Nucleic Acids Research Advance Access published October 16, 2007

Nucleic Acids Research, 2007, 1–6
doi:10.1093/nar/gkm729

Cyclebase.org—a comprehensive multi-organism online database of cell-cycle experiments

Nicholas Paul Gauthier[1], Malene Erup Larsen[1], Rasmus Wernersson[1], Ulrik de Lichtenberg[1], Lars Juhl Jensen[2], Søren Brunak[1,*] and Thomas Skøt Jensen[1]

[1]Center for Biological Sequence Analysis, BioCentrum-DTU, Technical University of Denmark, Building 208, DK-2800 Lyngby, Denmark and [2]European Molecular Biology Laboratory, Meyerhofstrasse 1, D-69117 Heidelberg, Germany

Received August 15, 2007; Accepted September 1, 2007

ABSTRACT

The past decade has seen the publication of a large number of cell-cycle microarray studies and many more are in the pipeline. However, data from these experiments are not easy to access, combine and evaluate. We have developed a centralized database with an easy-to-use interface, Cyclebase.org, for viewing and downloading these data. The user interface facilitates searches for genes of interest as well as downloads of genome-wide results. Individual genes are displayed with graphs of expression profiles throughout the cell cycle from all available experiments. These expression profiles are normalized to a common timescale to enable inspection of the combined experimental evidence. Furthermore, state-of-the-art computational analyses provide key information on both individual experiments and combined datasets such as whether or not a gene is periodically expressed and, if so, the time of peak expression. Cyclebase is available at http://www.cyclebase.org.

INTRODUCTION

The cell division cycle is one of the most fundamental processes of life, allowing cells to multiply and faithfully pass on their genetic information to future generations. The full complexity of this process became apparent a decade ago with the first genome-wide microarray studies of the mitotic cell cycle of budding yeast (1,2). Since then, numerous other microarray studies have been published on the cell cycle of the budding yeast *Saccharomyces cerevisiae* (3,4), the fission yeast *Schizosaccharomyces pombe* (5,7), human (8) and the plant *Arabidopsis thaliana* (9).

Accessing, analyzing and comparing these many datasets has unfortunately remained difficult for a variety of reasons. First, there is no single database from which one can download all the datasets in an unified file format. The expression profiles for each experiment are often stored on individual websites. Second, the same gene identifiers are not used across datasets, making it difficult to compare expression profiles from different studies on the same organism. Third, a variety of different methods have, with varying success, been used for identifying the significantly regulated genes (1–28). The use of many different algorithms has introduced uncertainty as to which is the correct set of cell-cycle regulated genes. Fourth, new experimental studies tend to disregard already existing expression data, and thus only evaluate cell-cycle regulation based on their own experiments. Finally, general microarray repositories, analysis methods and visualization tools have by nature not been designed to meet the specific needs of the cell-cycle community.

Here, we present Cyclebase.org, a database and web resource of cell-cycle microarray expression datasets (see Table 1 for an overview of the datasets included in Cyclebase). These datasets have been mapped to common gene identifiers and normalized onto a common timescale, facilitating direct comparison of expression profiles between all experiments within an organism. The web interface provides a good visual overview of all available expression data on a given gene, as well as the results from state-of-the-art computational analyses. This interface aids the user in interpreting the combined evidence on the cell-cycle regulation of a given gene.

PRESENTING CYCLEBASE

The interface of Cyclebase is designed to make it as simple as possible for users to find and browse the genes of interest. Searching for key terms such as standard gene names (e.g. HTA2), systematic names (e.g. YBL003C)

*To whom correspondence should be addressed. Tel: +011 45 45 25 24 77; Fax: +011 45 45 93 1585; Email: brunak@cbs.dtu.dk
Present address:
Ulrik de Lichtenberg, LEO Pharma, Industriparken 55, DK-2750 Ballerup, Denmark.

