The Fourth International Conference on Computational Systems Biology (ISB2010)
Suzhou, China, September 9–11, 2010

Predicting Cell Cycle Genes from E-MAP Profiles by Integrating Multiple Types of Data

Wang Lin[1] Hou Lin[1,2] Minping Qian[1,2]
Fangting Li[1,4,*] Minghua Deng[1,2,3*]

[1] Center for Theoretical Biology, Peking University, Beijing 100871
[2] LMAM, School of Mathematical Sciences, Peking University, Beijing 100871
[3] Center for Statistical Science, Peking University, Beijing 100871
[4] School of Physics, Peking University, Beijing 100871
[*]Corresponding authors: dengmh@pku.edu.cn, li_fangting@pku.edu.cn

Abstract Interactions between genes and proteins can be revealed by multiple experimental platforms. The derived interaction networks can be utilized to discover novel genes involved in specific biological process. E-MAP is an experimental platform to measure genetic interactions in a genome-wide scale, which successfully recovered known pathways and also revealed novel protein complexes in S. cerevisiae. However, E-MAP data can be quite noisy, and it is of great challenge to make reliable biological inference based on it. Here we propose a novel approach which aims to discover genes involved in the cell cycle process in S. cerevisiae by combining E-MAP data with other sources of data, such as gene expression, protein phosphorylation, and transcription factor (TF)–DNA binding. From an E-MAP screen with 35 query genes, we predict three unknown genes (YPL158C, YPR174C and YJR054W) as potential cell cycle genes. Our strategy can be applied to other biology processes as well.

Keywords: E-MAP; Microarray; Chip-Chip; Protein phosphorylation; Integrate; Cell Cycle

1 Introduction

Two genes are defined as "genetically interacting" when the phenotype generated as the result of mutations in both genes is unexpectedly not just a combination of the phenotypes of the two single mutants. Deciphering genetic interaction networks not only reveals the wiring diagram of biology processes, but also predict novel genes involved in certain processes. Recently, several high-throughput technologies have been developed to identify genetic interactions in the genome scale, including Synthetic genetic array (SGA) technology [18], diploid based synthetic lethality analysis on microarrays (dSLAM) [19], and Epistatic miniarray profile (E-MAP) [8]. The former two approaches aim to identify synthetic sick or lethal interactions (also known as negative interactions), which means the double mutant is lethal or sicker compared with the corresponding single mutants. Assuming that the expected phenotype of a double mutation is the additional effects of the single mutations, E-MAP, an adaptation of SGA, gains power by identifying negative interactions as

well as positive genetic interactions, which indicates the double mutant is healthier than expected.

Here, we exploited the E-MAP methodology to discover novel genes involved in the cell cycle process in the budding yeast. The distinct advantage of using E-MAP is the potential to discover functionally associated genes which are not physically interacting. These associations are unlikely to be revealed by physical interaction assays such as yeast two hybrid system and DNA-binding microarrays.

Despite the superiority of E-MAP, interpretation of the data is still challenging. First, genetic interactions occur both between and within functional modules. Thus, the function of a gene cannot be determined as the function of its interacting partners. Second, E-MAP suffers from high false positive rate and high false negative rate. In another word, the data is quite noisy, which makes it difficult for inference. In this case, integration of external information, such as gene expression, is necessary for identifying novel genes involved in the cell cycle process.

Several methods have been developed to integrate multiple types of data, including mRNA expression, chip-chip, physical interaction and protein phosphorylation, to infer a transcription regulatory network in eQTL analysis [20-23]. In this paper, we exploited a novo strategy to construct a specific network through integrating genetic interaction network with other genomic data, and applied it to the cell cycle process in the budding yeast. Our study provides new insights into genes and interactions involved in the cell cycle process.

2 Results

2.1 Construct a potential cell cycle gene set

Our strategy to integrate multiple types of data can be visualized by a flow-chart (Figure 1). In the first step, we want extend the known cell cycle gene set to include potential cell cycle gene sets. E-MAP method was adopted to identify genes which genetically interact with the known ones. We screened 35 known cell cycle genes (KCCGs) against a library of 1536 test strains in S. *cerevisiae*. Using the S-score cutoff S>2.5 or S<-3.5, 850 test genes with 1925 significant genetic interactions were selected. (See Methods and materials)

However, we cannot claim these 850 test genes are involved in cell cycle process for two reasons: (1) the known cell cycle genes could also take part in other biological processes; (2)genetic interactions also exist between genes participating different pathways [24]To concentrate on cell cycle genes, a possible solution is to use co-expression, especially co-expression in time course microarrays, as a filtering criterion, since gene pairs co-expressed in time course expression across one or two period are more likely to be co-functional in the cell cycle process.

