doi:10.1016/j.jmb.2006.06.024 J. Mol. Biol. (2006) 363, 589–610

JMB

Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Modularity of the Transcriptional Response of Protein Complexes in Yeast

Nicolas Simonis[1], Didier Gonze[1,2], Chris Orsi[3] Jacques van Helden[1] and Shoshana J. Wodak[1,3]*

[1]*Service de Conformation des Macromolécules Biologiques Centre de Biologie Structurale et Bioinformatique, CP 263 Université Libre de Bruxelles Bld. du Triomphe B-1050 Bruxelles, Belgium*

[2]*Unité de Chronobiologie Théorique, Service de Chimie Physique, CP 231, Université Libre de Bruxelles Bld. du Triomphe B-1050 Bruxelles, Belgium*

[3]*Structural Biology and Biochemistry Program, Hospital for Sick Children 555 University Avenue Toronto, Ontario M5G 1X8 Canada*

**Corresponding author*

A comprehensive study is performed on the condition-dependent expression of genes coding for the components of hand curated multi-protein complexes of the yeast *Saccharomyces cerevisiae*, in order to identify coherent transcriptional modules within these complexes. Such modules are defined as groups of genes within complexes whose expression profiles under a common set of experimental conditions allow us to discriminate them from random sets of genes. Our analysis reveals that complexes such as the cytoplasmic ribosome, the proteasome and the respiration chain complexes previously characterized as "stable" or "permanent" represent transcriptional modules that are coherently up or down-regulated in many different conditions. Overall however, some level of coherent expression is detected only in 71 out of the total of 113 complexes with at least five different protein components that could be reliably analyzed. Of these, 26 behave as coherently expressed transcriptional modules encompassing all the components of the complex. In another 15, at least half of the components make up such modules and in ten, few or no modules are detected. In an additional 20 complexes coherent expression is detected, but in too few conditions to enable reliable module detection. Interestingly, the transcriptional modules, when detected, often correspond to one or more known sub-complexes with specific functions. Furthermore, detected modules are generally consistent with transcriptional modules identified on the basis of predicted *cis*-regulatory sequence motifs. Also, groups of genes shared between complexes that carry out related functions tend to be part of overlapping transcriptional modules identified in these complexes. Together these findings suggest that transcriptional modules may represent basic functional and evolutionary building blocs of protein complexes.

Keywords: protein complexes; transcriptional regulation; mRNA expression; yeast

Introduction

Protein complexes play an essential role in all cellular processes.[1] The formation of such complexes is most likely dynamically regulated at different levels: transcriptional, post-translational modification and degradation, but the corresponding mechanisms are largely unknown.

The yeast *Saccharomyces cerevisiae* is one of the few organisms where some of these mechanisms, and in particular transcriptional regulation, can be investigated today. A large number of protein complexes have been characterized by affinity purification and mass spectrometry using both high throughput efforts[2,3] and case by case studies, with the results stored in databases such as MIPS/CYGD[4] and SGD.[5] Several genome-scale datasets on gene expression levels measured under many different experimental conditions are also publicly available.[6–8] In addition, information about transcription factor binding sites in yeast has been compiled in specialized databases such as TRANSFAC[9] and SCPD,[10] and genome-wide

Abbreviations used: UPC, uncentered Pearson correlation coefficient; RNAP, RNA polymerase.

E-mail address of the corresponding author: shoshana@sickkids.ca

localization experiments have yielded a wealth of information on the target genes of most yeast trans cription factors.[11,12]

In a previous study some of us investigated the transcriptional regulation of multi-protein complexes in yeast by mapping known sets of co-regulated genes onto these complexes and by identifying regulatory sequence motifs in the upstream regions of the genes coding for the components of complexes.[13] The results suggested that a portion of the analyzed protein complexes might be regulated *via* a combinatorial mechanism in which several transcription factors cooperate in controlling multiple sets of overlapping target genes, most likely in a time and condition-dependent fashion. But the derived picture was sketchy, mainly due to incomplete information on transcription factor–gene associations. Another study[14] investigated the correlation between the expression profiles of genes coding for components of yeast multi-protein complexes measured over a large set of conditions. Significant correlation was found for only a few complexes, such as the ribosome and proteasome, referred to as "permanent" by the authors. In a few other complexes some correlation was detected for subsets of the genes in the complex. But no clear picture emerged, since by considering the expression profiles over all the conditions, this study did not allow the detection of condition-specific responses. Recently, de Lichtenberg *et al*.[15] investigated the temporal gene expression pattern of components of protein complexes active during the cell cycle in yeast, revealing that several complexes were composed of dynamically regulated components affording a "just in time" assembly.

Here we investigate the condition-specific nature of the transcriptional response of the full repertoire of 243 hand-curated complexes stored in the MIPS/CYGD database[4] as revealed from the analysis of gene expression data from several authors,[6,7,16] measured on a total of 549 different DNA chips corresponding to different time points and conditions.

In a first step we identify the particular subsets of conditions (or chips) under which the components of individual complexes are coherently up or down-regulated relative to all other yeast genes. In a second step we use the expression levels of genes under those selected conditions to identify transcriptional modules within complexes. These modules are defined as gene sets corresponding to whole complexes or portions thereof, whose expression profiles under a common set of conditions allow us to discriminate them from random sets of genes. We test for the presence of such modules across the MIPS complexes. Modules identified in a selected set of complexes are described in detail and the role that these modules might play in the recruitment of groups of proteins into different complexes in a condition and time-dependent fashion is discussed.

Results

Cross-talk between multi-protein complexes

Inspection of the 243 multi-protein complexes of *S. cerevisiae* annotated in the MIPS database in terms of their component proteins or genes immediately reveals that individual complexes, as defined in the database entries, often share components. Among the 243 complexes, 128 (53%) are entirely included in some other larger complex and 59 (24%) are singleton complexes that share none of their genes with other complexes. Of the remaining 56 complexes 17 (7%) are "container" complexes, which are composed of two or more of the smaller complexes mentioned above, whereas 39 complexes (~16%) partially overlap with other complexes and hence also contain some genes that are unique to each complex.

This cross-talk can be represented using a graph where nodes depict individual MIPS complexes and two complexes sharing one or more genes are connected by arcs whose thickness is proportional to the number of shared genes (Figure 1). A commonly used force-directed layout algorithm[17] positions the nodes according to the number of genes that they share with other nodes, with "hubs" , the more highly connected nodes, positioned in the centre and weakly connected ones on the periphery. This yields a total of 22 clusters of complexes, containing two or more individual entries in the MIPS catalogue of yeast complexes, and the remaining 59 singleton complexes (see Table S1 of the Supplementary Data for a complete list of clusters).

The clusters of the cytoplasmic and mitochondrial ribosomes, the proteasome or the clathrin-associated complexes, represent typical cases where larger complexes are composed of several smaller building blocks that have also been isolated independently. In some cases such as the 26S proteasome, or the clathrin-associated complexes, the larger container complexes also include a few additional genes that are unique to the larger complex. Often the smaller building blocks carry out distinct functions, either in the context of the larger complex, or individually. They usually engage in physical interactions between themselves when they assemble into the full complex, but can also be stable on their own under specific conditions, and hence be isolated independently.

In a different category of cases, the sharing of genes between two or more complexes is not indicative of physical interactions but of the fact that a particular set of proteins is "recruited" by complex A under one set of circumstances (specific condition set or cell population), and by complex B under a different set. The functions of complexes A and B may differ and the shared proteins may play different roles in these complexes. Researchers working on the purification of individual complexes or groups of complexes can often distinguish

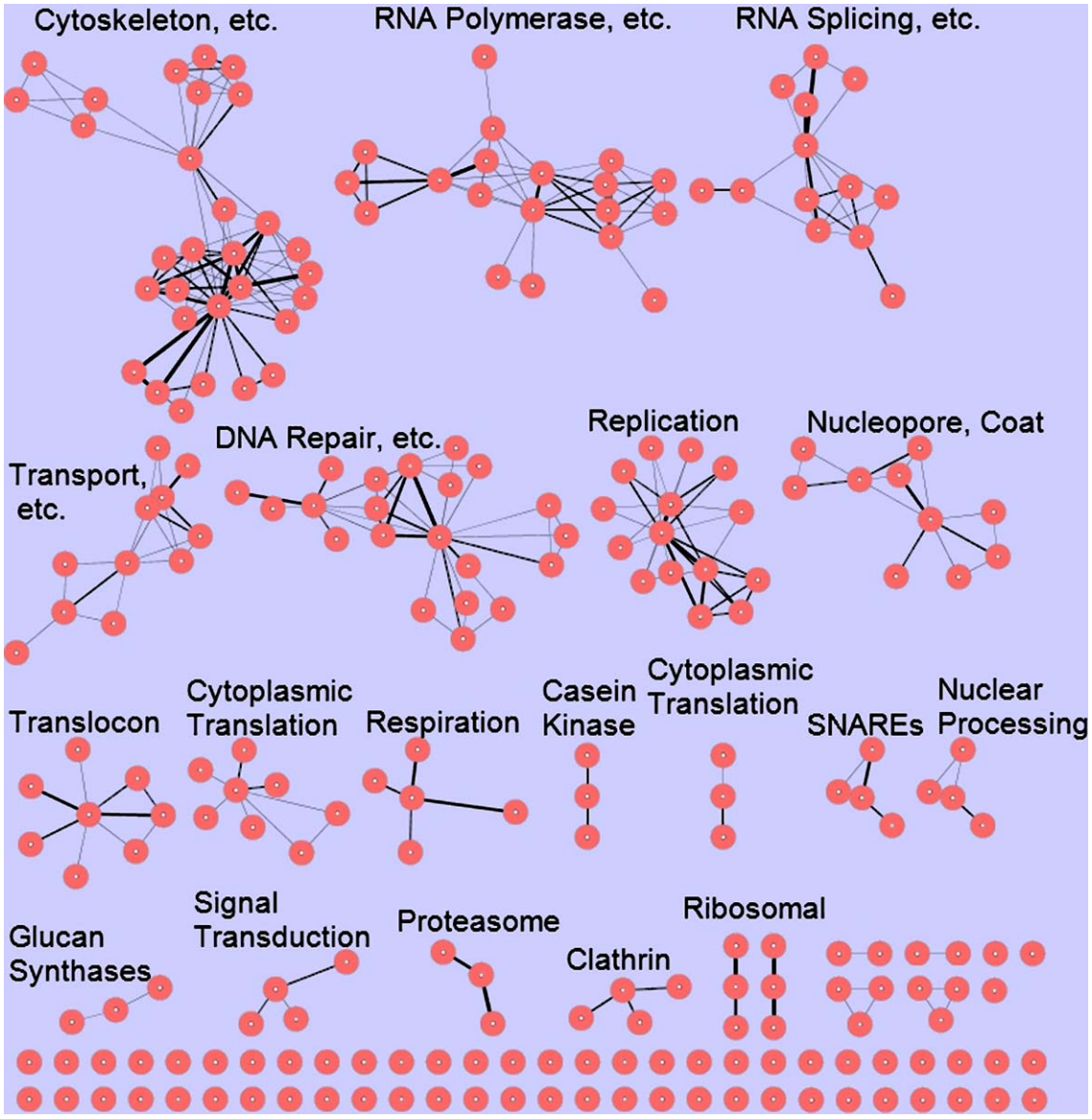


Figure 1. Network of the hand-curated multi-protein complexes of *S. cerevisiae* from the MIPS database. Graph representing the network formed between the 243 hand-curated multi-protein complexes retrieved from MIPS database (version 2003). Each node represents one entry (complex) in the MIPS catalogue, and two complexes are connected by an arc whenever they share one or more genes. The thickness of the arc is proportional to the number of shared genes. The network is composed of 22 distinct clusters of two or more interconnected complexes. Each of these clusters is labelled by the name of the complex that best characterizes its function. The 59 complexes that do not share genes with any other complex appear as singletons, which are arbitrarily aligned on the bottom of the Figure. This Figure is generated using the GenePro plugin for the Cytoscape software.[27,28]

between these two types of cases (physical interaction *versus* recruitment).[18] This information being unfortunately unavailable in the MIPS catalogue however, the cross-talk graph of Figure 1 does not make this distinction, but the ensuing analysis on the conditions-specific expression of complexes will address this issue.

