

---

# Global Transcription Profiles of *C. albicans* and their Comparison with Other Yeast Species

6

Sven Bergmann, Jan Ihmels, and Judith Berman

## Abstract

In addition to providing an inventory of genes and their regulatory regions, whole-genome sequencing has paved the way for high-throughput technologies that take snapshots of the regulatory programs governing the expression of these genes. Analyzing many different *C. albicans* transcription profile experiments revealed global patterns of gene expression that would have been difficult to glean from individual studies. A next step is to compare these global patterns between organisms; comparison to *S. cerevisiae* global expression patterns is particularly useful because of the large datasets available. The Differential Clustering Algorithm (DCA) is a flexible algorithm that systematically analyzes differences, as well as similarities, in gene expression patterns. It was used to identify important aspects of the transcriptional networks that have been rewired through the evolution of *S. cerevisiae* and *C. albicans*. Specific examples of changes in the patterns of expression of cell cycle genes and amino acid biosynthesis genes are detailed. Analysis of the evolution of a promoter sequence motif associated with rapid growth metabolism is also considered.

---

## Introduction

Approaches to global analysis of microarray data from a single organism DNA microarrays are firmly established as a standard tool in biological and biomedical research. Together with the rapid advancement of genome sequencing projects, microarrays and related high-throughput technologies have been key factors in the study of the more global aspects of biological systems (Kitano, 2002). While genomic sequence provides an inventory of parts, a proper organization and eventual understanding of these parts and their functions also requires global views of the regulatory relations between them (Lander, 1999). Genome-wide expression data offer such a global view by providing a simultaneous read-out of the mRNA levels of all (or many) genes of the genome.

In particular, DNA microarrays have become routine tools for the study of *C. albicans* biology. Several platforms for this analysis have now been used. These include: cDNA arrays (published first by Whiteway and co-workers (Cowen *et al.*, 2002; Nantel *et al.*, 2002) and soon thereafter by Eurogentec in collaboration with the European Galar Fungail Consortium ([www.pasteur.fr/recherche/unites/Galar\\_Fungail/](http://www.pasteur.fr/recherche/unites/Galar_Fungail/)) (Fradin *et al.*, 2005; Garcia-Sanchez *et al.*, 2004; Rogers and Barker, 2003)), by the Fink and Johnson groups

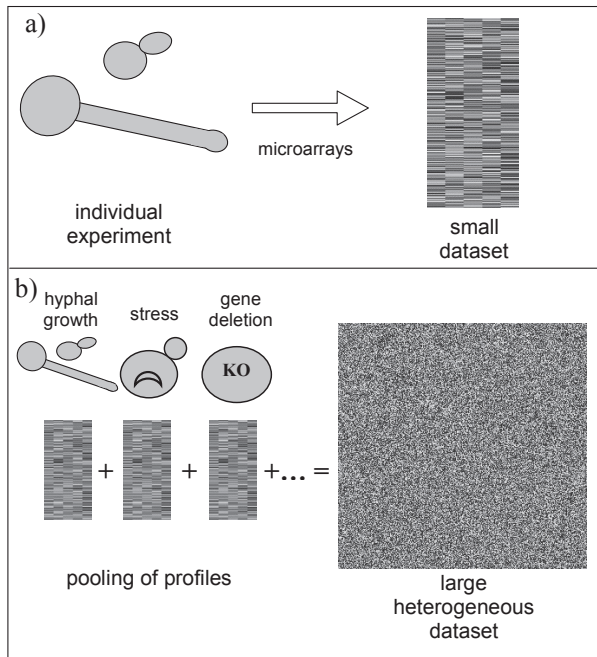
(Bennett *et al.*, 2003; Lorenz *et al.*, 2004; Tsong *et al.*, 2003) and by a consortium of labs led by Berman and Hoyer (Bensen *et al.*, 2004; Zhao *et al.*, 2005b)); Affymetrix-type arrays utilized by Agabian and co-workers (Lan *et al.*, 2002; Murillo *et al.*, 2005), and 70-mer oligonucleotide arrays made available commercially by Qiagen/Operon (Cao *et al.*, 2006; Zhao *et al.*, 2005a). Many of these arrays were designed prior to publication of the diploid assembly of the *C. albicans* genome sequence (Jones *et al.*, 2004) and thus utilized different sets of gene names that now can all be connected to Assembly Version 19 gene names that have been annotated by the *C. albicans* community (Braun *et al.*, 2005) and are curated by the Candida genome database (<http://www.candidagenome.org/>). Information on *C. albicans* genes is also available at the CandidaDB website (<http://genolist.pasteur.fr/CandidaDB/>) (Arnaud *et al.*, 2005; d'Enfert *et al.*, 2005).

The *C. albicans* microarray experiments conducted to date were designed to address specific biological issues such as hyphal growth, biofilm formation or drug resistance, mating behaviors, host–pathogen interactions, environmental or chemical stress; in addition, the role of specific transcription factors in these processes has been explored using mutant strains (Bachewich *et al.*, 2005; Bennett *et al.*, 2003; Bensen *et al.*, 2004; Cowen *et al.*, 2002; Enjalbert *et al.*, 2003; Enjalbert and Whiteway, 2005; Fradin *et al.*, 2005; Garcia-Sanchez *et al.*, 2004; Harcus *et al.*, 2004; Karababa *et al.*, 2004; Lee *et al.*, 2004; Lorenz *et al.*, 2004; Nantel *et al.*, 2002; Rogers and Barker, 2002; Rogers and Barker, 2003; Tsong *et al.*, 2003).

Even for such individual studies of gene expression under a small number of different conditions, or in a small number of different strains, the sheer amount of data points necessitates computational tools to analyze the relevant biological information (reviewed in Brazma and Vilo, 2000; Slonim, 2002). These analyses are usually performed with standard clustering algorithms that partition genes into disjointed clusters based on their correlations to all the other genes over all experimental conditions. Such a partitioning of the data is commonly presented using a hierarchical tree (Figure 6.1a).

In addition to specific biological questions probed in focused experiments, a wealth of additional information can be retrieved from large and heterogeneous datasets containing transcription profiles from a variety of different conditions (Lander, 1999). Such comprehensive data, primarily from *S. cerevisiae*, have been used to provide functional links for unclassified genes (Hughes *et al.*, 2000b; Ihmels *et al.*, 2002; Kim *et al.*, 2001; Tavazoie *et al.*, 1999; Wu *et al.*, 2002), to predict novel *cis*-regulatory elements (Bussemaker *et al.*, 2001; Hughes *et al.*, 2000a; Ihmels *et al.*, 2002; Tavazoie *et al.*, 1999; Wang *et al.*, 2002) and to elucidate the structure of the transcriptional program (Gasch *et al.*, 2000; Wang *et al.*, 2002).

While standard clustering methods reveal the major patterns of gene expression in small datasets, the utility of these tools for the analysis of large heterogeneous datasets is limited, primarily because gene co-regulation is context dependent: genes may participate in more than one function, resulting in one regulation pattern in one context and a different pattern in another. Usually genes are coordinately regulated only in specific experimental contexts, corresponding to a subset of the conditions in the dataset. Such conditions could be different growth environments (*external* conditions, e.g. temperature, presence of serum or of antifungal drugs), as well as distinct genetic or developmental states (*internal*



**Figure 6.1** Individual versus global approaches to study gene expression. (a) Individual experiments that address a specific biological question (like the one depicted here for transcription regulation of hyphal growth in *C. albicans*) usually employ a relatively small number of microarrays. The resulting “small” dataset is shown here as a grayscale table for the expression levels with thousands of rows corresponding to the genes, but only a few experimental conditions corresponding to the columns. Patterns of co-expression are readily identified using standard hierarchical clustering algorithms. (b) Global analysis aims to integrate data by pooling a large number of microarray datasets including many unrelated experimental conditions. For these large and heterogeneous datasets simple hierarchical clustering is of limited use in discerning the many (and often overlapping) subgroups of co-expressed genes and the conditions that induce this co-expression.

conditions, e.g. specific mutants or stages in the mating process) (Figure 6.1b). Thus, combinatorial regulation necessitates the assignment of genes to several context-specific and potentially overlapping modules. Furthermore, conditions irrelevant for the analysis of a particular regulatory context contribute only noise, hampering the identification of correlated behaviors that exist only over specific subsets of conditions.

Identifying modular units (*transcription modules*) within the data set reduces its complexity and facilitates understanding of the intricate behaviors of thousands of genes under a variety of internal and external conditions. The comparison and integration of data from different sources requires conceptual and computational approaches that can generate simple, accessible models and user-friendly tools from the massive information generated by hundreds of array experiments. Ultimately such models should be not only descriptive, but also predictive, provide insight into the design features of the organism of interest and facilitate the comparison of such features between related organisms.

## Approaches to comparative analysis of global gene expression patterns

Integrative studies across diverse organisms and conditions have the potential to provide exciting new insights into biological and evolutionary principles. Such comparative analysis may use conserved regulatory patterns to identify true biological signals and key mechanisms. In particular, *C. albicans* and *S. cerevisiae* are both from the hemiascomycete lineage, yet their lineages diverged ~150 million years ago. A straightforward approach is to study the *conservation* of coexpression and gene modules across organisms (Bergmann *et al.*, 2004; Ihmels *et al.*, 2005b; McCarroll *et al.*, 2004; Stuart *et al.*, 2003). In general, conserved coexpression relationships tend to correspond to core functions such as ribosome assembly, which have been well studied.