Hence, we filtered the 1925 gene pairs derived from the previous step by the co-expression constrain. Using 8 groups of time-course expression data sets from 4 publications, we defined the expression correlation of two genes (See Methods and materials). We have found that when a higher cutoff of co-expression level is applied, cell cycle related functions are more likely to be enriched in the remaining genes after filtering (Figure 2). When the cutoff is higher than 0.85, there is at least one cell cycle

related functional category enriched. Thus genes with correlation greater than 0.85 are selected as the potential set of cell cycle genes (PCCGs). 178 test genes with 257 significant interactions passed the two step filtering.

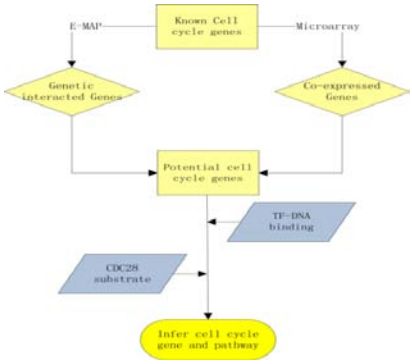


Figure 1: Overview of our strategy to integrate multiple types of data to identify cell cycle gene and infer regulatory pathways.

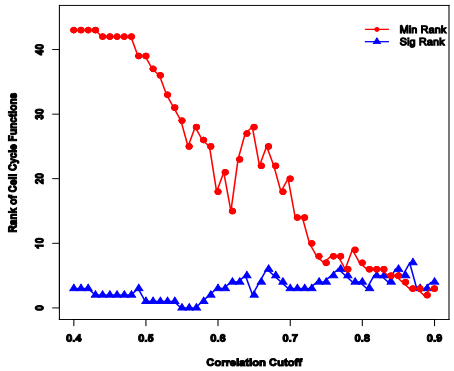


Figure 2: In order to demonstrate that co-expression is a filtering criterion for identifying cell cycle genes, we studied the enrichment of cell cycle genes at different level of co-expression. First, gene pairs with S score under certain cutoff are excluded from further analysis. Then, among the remaining gene pairs, we extracted the sub-sets in which genes are co-expressed at certain level. Varying the level of co-expression, gene set enrichment analysis were performed on all biological process terms in GO, and we investigated the rank of cell cycle related GO terms in the derived p-values (Fig 2). As the co-expression level increases, the minimum rank of the cell cycle related GO terms decreases. At correlation above 0.85, the filtered sub-sets are enriched in at least one cell cycle related GO terms.

2.2 A cell cycle transcriptional network based on the PCCGs and KCCGs

In the next step, we would search for main TFs which regulate both the PCCGs and the KCCGs, and then construct the transcriptional regulatory network. In previous studies, TF-DNA binding data (Chip-Chip data) is usually combined with expression information to construct the regulatory network. However, it is unnecessary for TF-target pairs to be co-expressed at a high level. Instead of setting constrains on the co-expression level, we required periodical expression for a TF to be included. Since genes involved in the cell cycle process are expressed periodically, it is reasonable to assume periodicity of their transcriptional regulators. In addition, we also assume that the regulatory targets of a TF involved in the cell cycle should be enriched for the known cell cycle gene. Hence, with known cell cycle genes combined from PCCGs and KCCGs, TFs which are enriched for cell cycle neighbors in transcription network were identified.

The significance of periodicity and enrichment of cell cycle genes are calculated (Methods and materials). Both approaches tend to select TFs which are known to be involved in cell cycle regulation according to MIPS functional annotation (Figure 3).

The two methods are consistent since most of the known cell cycle TFs rank top in both cases. In the meanwhile, some TFs are ranked differently (Table S1). For example, Mcm1 has the rank 5/130 in the enrichment test (ET), however it ranks 124/130 in periodic test (PT) which means its expression does not show periodicity. As we know, Mcm1 regulates different phases during the cell cycle [1,2], and its expression will not be periodic. However, many of its neighbor genes in the transcriptional network are cell cycle genes, so that we could identify it in ET. Similar to Mcm1, Skn7 ranks 22 and 114 respectively in ET and PT. In contrast, Yox1 have the rank 4 in PT but rank 86 in ET. One possible explanation is that the PCCGs and KCCGs cover a limited part of cell cycle related genes and some targets of Yox1 are missing in this set. Other examples like Hcm1, Yhp1 have similar situations.