Experimental conditions in which individual complexes show a coherent transcriptional response

As a first step towards unravelling the transcriptional regulation program of protein complexes, we set out to identify the subsets of conditions in which individual complexes are coherently up or down-regulated relative to other yeast genes from the total of 549 experimental conditions in the collated gene expression dataset. To that end we compute the mean and standard deviation of the expression ratios of,respectively, the genes coding for the components of each complex, and all other genes in the yeast genome. These quantities are used to compute the probability (*P*-value) that the mean expression level of the components of a given complex differs from that of all yeast genes under a given condition, using Welsh's *t*-test. A correction for multi-testing is applied in order to compute the expected number

of false positives (*E*-value = *C*x*P*-value, with *C* = 549, being the number of considered conditions). All the experimental conditions in which the components of a given complex are coherently expressed relative to all yeast genes are then selected by setting an upper limit of 0.05 for the *E*-value.

Table 1A. Results on selected conditions under which protein complexes are coherently up- and down-regulated, on the co-regulation of components and transcriptional modules

		t-Test			UPC		Discriminant analysis	
Complex	No. of proteins	Selected conditions	Up	Down	Mod	All	Coverage	PPV
A. *Results for the 51 complexes with five or more components, for which five or more conditions were selected by the t-test*								
Nucleosomal-protein-complex	8	135	55	80	0.96	0.96	1.00	1.00
RNA-polymerase-III	13	104	24	80	0.86	0.86	1.00	0.92
20 S-proteasome	15	87	76	11	0.87	0.87	1.00	0.78
F0-F1-ATP-synthase	14	83	59	24	0.87	0.87	1.00	0.88
Cytochrome-*bc*1-complex	9	51	44	7	0.93	0.93	1.00	0.68
H+-transporting-ATPase-vacuolar	13	38	18	20	0.77	0.77	1.00	0.92
Cytochrome-*c*-oxidase	8	37	29	8	0.94	0.94	1.00	0.80
Chaperonine-containing-T-complex-TRiC	8	23	11	12	0.93	0.93	1.00	0.92
Exosome-complex	7	21	6	15	0.95	0.95	1.00	0.89
Oligosaccharyltransferase	9	20	6	14	0.89	0.89	1.00	0.88
eIF2B	5	14	2	12	0.99	0.99	1.00	0.87
TOM-transport-across-the-outer-membrane	8	13	8	5	0.87	0.87	1.00	0.89
Signal-recognition-particle	6	10	5	5	0.97	0.97	1.00	0.88
eIF3	7	9	4	5	0.95	0.95	1.00	0.86
Tim22p-complex	5	9	8	1	1.00	1.00	1.00	0.86
Arp2p-Arp3p-complex	6	8	8	0	0.97	0.97	1.00	0.84
Cytoplasmic-ribosomal-small-subunit	57	304	89	215	0.74	0.72	0.98	0.96
Cytoplasmic-ribosomes	138	385	120	265	0.74	0.72	0.97	0.81
Mitochondrial-ribosomal-large-subunit	32	173	90	83	0.64	0.62	0.97	0.96
19-22 S-regulator	18	101	79	22	0.77	0.76	0.94	0.96
Cytoplasmic-ribosomal-large-subunit	81	334	109	225	0.74	0.68	0.94	0.97
Mitochondrial-ribosomes	48	213	116	97	0.59	0.54	0.94	0.93
RNA-polymerase-I	14	92	20	72	0.86	0.83	0.93	0.91
Mitochondrial-ribosomal-small-subunit	14	28	15	13	0.86	0.80	0.93	0.89
Respiration-chain-complexes	36	211	154	57	0.74	0.71	0.92	0.99
COPII	11	15	2	13	0.87	0.85	0.91	0.87
26 S-proteasome	36	146	110	36	0.71	0.65	0.89	0.95
Cytoplasmic-translation-initiation	27	170	47	123	0.76	0.70	0.81	0.88
Cytoplasmic-translation-elongation	9	12	0	12	0.94	0.91	0.78	0.87
Nuclear-pore-complex	24	54	16	38	0.68	0.55	0.75	0.86
rRNA-processing-complexes	18	45	15	30	0.81	0.72	0.72	0.86
Coat-complexes	25	45	9	36	0.70	0.44	0.72	0.83
COPI	7	8	0	8	0.96	0.95	0.71	0.78
RNA-polymerase-II	13	17	9	8	0.87	0.79	0.69	0.80
Mitochondrial-translocase-complex	16	37	16	21	0.74	0.65	0.69	0.87
Spindle-pole-body	32	35	5	30	0.57	0.38	0.69	0.86
Translocon	28	81	19	62	0.62	0.54	0.64	0.84
Replication-complex	19	25	0	25	0.71	0.58	0.58	0.79
Pre-replication-complex	14	12	0	12	0.82	0.75	0.57	0.76
Replication-fork-complexes	30	47	14	33	0.77	0.43	0.53	0.90
Actin-filaments	32	17	8	9	0.61	0.47	0.53	0.76
Actin-associated-proteins	24	16	10	6	0.76	0.53	0.46	0.75
Replication-complexes	49	88	14	74	0.56	0.29	0.45	0.88
Microtubules	32	5	1	4	0.79	0.33	0.34	0.73
mRNA-splicing	38	30	16	14	0.58	0.35	0.32	0.70
SNAREs	19	6	4	2	0.81	0.57	0.21	0.56
RNA-polymerase-II-holoenzyme	35	22	9	13	0.69	0.33	0.20	0.61
Nuclear-splicing-complexes-spliceosome	66	63	28	35	0.81	0.23	0.14	0.60
RSC-complex	10	7	2	5	NaN	0.85	0.10	0.46
Cytoskeleton	73	5	2	3	0.72	0.14	0.07	0.55
DNA-repair-complexes	33	6	5	1	NaN	0.39	0.03	0.29

Column 1 lists the names of the complex entries in the MIPS catalogue. The number of components (proteins) in each complex is given in column 2, and the number of conditions selected by the *t*-test is given in column 3 (see Materials and Methods for details). The number of selected conditions under which the complex is coherently up and down-regulated is given in columns 4 and 5, respectively. These conditions were selected from amongst the total of 549 experimental conditions considered here (see Materials and Methods). Columns 6 and 7 list the average pairwise UPC of the components/genes belonging to the transcriptional module (Mod), as identified by the discriminant analysis, and of all the components of the complex (All), respectively. The coverage (percent of the components of each complex assigned to a complex by discriminant analysis) and the Positive Predictive Value (*PVV*) (percent of the genes assigned to a complex by the discriminant analysis, which actually belong to it; the larger this fraction, the smaller the fraction of false positives) are given in columns 8 and 9.

Since the *t*-test assumes a normal distribution of the data and its power depends on the sample size, we test its behaviour on two random models comprising, respectively, normally distributed random values, and expression profiles of randomly selected gene sets having the same size as the complexes. Analysis of the results reveals that the rate of false positives is as expected for the background model (<5%) only for complexes containing at least five components that are coherently expressed under at least two experimental conditions selected by the *t*-test (see Supplementary Data for details).

Our subsequent analysis is therefore restricted to those complexes. They number 71, representing only 29% of all the annotated MIPS complexes and 63% of those with five or more components. It is noteworthy that all remaining complexes with at least five components (41 in total) were not analysed, either due to unavailable expression data or because coherent expression could not be detected on at least two DNA chips. The complete list of selected conditions for each complex can be found on our Web site†.

Up and down-regulation of protein complexes

Table 1summarises the results obtained for the 71 complexes identified as described above, which display a coherent transcriptional response in at least five conditions (Table 1A) and in two to four experimental conditions (Table 1B). This Table lists for each complex the total number of conditions under which it is coherently expressed and the number of conditions among those, where the coherent expression corresponds to up and down-regulation, respectively.

A pictorial illustration of the condition-dependent expression of these complexes is presented in Figure 2(a). This Figure displays a two-dimensional clustering of complexes and condition groups, considering conditions from the Gasch and Spellman studies. Conditions from the Hughes dataset were also analysed but the corresponding results are not represented in this Figure (see Materials and Methods and legend to Figure 2 for details). Inspection of Figure 2(a) and Table 1 reveals that complexes such as the cytoplasmic and mitochondrial ribosomes and their subunits, the respiration chain complexes and their subunits, the proteasome, the nucleosomal-protein complex and the RNA-polymerases I and II, have their components coherently expressed under a larger number of experimental conditions ranging from over 54 conditions (nuclear pore complex) to more than 300 conditions (cytoplasmic ribosome). Other complexes, such as the mRNA splicing complex, RNA polymerase II and the complex of the actin-associated proteins, are coherently expressed under fewer conditions (respectively, 38, 17 and 16 conditions only) and some such as the cyclin-CDK-complex and the kinetochore-protein-complexes, are coherently expressed under none of the conditions analysed here.