Some comparative analyses have provided a global perspective across multiple organisms (Bergmann *et al.*, 2004; Stuart *et al.*, 2003), while other more recent analyses highlighted functional modules that have a key role in a defined process of interest (Ihmels *et al.*, 2005b; McCarroll *et al.*, 2004; Tanay *et al.*, 2005). For example, Tanay *et al.* (Tanay *et al.*, 2005) explored the evolution of cis-regulatory programs associated with conserved modules by integrating expression profiles for *S. cerevisiae* and *S. pombe*.

Analysis of only the conserved aspects of global expression data ignores the majority of the data, pertaining to genes that are expressed in patterns that are partially different or that have different substructures. For example, in comparing *C. albicans* to *S. cerevisiae* expression patterns, there were obvious similarities in the expression of aspects of core functions such as protein synthesis or carbohydrate metabolism. Yet most transcription modules, when matched across the two species according to sequence similarity of the respective genes, contained no more than 20% orthologs. Similarly, when we analyzed the co-expression of groups of genes with the same Gene Ontology category (GO-term), there were several “core functions” especially related to nucleic acid metabolism and translation (e.g. GO-terms “RNA processing,” “tRNA metabolism,” “nucleocytoplasmic transport,” and “mRNA metabolism”) that were significantly correlated in both organisms. Several *S. cerevisiae* GO-terms were also correlated only in *S. cerevisiae* and not in *C. albicans*. These included “amino acid biosynthesis,” “proteolysis,” “mitotic cell cycle,” and “lipid biosynthesis” as well as “endoplasmic reticulum organization.” In contrast, there were two GO-terms (“protein mitochondrial targeting” and “protein folding”) whose co-expression was more tightly correlated in *C. albicans* than in *S. cerevisiae*. The smaller number of *C. albicans*-specific co-expression groups may be due to the fact that the majority of *C. albicans* genes have not been studied directly and were assigned GO-terms based upon the GO-terms associated with the orthologous *S. cerevisiae* genes. While the similarity between the genes makes this approach reasonable, for many proteins (e.g. transcription factors), minor changes in sequence can result in significant changes in the roles they play in a process and thus the GO-term assignments must be considered to be only a rough prediction.

Even in *S. cerevisiae*, where most GO-terms are annotated based upon some experimental data, the genes within a particular GO category are usually not all co-expressed. Rather, a subgroup of the genes may be co-expressed, or the genes may be expressed in several subgroups that have correlated expression within, but not between, the subgroups. For example, as discussed below, in *S. cerevisiae* genes annotated to the GO-term “gluconeogenesis” are coherently expressed in two separate groups (see [Figure 6.4](#)).

The DCA addresses this issue by providing an intuitive visual representation of the patterns of expression of groups of genes that are all co-expressed in one organism and exhibit one of four patterns in the other organism: full, partial, split or no co-expression. Such an approach permits analysis of differences in metabolic strategies, structural features and growth properties of the two different organisms.

Once different patterns of expression are identified in two distantly related organisms such as *C. albicans* and *S. cerevisiae*, sequence motifs associated with the differential patterns of expression can be identified and the mechanisms by which these changes evolved can be analyzed. The availability of whole genome sequence for a large number of hemiascomycete yeast species (Cliften *et al.*, 2003; Kellis *et al.*, 2003) also allows one to analyze evolutionary aspects of these changes. Variations on this approach were used by Tanay *et al.* and by Gasch and co-workers (Gasch and Eisen, 2002; Tanay *et al.*, 2005) to analyze the evolution of particular transcription factor binding sites in yeast species. These analyses focused on conserved units of co-expression. Such modules, e.g. those consisting of ribosomal proteins, rRNA processing genes or stress response genes, can be viewed as fundamental building blocks of the transcription network. While several of these modules were conserved between *S. cerevisiae* and *C. albicans*, application of the DCA approach revealed that the connections between some of them had diverged between the two species. These differences in the higher order organization of the transcription networks was associated with the different growth phenotypes of the two yeast species (preferred aerobic vs. anaerobic growth) and could be traced back to the evolution of a promoter sequence motif present throughout the hemiascomycete lineage (Ihmels *et al.*, 2005a; Ihmels *et al.*, 2005b). By comparing the appearance of this motif across the different yeasts it was possible to determine that the transcription pattern was rewired within the time interval associated with the ancient genome duplication that occurred in the hemiascomycete lineage.

In this chapter we will review the first global analysis of *C. albicans* microarray data. First, we will discuss global modular analysis of the *C. albicans* expression data, which used an unbiased approach to determine the groups of genes and conditions that are co-regulated and revealed regulatory sequence motifs and functional groups of genes that are co-expressed. We will also discuss more directed approaches to look at specific subsets of the data. Second, we will review our comparison of the global patterns of *C. albicans* transcription with those of *S. cerevisiae* using the DCA. Third, we will discuss insights into the evolution of biological regulation between these two organisms and in the hemiascomycete yeasts in general.

---

## Global patterns of gene expression in *C. albicans*

### Data collection

Individual genome-wide experiments are global only in the sense that the genes probed with one microarray span all or most of the genome. Global expression data analysis aims to include a large variety of conditions in order to span the space of transcriptional states of the cell. Such data present new and serious challenges to the computational tools used to analyze them and thus a number of different approaches have been proposed for these analyses (for review see Ihmels and Bergmann, 2004).

A large dataset of *C. albicans* array experiments was assembled by collecting published expression profiles and combining them into one comprehensive database (Figure 6.1b). The analysis described herein was performed in spring 2004 with datasets published, or made available prior to publication, at that time. The data were generated by seven different laboratories, using four independently designed microarrays and four different gene-naming schemes. All data were collected into a unified format (*orf19*), which included a total of 6167 open reading frames (ORFs). These included experiments with strains grown under different yeast or hyphal growth conditions (Nantel *et al.*, 2002), in biofilms (Garcia-Sanchez *et al.*, 2004), exposed to blood components (Fradin *et al.*, 2005; Lorenz *et al.*, 2004), altered pH (Bensen *et al.*, 2004), or signaling molecules (Harcus *et al.*, 2004; Lee *et al.*, 2004). Expression data from drug resistant strains (Cowen *et al.*, 2002; Karababa *et al.*, 2004; Rogers and Barker, 2003), strains treated with different stresses (Enjalbert *et al.*, 2003), strains with mutations in combinations of genes in the mating type-like locus (Tsong *et al.*, 2003), strains responding to mating pheromone (Bennett *et al.*, 2003) as well as some other laboratory strains and clinical isolates (Bensen *et al.*, 2004; Cowen *et al.*, 2002; Garcia-Sanchez *et al.*, 2004) were also included. Altogether, ~350 experiments were collected. These were condensed to 244 expression profiles by averaging dye swap data, when available, because preliminary analysis identified modules whose only common feature was that they had differential expression when measured with Cy3 vs. Cy5.

### The modular concept

Whenever a large number of individual elements with heterogeneous properties are analyzed, one can obtain a clearer understanding of the entire ensemble if elements with similar properties are grouped together. Individual genes can be categorized according to their properties to obtain a global picture of their organization in the genome using their predicted molecular function, biological process or cellular component (as defined using Gene Ontology (GO) terms (<http://www.geneontology.org>)), their predicted biochemical pathways (using the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (<http://www.genome.ad.jp/kegg>)), or using other criteria such as the presence of promoter sequence motifs, or in the case of microarray data, criteria such as correlation of expression patterns. A major advantage of studying the properties of modules, rather than individual elements, relies on a basic principle of statistics: The variance of an average decreases with the number  $N$  of (statistical) variables used to compute its value like  $1/N$ , because fluctuations in these variables tend to cancel each other out. Thus mean values over the elements of a module or between the elements of different modules are more robust measures than the measurements of each single element alone. This is particularly relevant for the noisy data produced by current chip-based high-throughput technologies.

Several examples for combinatorial regulation have been discussed in the literature. Yuh *et al.* (Yuh *et al.*, 1998) analyzed the combinatorial logic in the control element of a sea urchin gene. Barkai and co-workers elucidated the co-regulation of the TCA cycle in *Saccharomyces cerevisiae* and identified two subparts of the cycle that are autonomously co-regulated under different sets of conditions (Ihmels *et al.*, 2002). Several examples of condition-specific regulation in yeast and the correlation with transcription factor binding sites were given by Gasch *et al.* (Gasch and Eisen, 2002). Pilpel *et al.* (Pilpel *et al.*, 2001; Sudarsanam *et al.*, 2002) pursued a systematic approach to characterize motif combinations and their synergistic effect on expression patterns at the genomic level.