Table 1. Summary of cell-cycle microarray experiments in Cyclebase

Organism	Group	Microarray	Samples	Cycles	Experiment name
Saccharomyces cerevisiae	Cho *et al.* (1)	Affymetrix	17	2	Cho-cdc28
	Spellman *et al.* (2)	Spotted	18	2	Spellman-alpha
			24	2.5	Spellman-cdc15
	de Lichtenberg *et al.* (3)	Geniom one	16	2	de Lichtenberg-cdc15
	Pramilla *et al.* (4)	Spotted	25	2	Pramilla-alpha30
		Spotted	25	2	Pramilla-alpha38
Schizosaccharomyces pombe	Rustici *et al.* (5)	Spotted	20	2	Rustici-cdc25-1
			18	2	Rustici-cdc25-2
			20	2	Rustici-elu1
			20	2	Rustici-elu2
			20	2	Rustici-elu3
	Peng *et al.* (6)	Spotted	37	2	Peng-cdc25
			32	2	Peng-elu
	Oliva *et al.* (7)	Spotted	52	3	Oliva-cdc25
			50	2.5	Oliva-eluA
			33	3	Oliva-eluB
Homo sapiens	Whitfield *et al.* (8)	Spotted	11	2	Whitfield-thythy1
			26	3	Whitfield-thythy2
			47	3	Whitfield-thythy3
			19	2	Whitfield-thynoc
Arabidopsis thaliana	Menges *et al.* (9)	Affymetrix	10	1	Menges-aph
			6	0.5	Menges-suc

The table summarizes the experiments currently in Cyclebase. Group refers to the original publication on the data. Microarray lists the technology platform used; either single-channel Affymetrix GeneChips ('Affymetrix'), two-channel spotted cDNA microarrays ('Spotted'), or *in-situ* synthesized arrays using the Geniom one platform ('Geniom one'). Samples denotes the number of samples or time points included in the experiments. Cycles is an estimate of the number of full cell cycles covered by the experiment. Experiment name refers to the label used in Cyclebase for the experiment in question. Please note that the technical replicates by Pramilla *et al.* (4) are treated as independent experiment, because this leads to better overall performance of the analysis methods.

or descriptions (e.g. histone) will produce a list of candidate genes for inspection. Genes in this list are initially sorted by their match to the search criteria and then in ascending order on the cell-cycle rank score (most periodic genes at the top). The list can be sorted on any of the other columns simply by clicking them. In addition, an advanced search page allows the user to browse for genes that match certain criteria; for example, it allows researchers to find among the 100 most periodic human genes, those that peak in S-phase.

When a gene of interest has been selected, or if a query is entered that matches only a single gene, the user is taken to the Gene Details page (Figure 1). This page is the primary interface for viewing expression profiles, key results from statistical analyses and general information about the gene in question. By default, the statistical results are based on all available experiments. Expression profiles and analysis results for the individual experiments can be accessed by clicking on a single experiment in the experiments list (Figure 1A).

To allow for inspection of the accumulated evidence for transcriptional regulation during the cell cycle, all available expression data for a gene of interest are depicted in the expression profile chart (Figure 1B). Easy comparison of different experiments is obtained by placing each profile onto a common time scale, which we have chosen to be in percent of the cell division cycle with zero corresponding to cytokinesis (M/G_1-transition) (16,29,30). Such normalization is necessary as the individual experiments vary greatly in their absolute interdivision times, depending on the experimental conditions. Subsequent alignment of the timescales is also necessary, because different experiments release the cells from different points in the cell cycle. Finally, the expression values have been normalized to a standard deviation of one over the entire experiment to further aid comparison across experiments.

To provide an unbiased and comparable assessment of the expression data, a common computational analysis framework has been applied to all datasets in the database. For every expression profile, two *P*-values are calculated that assess the significance of periodicity and regulation (16). The *P*-values are summarized across all experiments in an organism and combined to a final score, which is used to rank all genes in the genome (16) (Figure 1C). A brief explanation of the algorithms is provided in the Methods section of Cyclebase.

Based on independent benchmarking, this methodology has previously been proven to be as good as or superior to all other published methods for identifying periodically expressed genes (16,29,30). We have expanded this benchmark to also include recent methods (1,2,5–28) and experiments (Figure 2). Benchmark sets were compiled that are enriched in cell-cycle regulated genes from targets of known cell-cycle transcription factors (16,29,30). We benchmarked each method's ability to retrieve genes in these sets. Figure 2 displays the benchmarking results, which shows that the method used in Cyclebase provides clear improvements over other methods and that combining all data for an organism is, not surprisingly, superior to any single dataset analyzed on its own. Based on the benchmarks, we have selected a set of significantly periodically expressed genes within each organism