Based on the analysis above, a TF that is significant in either test should be included. Hence, we use the multiplication of the two ranks as an index, and use its rank to evaluate the priority order (see Additional material for details).

To determine how many TFs should be involved, we examined the coverage rates of TFs. The coverage rate is evaluated at two levels: the fraction of genes which are neighbors of the selected TFs (STs) in the PCCGs and KCCGs; the fraction of gene pairs which are co-regulated by any one of the STs . 176/202 genes in the PCCGs and KCCGs are involved in the chip-chip data set (at least one TF can bind to them), and 100 gene pairs which are both genetically interacting and co-expressed can be simultaneously bound by the same TF.

We noticed that when the top 25 TFs are selected, most of the 176 genes and 100 gene pairs (85% and 94%) could be covered (Figure 4). The cover rate increases quite slowly when more TFs are selected. Thus we used these 25 TFs to construct the transcriptional network based on the PCCGs and KCCGs.

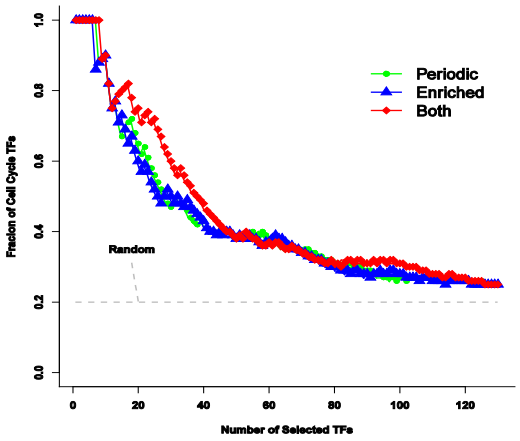


Figure 3: Comparison of the fraction of cell cycle TFs selected by different standards： enriched rank, periodic rank, and their combination (the rank of multiplication of the two ranks). All the three methods we adopted tend to select more cell cycle TFs than random (Fraction=37/183=0.2). The enriched test and periodic test show similar power, while the combination of them can increase the power.

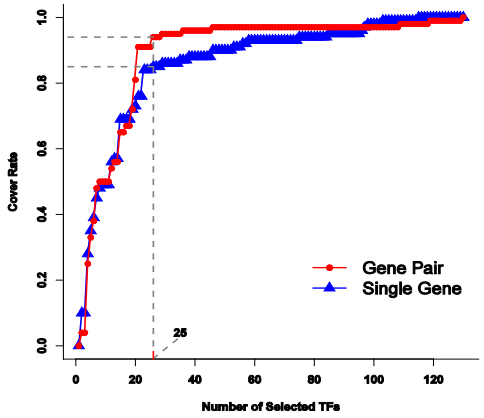


Figure 4: The percentage of single gene and gene pair coverage corresponding to the number of TFs selected.

2.3 Recover of known genetic interactions with our E-MAP

Akin to previous work**[3]**, we first tested the sensitivity and precision of the E-MAP data (see Additional materials). When compared to genetic interactions in BIOGRID, both the positive and negative interactions are of good precision ($\text{p-value} < 10^{-50}$). However, the significance level of precision didn't increase when

the co-expression test was applied (Table S2). It indicates that co-expression does not provide extra information on genetic interaction. This can also be proved when the distribution of S-score of all gene pairs and those of highly co-expressed pairs are compared (Figure S1). It means that genetic interaction and co-expression are conditionally independent, and they play different roles in our strategy. E-MAP provides the causal relationship, while co-expression ensures the genes are co-functional, here in our assay to be cell cycle related. Hence our strategy can find potential cell cycle genes and their relationships with other cell cycle genes, and we can construct a reliable network.

We also compared our S-score with previously published SGA data **[3]**. Significantly interacting gene pairs show obvious correlation between the two data sets (r=0.64, Figure S2). All these validated the accuracy of our measurements.

2.4 Enrichment for CDC28 substrates

Since cell-cycle events are controlled by cyclin-dependent kinases (CDKs), we checked if the Cdk1(CDC28) substrates are enriched in our PCCGs and STs. As expected, both the PCCGs and STs are enriched with CDC28 substrates ($p_G = 2 \times 10^{-4}, p_T = 5 \times 10^{-6}$, Table S4). This also supports that the PCCGs and STs are both cell cycle related with high confidence.