Table 1B. Results for the remaining 20 complexes with five or more components identified as coherently expressed by the *t*-test in two to four conditions only

		t-Test			UPC
No. of ORFs	No. of proteins	Selected conditions	Up	Down	All
Nucleotide-excision-repairosome	16	4	4	0	0.56
Mitochondrial-splicing-complexes	14	4	0	4	0.59
SPB-associated-proteins	14	4	0	4	0.73
Gim-complexes	5	4	0	4	0.97
SPB-components	16	3	0	3	0.55
ER-protein-translocation-complex	9	3	1	2	0.90
NEF3-complex	9	3	2	1	0.90
TFIIH	9	3	2	1	0.89
HAT-A-complexes	15	2	0	2	0.58
rRNA-splicing	15	2	1	1	0.77
Tubulin-associated-motorproteins	14	2	1	1	0.76
Clathrin-associated-protein-complex	13	2	2	0	0.70
Cdc28p-complexes	10	2	2	0	0.91
RNase-P	9	2	1	1	0.97
TIM-transport-across-the-inner-membrane	9	2	0	2	0.97
Casein-kinase	8	2	1	1	0.88
Replication-initiation-complex	8	2	0	2	0.91
RNase-MRP	8	2	2	0	0.96
Exocyst-complex	7	2	0	2	0.94
Non-homologous-end-joining-apparatus	7	2	0	2	0.96

Columns 1–5 are as in A. Column 6 lists the average pairwise UPC value computed for the components of the complex and the identified conditions. Further details can be found in the Supplementary Data.

To illustrate the extent to which the expression levels of components of individual complexes may vary under two experimental conditions these levels are mapped directly onto the network of Figure 1, with the result shown in Figure 3(a) and (b). Expression levels are represented as spikes coloured in red (up-regulation) or green (down-regulation). We see for example, that in the *heat.shock.005.minutes.hs.2* condition (5 min after heat shock onset) all the genes in the cytoplasmic ribosome cluster are up-regulated (red spikes), whereas those of the proteasome cluster are down-regulated (green spikes) (Figure 3(a)). Whereas 15 min after heat shock onset (*heat.shock.015.minutes.hs.2*), the cytoplasmic ribosome genes are all uniformly repressed

† http://ftp.scmbb.ulb.ac.be/pub/nicolas/html_upc_daexpr_05se/mips_synthetic_table.html

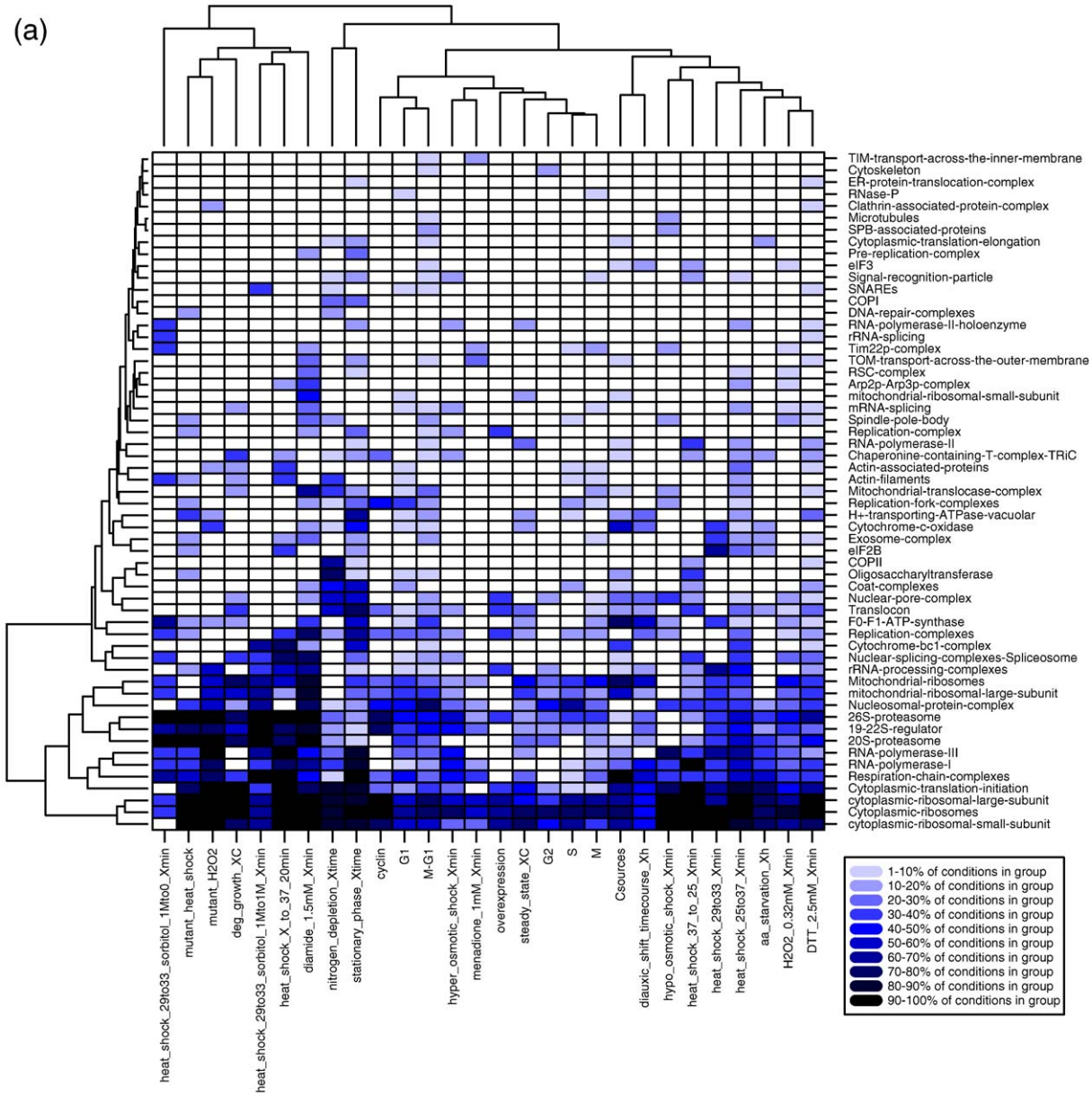


Figure 2. 2D clustering of complexes and experimental conditions in which their components show a coherent transcriptional response. Complete linkage hierarchical clustering performed in two dimensions in order to illustrate groupings of complexes and conditions selected by the *t*-test. Rows represent complexes ranked according to their behaviour with regards to the expression pattern of their components in the considered condition groups; complexes for which the *t*-test selects a large number of conditions in which their components are differentially co-expressed are at the bottom of the graph. Columns represent condition groups, clustered according to how well the considered complexes are differentially co-expressed in each group. The clustering was performed with the R package (http://www.r-project.org/), using as metric the Euclidian distance between the vectors representing the fraction of the conditions in each condition group that passes the *t*-test for a given complex. The size of this fraction is represented by colour shades using the scale shown on the right-hand side of each panel. Note however, that the number of conditions in each group varies widely, with some groups containing only five conditions (DNA chips), whereas others contain in excess of 20. Thus a lower shade in a condition group of 20 may in fact represent a greater proportion of all the conditions selected by the *t*-test for a given complex, than a darker shade for a group with five conditions only. The full list of conditions selected for a complex and the conditions groups to which they belong can be found at: http://ftp.scmbb.ulb.ac.be/pub/nicolas/html_upc_daexpr_05se/mips_synthetic_table.html. The condition groups are defined as explained in Materials and Methods. Only the 248 experimental conditions analysed by Gasch[6] and Spellman,[7] grouped into 28 condition groups were considered here, as those from Hughes *et al.*[16] could not be conveniently grouped. (a) 2D clustering results, with conditions selected by the t-test in which components of complexes are both up-regulated and down-regulated. (b) 2D clustering results as in (a), but colouring the condition groups according to the fraction of conditions in which the complexes are only up-regulated. (c) 2D clustering results as in (a), but colouring the condition groups according to the fraction of conditions in which the complexes are only down-regulated.

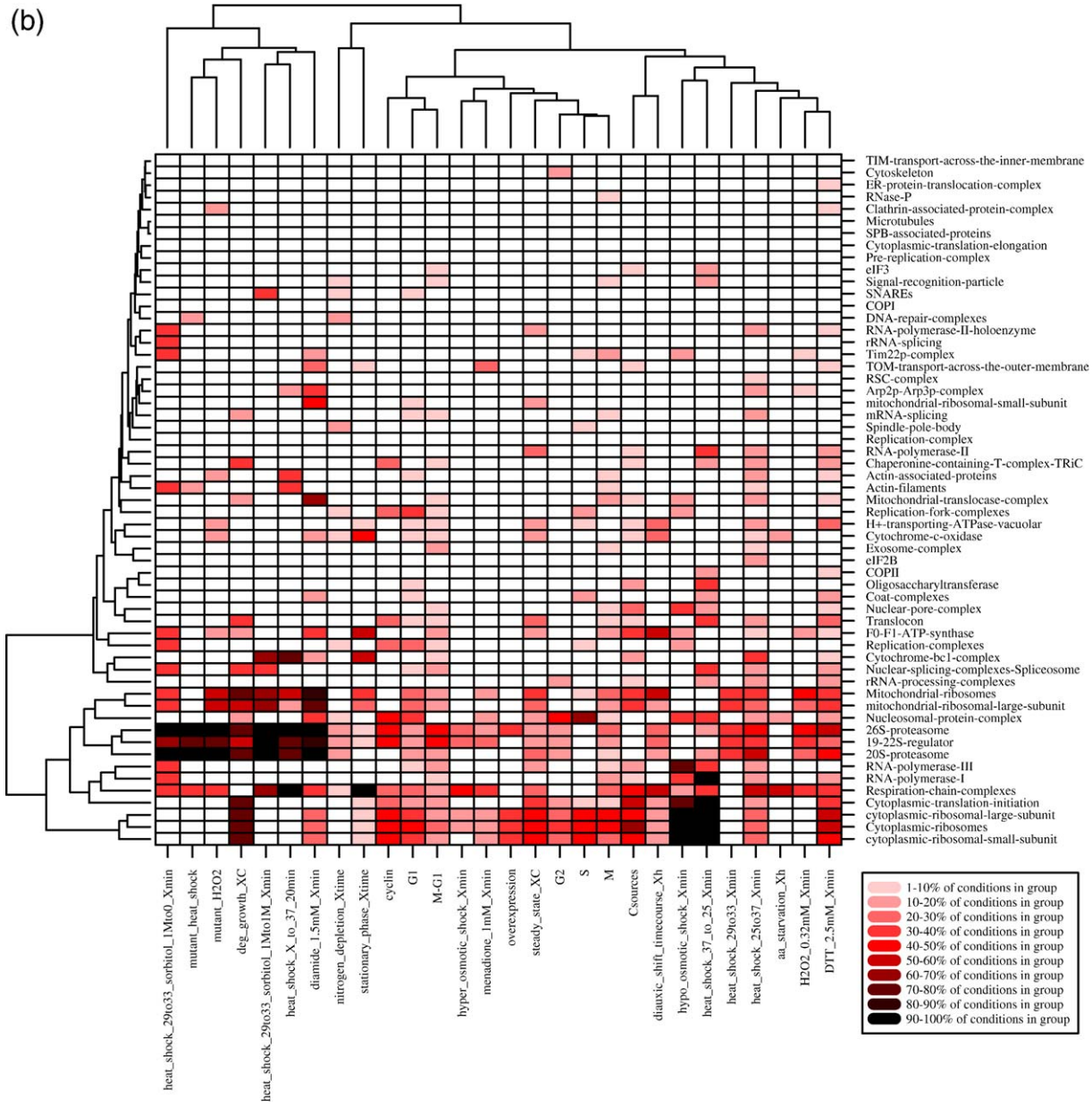


Figure 2 (*legend on previous page*)

(green spikes), while those of the proteasome subunits are activated (red spikes) (Figure 3(b)).