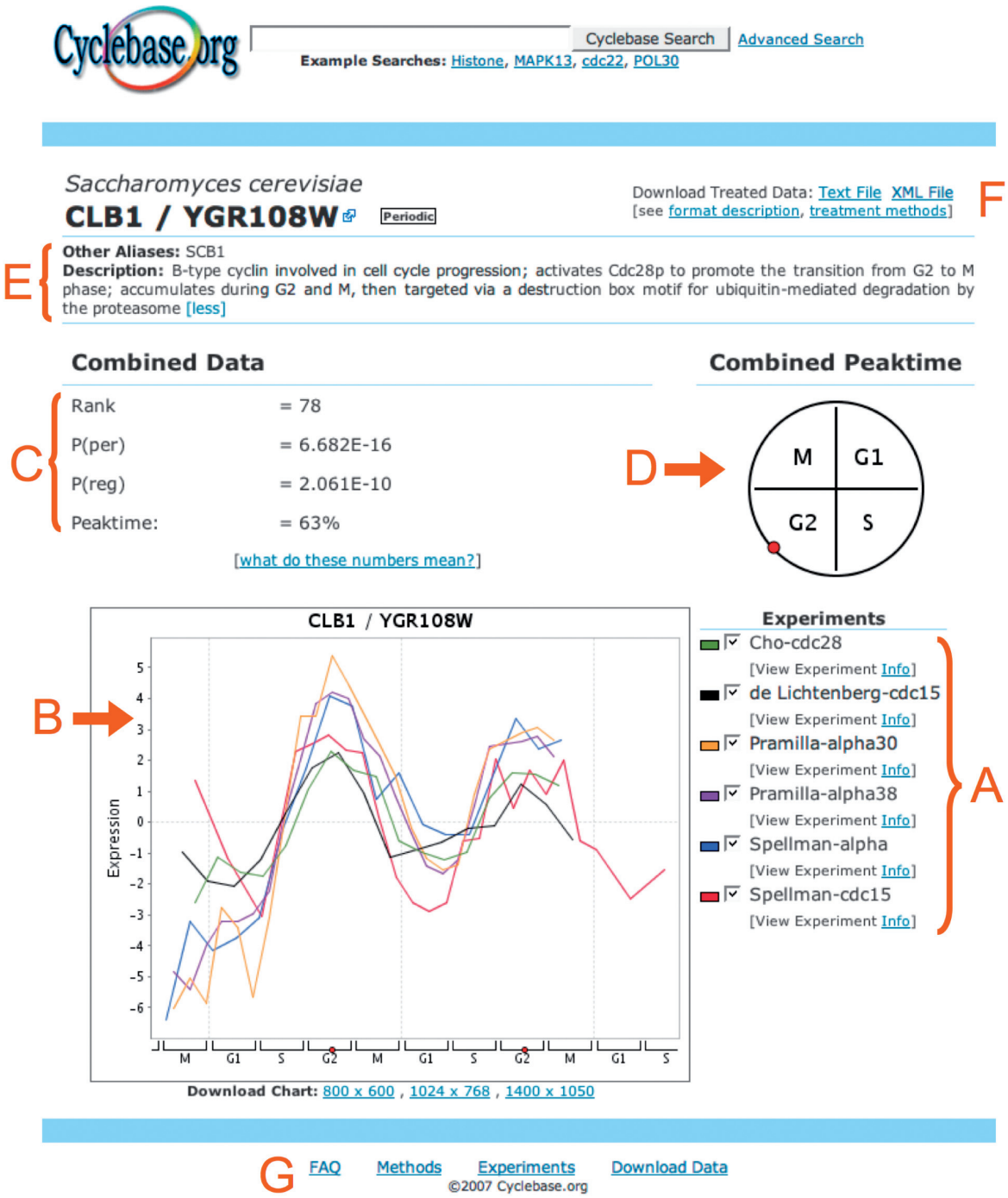

Figure 1. Screenshot for budding yeast CLB1. The figure shows the Gene Details Page for the gene CLB1 (a cyclin). (**A**) The list of experiments in which the gene is measured. Clicking any of these takes the user to another Gene Details Page with only data from that particular experiment. (**B**) Expression profile chart. The experiments are normalized and aligned onto a common time-scale (in percent of the cell cycle). The individual phases are marked along the time axis and the computationally determined peaktime is marked by a red dot. (**C**) Summary of the computational analysis based on all data available for this gene in Cyclebase. 'Rank' signifies that this is the 78th most periodic gene in budding yeast, 'P(per)' and 'P(reg)' are *P*-values that quantify the significance of periodicity and regulation, respectively, and 'peaktime' estimates how far into the cell cycle (from M/G_1) the gene is maximally expressed. (**D**) Schematic illustration of the peaktime (red dot) and phase duration. The gene CLB1 peaks 63 % into the cell cycle, corresponding to the middle of G_2 phase in budding yeast. (**E**) Gene aliases and description. (**F**) Download of data in various formats.(**G**) Database documentation and download.

