*D676–D683* *Nucleic Acids Research, 2013, Vol. 41, Database issue* *Published online 20 November 2012*
*doi:10.1093/nar/gks1064*

# dictyBase 2013: integrating multiple Dictyostelid species

**Siddhartha Basu, Petra Fey, Yogesh Pandit, Robert Dodson, Warren A. Kibbe and Rex L. Chisholm***

Biomedical Informatics Center and Center for Genetic Medicine, Northwestern Univesity, Feinberg School of Medicine, 750 North Lake Shore Drive, Chicago, IL 60611, USA

Received September 14, 2012; Accepted October 11, 2012

## ABSTRACT

**dictyBase (http://dictybase.org) is the model organism database for the social amoeba *Dictyostelium discoideum*. This contribution provides an update on dictyBase that has been previously presented. During the past 3 years, dictyBase has taken significant strides toward becoming a genome portal for the whole Amoebozoa clade. In its latest release, dictyBase has scaled up to host multiple Dictyostelids, including *Dictyostelium purpureum* [Sucgang, Kuo, Tian, Salerno, Parikh, Feasley, Dalin, Tu, Huang, Barry *et al* (2011) (Comparative genomics of the social amoebae *Dictyostelium discoideum* and *Dictyostelium purpureum*. *Genome Biol.*, 12, R20)], *Dictyostelium fasciculatum* and *Polysphondylium pallidum* [Heidel, Lawal, Felder, Schilde, Helps, Tunggal, Rivero, John, Schleicher, Eichinger *et al*. (2011) (Phylogeny-wide analysis of social amoeba genomes highlights ancient origins for complex intercellular communication. *Genome Res.*, 21, 1882–1891)]. The new release includes a new Genome Browser with RNAseq expression, interspecies Basic Local Alignment Search Tool alignments and a unified Basic Local Alignment Search Tool search for cross-species comparisons.**

## INTRODUCTION

The Amoebozoa possess an intriguing phylogenetic position, emerging as one of the first branches following the plant and animal split (1,2). Dictyostelids offer an elegantly simple model system to study early developmental switches. Under starvation stress, Dictyostelids undergo a life cycle between single-cell amoeba that develop into a multicellular fruiting body. This relatively complex 'social' behavior makes them useful in numerous fields of study, such as cell motility, signal transduction pathways, cell differentiation or interspecies interactions. Its unique niche in evolution also makes it highly suitable for comparative genomics to understand conserved functions and processes with higher eukaryotes. dictyBase (http://dictybase.org), the manually annotated model organism database, hosts the annotated genome of *Dictyostelium discoideum* (3). The database houses the 34-Mbp nuclear genome, the 55-kb mitochondrial genome (4), extrachromosomal ribosomal RNA (5) and >162 000 expressed sequence tags (EST) sequences (6).

In our last update (7), we reported the initial release of the *Dictyostelium purpureum* genome (8). *D. purpureum* is a Group 4 member, as described by the major taxonomic divisions (2) of the Dictyostelid clade. Here, we describe the expansion of dictyBase with two new genomes using a new underlying environment that is robust and scalable. This new environment uses a new genome browser that allows visualization of RNAseq expression data and interspecies genome alignments. The Basic Local Alignment Search Tool (BLAST) search has been unified and expanded with data sets of new genomes.

## dictyBase: A GATEWAY FOR DICTYOSTELID GENOMES

### Integration of new genomes

In this current release of dictyBase, we have updated the *D. purpureum* database and integrated two more genomes, those of *Dictyostelium fasciculatum* (9), a Group 1 member of the Dictyostelid clade, and *Polysphondylium pallidum*, (9) from Group 2. The two new genomes were incorporated based on their GenBank records along with their mitochondrial genomes (10). Each genome is currently available as a set of supercontigs, with 26 supercontigs for *D. fasciculatum* and 43 supercontigs

*To whom correspondence should be addressed. Tel: +1 312 503 3209; Fax: +1 312 908 5502; Email: r-chisholm@northwestern.edu