An illustration of the overall trends in the relative directions of the observed co-expression (up or down- regulation) is given in Figure 2(b) and (c), which display the same condition-dependent clustering of complexes as in Figure 2(a), but highlight separately the condition groups under which complexes are, respectively, transcriptionally activated and repressed.

Table 1 and Figure 2 illustrate clearly that more complexes tend to be down than up- regulated under the selected conditions. Amongst the 71 complexes considered here, 34 are repressed under at least 2/3 of the conditions, whereas only 16 complexes are activated under at least 2/3 of the conditions. The remaining complexes are activated/repressed under about half of the selected conditions. A majority of complexes are repressed under the considered conditions most likely because many of these conditions (in the Gasch and Hughes datasets) are stress inducing, causing the cellular machinery to shut down many vital processes. On the basis of this analysis, and with the expression data in hand, the set of multi-protein complexes in yeast, can be subdivided into complexes that display a coherent transcriptional response under many different experimental conditions, and those that display such response under only a limited number of conditions, or under none at all.

It is noteworthy that complexes such as the ribosomes, proteasome, respiration chain complex classified here in the first category, have previously been considered to be permanent or stable complexes by several authors,[14,19,20] implying that these complexes are present under a large number of

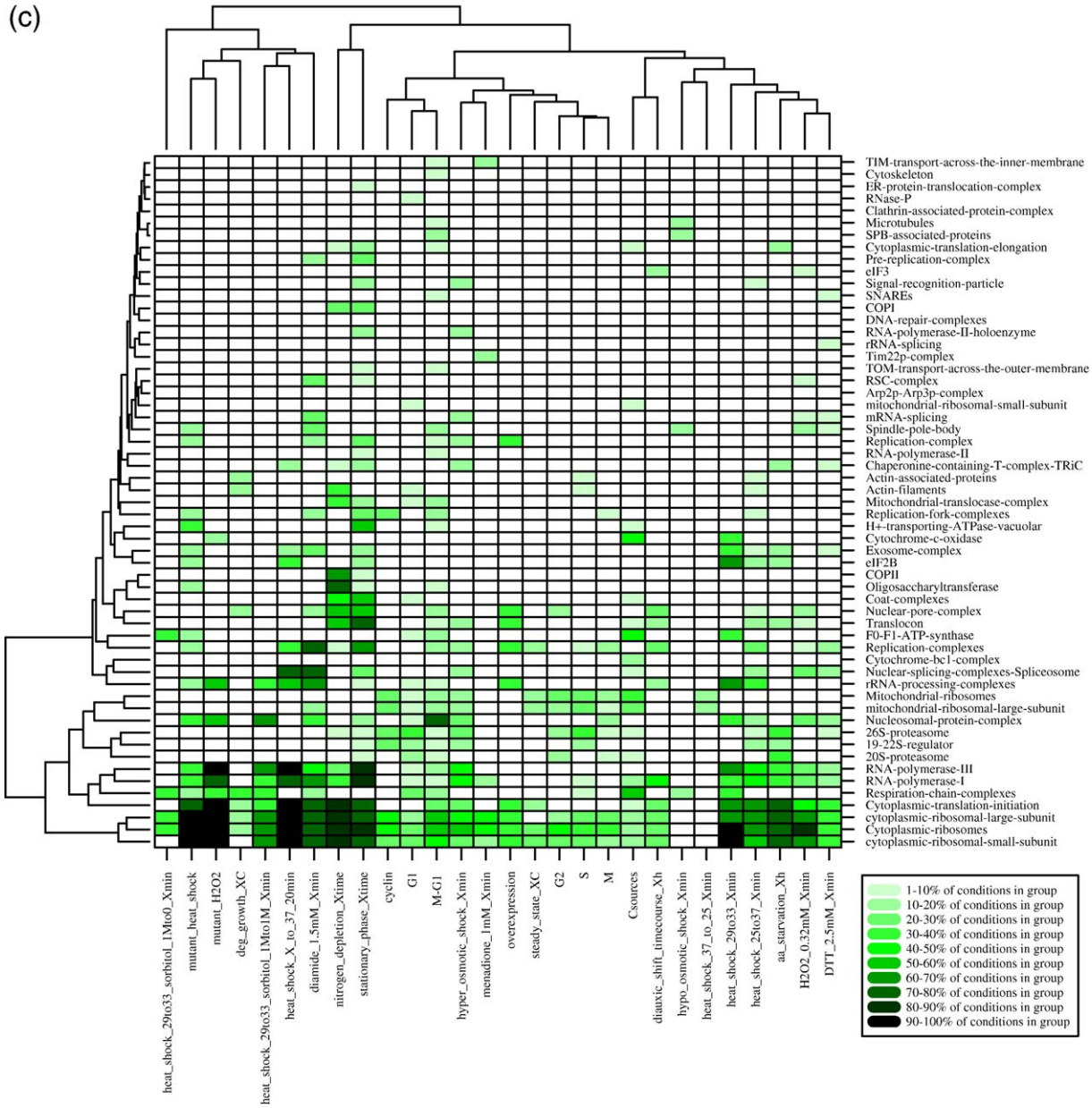


Figure 2 (*legend on page 594*)

experimental conditions. In contrast, "transient" complexes were defined as those that are present (or expressed) under a small number of conditions only.

Our analysis indicates that the so-called permanent complexes are in fact not permanent but have their component genes coherently up or down-regulated in a dynamic fashion under many different experimental conditions, whereas other complexes corresponding to the majority of the MIPS entries seem to lack such collective co-regulation of their components.

Condition-specific co-regulation of complex components

The level of coherence in the transcriptional response of components within complexes can be evaluated by computing for each complex the mean pairwise correlation coefficient between the expression profiles of its components.[14,21] This was done for the 71 complexes analysed here and considering, respectively, all the 549 experimental conditions in the dataset, and only the set of conditions selected by the *t*-test for each complex.

Figure 4 displays the two distributions of the average pairwise uncentered Pearson correlation coefficients (UPC) for the considered MIPS complexes. The UPC distribution computed using all the experimental conditions (purple bars) is clearly bimodal. It has a large peak around 0.1 (very poor correlation), a smaller peak around 0.4–0.5 (higher correlation) and a mean of 0.22.

A similar observation was made previously[21] from the distribution of the mean pairwise correlation coefficient between the expression levels of genes

and their interacting neighbours in the high confidence yeast protein–protein interaction network.

As expected, the distribution of the mean UPC values computed considering only the set of conditions under which the components of individual complexes are coherently expressed is shifted towards much higher values (yellow bars). It has a mean average UPC of 0.73, a large peak close to 1 (reflecting tightly co-regulated complexes) and several smaller ones at lower values. Overall however, the rather wide spread of this curve suggests that the level of coherence may differ between complexes as well as among components within the same complex, with the latter being a direct consequence of the noise level tolerated by the *t*-test.

Identifying transcriptional modules within complexes

Complexes with a moderate average UPC value might have either all their components moderately co-regulated or contain simultaneously tightly co-regulated subsets of components as well as non-co-regulated ones. To identify subsets of more tightly co-regulated components within complexes we applied a cross-validated linear discriminant analysis. The variables used for discrimination were the standardized log expression ratios (*Z*-scores) of individual genes of a given complex under the experimental conditions selected by the *t*-test (*E*-value ≤ 0.05) for that complex. To minimize the effect of noise this analysis was restricted to complexes for which at least five conditions were reliably selected by the *t*-test, which numbered 51, out of the total of 71 with at least two selected conditions. For each complex we looked for a combination of experimental conditions, which enables an optimal classification of each of its components (genes) into two groups: the complex and 100 independent draws of a random group of genes of the same size as the complex ([13,19] and Materials and Methods) The probability P that a component belongs to the complex was computed as the average over the posterior probabilities evaluated in all the trials. The component was assigned to the complex when $P \geq 0.5$.

The results of this analysis in terms of coverage and Positive Predictive Value (*PPV*) are listed in columns 6–9 of Table 1A. Coverage (or sensitivity) is defined as the fraction of the components of a complex that were reassigned to it by the discriminant analysis. The *PPV* is the fraction of the components classified as belonging to the complex by the discriminant analysis, which are actually part of it. Both fractions are expressed on the scale of 0 to 1.

Sixteen of the 51 analysed MIPS complexes have a coverage of 1, meaning that a set of experimental conditions can be found from amongst those that were selected by the *t*-test, which distinguishes between the particular components of each complex and sets of random genes.

A further ten complexes in Table 1A display a coverage higher than 0.9. A closer inspection of those reveals that the few components misclassified by the discriminant analysis, were so mainly because of missing data in the expression profiles.

Thus, in total, 26 of the considered complexes exhibit a characteristic pattern of condition-dependent expression, which allows us to reliably distinguish between all their components and sets of random genes. These complexes include the cytoplasmic and mitochondrial ribosomes, the proteasome subunits, RNA polymerases I and III, as well as other complexes such as the respiration chain complexes, F0-F1-ATP-synthase, H^+ transporting ATPase-vacuolar.

In general, these complexes also exhibit high average UPC values (0.85 −1.0) (Table 1A), although this is not always the case, as will be discussed below. These observations taken together, suggest that these complexes represent transcriptional modules, or portions of such modules (one may indeed expect that additional genes outside the complex might be transcriptionally co-regulated with those in the complex).

Different complexes behave as transcriptional modules under different conditions, as indicated by the differences between the number and types of conditions selected for individual complex by the *t*-test (column 3 of Table 1A, and Supplementary Data).

At the other side of the spectrum, we find complexes such as the DNA-repair complexes, the cytoskeleton complex, the chromatin structure remodelling-complex (RSC), spliceosome, and SNAREs, which display low coverage (<0.5) in the discriminant analysis. These ten complexes, each taken as a whole, do not represent transcriptional modules under any of the conditions selected by the *t*-test, nor do they seem to contain such modules, as confirmed by their low pairwise UPC.