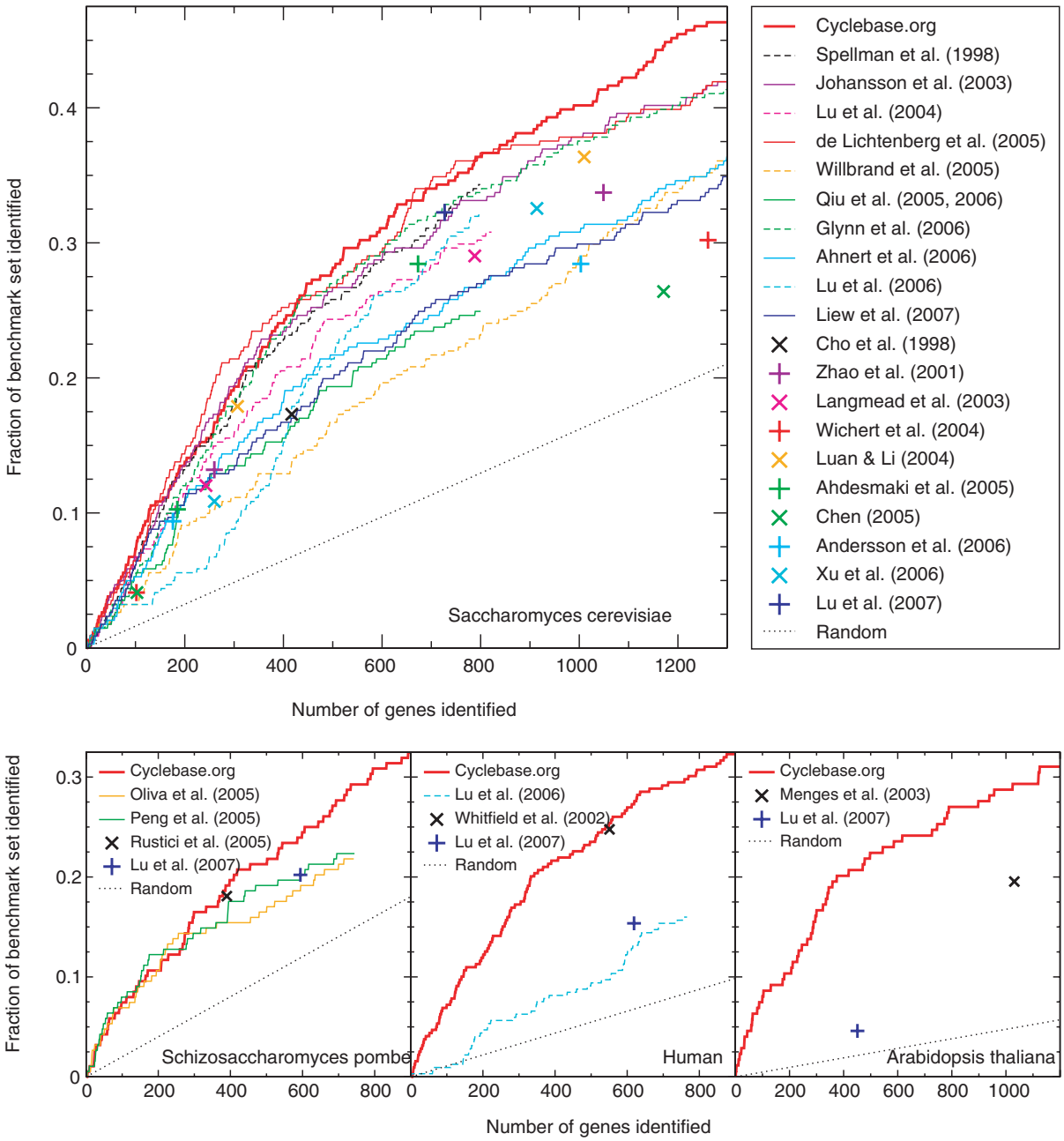

Figure 2. Benchmark of methods for identifying cell-cycle regulated genes. For each of the four organisms, a benchmark set was compiled of genes whose promoters are bound by known cell-cycle transcription factors (16,29,30), under the assumption that these genes should be highly overlapping with those that display cell-cycle regulation at the transcriptional level (i.e. periodic expression). The panels show the fraction of a benchmark set retrieved as a function of the number of genes suggested for each individual method (1,2,5–28). Better methods should therefore be towards the upper left corner of the plot. Methods which provide a ranked list of genes are displayed as a line, whereas those that only supply an unranked set of genes appear in the plots as cross mark/plus sign. The black dotted line corresponds to picking genes randomly. In all four organisms, the combined analysis of all data within an organism presented by Cyclebase outperforms all existing methods or suggested sets of periodically expressed genes. In all organisms, the curves eventually display the same slope as the random performance curve (black dotted), indicating that including more genes from this point on yields no enrichment in genes from the benchmark set.

(labeled with a small 'Periodic' icon). We found 600 periodic genes in budding yeast, 500 in fission yeast, 600 in human and 400 in the plant *A. thaliana*. For these periodic genes, we compute the 'peaktime' based on all available expression profiles (16).

The peaktime is a measure of when in the cell cycle a given gene is maximally expressed, and represents a summary of all the expression data (16). The peaktime is given as percent into the cell cycle (from when the new cell is born in cytokinesis) and is depicted as a red dot in both the expression profile chart (Figure 1B) and the peaktime chart (Figure 1D). The phase length can vary widely from organism to organism (e.g. G_2-phase occupies ~60–70% of the cell cycle in fission yeast versus

only ~25% in budding yeast), and the peaktime chart is therefore drawn differently for each species. Consequently, the peaktime values cannot be directly compared across organisms, since a specific percent (e.g. 60%) into the cell cycle may correspond to different phases in different organisms. The peaktime is only computed for genes that display periodicity and the remaining genes are labeled with 'uncertain' for the peaktime value. This label is also used if the different experiments disagree too much for a peaktime to be reliably assigned (16).

