Supplementary Information for:

The More the Merrier: Comparative Analysis of Microarray Studies on Cell Cycle-Regulated Genes in Fission Yeast

Samuel Marguerat[1], Thomas Skøt Jensen[2], Ulrik de Lichtenberg[2], Brian T. Wilhelm[1], Lars Juhl Jensen[3] and Jürg Bähler[1*]

[1]Fission Yeast Functional Genomics Group, Wellcome Trust Sanger Institute, Hinxton, Cambridge, CB10 1SA, UK. [2]Center for Biological Sequence Analysis, BioCentrum, Technical University of Denmark, Building 208, DK-2800 Lyngby, Denmark. [3]European Molecular Biology Laboratory, Heidelberg, Germany. *Corresponding author: Tel: +44 (0)1223-494861, Fax: +44 (0)1223-494919, Email: Jurg@sanger.ac.uk

Abstract

In this supplementary information, we describe in detail the analyses presented in the main paper. We describe how the raw data was processed and analyzed and how the benchmark sets were constructed. We advise to also visit our website: `http://www.cbs.dtu.dk/cellcycle`, where lists and data is made available for download.

Introduction

This document contain additional information and details regarding the papaer entitled "The More the Merrier: Comparative Analysis of Microarray Studies on Cell Cycle-Regulated Genes in Fission Yeast".

About the gene expression data

Rustici *et al.*

Rustici *et al.* (2004) performed five experiments in which samples were taken from synchrounsly growing cultures, labeled and hybridized to cDNA arrays along with sample from a reference asynchrounsly growing culture. Three experiments were performed using centrifugal elutriation, where cells of similar size were isolated and grown. Two experiments were performed using a cdc25-temperature sensitive mutant that arrested the cells at high temperature. By lowering the temperature a synchrounsly growning culture was obtain. In the three elutriation experiments, samples were taken at 15 minutes interval for 285 minutes, while samples were taken at 15 minutes interval for 270 minutes and 255 minutes in the two cdc25 experiments. A technical replicate of one cdc25-arrest based experiment was performed using dye-swapping. The data were normalized and the signal ratio between synchonized and unsynchronized were reported. These datasets can be downloaded from the authors webpage `http://www.sanger.ac.uk/PostGenomics/S_pombe/projects/cellcycle/` (Rustici *et al.*, 2004). Rustici *et al.* (2004) normalizeed the expression profile to an average log ratio of zero and calculated a Fourier score for each gene. Based on random shuffling of data points within an expression profile, they estimated the propability for the oscillation to occur by random. They selected genes with a p-value below 0.01 and filtered out genes with only subtle changes in expression. Based on visual inspection of the expression profiles for the remainging genes a set of 407 genes were identified as cell cycle regulated.

Peng *et al.*

Peng *et al.* (2005) performed two experiments, one based on centrifugal elutriation and one based on a cdc25-temperature sensitive mutant. In these experiments, samples were taken at 10 minutes interval for 310 minutes and 360 minutes for the elutriation- and cdc25 mutant based experiment, respectively. The samples were hybridized to cDNA micorarray and an asynchrously growing culture was used as reference. The log-signal ratio between sample and reference was reported. Each array was normalized to a median log-ratio of zero and each expression profile subjected to Gaussian smoothing. Afterwards, the time series was subjected to local zero-mean normalization, i.e. at each time point the average expression for a cell cycle was subtracted. These data can be downloaded from the Journal homepage: `http://www.molbiolcell.org/cgi/content/full/E04-04-0299/DC1`.
Genes were ranked based on a scoring scheme inspired by Spellman *et al.* (1998) (see Peng *et al.* (2005) for details). Based on random shuffling of data and etimation of the false discovery rate a set of 747 genes were found to be cell cycle regulated.