Complexes with intermediate coverage values ($0.5 \leq x < 0.9$) are particularly interesting, as they represent cases where a sizable portion of the complex, containing half of the components or more, might represent a transcriptional module, whereas the remaining genes are not part of the module. All 15 complexes in Table 1A with this coverage range have therefore been analysed in detail to determine the relation between the components that were assigned to the complex by the discriminant analysis and the pairwise UPC computed over the condition set selected by the *t*-test.

Several interesting examples are illustrated in Figure 5 and described in detail below. Results for additional complexes can be found in the Supplementary Data, or on the Web site‡.

Coat complexes

The MIPS entry for this assembly lists 25 genes, which are organised into several sub-complexes

‡ http://ftp.scmbb.ulb.ac.be/pub/nicolas/html_upc_daexpr_05se/mips_synthetic_table.html

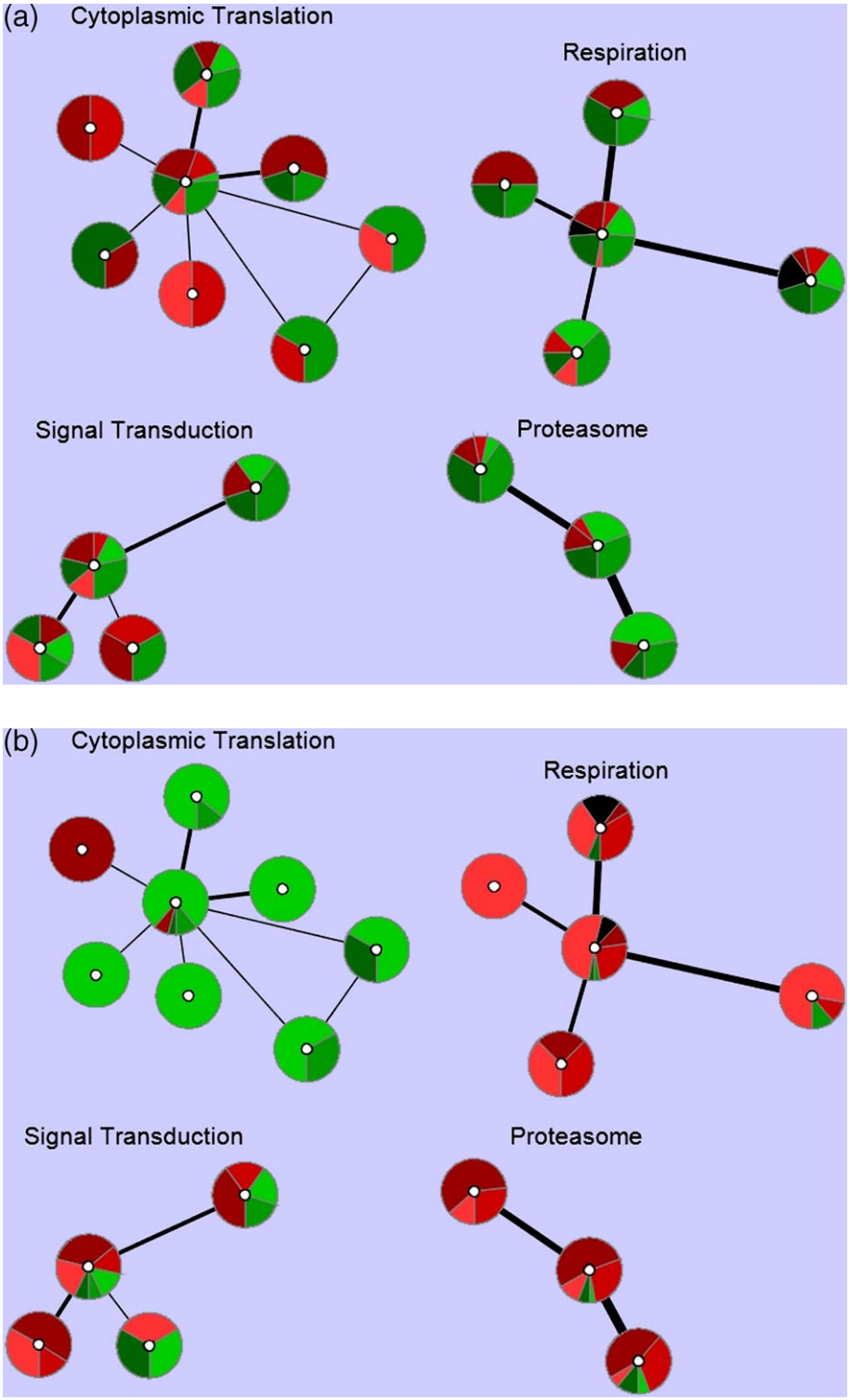


Figure 3. Expression levels of components of individual complexes in two specific experimental conditions, mapped onto the network of the MIPS complexes. The network of complexes is the same as in Figure 1, except that a close up view is given of 4 groups of complexes: the group involved in Cytoplasmic translation, the Respiration group, the group of Signal transduction complexes and the proteasome complexes. The name of each group is that of the most centrally located MIPS complex in the group. Each complex is represented by a pie chart node, and two nodes are linked whenever the complexes share at least one gene, with the thickness of the arc being proportional to the number of shared genes. The wedges of each pie chart represent the fraction of the proteins in the complex with an expression Z-score within a given range at the considered condition. Five Z-score ranges were considered, 3 for up-regulated expression: $Z>1$ (red), $0.5\leq Z<1$ (medium red), $0\leq Z<0.5$ (dark red), and 3 for down-regulated expression: $Z<-1$ (bright green), $-0.5\geq Z<-1$ (medium green), $-0.5\leq Z<0$ (dark green). Black wedges represent genes/proteins with no available expression data (NA).(a) Expression levels of components of complexes in the "heat.shock005.minutes.hs.2" conditions form Gasch.[6] (b) Expression levels of components of complexes in the "heat.shock015.minutes.hs.2" conditions form Gasch.[6]

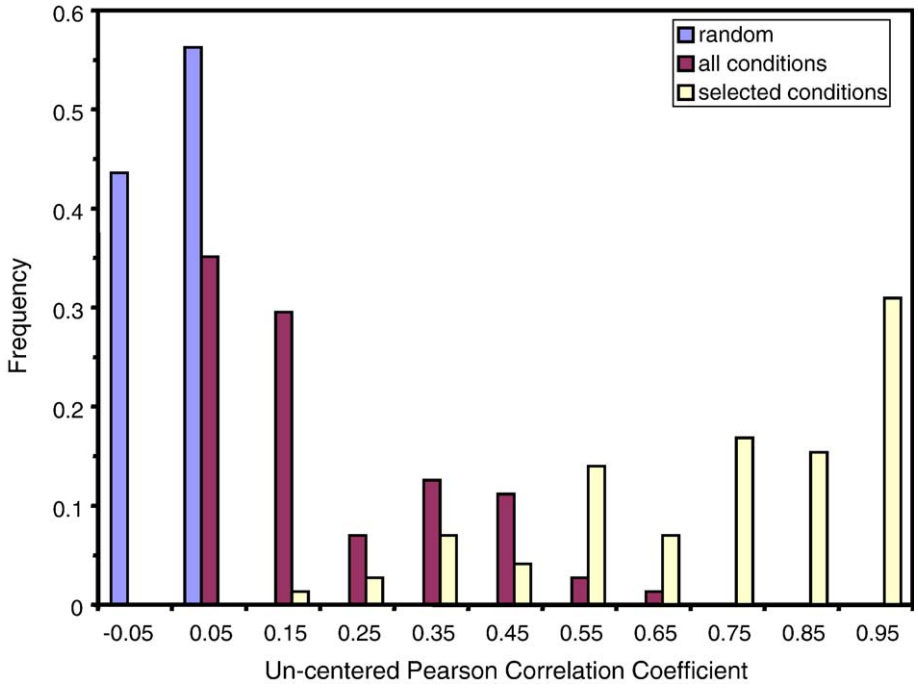


Figure 4. Distributions of the mean pairwise UPC between the expression levels of genes within the MIPS complexes. Three distributions are plotted, considering the complexes where at least two conditions were identified by the *t*-test using the *E*-value threshold $\geq$0.05: (1) mean UPC between pairs of genes within each complex (see Materials and Methods) considering all the experimental conditions in which gene expression has been measured (black); (2) mean pairwise UPC considering only the conditions selected by the *t*-test for each complex (red); (3) mean pairwise UPC between genes pairs for random corresponding complexes (blue).

(COPI, COPII, the retromer and smaller assemblies). These complexes direct membrane trafficking between early compartments of the secretory pathway in eukaryotic cells.[22] The *t*-test selects 45 conditions, the largest fraction of which is in the Hughes set (17), the stationary phase (13) and nitrogen depletion (5) groups (see Figure 2(a)), in which the 25 protein coding genes of these complexes are coherently expressed relative to other genes. The complex is down-regulated under 36 of the conditions (different time points of the above-mentioned condition groups) and up-regulated under nine conditions (spread over several condition groups). The discriminant analysis on the basis of the selected conditions assigns 18 of the genes to the complex (72% coverage, Table 1A), but fails to assign the remaining seven.

The 18 correctly assigned genes also display the highest pairwise UPCs between their expression profiles under the selected conditions, whereas the seven unassigned genes feature much lower UPCs between themselves and with the remaining genes of the complex (Figure 5(a)).

The unassigned genes comprise all the five retromer genes, one gene from COPII, and a gene annotated as "other" in the MIPS catalogue. We could verify that the *t*-test selects only one experimental condition for the retromer complex, suggesting that its five genes are not co-regulated. On the other hand eight and 15 conditions are selected, respectively, for the sub-complexes COPI and COPII. The pairwise UPC computed considering these complexes and their selected conditions individually, or when they are part of the larger assembly and its selected conditions, remains high. Our results thus suggest that the COPI and COPII complexes taken together have nearly all their components co-regulated at the transcriptional level, under the selected experimental conditions, whereas those of the retromer complex are not co-regulated under the experimental conditions analysed here.