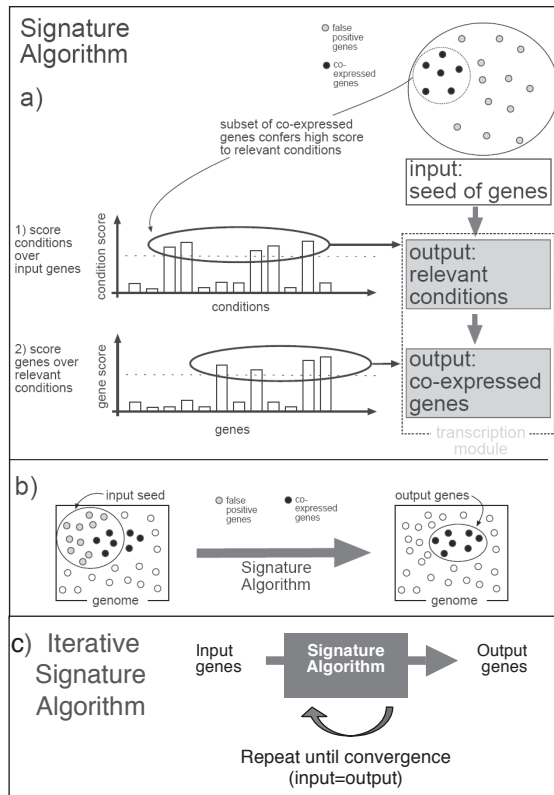
A successful approach to effectively deal with context-specific co-expression is to generate “transcription modules” (also known as “biclusters”), which are sets of co-regulated genes together with the conditions over which the co-regulation is observed. Transcription modules provide not only the basic building blocks that characterize the structure of the genome-wide transcription program under a variety of conditions, they also supply a map for a more interpretable characterization of transcriptional changes induced by novel experiments. In particular, when searching for coherent changes in expression in larger modules, one may identify patterns that are too weak to discern when considering each of its genes alone. For example, Mootha *et al.* (Mootha *et al.*, 2003) showed that the coordinate expression of a set of functionally related genes was significantly altered in human diabetic muscle, even though this effect was too subtle to be apparent at the single gene level.

### Signature Algorithm

Barkai and co-workers developed two important algorithms to analyze expression patterns with respect to specific subsets of genes and conditions. The *Signature Algorithm* is the core algorithmic unit of all subsequent work (Ihmels *et al.*, 2002). It offers a general framework to integrate large-scale expression profiles with other biological data. The algorithm requires as input (or *seed*) a set of genes, at least some of which are expected to be co-regulated (Figure 6.2a). These genes may be chosen according to some common features (e.g. membership in a GO category). The algorithm proceeds in two steps. First, the input seed is used to identify the subset of experimental conditions (i.e. microarray experiments) that is relevant to their coexpression. To this end, each condition in the dataset is scored by the average expression change over the input genes. Conditions that induce a coherent change for at least some of the input genes receive higher scores and are selected according to a cut-off parameter (the *condition threshold*). In the second stage of the algorithm, genes that are highly and consistently expressed over the conditions identified in the first step are selected according to a second cut-off parameter (the *gene threshold*). The final output of the algorithm (Figure 6.2b) is a *transcription module*, consisting of a set of co-regulated genes together with the regulating conditions. In general, the output set of genes contains the co-regulated part of the input seed, as well as other genes that were not part of the original input but display a similar expression profile over the relevant conditions. Genes in the seed that are not co-expressed under the relevant conditions do not appear in the module. Individual modules are identified independently and can naturally overlap both in gene and condition content.

### Iterative Signature Algorithm

The signature algorithm was extended into an iterative scheme (the *Iterative Signature Algorithm*, ISA) that allows for an efficient modular decomposition of large-scale expression data also in the absence of any *a priori* information (Bergmann *et al.*, 2003; Ihmels *et al.*, 2004a). In the ISA, the output genes of the *Signature Algorithm* are repeatedly re-used as input for the algorithm, until a point of convergence is reached where output and input are identical (Figure 6.2c). The resulting transcription module of genes and conditions is a *fixed point* of the *Signature Algorithm* and satisfies a criterion termed *self-consistency*. The criterion states that from all genes in the dataset, the module genes are the most coherently co-expressed over the module conditions. The module conditions, in turn, are those conditions in the dataset that induce the most coherent expression in the module genes. Such



**Figure 6.2** The signature algorithm and iterative signature algorithm. (a) The signature algorithm requires as input a set of genes, some of which are expected to be co-regulated based on additional biological information such as a common promoter binding motifs or functional annotation. The algorithm proceeds in two steps: In the first step, this input seed is used to identify the conditions that induce the highest average expression change in the input genes. Only conditions with a score above some threshold are selected. In the second stage of the algorithm, genes that are highly and consistently expressed over these conditions are identified. The result consists of a set of co-expressed genes together with the conditions inducing this co-expression and is termed a transcription module. (b) The output contains only the co-expressed part of the input seed, as well as other genes that were not part of the original input but display a similar expression profile over the relevant conditions. (c) The iterative signature algorithm repeats the signature algorithm, reusing the output genes as new seeds until fixed points (where input genes equal output genes) are reached.

an explicit formulation of the defining property of a module distinguishes the ISA method from most clustering algorithms. Importantly, transcription modules can be identified in a heuristic search by iterating from a large number of random input seeds. Alternatively, the iterative scheme can be initiated with biologically motivated sets of genes (Ihmels *et al.*, 2004b).

#### Visualizing modules and online modular analysis

When performing the ISA, a range of gene threshold levels is used. They determine the degree of expression coherence required for genes to be considered part of a module. In

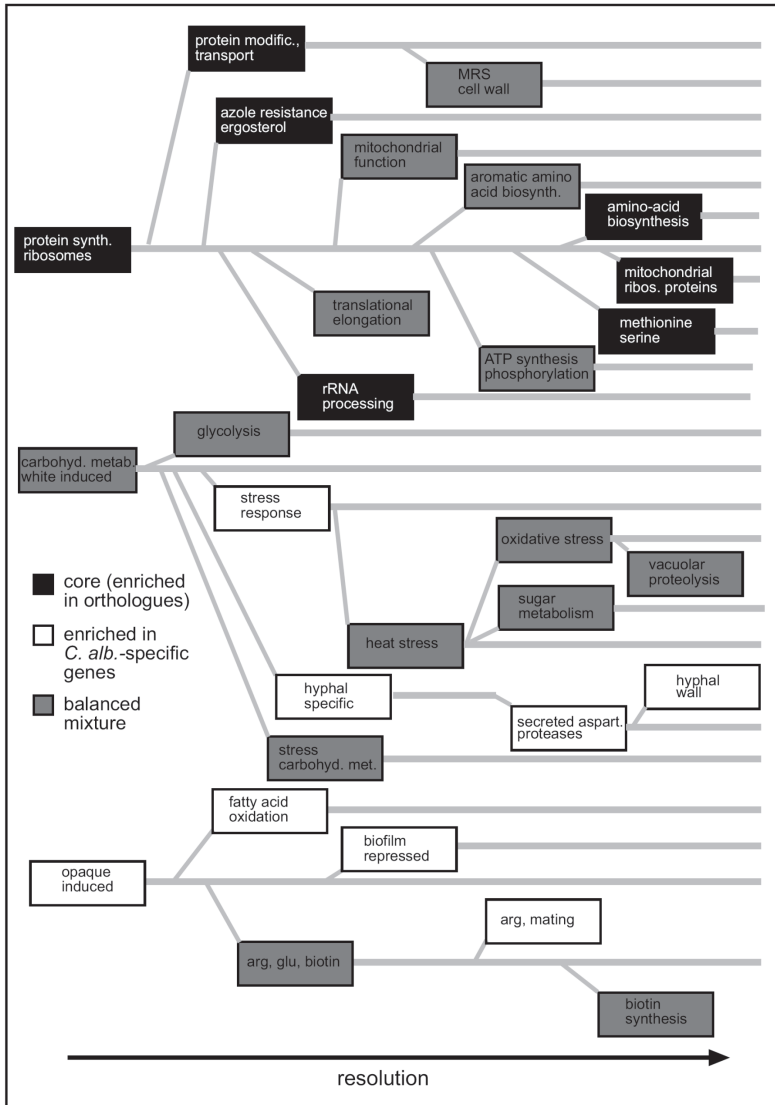
general, using a mild threshold yields few transcription modules that have many genes. At more stringent thresholds, more modules, each including fewer genes, are identified. Frequently these smaller modules correspond to subunits of the major transcription modules. A useful method to visualize the full modular structure for an organism is to use a hierarchical *module tree* (Bergmann *et al.*, 2003; Ihmels *et al.*, 2004a) (Figure 6.3) over a range of resolutions. Highly similar modules, in terms of genes and conditions that were identified at adjacent thresholds, are connected by lines and define the branches of the tree. Such connections simplify the biological interpretation of modules and reveal hierarchies of co-regulated units of varying expression coherence.