2.5 Indirect evidence supports the STs form a cooperative transcriptional network

From wild type vs. TF mutant microarrays data, we could get the indirect information of transcriptional regulation. Since this is independant evidence which describes transcriptional network, it is utilized to validate the network we constructed.

In ChIP-chip data, we found that a great part of TFs (170/183) could regulate each other within 10 steps. It means that any TF can be indirectly connected to the targets of the other TFs. Between our 25 TFs and the 176 target genes, there are 108 indirect TF-target pairs. By using all 183 TFs' transcriptional relations in the ChIP-chip data set, all these pairs can be connected within three steps (Figure 5).

Indirect evidence indicates connections in the transcriptional network, and we hyperthese the 25 TFs can form a sub-network that is able to explain their indirect connections with the 176 targets. If so, it will be a strong support for the authenticity of this sub-network.

We test the fraction of indirect TF-TG pairs which can be connected by using only the 25 TFs' transcriptional relationships. The result (Figure 5) shows that the sub-network can explain at most 94.4% (102/108) of the indirect relations, and it strongly support the conclusion above.

2.6 Clustering of the constructed transcriptional network

As mentioned above, genetic interactions may be missed in E-MAP analysis, which leads to the result that clustering of S-score profile cannot always show reasonable results. Hence, we adopted the transcriptional profile to do cluster analysis to understand the structure of the PCCGs and KCCGs (Figure S3).

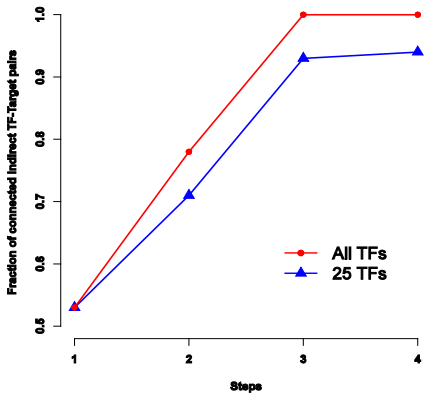


Figure 5: This figure shows all indirect interactions between the 25 TFs and their targets can be connected within 3 steps by using all TFs. If just using the 25 TFs, 94.4% (102/108) of these interactions can be connected. It illustrates that the 25 TFs can form a well connected sub-network.

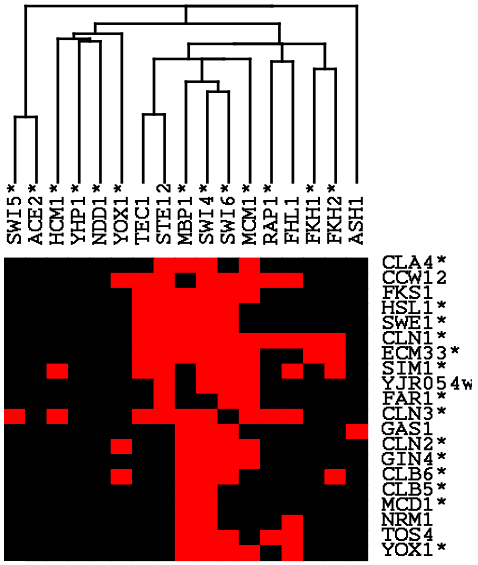


Figure 6: An example of sub-clusters in the transcriptional network. Red box means the corresponding TF/gene (Top) regulates or is regulated by the target gene/TF (right), and black box means there is no transcriptional connections between them. In this cluster, target set is enriched with known cell cycle genes (with * annotation).

The clustering result of the TFs are consistent with prior knowledge, and several cooperating TF clusters, such as Ace2/Swi5, Fkh1/Fkh2, Swi4/Swi6/MBP1, and Yhp1/Yox1 are recognized. This indicates TFs with similar function share neighbors

in the network.

In addition, some target clusters are significantly enriched with known cell cycle genes (Figure 6). These provide meaningful information to infer the function of genes.

2.7 Potential cell cycle related genes

Results of the analysis above are summarized in Table S6. Three genes with unknown function are identified as potential cell cycle genes.