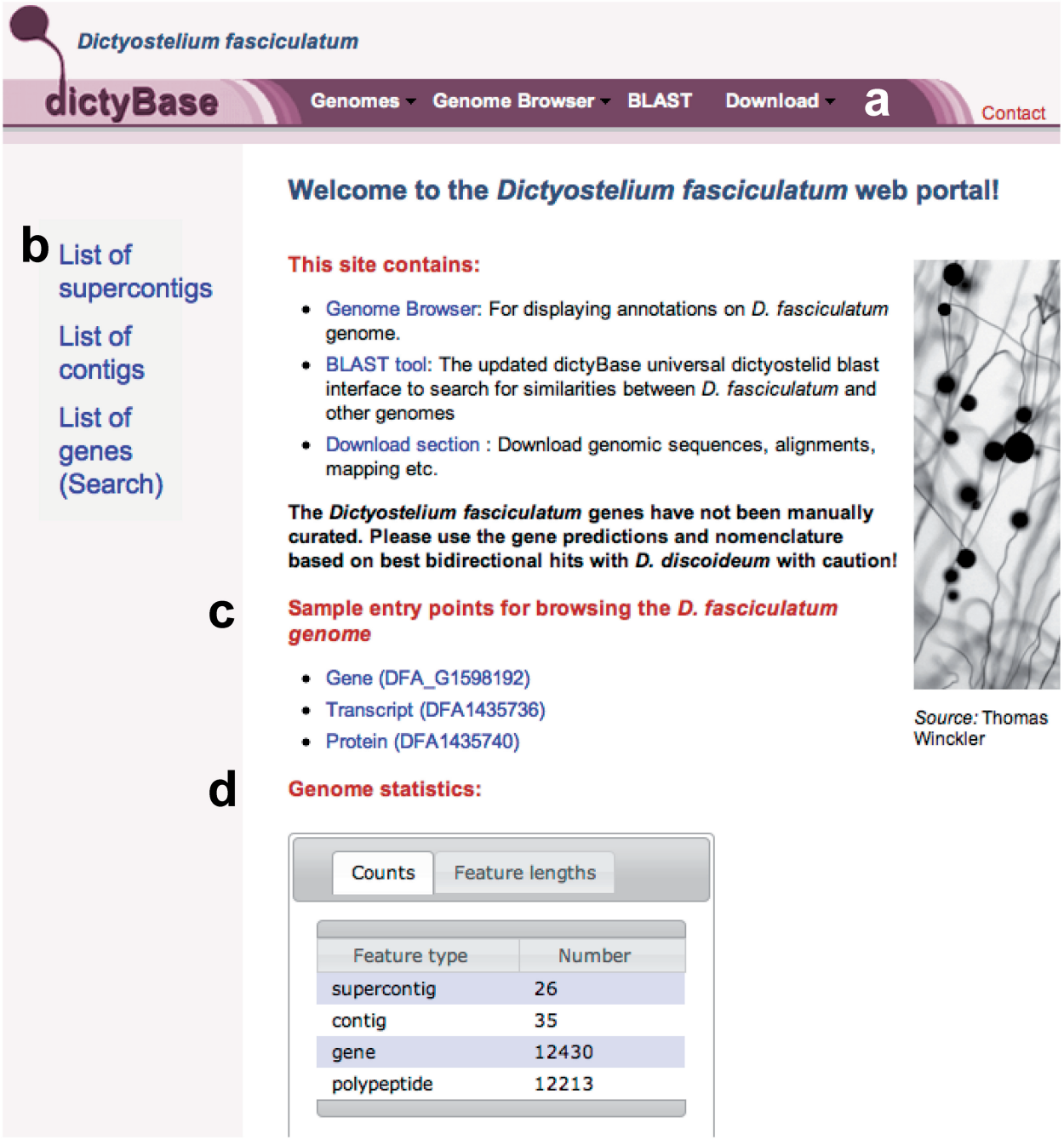


**Figure 1.** The front page of *D. fasciculatum*. (**a**) The top bar contains four drop-down menus (Genomes, Genome Browser, BLAST and Download) that provide a selection for all genomes (including *D. discoideum*), allowing the user to easily switch between genomes or access functions available for the currently displayed genome. (**b**) Quick links to lists of sequences are available on the left-hand side of the page. (**c**) The page also provides sample entry points for a gene, a transcript and a protein to help first time or occasional users. (**d**) Genome statistics display available statistics for the current genome. Note that features vary slightly for each genome depending on the available data.

for *P. pallidum*. Both genome assemblies also came with a basic set of gene predictions. The individual genomes are directly accessible through these links:

*D. purpureum*: http://genomes.dictybase.org/purpureum
*D. fasciculatum*: http://genomes.dictybase.org/fasciculatum
*P. pallidum*: http://genomes.dictybase.org/pallidum

## Uniform interface for browsing genomes

The front page (Figure 1) of every genome shares a similar look and feel to allow the dictyBase community to easily switch between genomes. All genomes are accessible from the url http://dictybase.org. For *D. fasciculatum*, it is http://dictybase.org/fasciculatum, for *P. pallidum*