The three major *C. albicans* modules corresponded to two core functions (*protein synthesis* and *carbohydrate metabolism*) and one *Candida*-specific module (*opaque-induced genes*) (Figure 6.3). The submodules of these major modules could be divided into modules primarily composed of core genes (those that have an *S. cerevisiae* ortholog), modules composed primarily of *C. albicans*-specific genes, and modules with a balanced mixture of core and *C. albicans*-specific genes. The submodules in the *protein synthesis* branch are all core or balanced mixture modules. The *carbohydrate metabolism* module includes core functions of sugar metabolism but was also associated with white-induced/opaque-repressed genes. The submodules contain either balanced mixtures of core and *C. albicans*-specific genes or primarily *C. albicans*-specific genes. This is consistent with the idea that the white/opaque growth forms are a *C. albicans*-specific growth property. The dominant influence of white/opaque expression may also be due to the large proportion of the data set (74 conditions) that was generated in two related studies addressing white/opaque and mating behaviors (Bennett *et al.*, 2003; Tsong *et al.*, 2003).

Besides the three major modules that had rich substructures, the algorithm also revealed two additional unrelated major modules which were associated primarily with data from a single, large set of experiments (Tsong *et al.*, 2003) and had less substructure. We did not consider these modules further, because they appear to be an artifact of the large proportion of the total data represented by many experiments that used only one of the array formats and a particular reference control RNA sample.

Web-based tools are available to analyze the genes, conditions and GO-terms that are enriched in particular *C. albicans* modules. A particularly useful feature is the gene-to-module search function (<http://barkai-serv.weizmann.ac.il/candida>). This feature allows one to identify all the modules with which a particular *C. albicans* gene of interest is associated and then to analyze those modules for the genes, conditions, overrepresented GO-terms and motifs associated with them. An interactive module tree (<http://barkai-serv.weizmann.ac.il/candida/figure5Interactive.html>) allows analysis of specifically named modules. A list of all the *C. albicans* transcription modules, together with information about whether they contain primarily genes with *C. albicans*-specific or “core” (highly conserved) functions can be found at <http://barkai-serv.weizmann.ac.il/candida/enrichmentList/enrichments.htm>. In addition, a list that simplifies the analysis by indicating those modules that have essentially similar sets of genes and conditions (because they merge with each other as resolution is decreased) is available at <http://barkai-serv.weizmann.ac.il/candida/representative-SetsCan.htm>.

A similar analysis of the *S. cerevisiae* data set, comprising more than 1 000 expression profiles (Ihmels *et al.*, 2004a; Ihmels *et al.*, 2004b), converged into five fixed points at low



**Figure 6.3** The module tree for *C. albicans* is composed of modules consisting mostly of “core” genes whose sequence is highly conserved (black), *C. albicans*-specific genes (white) as well as mixed modules (gray). At low resolution, there are three major modules: Protein synthesis (core), Carbohydrate metabolism/white-induced (balanced mixture) and opaque-induced (*C. albicans*-specific). At higher resolution, the modules become more refined and specific. For example, the protein synthesis module branch has submodules annotated as “amino acid biosynthesis,” “rRNA processing,” translation elongation,” and “mitochondrial ribosomal proteins.”

resolution. These correspond to fundamental functions of *S. cerevisiae* (i.e. protein synthesis, rRNA-processing, amino acid biosynthesis, stress response and mating). Information for probing the *S. cerevisiae* module analysis is available at <http://www.weizmann.ac.il/home/jan/NG/MainFrames.html>. Another nice tool for analyzing the motifs and GO-

terms overrepresented in groups of co-expressed genes is T-profiler (Boorsma *et al.*, 2005) which uses the *t*-test to score changes in the average activity of predefined groups of genes based upon GO categories, ChIP-chip experiments, transcription factor binding motifs or location on the same chromosome.

Our analysis of *C. albicans* global expression also showed that, as in *S. cerevisiae*, transcription modules tend to contain genes with related GO-terms and/or with similar sets of 6-mer and 7-mer sequence motifs in sequences 5' to the ORF (Ihmels *et al.*, 2005a). Furthermore, the modules have a good level of predictive value. When the *S. cerevisiae* transcription modules were compared to localization data in the global GFP tagging study of O'Shea and co-workers, genes that clustered together in transcription modules predicted from the *Signature Algorithm* (Ihmels *et al.*, 2002) also tended to localize to the same cellular compartment (Huh *et al.*, 2003). Similarly, a *C. albicans* gene of unknown function that was found within the ribosomal RNA processing module localized, as predicted, to the nucleus (Ihmels *et al.*, 2005a). Thus the transcription modules for *S. cerevisiae* and *C. albicans* modules can predict functions for unknown genes when they associate with genes of known function in a module.

---

## Comparative analysis of global transcription patterns

### Identification of orthologous genes

The *Signature Algorithm* can be used to compare the expression patterns of different organisms as well. Pairs of orthologous genes between *C. albicans* and *S. cerevisiae* were identified using the Inparanoid software (Remm *et al.*, 2001). Approximately 5% of the genes were matched to gene families that arose after the two lineages diverged from each other. In this case, the ortholog with the highest score to form the pair was selected. This yielded a set of 3619 orthologous gene pairs.

### Comparison of modules

Comparisons between the two sets of expression data can be done in several ways. One way is to first identify the transcription modules in each dataset and the organization of the transcription program in each organism using the methodologies outlined above and then to identify those that are primarily conserved. This has been done for a range of organisms. In general, several co-expressed sets of *ancient* genes, pertaining to fundamental cellular functions (like those coding for ribosomal proteins) have been conserved across all organisms from yeast to humans (Bergmann *et al.*, 2004) and, as expected, are well conserved between *C. albicans* and *S. cerevisiae* as well. While this identifies groups of genes with similar, basic cellular functions, it does not address the numerous changes in expression patterns that likely reflect interesting biological differences between the organisms.

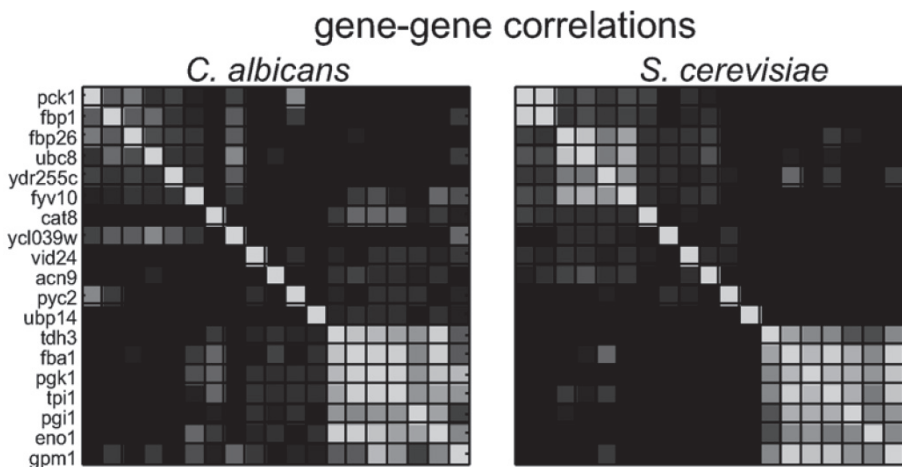
### Differential Clustering Algorithm

The *Differential Clustering Algorithm* (DCA) was developed to study systematically whether the co-expression of genes in one organism has been conserved fully, partially or not at all in another organism (Ihmels *et al.*, 2005a). The DCA is applied to a set of orthologous genes that are present in both organisms. The sets of orthologous genes can be defined according to different criteria such as GO-term, presence of a common motif in the pro-

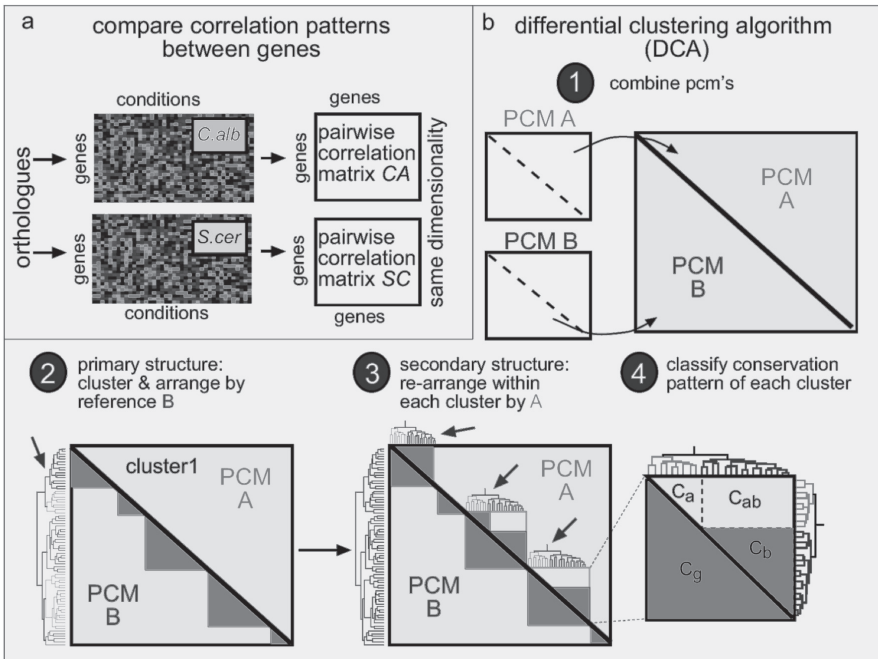
moter sequence, or presence in a transcription module in one of the datasets. The subset of expression data for these genes in each organism can be used to define two *pair-wise correlation matrices* (PCMs) that contain the correlations of the expression patterns between all possible pairs of genes. For example, the GO category “gluconeogenesis” contains two subgroups of genes that are independently co-expressed. Here the subdivision of the genes into the two subgroups is similar in *C. albicans* and *S. cerevisiae* (Figure 6.4). The DCA was designed to analyze efficiently the different possible patterns and degrees of conservation in co-expression between two orthologous sets of genes, and to provide an intuitive visual presentation in terms of the correlations.