The first one is YPL158C, which genetically interacts with PCL9, AMN1 and BUD4. All these four genes are regulated by known TFs in M phase (including G2/M and M/G1) (Figure 7A). The expression data shows that YPL158C, PCL9 and AMN1 are simultaneously expressed and their peak value of expression is later then ACE2 and SWI5 which are transcriptional regulators of them, and also BUD4 (Figure 7B). This is consistent with the regulatory network, because BUD4, ACE2, and SWI5 are mainly regulated by FKH1/2 and MCM1, while YPL158C is mainly regulated by ACE2 and SWI5. Table S7 lists the description of the functions of these genes in SGD. All of them are acting in M phase or early G1 phase. Based on these observations, YPL158C is possibly involved in M phase and co-operate with PCL9, AMN1 and BUD4.

The second is YPR174C, which is genetically interacts with CLN3 and potential substrates of CDC28. YPR174C and CLN3 are co-regulated by MBP1 and XBP1, another known cell cycle TF, which has not been selected in the procedure above because of low periodic rank (rank in PT: 120, rank in ET: 11, Final rank: 34). According to the description in SGA, XBP1 is a member of SWI4/MBP1 family. Since MBP1 and XBP1 do not have significant periodic expression, we compared the expression of SWI4 with CLN3 and YPR174C. We found SWI4 and YPR174C are significantly co-expressed, and CLN3 and YPR174C also showed a co-expression pattern, but with two time points lagged (Figure 7A, 7C). Now it's convincing to consider YPR174C to be involved in G1 phase in cell cycle process since all other related genes are mainly acting in this phase.

The last one we will introduce is YJR054W, which is genetically interacts with BUD4 and potential substrates of CDC28. As showed in figure 6A, BUD4 is regulated by MCM1. YJR054W is regulated by SWI4/SWI6 which are also regulated by MCM1. In the expression data (Figure 7A, 7D) we found that the expressions of SWI4 and BUD4 are highly negatively correlated. This can be explained by that MCM1 participates in the formation of both repressor and activator complexes, and SWI4 and BUD4 may be regulated by different complexes. The expression of YJR054W is similar to SWI4 and slightly lagged, which supports the regulation between them. Since BUD4 can influence the next round of budding and SWI4/6 mainly regulate the G1 phase, YJR054W may be involved in M/G1 phase and co-operate with BUD4.

3 Discussion

We developed a novel approach to discover genes participating in the cell cycle process in S. *cerevisiae*. Through integration of multiple sources of genomic data, we

Figure 7A

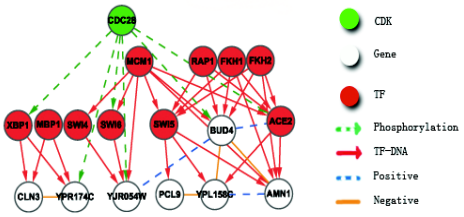


Figure 7D

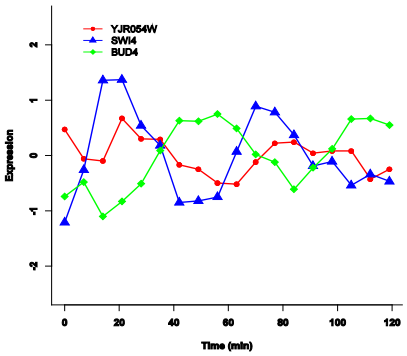
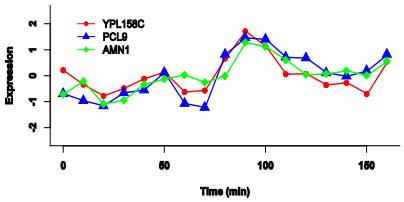


Figure 7B



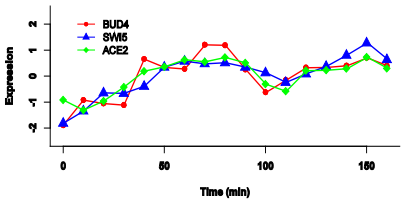
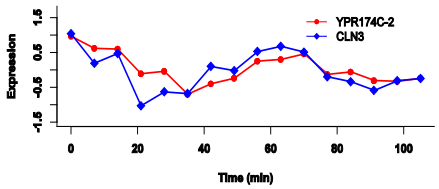