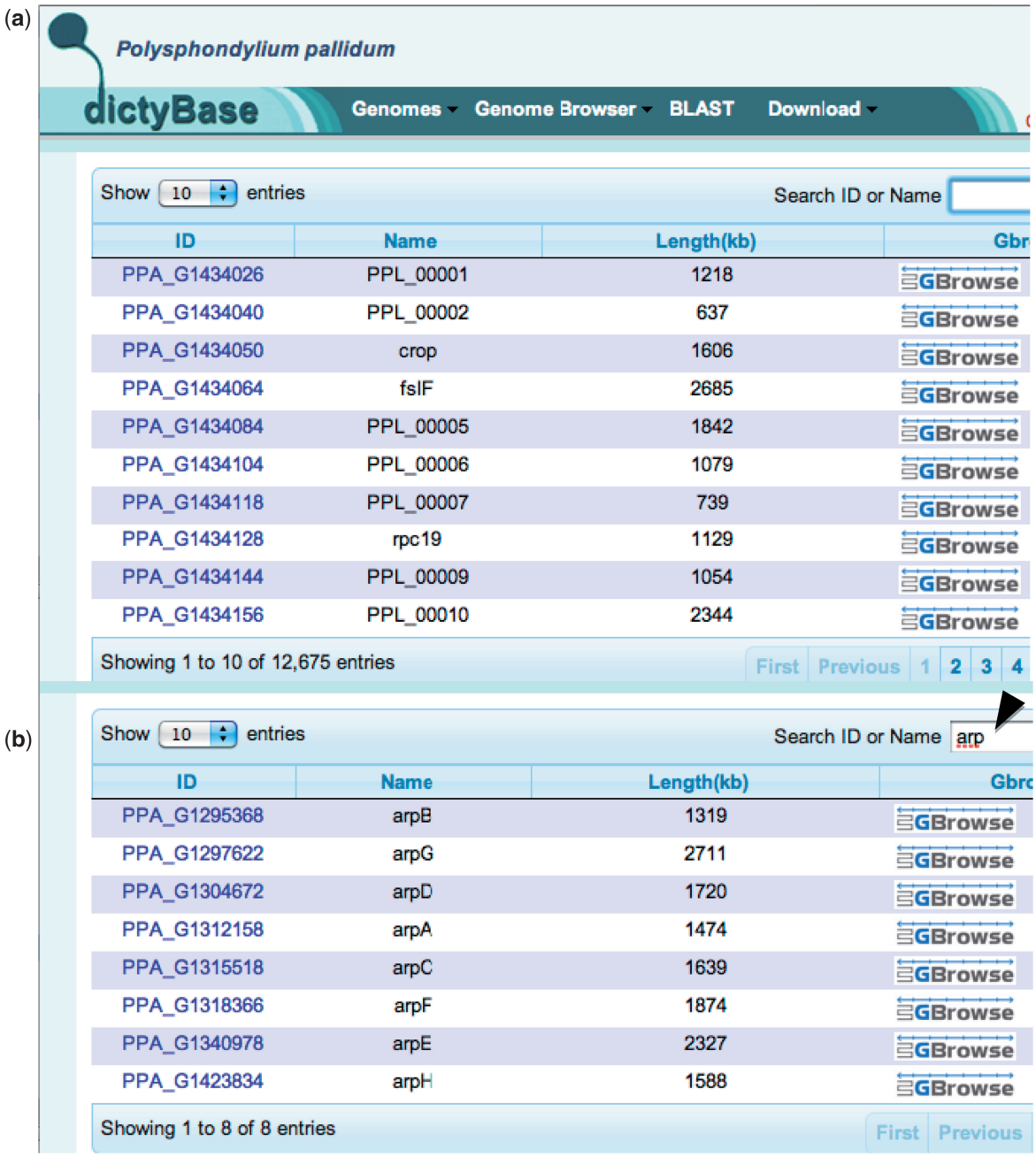


**Figure 2.** Table of genes with quick gene search for *P. pallidum*. (**a**) The table displays 10, 25, 50 or 100 entries at a time selectable by drop-down. The table includes the length of the gene and a link to that gene in the Genome Browser. (**b**) To search for a gene, start typing any gene name or gene ID and the list will be restricted instantly to those genes with that name or ID. (b) The search for 'arp' (arrow head) retrieves eight entries that contain those letters in their name.

http://dictybase.org/pallidum and for *D. purpureum* http://dictybase.org/purpureum. The four drop-down menus on top of each page (Figure 1a), Genomes, Genome Browser, BLAST and Download, provide selections to navigate among the different genomes. Below the toolbar, every front page comes with links to easily access lists and sample entry points (Figure 1b and c) for a variety of features such as supercontigs, contigs, genes, transcripts and the genome browser that may be viewed by individual record or by tabular display (Figure 2a). At the bottom of each genome home page is a display of 'Genome Statistics', which provides summary counts and statistics for genomic features (Figure 1d). It is primarily adapted from sequence ontology bioinformatics analysis (11), to provide a succinct overview of each genome. The tabular displays of various features include pagination and 'rows per page widgets' to facilitate browsing the collections (Figure 2a). The gene collection for every genome (e.g. http://dictybase.org/pallidum/gene) can be filtered dynamically using the top-right auto-complete widget

(Figure 2b). The widget accepts either the gene name or ID for filtering.

### Infrastructure and workflow changes

In addition to the uniform layout, the current release introduced numerous important changes to enable seamless genome addition while minimizing additional hardware and software requirements. The portal has been upgraded to a 64-bit linux operating system to take advantage of upgraded modern hardware and significantly increased memory. We are now able to load genomes from all standard data formats, e.g. GenBank or Generic Feature Format Version 3 (http://www.sequenceontology.org/gff3.shtml) (GFF3). We have also adopted a standard data model (Chado) for storing genomic features. This permits us to house current and future genomes in a single database instance, rather than requiring one instance per genome. We have also applied 'best practice' patterns, using web services to separate data from application logic. This has had the added benefit of enabling us to run the main dictyBase web application and genome browser servers on separate virtual machines. This allows us to more easily maintain our existing infrastructure and scale the resource by providing additional virtual machines with load balancing.