For each organism these correlations are computed separately over all microarrays in the respective dataset, resulting in two (symmetric) PCMs of the same size (i.e. the same genes are listed along the vertical and horizontal axes and the diagonal represents 100% correlation of each gene with itself (as in Figure 6.4 and Figure 6.5a).

The DCA has two major steps, and these are then performed reciprocally: once with the first organism (e.g. *S. cerevisiae*) treated as the “reference” and the second organism (e.g. *C. albicans*) as the “target” and then vice versa. It first orders the genes of the reference organism by clustering the data, assigning genes into subsets that are co-expressed in the reference organism (but not necessarily in the second organism) (Figure 6.5a). The orthologous genes in the target organism within each co-expressed subgroup are then re-ordered, by clustering according to patterns of correlated expression in the target organism. This procedure is then repeated reciprocally, such that each PCM is used once for the primary and once for secondary clustering, yielding two distinct orderings of the genes. The result of the two-way clustering is represented in terms of two re-arranged PCMs, one for each organism, where the genes are re-ordered according to the primary and secondary clusters



**Figure 6.4** Pairwise correlation matrices (PCMs) of *C. albicans* (left) and *S. cerevisiae* (right) orthologs that contain the GO term “gluconeogenesis.” In this case the genes are regulated by a split pattern in which the lower subgroup of seven genes is tightly co-expressed and the upper group of genes is more loosely co-expressed with a different pattern in each organism. Converted from a color figure: lighter shades indicate higher correlation coefficients between the expression profiles of pairs of genes than darker shades.

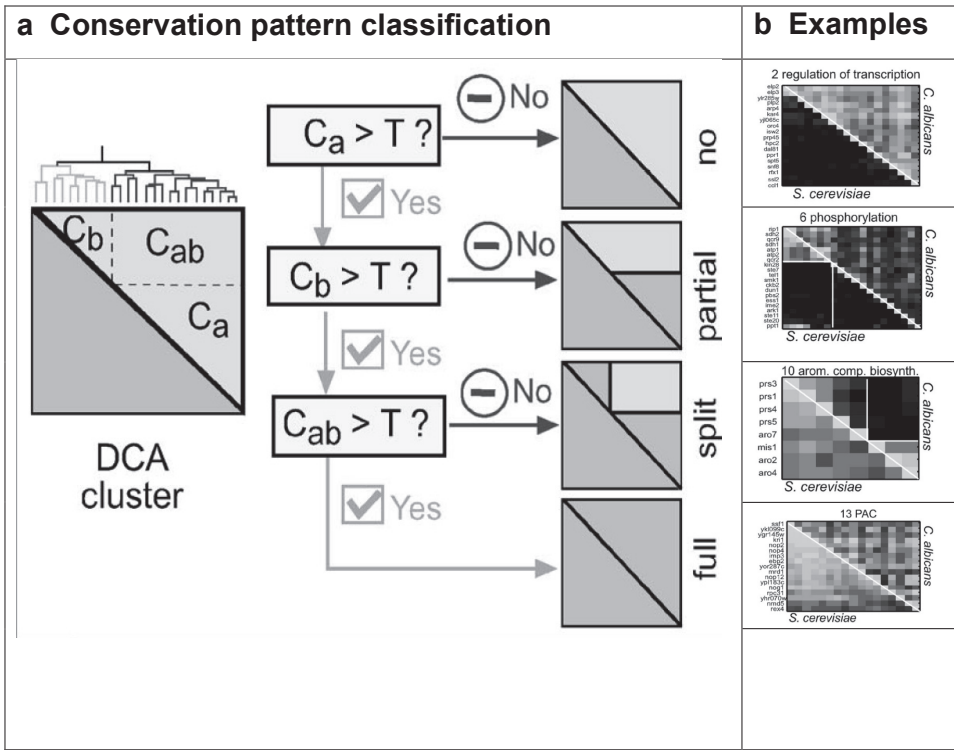


**Figure 6.5** Comparative analysis of global expression data using the differential clustering algorithm approach. (a) Pairwise correlation matrices (PCM) are calculated from the expression data in each organism. (b) The PCMs are combined into a single matrix, where each triangle corresponds to one of the PCMs (1). The genes are then ordered in two steps: First, genes are clustered and the PCMs are re-arranged according to the correlations in the reference organism (“B”) (2). Second, the genes assigned to each of the resulting primary clusters are re-clustered according to their correlations in the “target” organism “A” (secondary clustering) (3). Finally, the conservation patterns of each cluster are classified automatically into one of the four conservation classes (4).

and their correlation coefficient of expression is indicated by a color code. Since these matrices are symmetric and refer to the same set of orthologous genes, they can be combined into a single matrix without losing information. Specifically, the two PCMs are joined into one composite matrix such that the lower-left triangle depicts the pair-wise correlations in the reference organism, while the upper-right triangle depicts the correlations in the target organism (Figure 6.5b). Inspection of the rearranged composite PCM allows for an intuitive visualization of the differences and similarities in the co-expression pattern of the two organisms. The DCA also includes an automatic scoring method that classifies clusters into one of the four conservation categories: *full*, *partial*, *split*, or *no conservation* of co-expression (Figure 6.6a).

Examples of results obtained by applying the DCA to expression data from *C. albicans* and *S. cerevisiae*

GO-terms with high levels of conservation included those associated with *cytoplasmic ribosomes* and *monosaccharide metabolism*. In addition, the PAC motif (GATGAG), which is associated with genes involved in protein synthesis, is highly conserved in its association with genes involved in protein synthesis. Interesting GO-terms that were *not conserved* include



**Figure 6.6** The differential clustering algorithm identifies different patterns of conservation of gene co-expression. (a) Classification flow chart. Each primary cluster is subdivided into two secondary clusters  $a$  and  $b$  and characterized by three average correlation values, corresponding to correlation within ( $C_a$ ,  $C_b < C_a$ ) and between ( $C_{ab}$ ) these clusters. These correlations determine its assignment to one of four basic conservation patterns as depicted in the flow chart. The cut-off parameter  $T$  is chosen heuristically. (b) Examples of DCA patterns of unconserved, partially conserved, split and completely conserved conservation of co-expression.

“regulation of transcription,” which is consistent with the idea that many of the relationships between transcription factors and the genes they regulate have been rewired in the ~150 million years since the *C. albicans* and *S. cerevisiae* lineages diverged.

An important strength of the DCA is the intuitive visualization of partial and split conservation patterns. For example, in *C. albicans*, a group of genes within the “phosphorylation” GO category, which includes ATPase subunits as well as a number of known regulatory kinases, are all co-expressed (Figure 6.6b). Interestingly, only a small subgroup of these genes—primarily those involved in ATPase function—are co-expressed in *S. cerevisiae* as well. This suggests differences in the regulation of ATPase function relative to the regulation of these kinases and indicates that a series of kinase genes (e.g. *KIN28*, *STE7*, *TEL1*, *SMK1*, *CKB2*, *IME2*, *ARK1*, *STE11* and *STE20* (names correspond to the *S. cerevisiae* ortholog)) are coordinately expressed only in *C. albicans* and not in *S. cerevisiae*. A similar pattern of partial co-regulation was seen for genes involved in translation. In this case again, a subcluster of tRNA genes within the “translation” GO category are all co-expressed in *C. albicans*, while only a subgroup of these genes, all of which encode mitochondrial tRNA

synthases, are not co-expressed in *S. cerevisiae*. This likely reflects a major metabolic difference in the rapid growth of *C. albicans* and *S. cerevisiae*: in the presence of high levels of sugar *S. cerevisiae* catabolism is primarily fermentative irrespective of the presence or absence of oxygen, while *C. albicans* requires oxygen and utilizes respiratory pathways located in the mitochondria (Jones *et al.*, 2004). Thus, many core genes related to functions such as protein synthesis are coordinately expressed in the cytoplasm and the mitochondria of *C. albicans*. In contrast, *S. cerevisiae* genes encoding mitochondrial tRNAs are not coexpressed with those required for cytoplasmic protein synthesis. In fact, the expression of genes for cytoplasmic and mitochondrial protein synthesis is often anti-correlated.

The examples above are of expression patterns that are conserved in one organism and only partially conserved in the other organism. Another pattern detected by the DCA is “split conservation.” In this case, a group of genes is co-expressed in the reference organism, and these genes have two separate groups of genes with correlated expression in the target organism. An example is “aromatic compound biosynthesis.” In *S. cerevisiae*, several genes required for purine biosynthesis and aromatic amino acid biosynthesis are co-regulated; in *C. albicans* these two groups of genes are generally regulated separately (Figure 6.6b). Details of all of the DCA comparisons for *C. albicans* and *S. cerevisiae* are available at <http://barkai-serv.weizmann.ac.il/candida/geneVsGene/geneClusterList.htm>.

