

dictyBase, the model organism database for *Dictyostelium discoideum*

Rex L. Chisholm*, Pascale Gaudet, Eric M. Just, Karen E. Pilcher, Petra Fey, Sohel N. Merchant and Warren A. Kibbe

dictyBase, Center for Genetic Medicine, Feinberg School of Medicine, Northwestern University, Chicago, IL 60611, USA

Received September 14, 2005; Revised and Accepted October 13, 2005

ABSTRACT

dictyBase (<http://dictybase.org>) is the model organism database (MOD) for the social amoeba *Dictyostelium discoideum*. The unique biology and phylogenetic position of *Dictyostelium* offer a great opportunity to gain knowledge of processes not characterized in other organisms. The recent completion of the 34 MB genome sequence, together with the sizable scientific literature using *Dictyostelium* as a research organism, provided the necessary tools to create a well-annotated genome. dictyBase has leveraged software developed by the *Saccharomyces* Genome Database and the Generic Model Organism Database project. This has reduced the time required to develop a full-featured MOD and greatly facilitated our ability to focus on annotation and providing new functionality. We hope that manual curation of the *Dictyostelium* genome will facilitate the annotation of other genomes.

INTRODUCTION

dictyBase (<http://dictybase.org>) is the model organism database (MOD) for the social amoeba *Dictyostelium discoideum*. Like other MODs such as FlyBase (1), WormBase (2,3), Mouse Genome Informatics (4) and the *Saccharomyces* Genome Database (SGD) (5), dictyBase (6) uses the organism's genome sequence to organize the biological knowledge resulting from experimental studies using *Dictyostelium*. *Dictyostelium* is a model organism widely used for biomedical research. An amoeba during most of its life, starvation induces a very interesting developmental program during which individual cells stream together by chemotaxis to form a multicellular tissue (7). A morphogenetic process involving cell migration and cellular morphogenesis transforms a simple mound of cells into a slug or pseudoplasmodium establishing

a relatively simple developmental pattern. This structure then develops into a fruiting body consisting of multiple cell types including spores and terminally differentiated stalk cells. These features together with the efficient genetic manipulation by gene targeting and replacement as well as insertional mutagenesis and suppressor screens have made *Dictyostelium* a popular experimental system. Research using *Dictyostelium* has been critical in understanding fundamental processes such as cell migration (8), cell signaling (9,10), phagocytosis and morphogenesis. Recently *Dictyostelium* has also been used to help establish the mechanisms of action of medically important drugs such as cisplatin (11–14), used to treat various cancers, and lithium and valproic acid (15), used to treat depressive disorders. The recently completed *Dictyostelium* genome sequence is housed at dictyBase and has been integrated with other experimental data to provide investigators with a rich new resource that will facilitate future studies using *Dictyostelium* experimentally and for comparative genomics.

GENOME SEQUENCE COMPLETED

In May 2005 the sequence of the complete genome of *Dictyostelium* was reported (16). The A + T rich genome (77.6%) consists of 34 million bases of DNA sequence that are predicted to encode 13 573 genes—a number of genes comparable with *Drosophila*. dictyBase houses and maintains the genome sequence produced by the International Sequencing Consortium. As the first free living protozoan to be completely sequenced, the predicted proteome supports the hypothesis that *Dictyostelium* represents an early branch in the Eukaryotic Tree of Life that diverged after the split between animals, plants and fungi, with *Dictyostelium* and other amoebae more closely related to animals. Many *Dictyostelium* proteins are more similar to human orthologs than are those of *Saccharomyces cerevisiae*. The genome has several notable features. The sequence is very gene dense with an average gene spacing of 2.5 kb. The abundance of short sequence

*To whom correspondence should be addressed. Tel: +1 312 503 3209; Fax: +1 312 503 5603; Email: r-chisholm@northwestern.edu

repeats produces unusual runs of amino acids, such as poly-asparagine and polyglutamine tracts of 20 residues or more. These amino acid repeats occur in >2000 of the predicted proteins. The genome also contains a surprisingly large number of polyketide synthases, which may produce a previously unanticipated large complement of natural products. Consistent with this apparently large secondary metabolism, the genome contains many ABC transporters that may be involved in export of these natural products. Also encoded are a large number of proteins containing multiple EGF repeats that have been postulated to play roles in adhesion or cell recognition.