## NEW GENOME BROWSER AND INTEGRATION OF RNA SEQUENCE DATA

### Features of the new genome browser

In this release, we installed the latest version of the genome browser (GBrowse version 2.4) (12). This version of GBrowse runs from a separate GFF3 data instance (http://gmod.org/wiki/GBrowse_Adaptors) optimized for

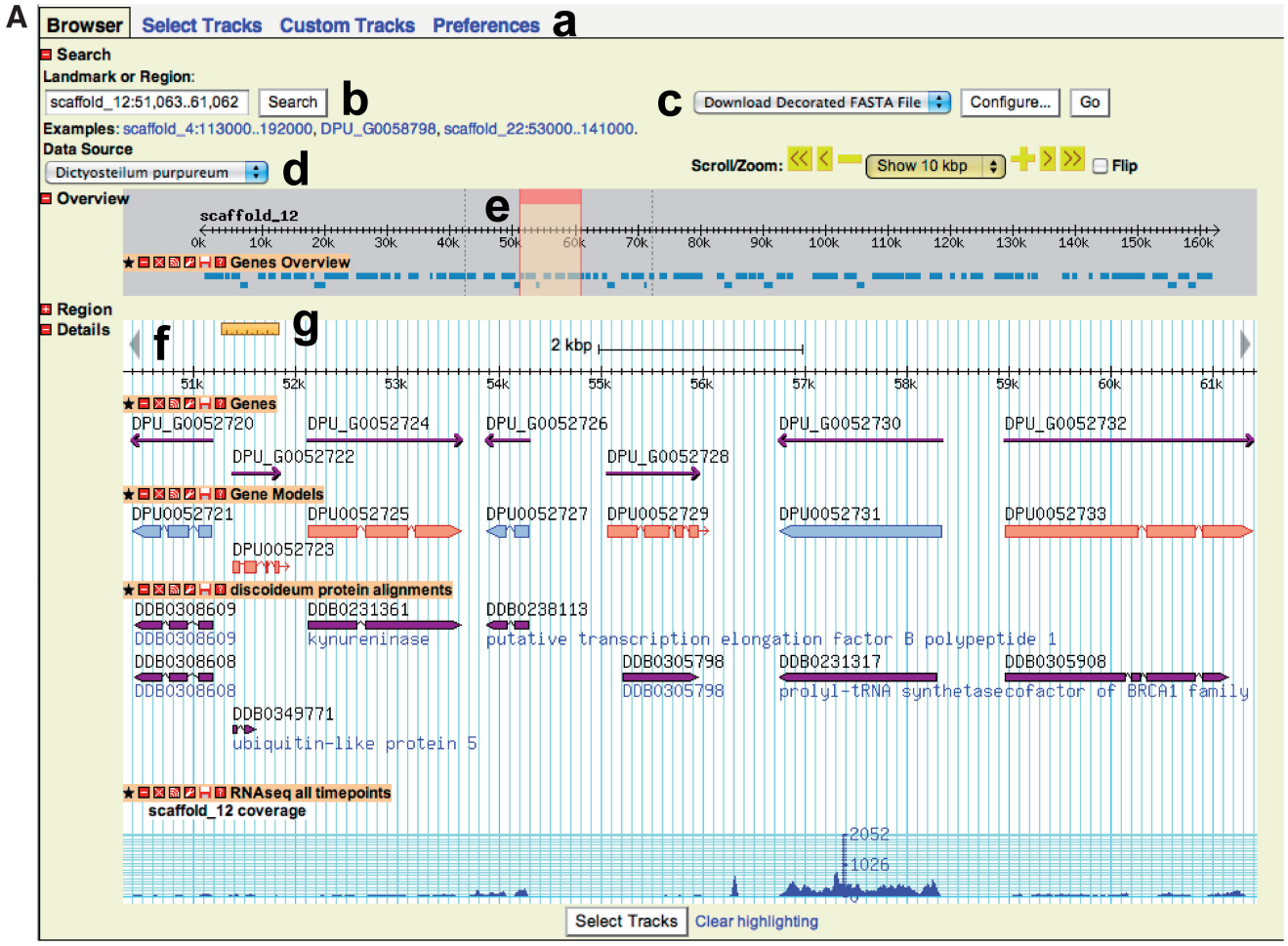


**Figure 3.** The Genome Browser. (**A**) A graphical display of 10 kb of the *D. purpureum* genome. (**a**) Tabs at the top, Select Tracks, Custom Tracks and Preferences, allow choosing specific genome features, add new ones and customize display preferences. (**b**) The search box allows to search for any gene name or ID using an auto-complete function, or to enter distinct coordinates. (**c**) The download menu contains options to display the DNA sequence or as a decorated Fasta file, the position of restriction sites or track data in GFF format. (**d**) The Data Source drop-down lets the visitor easily switch between genomes. (**e**) Overview section: this area shows a compressed view of the contig or chromosome. Moving the red box from side to side allows changing the Region Details area without reloading the page. (**f**) The Region Details area displays the individual chromosomal features in selectable tracks. Tracks shown from top to bottom are as follows: *D. purpureum* Genes, Gene Models, *D. discoideum* protein alignments and RNAseq (all time points). (**g**) A small ruler may be clicked and expanded to assess the exact alignment of sequences.