Nuclear pore complex

This is a transmembrane complex responsible for the nuclear transport of polypeptides. The MIPS entry has 24 genes for this complex. The *t*-test selects 54 experimental conditions, in which the components of this complex are coherently expressed. These conditions are principally in the stationary phase (12) and nitrogen depletion (six) groups , and a few (five) from the different cell cycle groups (see Figure 2(a)) and from the Hughes dataset (15). The complex is down-regulated in 38 of these conditions (stationary phase and nitrogen depletion groups) and up-regulated under 16 conditions (spread across the two condition groups mentioned above). The discriminant analysis assigns 18 genes to the complex, but fails to assign six. With one or two exceptions the assigned genes also display higher UPC than the unassigned set (Figure 5(b)). Interestingly, five out of the six

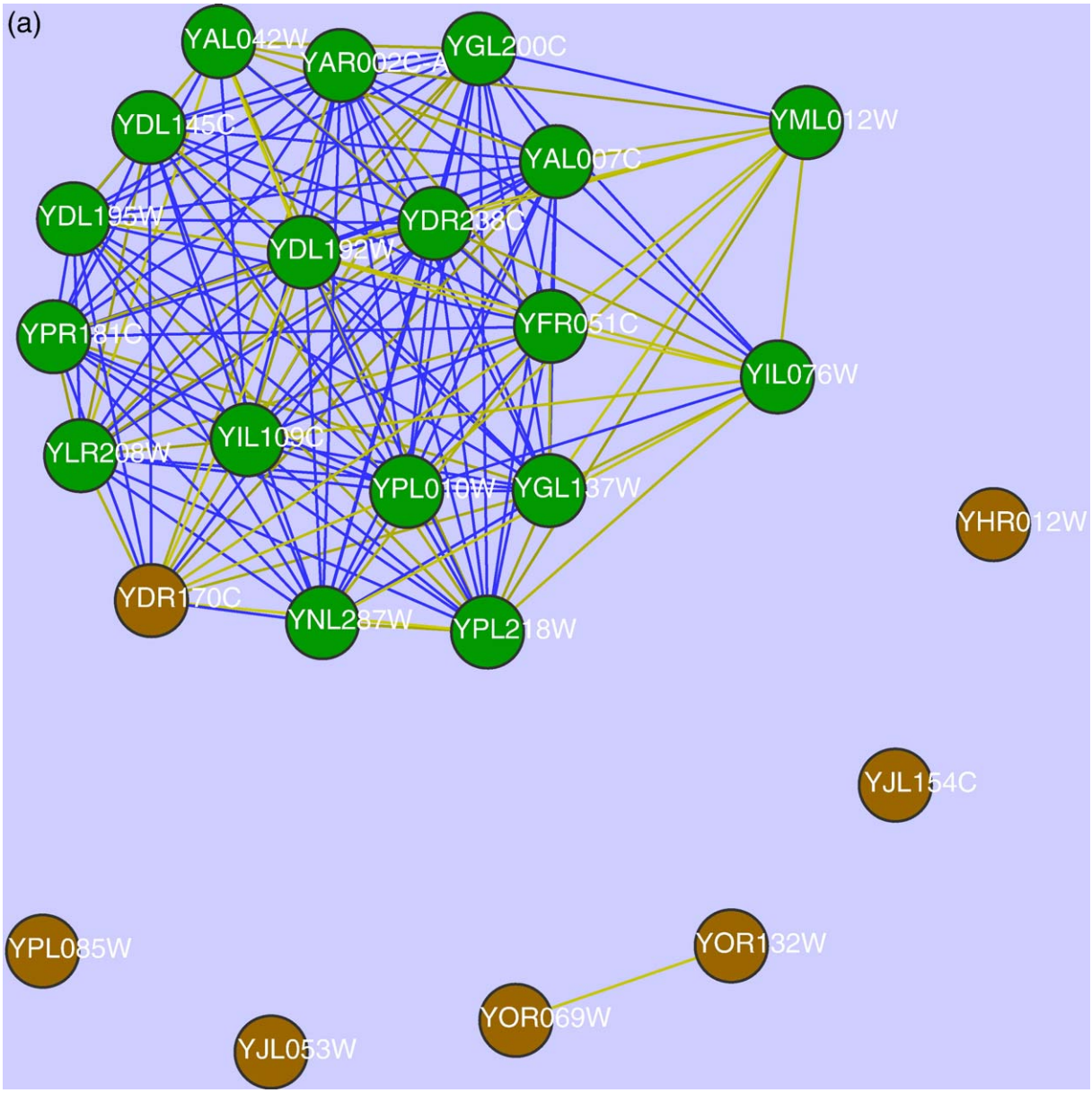


Figure 5. Examples of identified transcriptional modules in multi-protein complexes. Examples of transcriptional modules identified by combining the *t*-test and discriminant analysis. These examples are taken from complexes in which between 50%–90% of their components were found to belong to a transcriptional module (0.5–0.9 coverage by the discriminant analysis, see Table 2). For each complex, individual components (genes) belonging to the transcriptional module (assigned with a probability ≥ 0.5 to the complex by the discriminant analysis) are represented as filled green circles. Those not part of the module (assigned with a probability ≤ 0.5) are shown as filled brown circles. The genes names are displayed in full. Pairs of genes/proteins with a UPC ≥ 0.7 are linked by blue arcs, and pairs of genes with lower UPC ($0.5 \leq \text{UPC} < 0.7$) are linked by yellow arcs and their position on the Figure is given by the Force Directed layout algorithm of the Cytoscape package.[27] All Figures were generated with the GenePro Plugin.[28] (a) Coat complexes: 25 genes, 45 conditions selected by the *t*-test. 18 genes are part of the transcriptional module, seven are not. (b) Nuclear pore complex: 24 genes, 54 conditions selected by the *t*-test. 18 genes are part of the transcriptional module, six are not. (c) RNA polymerase I: 14 genes, 92 conditions selected by the *t*-test. 13 genes are part of the transcriptional module, one is not. Five genes are shared with RNAP II and III (circled in red). Two additional genes are shared with RNA polymerase III (circled in purple). (d) RNA polymerase II: 13 genes, 17 experimental conditions selected by the *t*-test. Nine genes are part of the transcriptional module, four are not. The genes common with RNAP-I and III are circled in red. (e) RNA polymerase III: 12 genes, 104 conditions selected by the *t*-test. All 12 genes are in the transcriptional module. The genes shared with RNAP-II and I are circled in red, those shared with RNAP I only are circled in purple. (f) Replication fork complexes: 30 genes, 47 experimental conditions selected by the *t*-test. 16 genes are part of the transcriptional module, 14 are not. The same 16 genes have been found to represent a transcriptional module on the basis of inferred *cis*-regulatory patterns (see Figure 6).

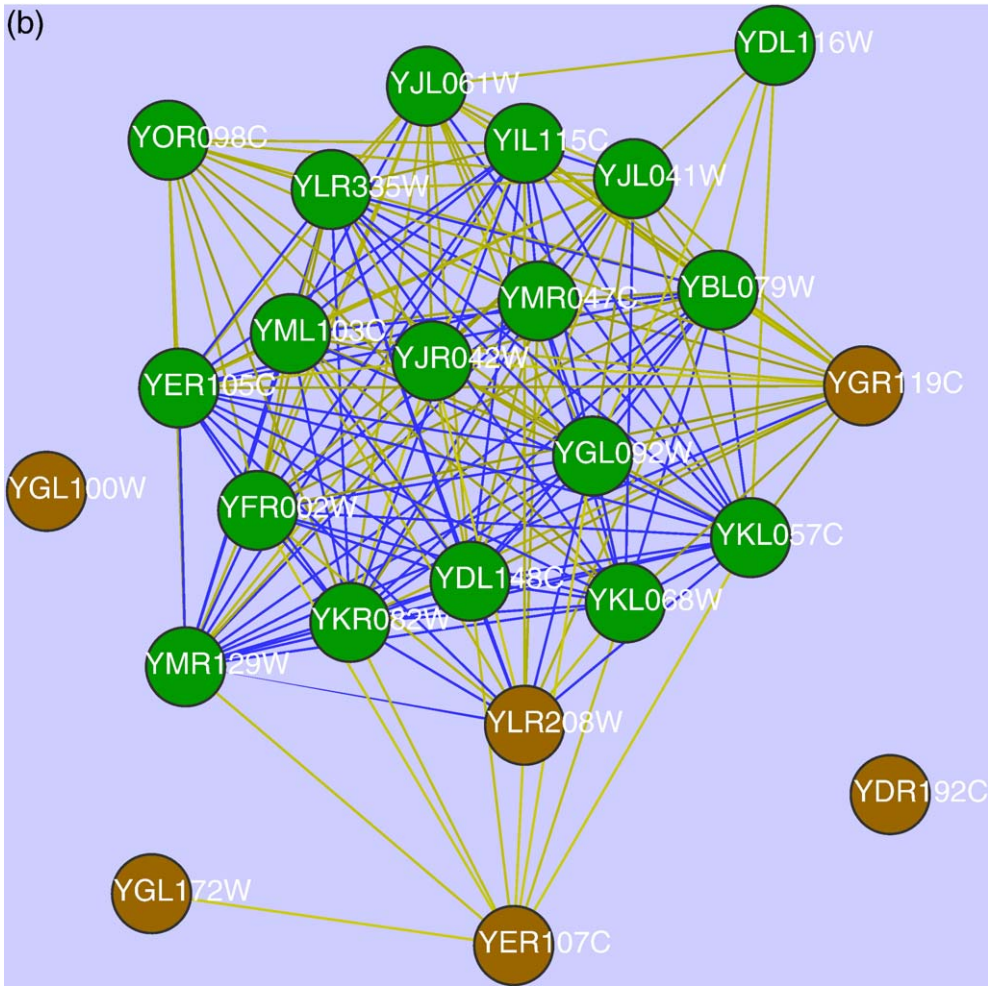


Figure 5 (*legend on previous page*)

unassigned and more loosely co-regulated genes are part of the cytoplasmic and nuclear peripheries of the central pore, respectively.

RNA polymerase II and related complexes

The RNA polymerase II and its related complexes RNA polymerase I and III (RNAP-I, II and III), are responsible for transcription of mRNA, ribosomal RNAs and smaller RNAs such as tRNA or snRNA, respectively. They share some of their genes but not all, and illustrate well how the dynamic modularity unravelled here operates when some components are associated with several complexes that carry out distinct but related functions.

The three polymerases share five genes YBR154C (RPB5), YPR187W (RPB6),YHR143W-A(RPB12), YOR224C(RPB8), YOR210W (RPB10). RNAP-I and III share two additional genes, YNL113W (RPC19, a homolog of RPB11)) and YPR110C (RPC40, a homolog of RPB3); given in parentheses are the common names for these genes.

Our analysis indicates that the 14 components of RNAP-I are coherently expressed in 92 experimental conditions (up-regulated in 20 and down-regulated in 72). The discriminant analysis assigns all 14 components, except one (YDR156W (RPA14)) to the complex with high probability and the 13 genes assigned to the complex also have highly correlated expression profiles as indicated by their pairwise UPC (see Figure 5(c)). The five genes shared between the three RNA polymerases are thus part of this tightly co-regulated module of RNAP-I.