These fascinating genomic features together with the unique biology of *Dictyostelium* and the rich body of *Dictyostelium* literature (nearly 8000 publications) provide a valuable opportunity to annotate many of the genes with functional information that will widen the spectrum of characterized proteins in public databases. This emphasizes the need for a well-curated database for *Dictyostelium*—a need that dictyBase strives to fulfill.

dictyBase LINKS GENOME SEQUENCE TO FUNCTIONAL INFORMATION

As the MOD for *D. discoideum*, dictyBase provides highly curated information that facilitates research by providing a searchable, structured repository of *Dictyostelium* experimental results. In addition to the complete genome sequence, groups performing high-throughput experiments such as large-scale mutagenesis and microarray-based gene expression studies are depositing their data in dictyBase for integration and distribution to the research community. dictyBase uses the genomic DNA sequence as a scaffold on which to organize and display the biological knowledge and experimental evidence derived from *Dictyostelium* research. dictyBase is the most comprehensive source of information regarding *Dictyostelium*, and the database presents the highest quality annotations of *Dictyostelium* genes. Table 1 presents the data content of dictyBase as of September 2005. Nearly all of the data in dictyBase is available for bulk download through our download center (<http://dictybase.org/downloads/>).

The heart of dictyBase is the Gene Page. A Gene Page is available for each gene that has been shown or predicted to exist in the genome. The Gene Page collects and organizes all of the available information related to that gene into several sections. Figure 1 shows an example of a typical Gene Page. A General Information section provides the official gene name together with any other names that have been used to refer to that gene. This assures that a search of the database will allow retrieval of a gene using any name that has been used to refer to that gene. Scientific curators verify published and proposed gene names to encourage use of the *Dictyostelium* Nomenclature Guidelines (<http://dictybase.org/nomenclatureguidelines.htm>) and to ensure that names are not duplicated. When appropriate, curators communicate with authors to discuss and modify gene names. Also in this section, gene product names and a brief description of the gene product are presented, as is the unique dictyBaseID. Next, the Chromosomal Coordinates section presents a graphical view of the gene and associated sequences such as expressed sequence

Table 1. Data and annotations in dictyBase (September 2005)

<ul style="list-style-type: none"> • 13 573 Automated Gene Predictions • 1387 GenBank records • 155 032 ESTs • 6011 PubMed references • 1131 Colleagues • 2445 Curated Models • Nine alternative transcripts • Gene products for 5228 genes • Brief descriptions for 1657 genes • Mutant phenotypes for 322 genes • GO annotations for 5697 genes • Summary paragraphs for 502 genes • External data: <ul style="list-style-type: none"> ○ 6801 Microarray expression profiles (BCM and UCSD) ○ <i>In situ</i> hybridization: 150 images (Tsukuba Atlas) ○ Insertional mutants: 817 links (BCM) • DSC: <ul style="list-style-type: none"> ○ 728 strains ○ 153 plasmids

tags (ESTs), cDNAs and other GenBank entries. An Associated Sequences section provides links to sequences corresponding to each of the sequence features known for that gene, including gene predictions as well as ESTs, cDNAs and genomic sequences submitted by individual researchers to other public databases. The Protein Information section contains the size, molecular weight and protein domains of the predicted protein.

Functional information is presented in sections for Gene Ontology (GO) annotations and phenotypes whenever this information is available. Phenotypic data is represented using a restricted vocabulary that describes the consequences of mutations reported for that gene. The Gene Ontology section lists terms from each of the function, process and component ontologies associated with the gene (17,18). Following links in each of these sections display additional details for these annotations. Clicking on a GO term or a phenotype will list all of the other genes annotated with the same term. The Expression section leads to developmental profiles for the gene that have been generated by large-scale microarray based studies performed at the Baylor College of Medicine and the University of California, San Diego. The Links section on the Gene Page provides links to GenBank, UniProt, GeneDB at the Sanger Institute and, where appropriate, to signaling pathways at the Signal Transduction Knowledge Environment at Science (19,20). This section also links to researchers in our colleague database who are investigating that gene or gene product. A Literature Section lists the most recent publications referring to the gene and a link to a Literature Guide that contains a complete listing of papers relevant to that gene. As of the fall of 2005, >500 Gene Pages had a Summary paragraph written by a dictyBase curator that contains the current knowledge regarding that gene. dictyBase curators are in the process of manually reviewing all of the automated gene predictions in light of available data such as ESTs and cDNAs. Manually reviewed genes are indicated by the presence of a 'Curated Model' on the Gene Page. As of September 2005 dictyBase curators have reviewed nearly 2500 genes, which corresponds to ~20% of the total number of genes.