Oliva *et al.*

Oliva *et al.* (2005) made two experiments based on centrifugal elutriation and one based on a temperature sensitive cdc25-mutant. Samples were taken for 515 minutes at 10 minute intervals for the cdc25 experiment. Samples were taken for 489 minutes at 15 minute interval for one elutriation experiment, whereas samples were taken at 8/10 minutes interval for 406 minutes in the other. In all three experiments samples were hybridized to cDNA with an asynchrousnly growing culutre as reference. Standard linear normalization of total intensity was performed, except that the bottom 4000 regulated genes were normalized separately to avoid weak periodicity being induced by the normalization due to the strongly regulated genes. These data can be downloaded from `http://publications.redgreengene.com/oliva_plos_2005/`. For each experiment, a Fourier score was calculated. The profile of each gene was shuffled to produce randomized data. The observed score was compared to the distribution of random scores, and the number of standard deviations that the observed score was higher than the mean of random scores was reported as a z-score. These were combined across experiments and converted into p-values.

Reanalyzing the data

All 10 time-series experiments report the ratio or log-ratio of sample to control, i.e. signal intensity in synchronized cells compared to unsynchronized cells. The ratios reported by Rustici *et al.* (2004) were converted into log-ratios and in each time-series we centered the profiles around the mean by subtracting (in log-space) the mean expression value (this makes data analysis easier).

Identifying the interdivision time

For *S. pombe*, a set of 33 genes previously identified as periodic in small scale experiments and in the study by Rustici *et al.* (2004) were used to identify the interdivision time, *i.e.* the time it takes for a cell to go through the cell cycle. For each gene, a Fourier score was calculated:

$$F_i = \sqrt{\Big(\sum_t \sin(\omega t)\cdot x_i(t)\Big)^2 + \Big(\sum_t \cos(\omega t)\cdot x_i(t)\Big)^2}$$

where $\omega = \frac{2\cdot\pi}{T}$ and T is the interdivision time. The optimal interdivision time was found for each gene as the interdivision time that gave rise to the highest Fourier score. The distribution of the optimal interdivision times for the 33 previously identified cell cycle regulated genes were used to find the best interdivision time for each experiment.

Identifying periodically expressed transcripts

To identify periodically expressed transcripts, we then applied the permutation-based computational method described by de Lichtenberg *et al.* (2005) to each experiment individually, as well as to all data in combination. This method combines two permutation-based statistical tests in a combined score. The two tests are:

Statistical tests for regulation

The standard deviation can be easily calculated for each log-ratio profile, giving a measure of the spread of the samples around the mean. Heavily regulated genes will thus have large standard deviations, whereas genes without significant regulation display little deviation from the mean. To test for the significance of regulation, we therefore compare the observed standard deviation for each expression profile to a randomly generated background distribution. 1,000,000 random profiles were constructed by selecting at each time point the log-ratio from a randomly chosen gene. A p-value for regulation was calculated as the fraction of the simulated profiles with standard deviations equal to or larger than that observed for the real expression profile.

Statistical tests for periodicity

To estimate a p-value for periodicity, we compared the Fourier score of the observed gene expression profile for each gene to those of random permutations of the same gene. For each gene, i, a Fourier score, F_i, was computed as

$$F_i = \sqrt{\Big(\sum_t \sin(\omega t)\cdot x_i(t)\Big)^2 + \Big(\sum_t \cos(\omega t)\cdot x_i(t)\Big)^2}$$

where $\omega = \frac{2\cdot\pi}{T}$ and T is the interdivision time. Similarly, scores were calculated for 1,000,000 artificial profiles constructed by random shuffling of the data points within the expression profile of the gene in question. The p-value for periodicity was calculated as the fraction of artificial profiles with Fourier scores equal to or larger than that observed for the real expression profile.

The p-value for regulation is thus a comparison between individual genes and the global distribution, whereas the p-value for periodicity is a comparison involving only data from the gene in question.

Combined tests for regulation and periodicity

For each gene, a combined p-value of regulation was calculated by multiplying the separate p-values of regulation from each of the experiments. Analogously, a combined p-value of periodicity was calculated. Subsequently, the p-value of regulation and p-value of periodicity were multiplied to obtain the total p-value. An undesirable feature of the total p-value is that it may become very low (*i.e.* highly significant) due to only one of the tests. Genes that are strongly regulated but not periodic (or vice versa) will thus receive good scores. To address this, we multiply the total p-value with two penalty terms that weight down genes that are either not significantly regulated or not significantly periodic. The final score used for ranking is:

$$p_{total} \cdot \left[1 + \left(\frac{p_{regulation}}{0.001}\right)^2\right] \cdot \left[1 + \left(\frac{p_{periodicity}}{0.001}\right)^2\right]$$

The calculation was done for each experiment separately as well as for the combined experiments.