The MIPS entry for RNAP-II comprises 13 genes. The *t*-test selects a total of only 17 experimental conditions in which the components of this complex are coherently expressed. In nine of those (four in the heat shocks group) the complex is up-regulated and in eight (five of which are from the Hughes mutant strains), the complex is down-regulated. Ten of these conditions are common with those selected for RNAP-I and 11 are common to those selected for RNAP-III. We see however that the discriminant analysis achieves only partial coverage (69%) for this complex, as four of its components could not be assigned to the complex (Figure 5(e)) and that the probabilities with which the remaining components are assigned to the complex are generally lower than in RNAP-I (see Table 1A). This suggests that the components of this complex might be less tightly co-regulated or that their expression data are more

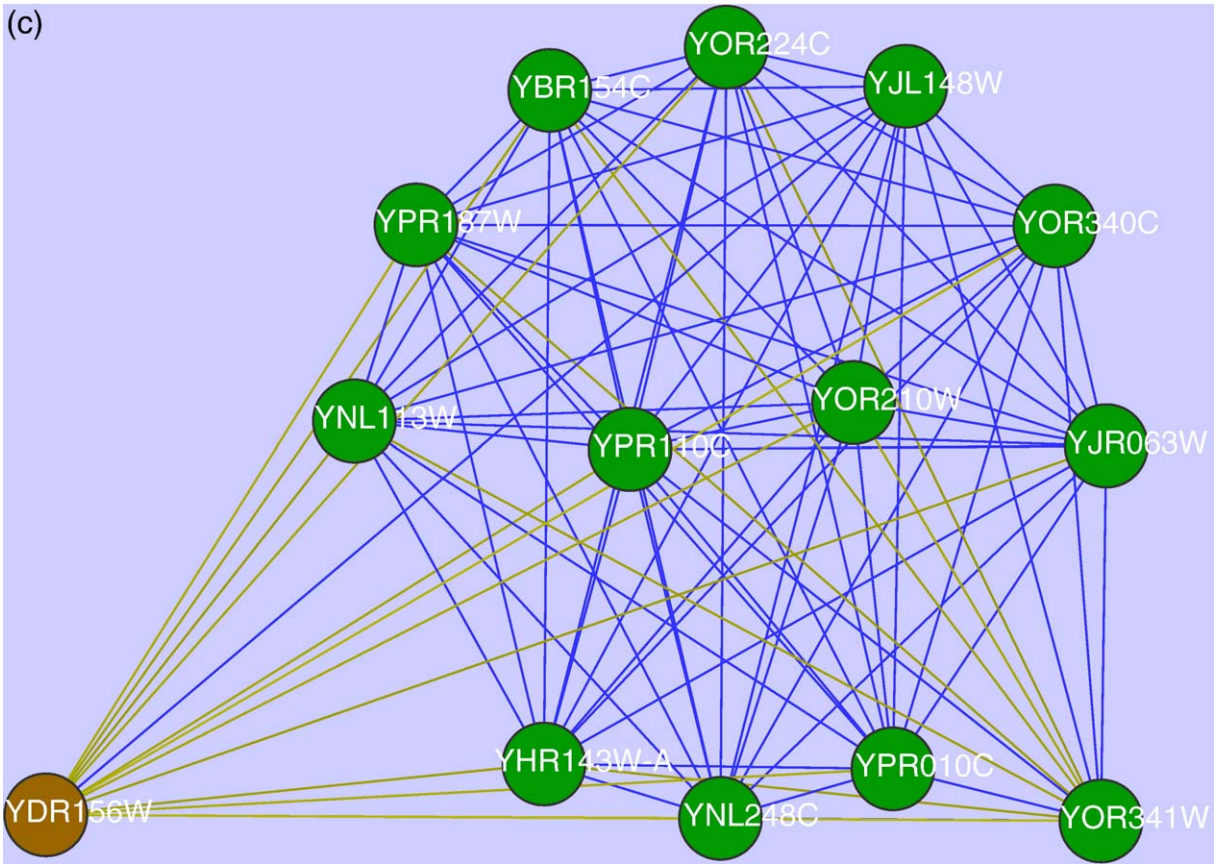


Figure 5 (*legend on page 600*)

noisy, as witnessed moreover by the lower pairwise UPC (Figure 5(d)). This notwithstanding, the above listed five genes shared with RNAP-I and III, are part of the co-regulated module of this complex.

RNAP-III contains 12 components according to MIPS. These components are coherently expressed under 104 experimental conditions (including 18 of the stationary phase, 24 of the heat shock group, 23 of the Hughes mutant dataset), 65 of which are common with RNAP-I and only ten with RNAP-II. Of the 104 selected conditions, up-regulation is observed under 24 conditions, the most relevant groups being cell cycle (phases M,M-G1,G1) and hypo osmotic shock, and down-regulation occurs under 80 conditions (heat shock and stationary phase groups accounting for 34 of those). The discriminant analysis indicates that the complex as a whole behaves as a transcriptional module, as all of its genes are assigned to the complex with high probability and display high UPC values (Figure 5(e)).

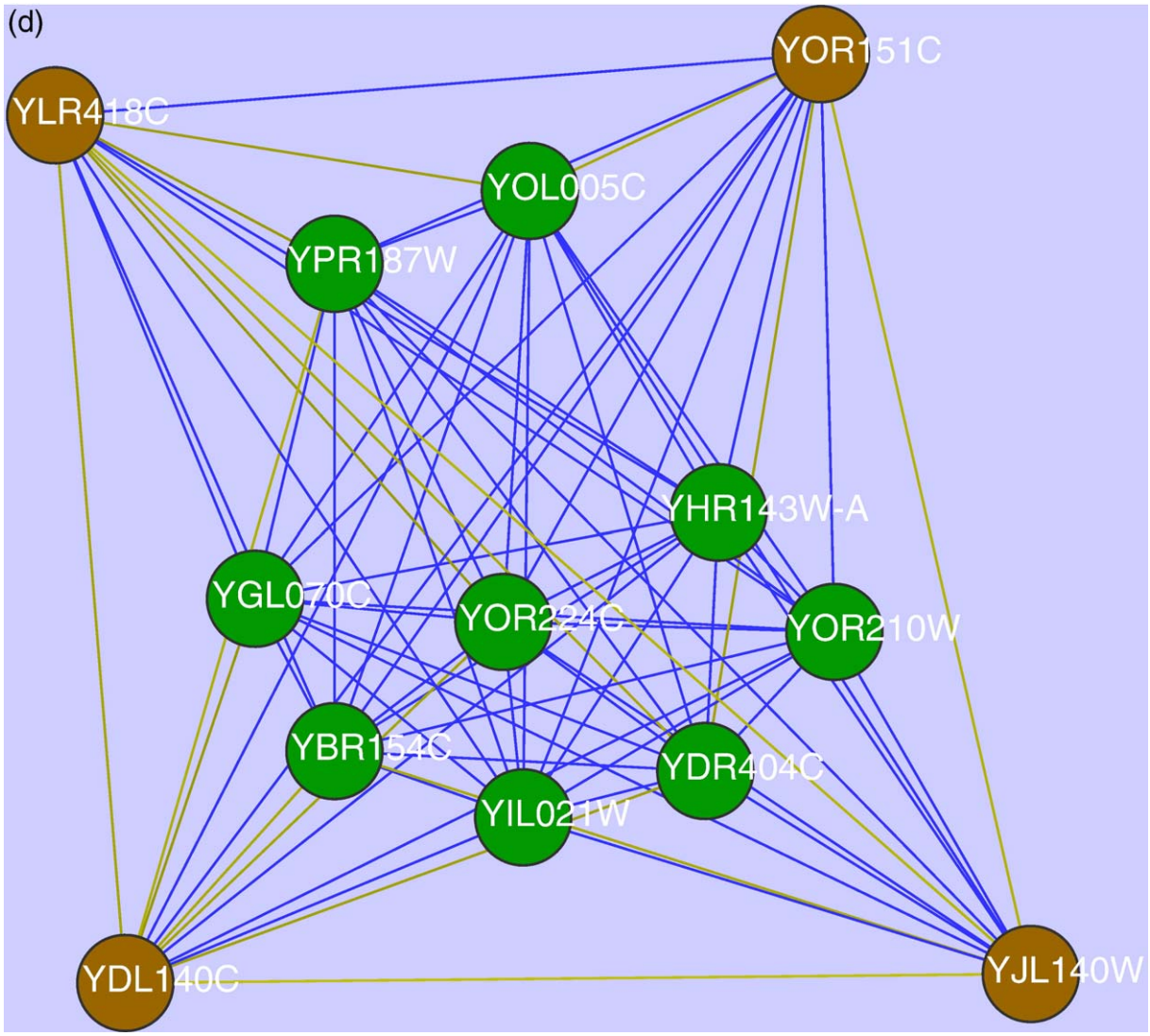
Thus, RNAP-I and III are each tightly co-regulated and coherently expressed under a large number of conditions, of which a good fraction is the same for both complexes, whereas RNAP-II is coherently expressed under a much more limited number of conditions and less tightly co-regulated. These differences might be rooted in specific functional requirements, some of which might relate to the life times of the corresponding mRNA molecules (Wang *et al.*[24]) (see Discussion).

We see in addition that the seven genes shared between RNAP-I and III, which represent about half of each complex, are part of transcriptional modules in both complexes as well.

