linker histones in plants. *Biochim. Biophys. Acta.*, **1769**, 330–345.
8. Morse,R.H. (2007) Transcription factor access to promoter elements. *J. Cell. Biochem.*, **39**, 1235–1244.
9. Rando,O. and Ahmad,K. (2007) Rules and regulation in the primary structure of chromatin. *Curr. Opin. Cell Biol.*, **19**, 250–256.
10. Geiman,T.M. and Robertson,K.D. (2002) Chromatin remodeling, histone modifications, and DNA methylation-how does it all fit together? *J. Cell Biochem.*, **87**, 117–125.
11. Berger,S.L. (2007) The complex language of chromatin regulation during transcription. *Nature*, **447**, 407–412.
12. Malik,H.S. and Henikoff,S. (2003) Phylogenomics of the nucleosome. *Nat. Struct. Biol.*, **10**, 882–891.
13. Bernstein,E. and Allis,C.D. (2005) RNA meets chromatin. *Genes Dev.*, **19**, 1635–1655.
14. Eisen,J.A., Sweder,K.S. and Hanawalt,P.C. (1995) Evolution of the SNF2 family of proteins: subfamilies with distinct sequences and functions. *Nucleic Acids Res.*, **23**, 2715–2723.
15. Flaus,A., Martin,D.M.A., Barton,G.J. and Owen-Hughes,T. (2006) Identification of multiple distinct Snf2 subfamilies with conserved structural motifs. *Nucleic Acids Res.*, **34**, 2887–2905.
16. Stein,L.D., Mungall,C., Shu,S., Caudy,M., Mangone,M., Day,A., Nickerson,E., Stajich,J.E., Harris,T.W. *et al.* (2002) The generic genome browser: a building block for a model organism system database. *Genome Res.*, **12**, 1599–1610.
17. Alonso,J.M., Stepanova,A.N., Leisse,T.J., Kim,C.J., Chen,H., Shinn,P., Stevenson,D.K., Zimmerman,J., Barajas,P. *et al.* (2003) Genome-wide insertional mutagenesis of *Arabidopsis thaliana*. *Science*, **5633**, 653–657.
18. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.