(continued)

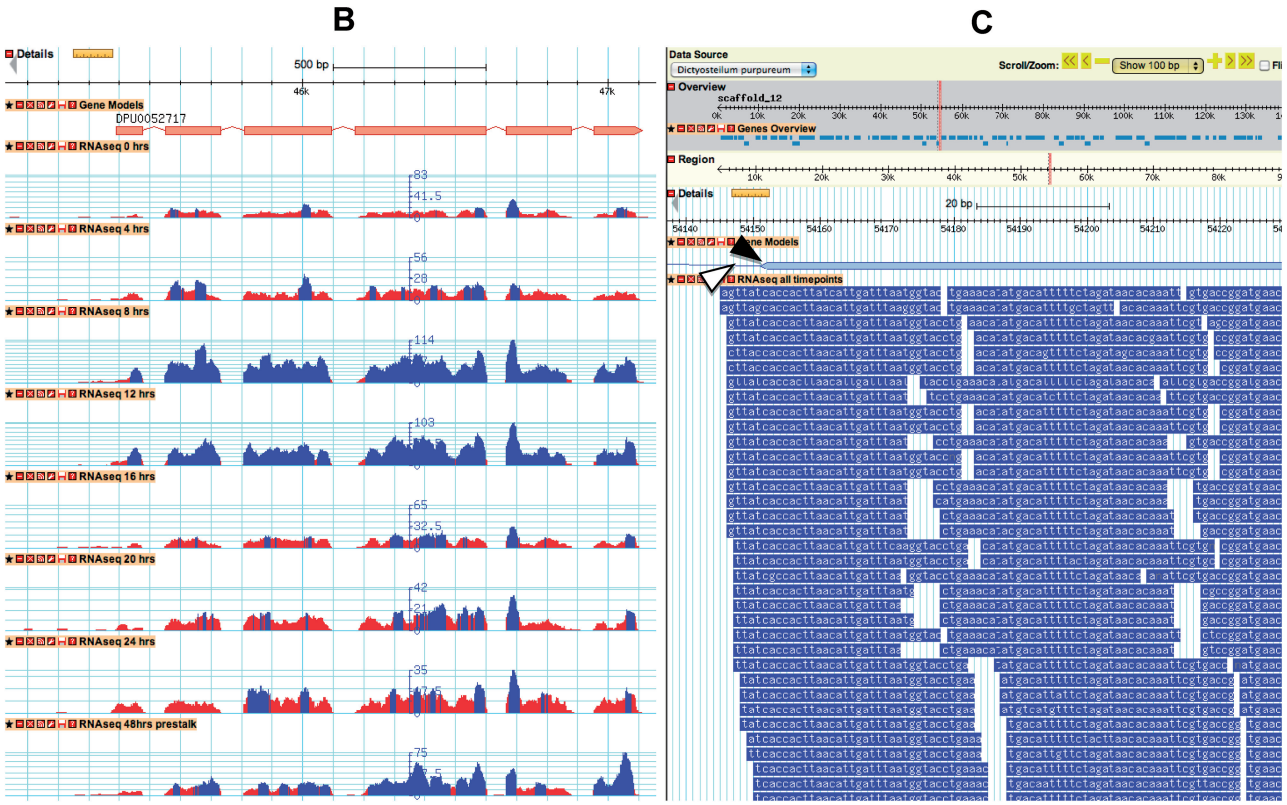


**Figure 3.** Continued.
(**B**) This is an expansion of the Region Details section, showing selected RNAseq data available for the sequence DPU0052717. The top track displays the gene model, all tracks below display RNAseq levels at different developmental time points: 0, 4, 8, 12, 16, 20, 24 and at 48 h after slug migration, pre-stalk cells only. The RNAseq data shows that this gene (the *D. purpureum* ortholog of the *D. discoideum* KsrA kinase) is up-regulated at 12 and 16 h, the stage when stalk formation begins. Pre-stalk expression is also confirmed in the 48-h prestalk track at the bottom. The blue color indicates expression above threshold, red indicates low expression below threshold. Also, keep in mind that the scale of the *y*-axis must be considered when assessing expression levels—GBrowse autoscales each track. (**C**) When the 'RNAseq all time points' are selected and zoomed in to <100 bp, all individual reads of the RNAseq data can be viewed. In this *D. purpureum* example, 100 bp at an intron/exon boundary are shown, and the RNAseq data indicate that the exon start from the automatic gene prediction (black arrow head) should be moved a few nucleotides upstream to the agT splice acceptor as indicated by the top RNAseq tracks (white arrow head).

faster lookups of genomic intervals. This has had an enormous positive impact on our ability to deliver data to the dictyBase community. Keeping the GFF3 data store in sync with the rest of dictyBase is relatively straightforward. The GFF3 data storage is populated in bulk from GFF3 dumps generated from each primary genomic database on a weekly basis, or *ad hoc* after a major genome update. The entire process is now automated.