Avoiding overestimation

The statistical tests assume independence between neighboring measurement in an experiment - an assumption that is not entirely fulfilled in some data sets. To avoid any overestimation of the significance of the p-values, we therefore normalized all p-values within each data set by the median p-value (prior). This corresponds to assuming that there is no significant regulation or periodicity of the average gene. For the experiments from Rustici *et al.* (2004) and Oliva *et al.* (2005), these prior p-values were between 0.07 and 0.24 for periodicity and between 0.7 and 0.8 for regulation. However, for the two experiments from Peng *et al.* (2005), the prior p-value for periodicity was close to our sampling resolution (10^{-6}). We speculate that this may result from a very high correlation between neighboring time-points, and therefore split the time-series into two, calculating our statistics for each and then combining the results. As expected, this lowered the prior considerably and we therefore believe our analysis of the Peng *et al.* (2005) data to be valid and comparable to the rest of the data.

Assigning the time of peak expression

Since we approximate each expression profile by a sine wave, the time of peak expression for a gene in a single experiment is trivially defined as the time where the sine wave is maximal. We refer to this as the *peak time*(de Lichtenberg *et al.*, 2005). Due to differences in experimental conditions, the time it takes the cell to complete a cycle (the interdivision time) varies greatly between elutriation and cdc25 experiments. In order to compare the timing of peak expression across experiments, we therefore transformed the time-scales from minutes to percent of the cell cycle in each experiment by dividing with the interdivision time.
Subsequently, differences in release point of the synchronization techniques were corrected for by aligning the time scales of the ten experiments. The optimal offsets for the experiments were determined by minimizing an error function, $E1 = \sum_i E1_i$, that measures the disagreement in the time of peak expression of the same gene in different experiments:

$$E1_i \quad = \quad \sum_{exp} (w_i^{exp1} w_i^{exp2} dist(t_i^{exp1}, t_i^{exp2})^2)$$

As weights the negative logarithm of the respective total p-values were used. The function *dist* refers to the shortest possible distance between two points on a circle. The error function was minimized using a simulated annealing algorithm. To reduce computation time, each experiment was shifted before running the simulated annealing algorithm, so time zero corresponded to the peak in distribution of genes annotated with a M/G1-phase related function (Rustici *et al.*, 2004). The simulated annealing algorithm was then executed ten times, and the shifts from the run that gave rise to the lowest error function was used to align the experiments. The shifts can be seen in Table .

Experiment	Relative off set of M/G_1
Rustici Cdc25-1	71
Rustici Cdc25-2	63
Rustici Elu-1	78
Rustici Elu-2	69
Rustici Elu-3	57
Oliva Cdc25	10
Oliva Elu-1	47
Oliva Elu-2	4
Peng Cdc25	55
Peng Elu	46

Table 1: ***The shifts in percent of division time for each experiment relative to*** M/G_1 ***phase***

Combining peak times from different experiments into one is a non-trivial task, since the assignment should not be trusted in those experiments where the expression profile is not sufficiently periodic. To compensate for this, we weighted the individual peak times when computing the global, combined peak time. For each gene, a combined peak time (t_i) was calculated from the individual peak times by minimizing the following error function:

$$E2_i(t_i) \quad = \quad \sum dist(t_i^{exp1}, t_i)^2 w_i^{exp1} / W$$

where $W = \sum_{exp} w_i^{exp}$ and the weights are defined as in $E1_i$.

Distributions of peaktimes was used to aid the visual inspection of Figure 6, where distributions of at least 75 % of the members in each group (Histones, ribosome biogenesis (Tanay *et al.*, 2005), cytokinesis (Ashburner *et al.*, 2000)) were included in the figure as vertical lines.

Benchmark sets

B1 40 genes previously identified as periodically expressed in small scale experiments. The set encompasses the 35 genes used by Rustici *et al.* (2004) adding five genes that have recently been reported to be cell cycle-regulated (Alonso-Nunez *et al.*, 2005) and the gene uvi31. One gene was removed as recent small scale studies could not conferm the gene as periodically expressed.