An Online Informatics Resource for *Dictyostelium*

Search dictyBase: use * as a wildcard character Include dicty Newsletter in Search

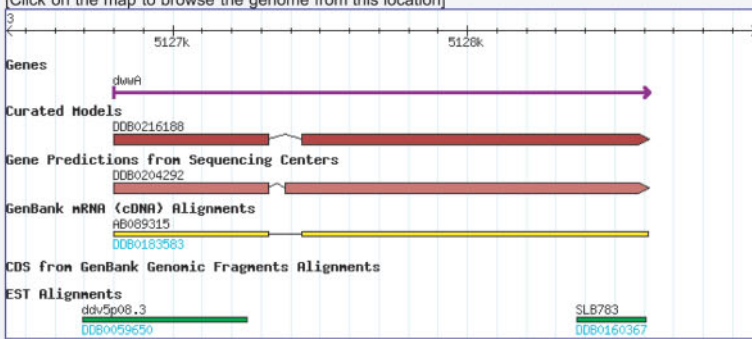
dictyBase Genome Browser BLAST Colleagues Stock Center Research Tools Help Links Contact Us

Gene Page for *dwwA* ? Help

[dictyBase Home](#) >> [Chromosome 3:5126436..5128980](#) >> Gene: *dwwA*

General Information [TOP]	
Gene Name	<i>dwwA</i>
Gene Product	WW domain-containing protein
Description	homologous to human KIBRA, a Dendrin binding protein; required for normal cytokinesis on solid surface
dictyBaseID	DDB0216188

Chromosomal Coordinates [TOP]	
Location	Chromosome 3 coordinates 5126799 to 5128617, Watson strand
Genome Browser Snapshot	[Click on the map to browse the genome from this location]



Notes
Note regarding this sequence: A 1 nt difference between the sequence from GenBank record AB089315 and the sequence from the Sequencing Center results in 1 amino acid exchange at position 208.

Associated Sequences (Click on the link to access Sequence Information Page) [TOP]	
Curated Model	BLAST at dictyBase DDB0216188 (Primary Model) BLAST at NCBI Derived from Sequencing Center Gene Prediction. Supported by mRNA.
Gene Prediction From Sequencing Center	BLAST at dictyBase DDB0204292
GenBank mRNA	BLAST at dictyBase DDB0183583
ESTs	DDB0059650 , DDB0160367

Protein Information View <i>dwwA</i> Sequence Information Page [TOP]	
Length	568 aa
Molecular Weight	64,752 Da
Protein Domain	Predicted Protein Domains at GeneDB

Gene Ontology Annotations for <i>dwwA</i> View evidence and references [TOP]	
Molecular Function	calmodulin binding (IDA)
Biological Process	cytokinesis (IMP), actin filament organization (IMP)
Cellular Component	nucleus (IDA), cell cortex (IDA)

Expression [TOP]	
UCSD Expression Profile BCM Expression Profile	

Phenotype View <i>dwwA</i> Phenotype details and references [TOP]	
Null	aberrant cytokinesis , aberrant F-actin organization , aberrant cell morphology

Links [TOP]	
dwwA Researchers GeneDB Entrez Nucleotide Entrez Protein	

Summary [TOP]	
The <i>dwwA</i> gene is required for cytokinesis of <i>Dictyostelium</i> cells on solid surfaces. Its product, DwwA, contains two WW domains, an IQ motif, a C2 domain and a proline-rich region. DwwA is localized to the cell cortex and the nucleus ; the N-terminal half of the protein, which contains the C2 domain, is required for the cortical localization of DWWA. <i>dwwA</i> -null cells have higher levels of F-actin around the periphery of the ventral surface of the cell (actin filament organization). The IQ motif of DwwA binds <i>calmodulin in vitro</i> (calmodulin binding). DwwA has been proposed to function as an adaptor protein, participating in cytokinesis by mediating the interaction of Ca ²⁺ /calmodulin with the actin cytoskeleton (Nagasaki and Uyeda 2003).	