The available dictyBase genomes are accessible directly in GBrowse through a drop-down menu (Figure 3a, d). Track selection, customization and user settings are grouped into separate tabs on top of the GBrowse page (Figure 3a, a). The GBrowse display is updated without reloading (thanks to Web 2.0 goodness) following any changes or customization from the user. The search (Figure 3a, b) now includes auto-completion that queries the database and comes back with a list of possible choices as soon as few letters are entered (gene names and IDs). A 'details multiplier' option allows rapid mouse panning across the detail panel without having the overhead of loading the image from the server (Figure 3a, e). To zoom in or out, a window of any size up to 150 kb can be drawn in the upper 'Overview' pane by holding down the mouse button. The zoomed window size in kb is shown and provides immediate feedback. The ruler in the detail panel provides precise measurement of feature locations (Figure 3a, g) and is also draggable, allowing another way to move along the chromosome/contig. The sequence displayed in the window may be downloaded as a simple Fasta file or as a decorated Fasta file (Figures 3a, c and 4) by clicking the 'Download FASTA File' button (Figure 3a, c).

## Display of RNAseq data

Another benefit of implementing the latest version of GBrowse is the ability to integrate next-generation sequencing data, particularly short read alignments in sequence file format (BAM) (13). In the GBrowse display, we have incorporated RNAseq expression data from *D. purpureum* (14). The data cover nine developmental time points that can be included as separate tracks in the GBrowse display, including one extra track that aggregates all intervals (Figure 3b). If GBrowse is zoomed into a view window of <100 bp, GBrowse 'semantic zooming' triggers the display of individual reads of that particular region (Figure 3c).

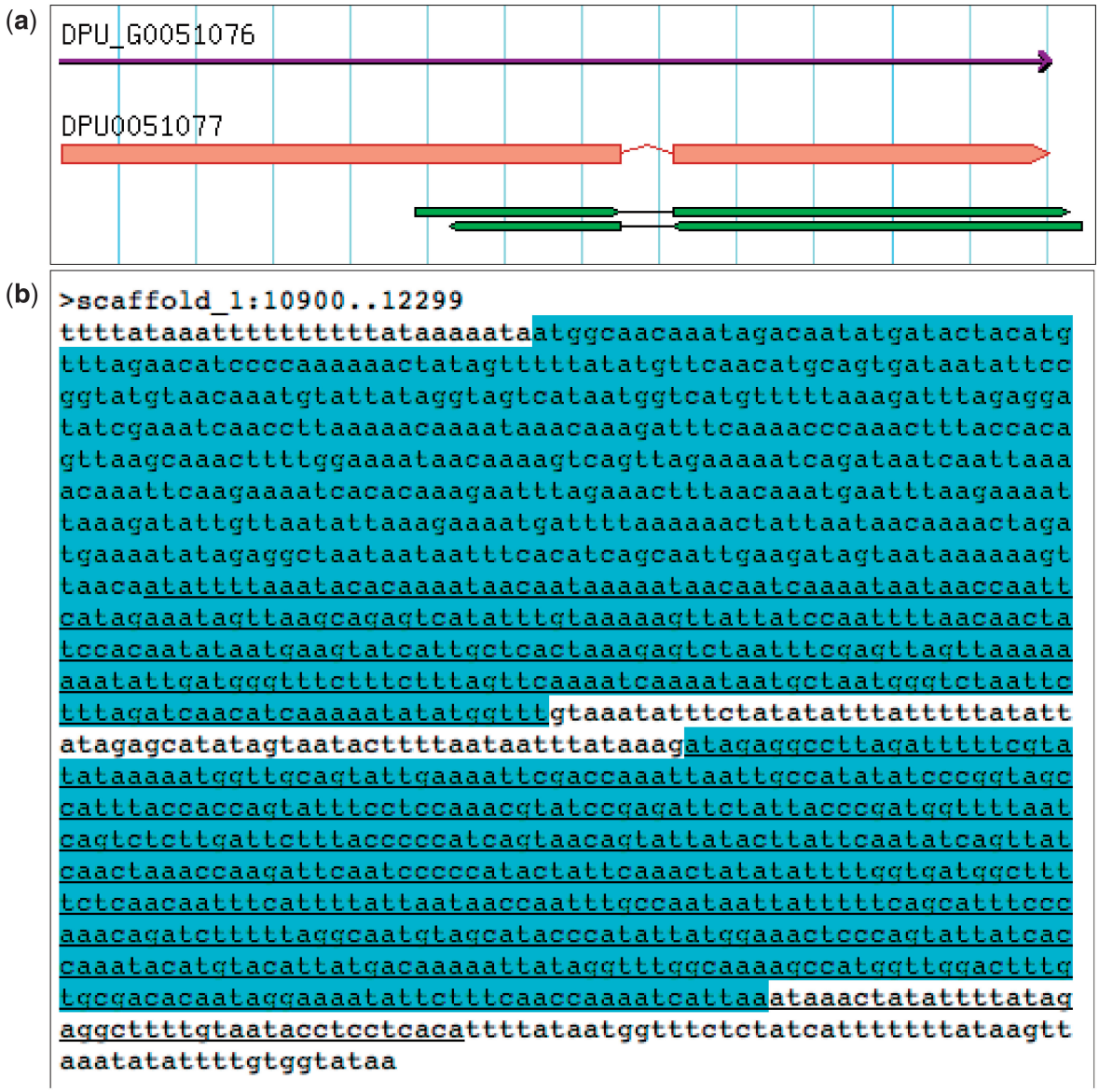


**Figure 4.** The decorated Fasta file. This decorated Fasta file is available at the 'download' drop-down (see Figure 3A). (**a**) Gene Model of gene DPU_G0051076. On top, gene track and below, gene prediction; the last track represents BLAST-aligned ESTs. (**b**) Decoration needs to be configured. In this example, the exons are highlighted and the ESTs are indicated by the underline. Note that the EST alignment supports the gene model exon/intron structure.