B2 Genes whose promoters are bound by at least one of the transcription factors cdc10, res1, res2 or fkh2 based on ChIP-chip experiments performed by Brian Wilhelm (unpublished data). In case of divergently transcribed genes, where the binding is observed between the genes, both are included in the set. To obtain a benchmark set that is independent of B1 (and all other sets), we removed all genes included in B1 (50). The resulting benchmark set, B2, consists of 352 genes of which many should be expected to be cell cycle regulated, since their promoters are associated with known stage specific cell cycle transcription factors.

B3 Genes that are differentially regulated in response to knock-out or over-expression of ace2, cdc10, sep1, as well as in a hydroxyurea block experiment (Rustici *et al.*, 2004). Details on these experiments can be found at `http://www.sanger.ac.uk/PostGenomics/S_pombe/projects/cellcycle`. To avoid overlap between the benchmark sets all genes already contained in B1 and B2 were removed. This left 188 genes, of which many should be expected to be transcriptionally regulated during the cell cycle.

Systematic gene names

Gene names in each experiments, previously proposed lists of cell cycle regulated genes, and benchmark sets were converted into systematic names from geneDB (Hertz-Fowler *et al.*, 2004) to allow for systematic comparison and benchmarking. The gene name mapping file was downloaded from geneDB and synonyms gene names were changed into the corresponding systematic name. In the case were there were multiple expression profiles for the same gene in an experiment, the best score was reported. In benchmark sets and lists of previously proposed genes, genes were only included if a systematic name from geneDB could be found.

Avialability

Results and benchmark sets are available at `http://www.cbs.dtu.dk/cellcycle`.

References

Alonso-Nunez, M. L., An, H., Martin-Cuadrado, A. B., Mehta, S., Petit, C., Sipiczki, M., del Rey, F., Gould, K. L., & de Aldana, C. R. (2005). Ace2p controls the expression of genes required for cell separation in Schizosaccharomyces pombe. Mol. Biol. Cell *16(4)*, 2003–17.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., anmd J. T. Eppig, S. S. D., Harris, M. A., Hill, D. P., Issel-Tarber, L., Kasarskis, A., LEwis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene ontology: tool for the unification of biology. Nature Genetics *25*, 24–29.

de Lichtenberg, U., Jensen, L., ll, A. F., Jensen, T., Bork, P., & Brunak, S. (2005). Comparison of computational methods for the identification of cell cycle-regulated genes. Bioinformatics *21(7)*, 1164–71.

Hertz-Fowler, C., Peacock, C. S., Wood, V., Aslett, M., Kerhornou, A., Mooney, P., Tivey, A., Berriman, M., Hall, N., Rutherford, K., Parkhill, J., Ivens, A. C., Rajandream, M. A., & Barrell, B. (2004). Genedb: a resource for prokaryotic and eukaryotic organisms. Nuclei Acids Res. *32(Database issue)*, D339–43.

Oliva, A., Rosebrock, A., Ferrezuelo, F., Pyne, S., Chen, H., Skiena, S., Futcher, B., & Leatherwood, J. (2005). The cell cycle-regulated genes of Schizosaccharomyces pombe. PLoS Biology *3(7)*, Epub.

Peng, X., Karuturi, R. K., Miller, L. D., Lin, K., Jia, Y., Kondu, P., Wang, L., Wong, L. S., Liu, E. T., Balasubramanian, M. K., & Liu, J. (2005). Identification of cell cycle-regulated genes in fission yeast. Molecular Biology of the Cell *16(3)*, 1026–1042.

Rustici, G., Mata, J., Kivinen, K., Lio, P., Penkett, C. J., Burns, G., Hayles, J., Brazma, A., Nurse, P., & Bahler, J. (2004). Periodic gene expression program of the fission yeast cell cycle. Nature Genetics *36(8)*, 809–817.

Spellman, P., Sherlock, G., Zhang, M., Iyer, V., Anders, K., Eisen, M., Brown, P., Botstein, D., & Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast *S. cerevisiae* by microarray hybridization. Mol. Biol. Cell *9*, 3273–3297.

Tanay, A., Regev, A., & Shamir, R. (2005). Conservation and evolvability in regulatory networks: the evolution of ribosomal regulation in yeast. Proc Natl Acad Sci U S A *102(20)*, 7203–8.