When comparing expression data across experiments, one issue is that different gene names for the same gene have been used in the different experiments. We have solved this problem by combining expression data and key results based on systematic gene identifiers. When they exist, a list of aliases is provided in the Gene Details page (Figure 1E), allowing the user to relate to the original experiment and to crosslink to external databases. The Gene Details page also contains a functional description (Figure 1E) populated from external databases (31–35) and is therefore not available for all genes.

All Cyclebase analysis results are available for download, both as values for individual genes and as whole-experiment datasets. XML and tab-delimited formats are available, both of which are fully documented on the website. Furthermore, where permission has been granted from the original authors, expression profile datasets are also available for download. Every page in Cyclebase also contains links to information about the database (FAQ and Methods), information about the individual experiments, and a link to the datasets available for download (Figure 1G).

OUTLOOK

Many more cell-cycle experiments may be performed in the future, and we encourage researchers to contact us, so that new cell-cycle experiments are analyzed consistently, and can be included in Cyclebase. As other types of large-scale experiments (e.g. metabolite information, kinase activity or protein expression) become available, it will become imperative that researchers integrate and analyze these data together with existing datasets. Cyclebase has been designed to store diverse data types from time-series experiments and we intend for Cyclebase to become a standard interface and tool for combining cell cycle datasets beyond transcriptional regulation. This would give researchers a one-stop shop for visualizing and downloading time-series events from the cell-cycle.

ACKNOWLEDGEMENTS

The authors wish to thank Hans-Henrik Stæfeldt, Kristoffer Rapacki and Peter W. Sacket for technical help with the database. This work was supported by grants from the Villum Kahn Rasmussen Foundation, the Danish Technical Research Council, as well as the BioSapiens Network of Excellence (LSHG-CT-2003-503265) funded by the European Commission FP6 Programme. Funding to pay the Open Access publication charges for this article was provided by the Villum Kahn Rasmussen Foundation.

Conflict of interest statement. None declared.