## CROSS-SPECIES ALIGNMENTS FOR COMPARATIVE GENOMICS

### TBLASTN alignment with *D. discoideum* proteins

We have used a genome-to-genome pairwise TBLASTN (15) to align the *D. discoideum* proteome with each of the other three genomes in the GBrowse portal (Figure 3a, e). The *D. discoideum* genome has a full set of manually curated gene models and is the high-quality reference proteome for Ameobozoa cross-species alignments. The primary goal for the proteome alignment is to provide support for predicted gene models of uncurated genomes. For the alignment, we used a moderately conservative strategy. In this approach, we soft masked the genome with the 'dustmasker' program to avoid repeats, used an *e*-value of $1 \times 10^{-10}$, added a high value for neighborhood word (999) to ensure a stringent seeding phase and lastly only included the top five alignments of every protein. The high-scoring pair of every hit then segregated into the proper genomic strand to generate the final 'best alignment' for display. The TBLASTN alignments to the three Dictyostelid genomes can be visualized in a GBrowse track named 'discoideum protein alignment' (Figure 3a, e). By default, the alignments are displayed in a segmented glyph, where all high-scoring pairs in a group are connected to represent a contiguous alignment. The direction of the track tells the genomic strand of the alignment. Any zoom level <30 kb in GBrowse will display the dictyBase ID and the gene product name of the *D. discoideum* protein if available.

**Table 1.** The dictyBase BLAST databases

| BLAST database | *D. discoideum* | *D. fasciculatum* | *D. purpureum* | *P. pallidum* | All |
|---|---|---|---|---|---|
| Protein sequences | X | X | X | X | X |
| Coding sequences | X | X | X | X | X |
| Genomic sequences[a] | X | | | | |
| EST sequences | X | | X | | |
| Non-coding sequences | X | | | | |
| Mitochondrial DNA | | | X | X | |
| Chromosomal DNA[b] | X | X | X | X | |

Every sequence can be blasted against a single species or all four available species simultaneously. Access to the BLAST search is at http://dictybase.org/tools/blast.
[a]Genomic sequences for *D. discoideum* are currently defined as coding sequences plus 1 000-bp flanking sequence on each side.
[b]Chromosomal DNA includes all six chromosomes, the mitochondrial genome and floating contigs for *D. discoideum*; for *D. purpureum*, it is all sequence scaffolds, and for *D. fasciculatum* and *P. pallidum*, it contains chromosomal supercontigs plus the mitochondrial genome.

**Table 2.** Downloads for Dictyostelids

| Downloadable items | *D. purpureum* | *D. fasciculatum* | *P. pallidum* |
|---|---|---|---|
| Nuclear chromosomal | X | X | X |
| Nuclear coding sequence | X | X | X |
| Nuclear protein sequences | X | X | X |
| Nuclear genome annotations—GFF3 | X | X | X |
| Mitochondrial chromosomal | | X | X |
| Mitochondrial genome annotations—GFF3 | | X | X |
| EST sequences | X | | |
| DPU_G–JGI (Joint Genome Institute) mapping | X | | |
| Ortholog information | X | | |

The downloads are standardized across the three species, with species-specific download pages added individually.

### Unified interface for BLAST search

The BLAST search interface in dictyBase has become a unified portal tool for doing pairwise alignments between genomes in our database. It provides all flavors of BLAST (BLASTN, TBLASTN, TBLASTX BLASTX and BLASTP), with all available types of databases (Table 1). The 'All' database combines all genomes for launching a multi-genomic pairwise comparison. As we release a new genome, three core databases (genomic, coding and protein) become available for that genome, and they are included in the 'All' data sets. In our current release, the mitochondrial and chromosomal databases (supercontigs) are available for the *D. fasciculatum* and *P. pallidum* genomes. The search is available as a stand-alone web tool, as well as being integrated into the 'Gene Pages' of every genome. The query sequence can either be entered directly or may be retrieved using a sequence identifier. The BLAST output shows a graphical summary of alignments followed by detailed information about the individual entries.

## DATA AVAILABILITY

Each genome in dictyBase has its own Download page, which is directly accessible from a drop-down in the top bar of the front page for each genome (e.g. *D. fasciculatum*: http://genomes.dictybase.org/fasciculatum/downloads). Downloadable items are governed by availability as shown in Table 2 and available in either Fasta or Generic Feature Format Version 3 (GFF3) format.

