Genomic database resources for *Dictyostelium discoideum*

Lisa Kreppel and Alan R. Kimmel*

Laboratory of Cellular and Developmental Biology (50/3351), National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD 20892-8028, USA

Received September 5, 2001; Revised and Accepted October 16, 2001

ABSTRACT

*Dictyostelium* is an attractive model system for the study of mechanisms basic to cellular function or complex multicellular developmental processes. Recent advances in *Dictyostelium* genomics have generated a wide spectrum of resources. However, much of the current genomic sequence information is still not currently available through GenBank or related databases. Thus, many investigators are unaware that extensive sequence data from *Dictyostelium* has been compiled, or of its availability and access. Here, we discuss progress in *Dictyostelium* genomics and gene annotation, and highlight the primary portals for sequence access, manipulation and analysis (http://genome.imb-jena.de/dictyostelium/; http://dictygenome.bcm.tmc.edu/; http://www.sanger.ac.uk/Projects/D_discoideum/; http://www.csm.biol.tsukuba.ac.jp/cDNAProj.html).

INTRODUCTION

*Dictyostelium discoideum* is a member of a unique grouping of organisms that exists at the transition of multicellularity. It has proven a powerful system for studying molecular mechanisms that underlie fundamental cellular processes, including cytokinesis, motility, phagocytosis, chemotaxis and signal transduction. In addition, many developmental pathways that regulate cell sorting, addition, many developmental pathways that regulate cell sorting, chemotaxis and morphogenesis, can be found at DictyBase (R. Chisholm, Northwestern University Medical School, IL; http://dictybase.org/) or DictyDB, an ACeDB database for *Dictyostelium* (D. Smith and W. Loomis, University of California, San Diego, CA; http://www-biology.ucsd.edu/others/dsmith/dictydb.html). Additional web sites can be found as Supplementary Material at NAR Online.

THE GENOME CONSORTIUM

The sequencing and analyses of the *Dictyostelium* genome was established as a collaborative effort among the Institute of Molecular Biotechnology at Jena (Germany), the University of Cologne (Germany), the Baylor College of Medicine (USA), the Sanger Centre (UK) and the Pasteur Institute (France). To facilitate assembly, sequencing was organized on a chromosome by chromosome shotgun basis. Individual chromosomes were separated by Edward Cox, Princeton University (USA), and the enriched chromosomal-specific DNA preparations were randomly sheared to 1–4 kb fragments and cloned into plasmid-based libraries. The libraries were distributed among the sequencing centers, Jena/Cologne for chromosomes 1, 2 and 3 and Baylor/Sanger for chromosomes 4, 5 and 6. Initial focus was on the chromosomes most easily resolved, the largest, chromosome 2, and the smallest, chromosome 6. Since all the chromosomes are of similar size, none could be purified to homogeneity. Thus, the chromosomal ‘specific’ reads were inevitably ‘contaminated’ with sequences from other chromosomes, but by continuously exchanging all primary data and clones, complete coverage has been facilitated. Sequence assembly is anchored using data obtained from a combination of YAC and HAPPY mapping. An overlapping set of ordered YACs exists for each of the *Dictyostelium* chromosomes (2). Skimmed sequences derived from the ordered YACs define landing markers to identify and assemble linked sequences. However, YACs are subject to severe artifacts of chimerism that can yield false linkage information. HAPPY mapping (3) is a completely independent in vitro approach that is functionally analogous to classical genetic linkage mapping. The HAPPY maps are being used to identify chromosomal (STS equivalent) markers at ~10 kb spacing and to eliminate chimeric and incorrectly mapped YACs (4). These

---

*To whom correspondence should be addressed. Tel: +1 301 496 3016; Fax: +1 301 496 5239; Email: ark1@helix.nih.gov*
The clustering of a new tiling set of YACs for seeding and gap-filling.

Chromosomes 1, 2 and 6 have been sequenced to an approximate overall depth of 6-fold. However, due to the non-random distribution of A+T-rich stretches that bias cloning and sequence representation, the protein coding regions have been sequenced to an 8-fold depth, with intergenic regions represented at <4-fold. Sequence analysis of the dispersed, complex, chromosomal repeat families is complete (5), as is the that of the 55 kb circular mitochondrial (mtDNA) genome (6) and the 88 kb linear, extrachromosomal rDNA palindrome (A. Kuspa, see below). Assembly of chromosomes 1, 2 and 6 is nearing completion. Sequencing of chromosomes 3, 4 and 5 is proceeding.