Figure 7C



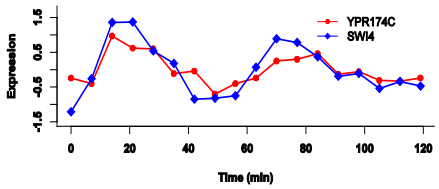


Figure 7 Analysis of potential cell cycle related genes. (A) The sub-network we constructed which contains three genes with unknown function. (B, C) The expression of YPL158C related genes in Cho-cdc28 experiment. The genes can be separated into two groups. The first includes YPL158C, PCL9 and AMN1 (upper), and the rest forms the other (below). Genes in each group are co-expressed, and the peak value of the first group is lagged comparing to the second one. (C) The expression of YJR054W, SWI4 and BUD4 in Spellman-alpha experiment. SWI4 and BUD4 are significantly negatively correlated (upper), and YJR054W looks similar to SWI4 but is appropriately one time point lagged. (D) The expression of YPR174C, CLN3 and SWI4 in Spellman-alpha experiment. It's obvious that YPR174C is co-expressed with SWI4. When YPR174C is left shifted for 2 time points (Noted as YPR174C-2), it's also co-expressed with CLN3.

reconstructed a sub-network from known cell cycle genes and predicted genes in the sub-network as potential cell cycle genes. Furthermore, in order to find how these genes are regulated, a transcriptional network was constructed based on ChIP-chip experiments. A transcription factor is considered to regulate the cell cycle process, if it is periodically expressed or its neighbors in the transcriptional network are enriched with cell cycle genes. Three genes with unknown function are identified as novel cell cycle genes.

Our approach integrates the genetic interaction network, co-expression network, and transcriptional network, and performed well in predicting cell cycle genes. However, there are other types of networks, such as protein physical interaction networks, which are informative for the prediction of gene function. We believe the efficiency of prediction can be increased when such data are integrated in a reasonable framework.

Although the current study focused on the cell cycle process, our approach is not limited and can be easily applied to other biological processes with data available.

4 Methods and materials

4.1 E-MAP experiment and data analysis

The S-score in E-MAP was derived as previously described [8]. We used 35 cell cycle query genes (Table S5) against a library of 1536 test genes. The query genes are selected because they play important roles in different phases in the cell cycle process. However, the analyzing framework we developed is not affected by the selection of the queries genes, and it can be applied to other processes as well. Finally 1387 test genes remained in our analysis, and there were 48545 S-scores with 613 missing values. To define the strong genetic interactions, we introduced a cutoff of $S<-3.5$ for negative interactions and $S>2.5$ for positive interactions. This will select 636 positive and 1278 negative interacting gene pairs. We adopted this cutoff to highlight the top 5% significant interactions.

4.2 Time course expression data and definition of correlation

We use eight time course microarray experiments from four previously published work to do the co-expression analysis [4,5,6,7]. The data is downloaded from the supplementary or the authors' website.

To measure the similarity between the time course expression profile of two genes, we used the time-lagged correlation [9]. For multiple experiments, we adopt a loose definition for correlation between two genes - the maximum time lag correlation score in all the eight experiments. It means if two genes show high correlation in one experiment, they are considered as co-expressed. The reason we use this loose is that as we have already had a stringent constrain in E-MAP analysis, the interactions are reliable even if two genes only show co-expression in one experiment.

4.3 Transcriptional Regulation and CDC28 substrate data sets

The Chip-Chip data and indirect wild type vs. TF mutant microarrays data are downloaded from YeastRact (http://www.yeastract.com, [11, 12]). Among 183 TFs in our data set, 37 are annotated as cell cycle related. Since only 130/183 TFs have at least one target in the PCCGs and KCCGs, we just did the relative analysis to these ones.

The CDC28 substrate data set was downloaded from the supplementary of previous published work which identified around 200 substrates [15].

4.4 Enrichment for cell cycle genes and TFs

In order to check whether the PCCGs are enriched with known cell cycle genes, we calculated the proportion of genes which are annotated to participate in the cell cycle process (in MIPs) in the PCCGs, and use the cumulative hypergeometric distribution to define the p-values. About 1/3 of the PCCGs (57/178) are known cell cycle genes ($p = 3 \times 10^{-4}$, Table S3). We performed the same test to the STs. 18 of them are known to be cell cycle TFs ($p = 5 \times 10^{-11}$, Table S3).