The modular analyses discussed above, as well as several DCA subclusters suggested that the transcriptional regulation of cell cycle genes is very different in *C. albicans* and *S. cerevisiae*. Others also found that regulation of cell cycle genes in *S. pombe* is very different from that in *S. cerevisiae* (Oliva *et al.*, 2005; Peng *et al.*, 2005). We performed multiple DCA analyses of genes annotated to the GO category “cell cycle” and that have orthologs in all three of these distantly related yeast species. In each case, one of the three species was used as the reference organism and was then compared, pairwise, with each of the other two target organisms. Each of the subclusters was, at best, partially conserved in the other two species, and usually different orthologs were partially co-expressed in the two target organisms. As an example, the clusters containing the major cyclin-dependent kinase gene *CDC28* in *S. cerevisiae* and *Cdc2* in *S. pombe*, were analyzed in detail. In general, very few of the genes in any of these subclusters are co-regulated similarly in any two of the organisms (Ihmels *et al.*, 2005a).

The genes subjected to a DCA analysis can be based on GO-terms, over-represented promoter sequence motifs, KEGG pathway terms or other features. In addition, the DCA approach can be used to analyze higher order relationships between properties of two organisms. For example, one can ask if there are specific groups of GO categories that tend to be correlated in conserved, partial, split or non-conserved ways in the two organisms. This Higher Order Connectivity Analysis (HOCA) employs PCMs describing the pairwise correlations between gene groups, rather than single genes (Ihmels *et al.*, 2005a). Although the way these correlations are defined is more subtle (employing network theory; for details see Ihmels *et al.*, 2005a; Ravasz *et al.*, 2002), the DCA analysis is performed exactly in the same manner as described above for the classification of the conservation patterns for pairwise gene–gene correlations. From such an analysis, new relationships can be revealed. For example, in *S. cerevisiae*, oxidative stress is coordinately regulated with carbohydrate metabolism, while in *C. albicans* only the expression of carbohydrate metabolism genes in this cluster is correlated (Ihmels *et al.*, 2005a).

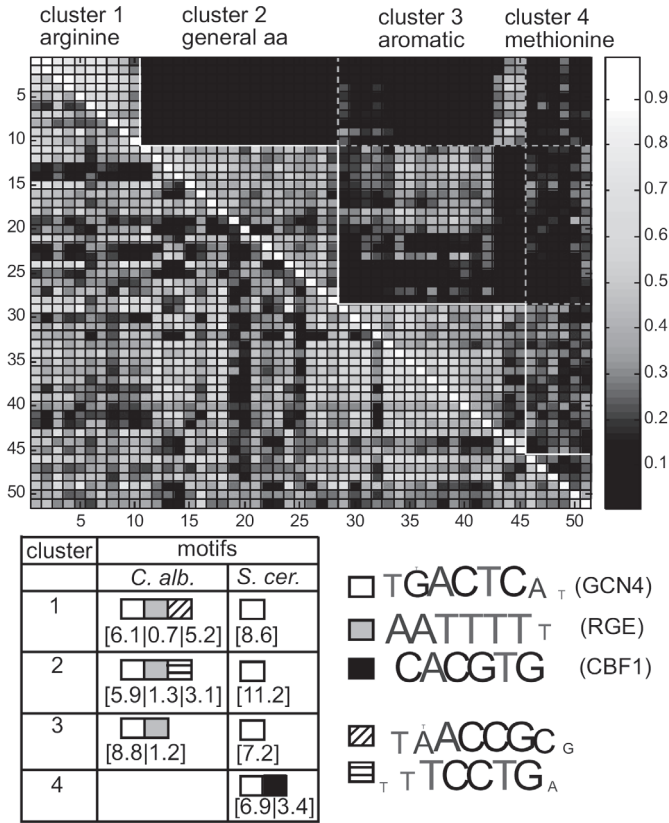
A similar higher order analysis using promoter sequence motifs to define gene groups revealed that in *S. cerevisiae*, genes carrying the sequences TGAAAAT or AAAA(T/A)TT, hereafter termed rapid growth elements (RGE), along with either the HOMOL1 (CCGTACA) or the PAC (GATGAG) motif are all coordinately expressed (Ihmels *et al.*, 2002). In contrast, in *C. albicans*, expression of genes with both RGE (TGAAAAT) and HOMOL1 motifs are not co-expressed while genes with the RGE (AAAA(T/A)TT) and PAC motifs are co-expressed with high levels of correlation. This is consistent with the report from Tanay *et al.* (Tanay *et al.*, 2005) that *C. albicans* ribosomal protein genes lack the Homol-D motif found in *S. cerevisiae*.

Another difference in transcriptional regulatory patterns became evident from modular analyses as well as from the simple and higher order DCA analysis. In *S. cerevisiae*, virtually all genes in pathways of amino acid biosynthesis are coordinately regulated, in large part due to the activity of the Gcn4p transcription factor. In contrast, in *C. albicans*, different groups of genes involved in amino acid biosynthesis were expressed together in subclusters that were not coordinately expressed with one another. In particular, arginine biosynthesis was independently coexpressed, as were aromatic amino acids and a group of genes annotated as “general” amino acid synthesis. Interestingly, analysis of the promoter sequence motifs in these groups of genes revealed a change in the role of the Gcn4p binding motif. This motif is present in the promoter of almost all of the *S. cerevisiae* amino acid synthesis genes, as well as the *C. albicans* orthologs with the exception of methionine biosynthesis genes, which have lost their co-regulation in *C. albicans*. In addition to this global regulator, two new motifs were found to be associated specifically with the *C. albicans* arginine and general amino acid subclusters (Figure 6.7). Thus, *C. albicans*, which grows in contact with the human host, has a more differential regulation of pathways required for the synthesis of different amino acids. This suggests that some subgroups of amino acids may be more available in the environmental niches occupied by *C. albicans* within the human host than are available in the environments occupied by *S. cerevisiae*.

---

## Evolution of the transcriptional program for mitochondrial protein synthesis

The availability of whole genome sequence from a large group of hemiascomycete yeasts, which ranges from *S. cerevisiae* to *C. albicans* and beyond, facilitates the analysis of evolutionary events that occurred in this lineage. A major event was an ancient genome duplication, followed by the loss of many genes and the retention of ~500 duplicate copies of genes in the species more closely related to *S. cerevisiae* (Wolfe, 2004). When glucose is present, *S. cerevisiae* and closely related species grow rapidly via fermentation, regardless of whether oxygen is present. The genome duplication apparently enabled the development of anaerobic as well as aerobic fermentation in this group of yeasts. In fact, a number of the duplicated genes that remain in the genomes of the yeasts that underwent the genome duplication are important for glucose sensing, hypoxic growth and the response to glucose (Wolfe, 2004). Furthermore, *S. cerevisiae* does not require a mitochondrial genome when grown in hypoxic conditions. In contrast, *C. albicans* and other related yeasts that did not undergo the ancient genome duplication utilize aerobic respiration for rapid growth and cannot dispense with their mitochondrial genome functions.



**Figure 6.7** Differential clustering algorithm analysis of amino acid biosynthesis genes. (a) Gene-gene correlation matrix for genes assigned to the *S. cerevisiae* amino acid biosynthesis module. Lower triangle corresponds to the *S. cerevisiae* data, while the upper triangle depicts the *C. albicans* correlations. Converted from a color figure: lighter shades indicate higher correlation coefficients than darker shades; negative correlations present between clusters 1 and 2 are not visible. (b) Sequence motifs over-represented in the different DCA clusters.

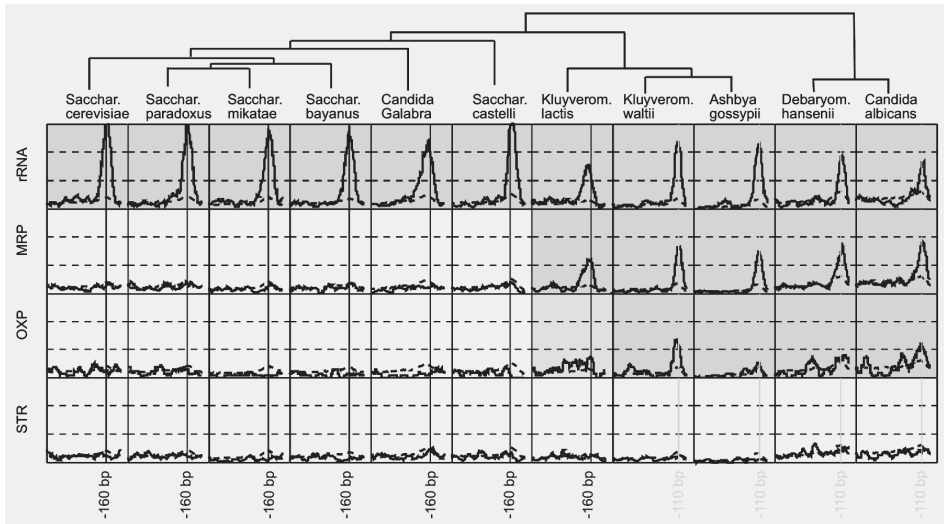
Comparing the transcription program of *C. albicans* with that of *S. cerevisiae*, many of the obvious differences are related to the differential requirement for mitochondrial function in the two yeasts. For example, cytoplasmic and mitochondrial tRNA synthases are coexpressed in *C. albicans* and expressed as two separate groups in *S. cerevisiae*. Similarly, cytoplasmic and mitochondrial ribosomal proteins are coordinately expressed in *C. albicans* and expressed independently in *S. cerevisiae*. Interestingly, the AATTTT sequence motif was present in the promoter sequences of groups of genes that showed this split pattern of regulation in the two organisms. That is, the sequence was present in the promoters of mitochondrial and cytoplasmic genes in *C. albicans* but present only in the promoters of the cytoplasmic genes in *S. cerevisiae*. This suggested that the AATTTT sequence is associated with genes necessary for coordinated protein synthesis, and hence rapid growth.