## LOOKING FORWARD

In our latest release, we have incorporated genomes from two major Dictyostelid taxonomic groups (Group 1 and Group 2), as well as setting up a scalable infrastructure for future new genomes to come. Our future plan is to complete the taxonomic coverage and add genomes from Group 3. In the process, we will also leverage the infrastructure for integrating more genomes as they are sequenced by the community. However, in recent years, next-generation sequencing technology has forced a paradigm shift in genome annotations. Unlike first-generation sequencing projects with well-studied model organisms, new genomes often lack 'gold standard' gene models and do not come with verifiable annotations. To alleviate this problem, MAKER (16) is a comprehensive tool designed for genome annotation of second-generation genome projects. We plan to use 'gold standard' *D. discoideum* annotations with MAKER to run cross-species alignments and use them as evidence to run hint-based gene predictions. The annotations will be merged with existing predictions and finally synthesized to produce evidence based quality values for further downstream annotations. We also will integrate the annotations with WebApollo (http://gmod.org/wiki/WebApollo), a web-based genome editor for opening up community-based genome annotations.

## FUNDING

National Institutes of Health (NIH) [GM64426, GM087371, HG0022]. Funding for open access charge: NIH [GM64426].

*Conflict of interest statement*. None declared.

## REFERENCES

1. Eichinger,L., Pachebat,J.A., Glöckner,G., Rajandream,M., Sucgang,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.
2. Schaap,P., Winckler,T., Nelson,M., Alvarez-Curto,E., Elgie,B., Hagiwara,H., Cavender,J., Milano-Curto,A., Rozen,D.E., Dingermann,T. *et al*. (2006) Molecular phylogeny and evolution of morphology in the social amoebas. *Science*, **314**, 661–663.
3. Chisholm,R.L., Gaudet,P., Just,E.M., Pilcher,K.E., Fey,P., Merchant,S.N. and Kibbe,W.A. (2006) dictyBase, the model organism database for Dictyostelium discoideum. *Nucleic Acids Res*., **34**, D423–D427.
4. Ogawa,S., Yoshino,R., Angata,K., Iwamoto,M., Pi,M., Kuroe,K., Matsuo,K., Morio,T., Urushihara,H., Yanagisawa,K. *et al*. (2000) The mitochondrial DNA of Dictyostelium discoideum: complete sequence, gene content and genome organization. *Mol. Gen. Genet*., **263**, 514–519.
5. Sucgang,R., Chen,G., Liu,W., Lindsay,R., Lu,J., Muzny,D., Shaulsky,G., Loomis,W., Gibbs,R. and Kuspa,A. (2003) Sequence and structure of the extrachromosomal palindrome encoding the ribosomal RNA genes in Dictyostelium. *Nucleic Acids Res*., **31**, 2361–2368.
6. Urushihara,H., Morio,T. and Tanaka,Y. (2006) The cDNA sequencing project. *Methods Mol. Biol*., **346**, 31–49.
7. Gaudet,P., Fey,P., Basu,S., Bushmanova,Y.A., Dodson,R., Sheppard,K.A., Just,E.M., Kibbe,W.A. and Chisholm,R.L. (2011) dictyBase update 2011: web 2.0 functionality and the initial steps towards a genome portal for the Amoebozoa. *Nucleic Acids Res*., **39**, D620–D624.
8. Sucgang,R., Kuo,A., Tian,X., Salerno,W., Parikh,A., Feasley,C.L., Dalin,E., Tu,H., Huang,E., Barry,K. *et al*. (2011) Comparative genomics of the social amoebae Dictyostelium discoideum and Dictyostelium purpureum. *Genome Biol*., **12**, R20.
9. Heidel,A.J., Lawal,H.M., Felder,M., Schilde,C., Helps,N.R., Tunggal,B., Rivero,F., John,U., Schleicher,M., Eichinger,L. *et al*. (2011) Phylogeny-wide analysis of social amoeba genomes highlights ancient origins for complex intercellular communication. *Genome Res*., **21**, 1882–1891.
10. Heidel,A.J. and Glöckner,G. (2008) Mitochondrial genome evolution in the social amoebae. *Mol. Biol. Evol*., **25**, 1440–1450.
11. Moore,B., Fan,G. and Eilbeck,K. (2010) SOBA: sequence ontology bioinformatics analysis. *Nucleic Acids Res*., **38**, W161–W164.
12. Stein,L.D., Mungall,C., Shu,S., Caudy,M., Mangone,M., Day,A., Nickerson,E., Stajich,J.E., Harris,T.W., Arva,A. *et al*. (2002) The generic genome browser: a building block for a model organism system database. *Genome Res*., **12**, 1599–1610.
13. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G. and Durbin,R.; 1000 Genome Project Data Processing Subgroup. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
14. Parikh,A., Miranda,E.R., Katoh-Kurasawa,M., Fuller,D., Rot,G., Zagar,L., Curk,T., Sucgang,R., Chen,R., Zupan,B. *et al*. (2010) Conserved developmental transcriptomes in evolutionarily divergent species. *Genome Biol*., **11**, R35.
15. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *Mol. Biol*., **215**, 403–410.
16. Holt,C. and Yandell,M. (2011) MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics*, **12**, 491.