4.5 Definition of p-values for periodicity and enriched for cell cycle genes

The significance of periodicity is defined as previously published work [13,14]. The data is downloaded from their website http://www.cyclebase.org. The definition of p-value for enrichment is similar to the GO enrichment test. For TF_i, the p-value is defined as

$$p_i = p(m_i, n, M_i, N) = \sum_{k>m_i} \frac{C_{M_i}^k C_{N-M_i}^{n-k}}{C_N^n}$$

Here, m_i is the neighbors in the PCCGs and KCCGs; n is the number of genes in PCCGs and KCCGs; M_i is TF_i's neighbors (TFs or targets) in the test genes and N is the number of test genes.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No.10871009, No.10721403, and No.10774009), the National High Technology Research and Development of China (No.2008AA02Z306), the National Key Basic Research Project of China (No.2009CB918503, No. 2006CB910706), and "The Fundamental Research Funds for the Central Universities" in China.

References

[1] Elble R and Tye BK. (1991) Both activation and repression of a-mating-type-specific genes in yeast require transcription factor Mcm1. Proc Natl Acad Sci USA 88(23):10966-70

[2] Lydall D, et al. (1991) A new role for MCM1 in yeast: cell cycle regulation of SW15 transcription. Genes Dev 5(12B):2405-19

[3] Michael Costanzo, et al. (2010) The Genetic Landscape of a Cell. Science 327, 425

[4] Spellman, P.T., et al. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol. Biol. Cell 9:

3273–3297.
[5] Cho, R.J., et al. (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. Mol. Cell 2: 65–73.
[6] de Lichtenberg, U., et al. (2005) New weakly expressed cell cycle-regulated genes in yeast. Yeast 22: 1191–1201.
[7] Pramila T, et al. (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
[8] Sean R Collins, et al. (2006) A strategy for extracting and analyzing large-scale quantitative epistatic interaction data. Genome Biology 2006, 7:R63
[9] Pierre R Bushel. Dissecting the fission yeast regulatory network reveals phase-specific control elements of its cell cycle. BMC Systems Biology 2009, 3:93
[10] Marguerat S, et al. (2006)The more the merrier: comparative analysis of microarray studies on cell cycle-regulated genes in fission yeast. Yeast (Chichester, England) 2006, 23(4):261-277.
[11] Miguel C. Teixeira, et al. (2006) The YEASTRACT database: a tool for the analysis of transcription regulatory associations in Saccharomyces cerevisiae, Nucl. Acids Res., 2006, 34: D446-D451, Oxford University Press
[12] Pedro T. Monteiro, et al. (2008) YEASTRACT-DISCOVERER: new tools to improve the analysis of transcriptional regulatory associations in Saccharomyces cerevisiae, Nucl. Acids Res., Jan. 2008, 36: D132-D136, Oxford University Press
[13] Gauthier, N. P., et al. (2009) Cyclebase.org: version 2.0, an updated comprehensive, multi-species repository of cell cycle experiments and derived analysis results. Nucleic Acids Research
[14] Gauthier, N. P., et al. (2008) Cyclebase.org—a comprehensive multi-organism online database of cell-cycle experiments. Nucleic Acids Research, 36(Database issue), D854-859
[15] Jeffrey A. Ubersax, et al. (2003) Targets of the cyclin-dependent kinase Cdk1. Nature 425, 859-864
[16] Dorothea Fiedler, et al. (2009) Functional Organization of the S. cerevisiae Phosphorylation Network. Cell 136, 952–963,
[17] Sourav Bandyopadhyay, et al. (2008) Functional Maps of Protein Complexes from Quantitative Genetic Interaction Data. PLoS Computational Biology, Volume 4, Issue 4.
[18] Tong, A.H.Y., et al. (2001) Systematic genetic analysis with ordered arrays of yeast deletion mutants. Science 294, 2364–2368.
[19] Pan, X., et al. (2004) A robust toolkit for functional profiling of the yeast genome. Mol. Cell 16, 487–496.
[20] Zhidong Tu, et al. (2006) An integrative approach for causal gene identification and gene regulatory pathway inference. Bioinformatics Vol. 22 no. 14 pages e489–e496
[21] Jun Zhu, et al. (2007) Increasing the Power to Detect Causal Associations by Combining Genotypic and Expression Data in Segregating Populations. PLoS Computational Biology, Volume 3, Issue 4
[22] Jun Zhu, et al. (2008) Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. Nature genetics, Volume 40, Number 7.
[23] Su-In Lee, et al. (2009) Learning a Prior on Regulatory Potential from eQTL Data. PLoS Genetics, Volume 5, Issue 1.
[24] Daniel Segre, et al. (2005) Modular epistasis in yeast metabolism. Nature genetics Volume 37: 77-83.