Indeed, the AATTTT sequence is necessary for rapid growth based on several criteria (Ihmels *et al.*, 2005b). First, genes with this sequence in the promoter are expressed at much

higher levels in log phase than in stationary phase, as measured using GFP reporter constructs. Second, when the AATTTT sequence in *C. albicans* mitochondrial ribosomal promoters was mutated (to either a GC-rich sequence or to an AT-rich sequence (ATATAT)), expression of a reporter GFP was reduced dramatically, especially during log phase growth. Similar results were seen at three different mitochondrial ribosomal protein promoters.

The AATTTT sequence was found upstream of ribosomal protein genes in *S. cerevisiae*, *C. albicans* and all of the other hemiascomycetes. Interestingly, this sequence was found upstream of the mitochondrial ribosomal protein genes only in yeast species that did not undergo the genome duplication (Figure 6.8). This suggests that the change in the regulation of mitochondrial ribosomal protein genes occurred soon after the genome duplication.

It is also interesting to note that the average position of the AATTTT is at  $\sim 160$  bp 5' to the start-codon in *S. cerevisiae* genes and at  $\sim 110$  bp 5' to the start codon in *C. albicans* genes. Furthermore, the difference in distance from the start codon tends to remain in most of the intermediate species: those that underwent the genome duplication have the sequence closer to the start codon than those that did not undergo the genome duplication. It will be interesting to determine if this is a general feature of motifs that regulate transcription in this lineage and whether it reflects a change in the basal transcription machinery.



**Figure 6.8** Frequency of occurrence of the rapid growth element (RGE motif in eleven yeast species (whose phylogeny is indicated by the dendrogram). Shown is the frequency of AATTTT (or its reverse complement), observed in a window between  $x$  and  $x + 50$ , where  $x$  is the position along the 600 bp upstream sequence for the genes of four groups of genes that are co-expressed in *S. cerevisiae*: rRNA processing (rRNA), mitochondrial ribosomal proteins (MRP), oxidative phosphorylation (OXP) and the stress-related genes (STR). Gene sets for the other species consist of orthologs of the *S. cerevisiae* genes. For the yeast species that emerged from the lineage that underwent a whole genome duplication event (six species on the left) the RGE has been lost in the MRP genes. The position peaks at  $\sim 160$  bp before the start codon (black lines). For the five other species (that prefer aerobic growth in rich media) the RGE is present in the MRP (and to some extent also in the OXP) genes, but located somewhat closer to the start-codon (gray lines).

Thus, it appears that the emergence of anaerobic growth capacity in yeast is associated with a global rewiring of its transcriptional network involving dozens of MRP genes which lost a specific regulatory motif. How such a large change in transcription regulation occurred, and how that change may be connected to changes in distance of the sequence from the start codon, remains to be determined.

---

## Conclusions

The availability of genome sequences has opened up new possibilities for the study of yeast biology. Global analysis of gene expression can yield new insights into metabolism, growth strategies and mechanisms of gene regulation. The comparison of global gene expression patterns is a very new field in which new approaches are necessary for handling the large datasets and revealing the complex relationships between co-expressed units within and between organisms. Tools like the Iterative Signature Algorithm, that take into account the context-specific nature of gene expression, allow for an efficient identification the modular units of each transcription program. The differential clustering algorithm facilitates interspecies comparison of expression data, revealing the extent to which the co-expression of such units has been conserved. Importantly, these tools include intuitive visualization schemes that make them useful for biologists with little bioinformatic training. Finally, analyses gleaned from expression studies in two distant yeast species can be extended to other hemiascomycete species to reveal important insights into the evolutionary processes that gave rise to the differences between them.

---

## References

- Arnaud, M.B., Costanzo, M.C., Skrzypek, M.S., Binkley, G., Lane, C., Miyasato, S.R., and Sherlock, G. (2005). The *Candida* Genome Database (CGD), a community resource for *Candida albicans* gene and protein information. *Nucleic Acids Res.* 33, D358–363.
- Bachewich, C., Nantel, A., and Whiteway, M. (2005). Cell cycle arrest during S or M phase generates polarized growth via distinct signals in *Candida albicans*. *Mol. Microbiol.* 57, 942–959.
- Bennett, R.J., Uhl, M.A., Miller, M.G., and Johnson, A.D. (2003). Identification and characterization of a *Candida albicans* mating pheromone. *Mol. Cell Biol.* 23, 8189–8201.
- Bensen, E.S., Martin, S.J., Li, M., Berman, J., and Davis, D.A. (2004). Transcriptional profiling in *C. albicans* reveals new adaptive responses to extracellular pH and functions for Rim101p. *Mol. Microbiol.* 54, 1335–1351.
- Bergmann, S., Ihmels, J., and Barkai, N. (2003). Iterative signature algorithm for the analysis of large-scale gene expression data. *Phys. Rev. E. Stat. Nonlin. Soft Matter Phys.* 67, 031902.
- Bergmann, S., Ihmels, J., and Barkai, N. (2004). Similarities and differences in genome-wide expression data of six organisms. *PLoS Biol.* 2, E9.
- Boorsma, A., Foat, B.C., Vis, D., Klis, F., and Bussemaker, H.J. (2005). T-profiler: scoring the activity of predefined groups of genes using gene expression data. *Nucleic Acids Res.* 33, W592–595.
- Braun, B.R., van Het Hoog, M., d'Enfert, C., Martchenko, M., Dungan, J., Kuo, A., Inglis, D.O., Uhl, M.A., Hogues, H., Berriman, M., *et al.* (2005). A human-curated annotation of the *Candida albicans* genome. *PLoS Genet.* 1, e1.
- Brazma, A., and Vilo, J. (2000). Gene expression data analysis. *FEBS Lett.* 480, 17–24.
- Bussemaker, H.J., Li, H., and Siggia, E.D. (2001). Regulatory element detection using correlation with expression. *Nat. Genet.* 27, 167–171.
- Cao, F., Lane, S., Raniga, P.P., Lu, Y., Zhou, Z., Ramon, K., Chen, J., and Liu, H. (2006). The Flo8 transcription factor is essential for hyphal development and virulence in *Candida albicans*. *Mol. Biol. Cell* 17, 295–307.
- Cliften, P., Sudarsanam, P., Desikan, A., Fulton, L., Fulton, B., Majors, J., Waterston, R., Cohen, B.A., and Johnston, M. (2003). Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science* 301, 71–76.