REFERENCES

1. Cho,R.J., Campbell,M.J., Winzeler,E.A., Steinmetz,L., Conway,A., Wodicka,L., Wolfsberg,T.G., Gabrielian,A.E., Landsman,D. *et al.* (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell*, **2**, 65–73.
2. Spellman,P.T., Sherlock,G., Zhang,M.Q., Iyer,V.R., Anders,K., Eisen,M.B., Brown,P.O., Botstein,D. and Futcher,B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast S. cerevisiae by microarray hybridization.*Mol. Biol. Cell*, **9**, 3273–3297.
3. de Lichtenberg,U., Wernersson,R., Jensen,T.S., Nielsen,H.B., Fausbøll,A., Schmidt,P., Hansen,F.B., Knudsen,S. and Brunak,S. (2005) New weakly expressed cell cycle-regulated genes in yeast. *Yeast*, **22**, 1191–1201.
4. Pramila,T., Wu,W., Miles,S., Noble,W.S. and Breeden,L.L. (2006) The forkhead transcription factor Hcm1 regulates chromosome segregation genes and fills the S-phase gap in the transcriptional circuitry of the cell cycle. *Genes Dev.*, **20**, 2266–2278.
5. Rustici,G., Mata,J., Kivinen,K., Lió,P., Penkett,C.J., Burns,G., Hayles,J., Brazma,A., Nurse,P *et al.* (2004) Periodic gene expression program of the fission yeast cell cycle. *Nature Genet.*, **36**, 809–817.
6. Peng,X., Karuturi,R.K., Miller,L.D., Lin,K., Jia,Y., Kondu,P., Wang,L., Wong,L.S., Liu,E.T. *et al.* (2005) Identification of cell cycle-regulated genes in fission yeast. *Mol. Biol. Cell*, **16**, 1026–1042.
7. Oliva,A., Rosebrock,A., Ferrezuelo,F., Pyne,S., Chen,H., Skiena,S., Futcher,B. and Leatherwood,J. (2005) The cell cycle-regulated genes of *Schizosaccharomyces pombe*. *PLoS Biol.*, **3**, e225.
8. Whitfield,M.L., Sherlock,G., Saldanha,A.J., Murray,J.I., Ball,C.A., Alexander,K.E., Matese,J.C., Perou,C.M., Hurt,M.M. *et al.* (2002) Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol. Biol. Cell*, 13, 1977–2000.
9. Menges,M., Hennig,L., Gruissem,W. and Murray,J.A.H. (2003) Genome-wide gene expression in an *Arabidopsis* cell suspension. *Plant Mol. Biol.*, **53**, 423–442.
10. Zhao,L.P., Prentice,R. and Breeden,L. (2001) Statistical modeling of large microarray data sets to identify stimulus-response profiles. *Proc. Natl Acad. Sci. USA,* **98**, 5631–5636.
11. Johansson,D., Lindgren,P. and Berglund,A. (2003) A multivariate approach applied to microarray data for identification of genes with cell cycle-coupled transcription. *Bioinformatics*, **19**, 467–473.
12. Langmead,C., Yan,T., McClung,C.R. and Donald,B.R. (2003) Phase-independent rhythmic analysis of genome-wide expression patterns. *J. Comput. Biol.*, **10**, 521–536.
13. Lu,X., Zhang,W., Qin,Z.S., Kwast,K.E. and Liu,J.S. (2004) Statistical resyncronization and Bayesian detection of periodically expressed genes. *Nucleic Acids Res.*, **32**, 447–455.
14. Wichert, S., Fokianos, K. and Strimmer, K. (2004) Identifying periodically expressed transcripts in microarray time series data. *Bioinformatics*, **20**, 5–20.
15. Luan,Y. and Li,H. (2004) Model-based methods for identifying periodically expressed genes based on time course microarray gene expression data. *Bioinformatics*, **20**, 332–339.
16. de Lichtenberg,U., Jensen,L.J., Fausbøl,A., Jensen,T.S., Bork,P. and Brunak,S. (2005) Comparison of computational methods for the identification of cell cycle regulated genes. *Bioinformatics*, **21**, 1164–1171.
17. Ahdesmaki,M., Lahdesmaki,H., Pearson,R., Huttunen,H. and Yli-Harja,O. (2005) Robust detection of periodic time series measured from biological systems. *BMC Bioinformatics*, **6**, 117.