Latest References View complete list of references for <i>dwwA</i> (1 paper) [TOP]	
dictyBase PubMed Access Full Text	Nagasaki & Uyeda (2004) 'DWWA, a novel protein containing two WW domains and an IQ motif, is required for scission of the residual cytoplasmic bridge during cytokinesis in <i>Dictyostelium</i> .' <i>Mol Biol Cell</i> 15:435-46

Figure 1. A typical dictyBase Gene Page. This sample Gene Page shows the types of information that are displayed on a dictyBase Gene Page.

GENERAL RESOURCES FOR *DICTYOSTELIUM*

dictyBase also serves as a clearinghouse for resource materials for students, researchers and educators using *Dictyostelium* in a wide range of classroom and laboratory settings. We maintain archives of the dictyNews, a weekly electronic newsletter that presents abstracts of papers available upon acceptance for publication, as well as a growing collection of protocols for working with *Dictyostelium*.

DICTY STOCK CENTER (DSC)

dictyBase also provides a direct portal to the DSC maintained at Columbia University. This resource provides access to *Dictyostelium* strains, including natural isolates, targeted mutants and GFP labeled strains. The collection also contains plasmids used to manipulate gene expression or create targeted gene disruptions. Currently the DSC has >700 strains and >150 plasmids all available to researchers. A shopping cart system allows Stock Center users to add strains and proceed to a checkout system familiar to any user of online ordering systems. For users who are also registered as colleagues in the dictyBase colleagues database, addresses for delivery and contact information are populated directly from the database. dictyBase holds all of the Stock Center data and provides the informatics support for the DSC. This will enable increased integration between gene information and strains available through the DSC.

DICTYBASE INCORPORATES SOFTWARE DEVELOPED BY OTHER MODS AND THE GENERIC MODEL ORGANISM DATABASE (GMOD) PROJECT

dictyBase was initially established as a clone of the software developed and generously provided by SGD. dictyBase developers have subsequently integrated several software packages from the GMOD project (<http://gmod.org>). These include the GBrowse tool for chromosome and map displays and, most recently, the Chado database schema for storing sequence and sequence feature information. This entailed porting of the Chado schema to Oracle since both dictyBase and SGD are implemented in Oracle. dictyBase has developed new code that separates communication with the database from code that generates the HTML interfaces. This object layer facilitated adoption of the Chado schema without significant impact on the presentation interface and enables good software practices such as unit testing.

CONCLUSION AND FUTURE DIRECTIONS

dictyBase strives to present a comprehensive, reliable, carefully curated dataset that integrates sequences, functional annotation, phenotypes, expression data and the research literature, while retaining the flexibility to integrate new data types as they become available. Our goal is to assure that the information in dictyBase is accessible in an effective and intuitive interface useful for biologists, while also maximizing its utility for the bioinformatics community to use computationally. Our operating principles are to respond

to users' needs, to capitalize on the vast efforts of the *Dictyostelium* research community, and to provide an important link in the network of MODs. As additional amoebae genome sequences become available dictyBase hopes to implement, and where necessary, develop interfaces and tools that facilitate comparative genomics of this diverse group of organisms.

ACKNOWLEDGEMENTS

dictyBase is supported by NIH Grants GM64426 and HG0022. Funding to pay the Open Access publication charges for this article was provided by the National Institutes of Health (GM64426).

Conflict of interest statement. None declared.