- Cowen, L.E., Nantel, A., Tessier, D., Whiteway, M., Thomas, D.Y., Kohn, L.M., and Anderson, J.B. (2002). Population genomics of drug resistance in experimental populations of *Candida albicans*. *Proc. Natl. Acad. Sci. USA.* 99, 9284–9289.
- d'Enfert, C., Goyard, S., Rodriguez-Arnaiveille, S., Frangeul, L., Jones, L., Tekaiia, F., Bader, O., Albrecht, A., Castillo, L., Dominguez, A., *et al.* (2005). CandidaDB: a genome database for *Candida albicans* pathogenomics. *Nucleic Acids Res.* 33, D353–357.
- Enjalbert, B., Nantel, A., and Whiteway, M. (2003). Stress induced gene expression in *Candida albicans*: absence of a general stress response. *Mol. Biol. Cell* 14, 1460–1467.
- Enjalbert, B., and Whiteway, M. (2005). Release from quorum-sensing molecules triggers hyphal formation during *Candida albicans* resumption of growth. *Eukaryot. Cell* 4, 1203–1210.
- Fradin, C., De Groot, P., MacCallum, D., Schaller, M., Klis, F., Odds, F.C., and Hube, B. (2005). Granulocytes govern the transcriptional response, morphology and proliferation of *Candida albicans* in human blood. *Mol. Microbiol.* 56, 397–415.
- Garcia-Sanchez, S., Aubert, S., Ibraqui, I., Janbon, G., Ghigo, J.M., and d'Enfert, C. (2004). *Candida albicans* biofilms: a developmental state associated with specific and stable gene expression patterns. *Eukaryot. Cell* 3, 536–545.
- Gasch, A.P., and Eisen, M.B. (2002). Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. *Genome Biol.* 3, RESEARCH0059.
- Gasch, A.P., Spellman, P.T., Kao, C.M., Carmel-Harel, O., Eisen, M.B., Storz, G., Botstein, D., and Brown, P.O. (2000). Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell* 11, 4241–4257.
- Harcus, D., Nantel, A., Marcil, A., Rigby, T., and Whiteway, M. (2004). Transcription profiling of cyclic AMP signaling in *Candida albicans*. *Mol. Biol. Cell* 15, 4490–4499.
- Hughes, J.D., Estep, P.W., Tavazoie, S., and Church, G.M. (2000a). Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J. Mol. Biol.* 296, 1205–1214.
- Hughes, T.R., Marton, M.J., Jones, A.R., Roberts, C.J., Stoughton, R., Armour, C.D., Bennett, H.A., Coffey, E., Dai, H., He, Y.D., *et al.* (2000b). Functional discovery via a compendium of expression profiles. *Cell* 102, 109–126.
- Huh, W.K., Falvo, J.V., Gerke, L.C., Carroll, A.S., Howson, R.W., Weissman, J.S., and O'Shea, E.K. (2003). Global analysis of protein localization in budding yeast. *Nature* 425, 686–691.
- Ihmels, J., Bergmann, S., and Barkai, N. (2004a). Defining transcription modules using large-scale gene expression data. *Bioinformatics* 20, 1993–2003.
- Ihmels, J., Bergmann, S., Berman, J., and Barkai, N. (2005a). Comparative gene expression analysis by a differential clustering approach: application to the *Candida albicans* transcription program. *PLoS Genetics* 1, 0380–0393.
- Ihmels, J., Bergmann, S., Gerami-Nejad, M., Yanai, I., McClellan, M., Berman, J., and Barkai, N. (2005b). Rewiring of the yeast transcriptional network through the evolution of motif usage. *Science* 309, 938–940.
- Ihmels, J., Friedlander, G., Bergmann, S., Sarig, O., Ziv, Y., and Barkai, N. (2002). Revealing modular organization in the yeast transcriptional network. *Nat. Genet.* 31, 370–377.
- Ihmels, J., Levy, R., and Barkai, N. (2004b). Principles of transcriptional control in the metabolic network of *Saccharomyces cerevisiae*. *Nat. Biotechnol.* 22, 86–92.
- Ihmels, J.H., and Bergmann, S. (2004). Challenges and prospects in the analysis of large-scale gene expression data. *Brief Bioinform* 5, 313–327.
- Jones, T., Federspiel, N.A., Chibana, H., Dungan, J., Kalman, S., Magee, B.B., Newport, G., Thorstenson, Y.R., Agabian, N., Magee, P.T., *et al.* (2004). The diploid genome sequence of *Candida albicans*. *Proc. Natl. Acad. Sci. USA.* 101, 7329–7334.
- Karababa, M., Coste, A.T., Rognon, B., Bille, J., and Sanglard, D. (2004). Comparison of gene expression profiling between *Candida albicans* azole-resistant clinical isolates and laboratory strains exposed to drugs inducing multidrug transporters. *Antimicrob. Agents Chemother.* 48, 3064–3079.
- Kellis, M., Patterson, N., Endrizzi, M., Birren, B., and Lander, E.S. (2003). Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* 423, 241–254.
- Kim, S.K., Lund, J., Kiraly, M., Duke, K., Jiang, M., Stuart, J.M., Eizinger, A., Wylie, B.N., and Davidson, G.S. (2001). A gene expression map for *Caenorhabditis elegans*. *Science* 293, 2087–2092.
- Kitano, H. (2002). Systems biology: a brief overview. *Science* 295, 1662–1664.

- Lan, C.Y., Newport, G., Murillo, L.A., Jones, T., Scherer, S., Davis, R.W., and Agabian, N. (2002). Metabolic specialization associated with phenotypic switching in *Candida albicans*. *Proc. Natl. Acad. Sci. USA.* 99, 14907–14912.
- Lander, E.S. (1999). Array of hope. *Nat. Genet.* 21, 3–4.
- Lee, C.M., Nantel, A., Jiang, L., Whiteway, M., and Shen, S.H. (2004). The serine/threonine protein phosphatase SIT4 modulates yeast-to-hypha morphogenesis and virulence in *Candida albicans*. *Mol. Microbiol.* 51, 691–709.
- Lorenz, M.C., Bender, J.A., and Fink, G.R. (2004). Transcriptional response of *Candida albicans* upon internalization by macrophages. *Eukaryot. Cell* 3, 1076–1087.
- McCarroll, S.A., Murphy, C.T., Zou, S., Pletcher, S.D., Chin, C.S., Jan, Y.N., Kenyon, C., Bargmann, C.I., and Li, H. (2004). Comparing genomic expression patterns across species identifies shared transcriptional profile in aging. *Nat. Genet.* 36, 197–204.
- Mootha, V.K., Lindgren, C.M., Eriksson, K.F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M., Laurila, E., *et al.* (2003). PGC-1 $\alpha$ -responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* 34, 267–273.
- Murillo, L.A., Newport, G., Lan, C.Y., Habelitz, S., Dungan, J., and Agabian, N.M. (2005). Genome-wide transcription profiling of the early phase of biofilm formation by *Candida albicans*. *Eukaryot. Cell* 4, 1562–1573.
- Nantel, A., Dignard, D., Bachewich, C., Harcus, D., Marcil, A., Bouin, A.-P., Sensen, C.W., Hogues, H., Hoog, M.v.h., Gordon, P., *et al.* (2002). Transcription profiling of *C. albicans* cells undergoing the yeast to hyphal transition. *Mol. Biol. Cell* 13, 3452–3465.
- Oliva, A., Rosebrock, A., Ferrezuelo, F., Pyne, S., Chen, H., Skiena, S., Fitcher, B., and Leatherwood, J. (2005). The cell cycle-regulated genes of *Schizosaccharomyces pombe*. *PLoS Biol.* 3, e225.
- 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., and Liu, J. (2005). Identification of cell cycle-regulated genes in fission yeast. *Mol. Biol. Cell* 16, 1026–1042.
- Pilpel, Y., Sudarsanam, P., and Church, G.M. (2001). Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat. Genet.* 29, 153–159.
- Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N., and Barabasi, A.L. (2002). Hierarchical organization of modularity in metabolic networks. *Science* 297, 1551–1555.
- Remm, M., Storm, C.E., and Sonnhammer, E.L. (2001). Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.* 314, 1041–1052.
- Rogers, P.D., and Barker, K.S. (2002). Evaluation of differential gene expression in fluconazole-susceptible and -resistant isolates of *Candida albicans* by cDNA microarray analysis. *Antimicrob. Agents Chemother.* 46, 3412–3417.
- Rogers, P.D., and Barker, K.S. (2003). Genome-wide expression profile analysis reveals coordinately regulated genes associated with stepwise acquisition of azole resistance in *Candida albicans* clinical isolates. *Antimicrob. Agents Chemother.* 47, 1220–1227.
- Slonim, D.K. (2002). From patterns to pathways: gene expression data analysis comes of age. *Nat. Genet.* 32 *Suppl.* 502–508.
- Stuart, J.M., Segal, E., Koller, D., and Kim, S.K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. *Science* 302, 249–255.
- Sudarsanam, P., Pilpel, Y., and Church, G.M. (2002). Genome-wide co-occurrence of promoter elements reveals a cis-regulatory cassette of rRNA transcription motifs in *Saccharomyces cerevisiae*. *Genome Res.* 12, 1723–1731.
- Tanay, A., Regev, A., and Shamir, R. (2005). Conservation and evolvability in regulatory networks: the evolution of ribosomal regulation in yeast. *Proc. Natl. Acad. Sci. USA.* 102, 7203–7208.
- Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J., and Church, G.M. (1999). Systematic determination of genetic network architecture. *Nat. Genet.* 22, 281–285.
- Tsong, A.E., Miller, M.G., Raisner, R.M., and Johnson, A.D. (2003). Evolution of a combinatorial transcriptional circuit: a case study in yeasts. *Cell* 115, 389–399.
- Wang, W., Cherry, J.M., Botstein, D., and Li, H. (2002). A systematic approach to reconstructing transcription networks in *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. USA.* 99, 16893–16898.
- Wolfe, K. (2004). Evolutionary genomics: yeasts accelerate beyond BLAST. *Curr. Biol.* 14, R392–394.
- Wu, L.F., Hughes, T.R., Davierwala, A.P., Robinson, M.D., Stoughton, R., and Altschuler, S.J. (2002). Large-scale prediction of *Saccharomyces cerevisiae* gene function using overlapping transcriptional clusters. *Nat. Genet.* 31, 255–265.

- Yuh, C.H., Bolouri, H., and Davidson, E.H. (1998). Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene. *Science* 279, 1896–1902.
- Zhao, R., Daniels, K.J., Lockhart, S.R., Yeater, K.M., Hoyer, L.L., and Soll, D.R. (2005a). Unique aspects of gene expression during *Candida albicans* mating and possible G(1) dependency. *Eukaryot. Cell* 4, 1175–1190.
- Zhao, X., Oh, S.H., Yeater, K.M., and Hoyer, L.L. (2005b). Analysis of the *Candida albicans* Als2p and Als4p adhesins suggests the potential for compensatory function within the Als family. *Microbiology* 151, 1619–1630.