18. Chen,J. (2005) Identification of significant periodic genes in microarray gene expression data. *BMC Bioinformatics*, **6**, 286.
19. Willbrand,K., Radvanyi,F., Nadal,J.-P., Thiery,J.-P. and Fink,T.M.A. (2005) Identifying genes from up-down properties of microarray expression series. *Bioinformatics*, **21**, 3859–3864.
20. Qiu,P., Wang,Z.J. and Liu,K.J.R. (2005) Tracking the herd: resynchronization analysis of cell-cycle gene expression d ata in *Saccharomyces cerevisiae*. *Conf. Proc. IEEE Eng. Med. Biol. Soc.*, **5**, 4826–4829.
21. Qiu,P., Wang,Z.J. and Liu,K.J.R. (2006) Polynomial model approach for resynchronization analysis of cell-cycle gene expression data. *Bioinformatics*, 22, 959–966.
22. Ahnert,S.E., Willbrand,K., Brown,F.C.S. and Fink,T.M.A. (2006) Unbiased pattern detection in microarray data series. *Bioinformatics*, **22**, 1471–1476.
23. Andersson,C.R., Isaksson,A. and Gustafsson,M.G. (2006) Bayesian detection of periodic mRNA time profiles without use of training examples. *BMC Bioinformatics*, **7**, 63.
24. Glynn,E.F., Chen,J. and Mushegian,A.R. (2006) Detecting periodic patterns in unevenly spaced gene expression time series using Lomb-Scargle periodograms. *Bioinformatics*, **22**, 310–316.
25. Lu,Y., Rosenfeld,R. and Bar-Joseph,Z. (2006) Identifying cycling genes by combining sequence homology and expression data. *Bioinformatics*, **22**, e314–e322.
26. Xu,H., Sung,W.-K. and Feng,L. (2006) PEM: a general statistical approach for identifying differentially expressed genes in time-course cDNA microarray experiment without replicates. In *Proc. IEEE Computer Society Bioinformatics Conference*. pp. 123–132.
27. Liew,A.W.-C., Xian,J., Wu,S., Smith,D. and Yan,H. (2007) Spectral estimation in unevenly sampled space of periodically expressed microarray time series data. *BMC Bioinformatics*, **8**, 137.
28. Lu,Y., Mahony,S., Benos,P.V., Rosenfeld,R., Simon,I., Breeden,L.L. and Bar-Joseph,Z. (2007) Combined analysis reveals a core set of cycling genes. *Genome Biol.*, **8**, R146.
29. Marguerat,S., Jensen,T.S., de Lichtenberg,U., Wilhelm,B.T., Jensen,L.J., and Bähler,J. (2006) The more the merrier: comparative analysis of microarray studies on cell cycle-regulated genes in fission yeast. *Yeast*, **23**, 261–277.
30. Jensen,L.J., Jensen,T.S., de Lichtenberg,U., Brunak,S. and Bork,P. (2006) Coevolution of transcriptional and post-translational cell-cycle regulation. *Nature*, **443**, 594–597.
31. Nash,R., Weng,S., Hitz,B., Balakrishnan,R., Christie,K.R., Costanzo,M.C., Dwight,S.S., Engel,S.R., Fisk,D.G. *et al.* (2007) Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from *Saccharomyces cerevisiae* and related sequences from other organisms. *Nucleic Acids Res.*, **35**, D468–D471.
32. Hertz-Fowler,C., Peacock,C.S., Wood,V., Aslett,M., Kerhornou,A., Mooney,P., Tivey,A., Berriman,M., Hall,N. *et al.* (2004) GeneDB: a resource for prokaryotic and eukaryotic organisms. *Nucleic Acids Res.*, **32**, D339–D343.
33. Hubbard,T.J.P., Aken,B.L., Beal,K., Ballester,B., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cunningham,F. *et al.* (2007) Ensembl 2007. *Nucleic Acids Res.*, **35**, D610–D617.
34. Rhee,S.Y., Beavis,W., Berardini,T.Z., Chen,G., Dixon,D., Doyle,A., Garcia-Hernandez,M., Huala,E., Lander,G. *et al.* (2003) The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. *Nucleic Acids Res.*, **31**, 224–228.
35. The UniProt Consortium (2007) The universal protein resource (UniProt). *Nucleic Acids Res.*, **35**, D193–D197.