REFERENCES

1. Drysdale, R.A. and Crosby, M.A. (2005) FlyBase: genes and gene models. *Nucleic Acids Res.*, **33**, D390–D395.
2. Chen, N., Lawson, D., Bradnam, K., Harris, T.W. and Stein, L.D. (2004) WormBase as an integrated platform for the *C.elegans* ORFeome. *Genome Res.*, **14**, 2155–2161.
3. Harris, T.W., Chen, N., Cunningham, F., Tello-Ruiz, M., Antoshechkin, I., Bastiani, C., Bieri, T., Blasiar, D., Bradnam, K., Chan, J. *et al.* (2004) WormBase: a multi-species resource for nematode biology and genomics. *Nucleic Acids Res.*, **32**, D411–D417.
4. Eppig, J.T., Bult, C.J., Kadin, J.A., Richardson, J.E., Blake, J.A., Anagnostopoulos, A., Baldarelli, R.M., Baya, M., Beal, J.S., Bello, S.M. *et al.* (2005) The Mouse Genome Database (MGD): from genes to mice—a community resource for mouse biology. *Nucleic Acids Res.*, **33**, D471–D475.
5. Christie, K.R., Weng, S., Balakrishnan, R., Costanzo, M.C., Dolinski, K., Dwight, S.S., Engel, S.R., Feierbach, B., Fisk, D.G., Hirschman, J.E. *et al.* (2004) *Saccharomyces* Genome Database (SGD) provides tools to identify and analyze sequences from *Saccharomyces cerevisiae* and related sequences from other organisms. *Nucleic Acids Res.*, **32**, D311–D314.
6. Kreppel, L., Fey, P., Gaudet, P., Just, E., Kibbe, W.A., Chisholm, R.L. and Kimmel, A.R. (2004) dictyBase: a new *Dictyostelium discoideum* genome database. *Nucleic Acids Res.*, **32**, D332–D333.
7. Chisholm, R.L. and Firtel, R.A. (2004) Insights into morphogenesis from a simple developmental system. *Nature Rev. Mol. Cell Biol.*, **5**, 531–541.
8. Parent, C.A. (2004) Making all the right moves: chemotaxis in neutrophils and *Dictyostelium*. *Curr. Opin. Cell Biol.*, **16**, 4–13.
9. Van Haastert, P.J. and Devreotes, P.N. (2004) Chemotaxis: signalling the way forward. *Nature Rev. Mol. Cell Biol.*, **5**, 626–634.
10. Weeks, G., Gaudet, P. and Insall, R. (2005) The Small GTPase Superfamily. In Loomis, W.F. and Kuspa, A. (eds), *The Dictyostelium Genome*. Horizon Bioscience, Norfolk, UK, pp. 173–210.
11. Li, G., Alexander, H., Schneider, N. and Alexander, S. (2000) Molecular basis for resistance to the anticancer drug cisplatin in *Dictyostelium*. *Microbiology*, **146**, 2219–2227.
12. Li, Z., Solomon, J.M. and Isberg, R.R. (2005) *Dictyostelium discoideum* strains lacking the RtoA protein are defective for maturation of the *Legionella pneumophila* replication vacuole. *Cell. Microbiol.*, **7**, 431–442.
13. Min, J., Traynor, D., Stegner, A.L., Zhang, L., Hanigan, M.H., Alexander, H. and Alexander, S. (2005) Sphingosine kinase regulates the sensitivity of *Dictyostelium discoideum* cells to the anticancer drug cisplatin. *Eukaryotic Cell*, **4**, 178–189.
14. Min, J., Van Veldhoven, P.P., Zhang, L., Hanigan, M.H., Alexander, H. and Alexander, S. (2005) Sphingosine-L-phosphate lyase regulates sensitivity of human cells to select chemotherapy drugs in a p38-dependent manner. *Mol. Cancer Res.*, **3**, 287–296.
15. Eickholt, B.J., Towers, G.J., Ryves, W.J., Eikel, D., Adley, K., Ylinen, L.M., Chadborn, N.H., Harwood, A.J., Nau, H. and Williams, R.S. (2005) Effects

- of valproic acid derivatives on inositol trisphosphate depletion, teratogenicity, glycogen synthase kinase-3beta inhibition, and viral replication: a screening approach for new bipolar disorder drugs derived from the valproic acid core structure. *Mol. Pharmacol.*, **67**, 1426–1433.
16. Eichinger,L., Pachebat,J.A., Glockner,G., Rajandream,M.A., Sugang,R., Berriman,M., Song,J., Olsen,R., Szafranski,K., Xu,Q. *et al.* (2005) The genome of the social amoeba *Dictyostelium discoideum*. *Nature*, **435**, 43–57.
17. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.*, **25**, 25–29.
18. Harris,M.A., Clark,J., Ireland,A., Lomax,J., Ashburner,M., Foulger,R., Eilbeck,K., Lewis,S., Marshall,B., Mungall,C. *et al.* (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, D258–D261.
19. Kimmel,A.R. and Parent,C.A. (2003) The signal to move: *D.discoideum* go orienteering. *Science*, **300**, 1525–1527.
20. Gough,N.R. (2002) Science's signal transduction knowledge environment: the connections maps database. *Ann. N Y Acad. Sci.*, **971**, 585–587.