Approximately 66% of the entire chromosomal genome can be displayed in contigs of >2 kb, with the largest scaffolds approaching 500 kb. Annotation (see below) predicts an average spacing of one gene per 3 kb, consistent with an estimate of approximately 10,000 genes in the Dictyostelium genome and one of the highest gene densities for any eukaryote.

GenBank is still not the primary resource for Dictyostelium genomic data. Sequences will only be deposited that are unequivocal and compiled to a level of extremely high quality. However, all sequence data can be accessed and searched through the various centers listed below. All sites offer access to genomic sequences from all of the centers. In addition, primary and contig data + GenBank + ESTs + mtDNA + rDNA can be searched by web-based modes for BLAST (or BLAST-variant) analyses and comparisons (7). However, note that since all primary data are available, some may be of poor quality or even of non-Dictyostelium origin. The contig sequences are filtered of these problematic data.

Dictyostelium genomic resource sites

The University of Cologne (A.A. Noegel and L. Eichinger) and GSC Jena (G. Glockner, M. Platzer and A. Rosenethal); http://genome.imb-jena.de/dictyostelium/.

The Baylor College of Medicine (A. Kuspa and R. Gibbs); http://dictygenome.bcm.tmc.edu/.

The Sanger Centre (B. Barrell, M.-A. Rajandream and M. Quail); http://www.sanger.ac.uk/Projects/D_discoideum/.

A basic BLAST server for all Dictyostelium sequences is available through the San Diego Supercomputer Center (N. Iranfar and W.F. Loomis) at UCSD; http://dicty.sdsc.edu/. The data at this site are updated within days of the appearance of new data at the other sequencing centers.

THE TRANSCRIPTOME

The Dictyostelium cDNA Project of Japan is a collaborative effort among the Universities of Tsukuba (Y. Tanaka, H. Urushihara, T. Morio, M. Katoh and H. Kuwayama), Hokkaido (H. Ochiai and T. Saito) and Osaka (M. Maeda) to identify developmental characteristics with comparative analyses of the extensive databases developed by the cDNA and genomic sequencing projects, prediction and assembly for any selected element can be easily performed by simple scanning. Processes have been automated with moderate success using variations of Glimmer (14) and GeneFinder (15). Each of the sequencing centers provides unique analyses. Additional classification is available through UCSD/SDSC (http://dicty.sdsc.edu/annot-blast.html); Dicty Workbench portal (T.B.K. Reddy; http://dictyworkbench.sdsc.edu/) is based on Oracle and provides BLAST, RPS-BLAST and PFAM analysis data.

More than 6000 protein sequences have, thus far, been annotated, and the depth of genomic sequencing predicts that ~98% of all Dictyostelium genes are at least partially represented in the various databases. These analyses have already led to the identification and functional characterization of numerous genes and gene families shared by the metazoa. Nonetheless, >50% of the genes appear unique to Dictyostelium. Interestingly, 11 of the 113 human genes that are absent from the genomes of C.elegans, S.cerevisiae or Drosophila melanogaster, but which share sequence identity with bacterial genes, are also present in Dictyostelium (16). Many others are not present in S.cerevisiae, but are shared by Dictyostelium and the metazoa.
FUTURE PERSPECTIVES

Completion of the *Dictyostelium* genome during the next year will enable superimposition of physical and gene maps on the chromosomes. Full annotation will facilitate the design of a unigene set and production of all-gene microarray platforms. Finally, the facile ability of high frequency targeted gene disruption in *Dictyostelium* will lead to the directed mutagenesis and functional studies of every predicted gene.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

We are indebted to all of our colleagues mentioned above, who have focused much of their research on the analysis of the *Dictyostelium* genome. We are grateful for the many long conversations with Drs R. Chisholm, E. Cox, P. Devreotes, L. Eichinger, R. Firtel, R. Kay, A. Kuspa, W. Loomis, M. Maeda, T. Morio, A. Noegel, G. Shaulsky, R. Sucgang, Y. Tanaka, H. Urushihara and J. Williams with regard to this subject. Finally, we wish to thank Drs J. Brzostowski, F. Comer, L. Kim, T. Khurana, C. Parent, D. Rosel and P. Schwartzberg for their helpful discussions and suggestions.

REFERENCES