Shared genes between the different polymerases hence tend to belong to transcriptional modules in these complexes. It might be that they actually represent an independent transcriptional module under a subset of the identified conditions. But further analysis under different subsets of the selected conditions is needed to establish this fact. It is remarkable that many of the genes that are specific for each of the three polymerases (those that are not shared) are related to one another (homologs). One can therefore surmise that a number of related transcription factors with partially overlapping target gene sets might be operating as regulators for these modules. Unfortunately only scant information is available on such transcription factors.[13]

Replication fork complexes

Interesting results were also obtained for this assembly, which is involved in the cell cycle-dependent DNA replication. The MIPS entry for this assembly comprises 30 genes, and the *t*-test selects 47 experimental conditions under which the genes of this complex are coherently expressed. These conditions are mainly from the Hughes



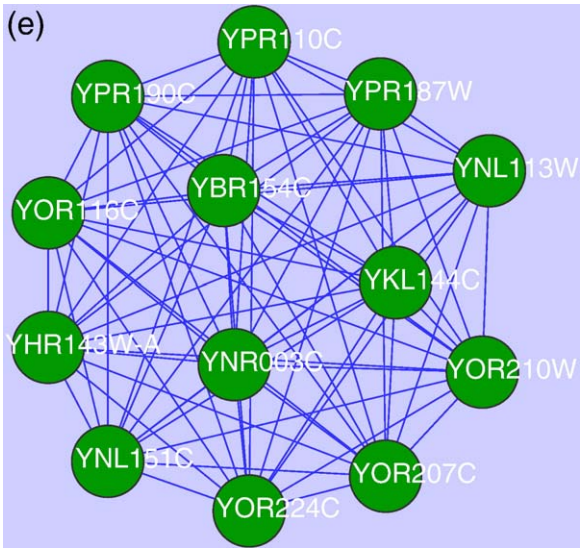


Figure 5 (*legend on page 600*)

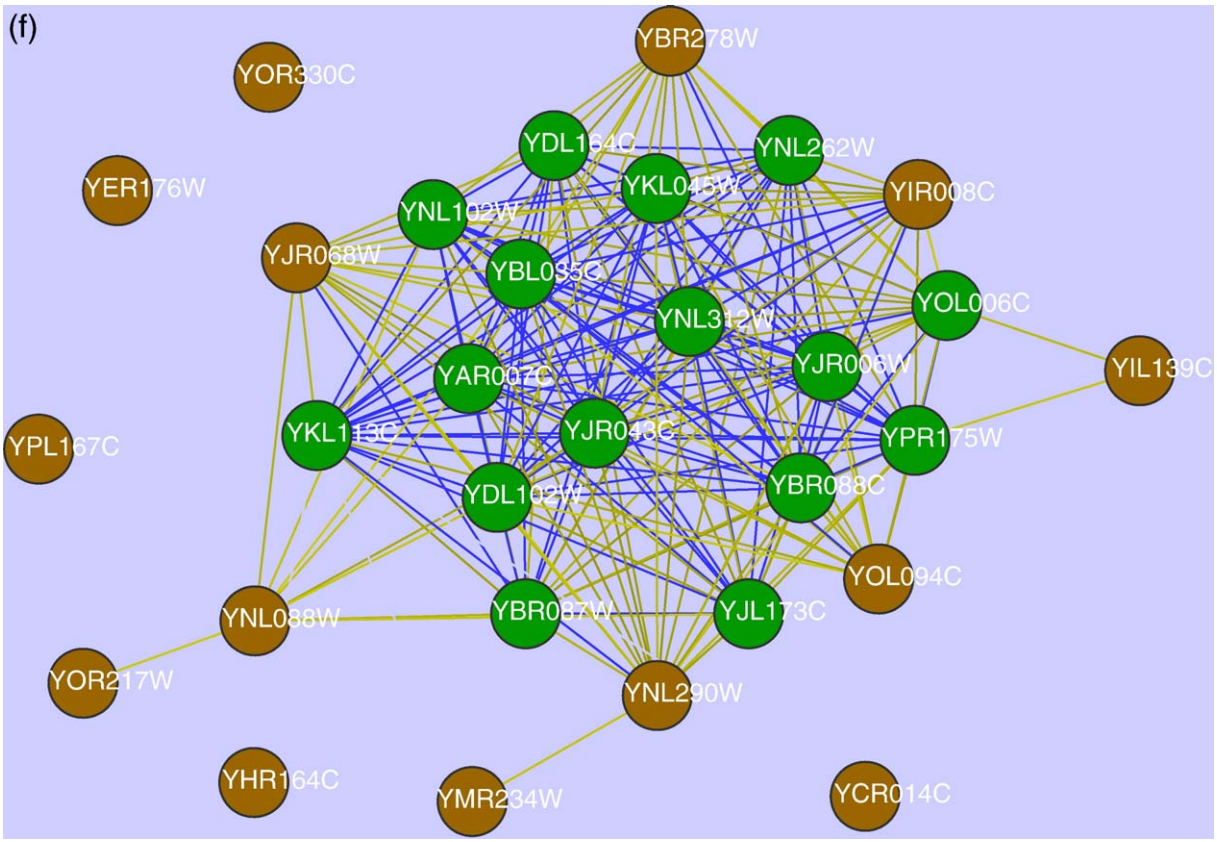


Figure 5 (*legend on page 600*)

mutant dataset (21), the cell cycle (13) and stationary phase(6) groups. The complex appears to be up-regulated in 14 of the conditions (nine of them in G1 and S), and down-regulated in 33 (mostly in the Hughes dataset (20) and stationary phase (6)). Figure 5(f) shows that the discriminant analysis assigns only 16 of the genes to the complex, leaving 14 genes unassigned (53% coverage in Table 1). The 16 assigned genes also feature much higher UPCs than their unassigned counterparts. With two exceptions all the genes assigned to this assembly on the basis of their expression patterns are part of well-defined sub-complexes. Those are the DNA polymerases I, II and III complexes, the replication factor A complex, the exonuclease RAD27, the topoisomerases TOP1 and 2 and PCNA (Pol30). Whereas all the unassigned genes belong to the DNA polymerases IV, γ and z, the DNA ligase (CDC9), the replication factor C complex, two DNA helicases and RNaseH1. Quite remarkably, the same sub-division was observed in a previous study, where some of us identified co-regulated genes in the replication fork complexes on the basis of *cis*-regulatory sequence motifs predicted in the non-coding regions of the corresponding genes.[19] Almost the same set of genes was predicted as co-regulated on the basis of the predicted regulatory motifs, as those assigned here to the complex on the basis of the set of conditions under which their genes are differentially co-expressed. This excellent correspondence is illustrated in Figure 6, which plots the relations between the probabilities of individual genes being assigned to the complex by the discriminant analysis performed here, and those computed by another discriminant analysis on the basis of regulatory patterns identified in the upstream regions of the corresponding genes.[21] These different observations strongly suggest that about half of the components of this complex, comprising the DNA polymerases I, II and II and the few functionally different genes mentioned above, are more tightly co-regulated, most likely by the same transcription factor or factors, than the remaining 14 genes.

Complexes coherently expressed under less than five conditions

Twenty complexes with at least five components were not amenable to the discriminant analysis, because coherent expression of their components was detected under too few conditions (two to four) to enable reliable classification. Results for these complexes are given in Table 1B. This Table lists for each complex the total number of conditions selected by the *t*-test, the number of conditions in which each complex is up and down-regulated relative to all other genes, respectively, and the average pairwise UPC value for the complex. For the majority of these complexes (12) only two conditions were selected by the *t*-test, and about half of the complexes have ten or more components. A large

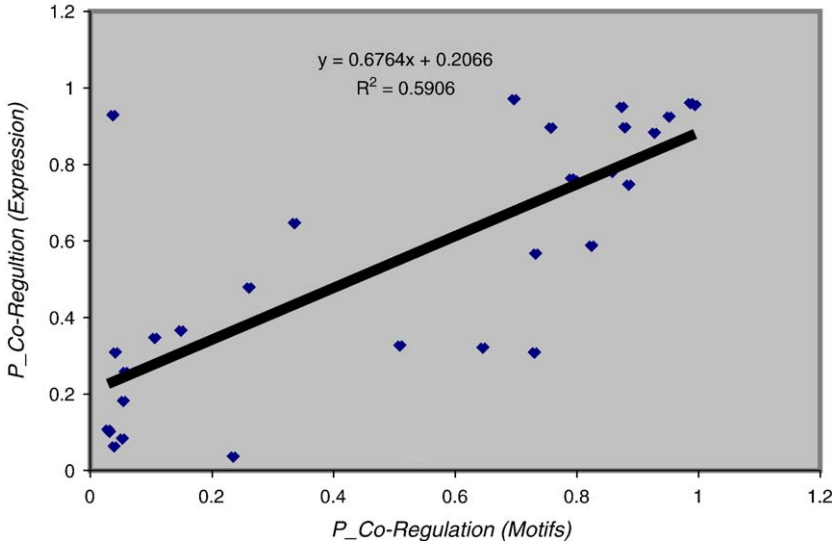


Figure 6. Correspondence between the transcriptional modules in the replication fork complexes identified on the basis of gene expression levels and inferred regulatory motifs, respectively. Individual genes (components) of the replication fork complex positioned according the posterior probabilities with which they were assigned to the complex by the discriminant analysis (*P*_Co-Regulation) performed here (Expression; ordinate) and that was performed previously on the basis of inferred *cis*-regulatory motifs by Simonis *et al.*[19] (Motifs; abscissa).

fraction of the complexes, many with more than five components, have average UPC values higher than 0.9, indicative of highly correlated expression profiles and hence a coherent transcriptional response of their components. Among the complexes in Table 1B with highly correlated expression profiles, are the Gim complexes and the RNase-P and RNase-MRP complexes. The latter two share most of their components, and several of their components are part of the MIPS entry for the rRNA processing complexes. In the latter complex, our analysis identified a transcriptional module including 13 of its 18 components (Table 1A). Interestingly, of the five components that are excluded from this module, four belong to the RNase-MRP (and RNase-P) complex (see Supplementary Data), indicating that this complex is under independent transcriptional control.

Correspondence between transcriptional modules and modules derived from predicted *cis*-regulatory motifs

For the replication fork complexes an excellent overlap was detected between the transcriptional module identified here and that identified previously on the basis of predicted *cis*-regulatory sequence motifs.[19] Table 2 summarizes the correspondence observed between the modules derived by both approaches for other complexes. As already pointed out,[19] modules derived on the basis of *cis*-regulatory motifs could be identified in only a small fraction of the MIPS complexes, 31 in all, compared to the 51 complexes in which modules were identified in this study. For six of those complexes, coherent expression was detected under too few conditions (≤3) to enable module detection on the basis of expression levels (Table S4 Supplementary Data). For the remaining 25, excellent agreement with the transcriptional modules predicted on the basis of *cis*-regulatory motifs could be observed mainly for the proteasome, the ribosomes, the nucleosomal protein complexes, cytochrome-*c* oxidase and some other complexes in which all the components were found to behave as a single transcriptional module in both studies (Table 2). With a few exceptions (such as the replication fork complexes), in the remaining complexes, the modules predicted on the basis of predicted *cis*-regulatory motifs contain fewer components and hence a smaller fraction of the components of the complex than the modules identified here. In general, these smaller *cis*-regulatory-based modules are completely contained in the larger expression-based modules (Table 2). Overall therefore, considering the difference in coverage of both types of analyses, it is quite encouraging to see that they do yield largely consistent information.

Discussion

This study presents a rigorous statistical approach for identifying experimental conditions under which components of individual multi-protein complexes display a coherent transcriptional response relative to other genes in the genome, and for identifying transcriptional modules within complexes. Such modules are defined here as groups of genes within complexes that could be discriminated from random sets of genes on the basis of their mRNA expression profiles under a common set of conditions.

This approach was applied to the full repertoire of hand-curated multi-protein complexes of the yeast *S. cerevisiae* stored in the MIPS database, considering the expression levels for genes coding for their components measured under 549 different experimental conditions (DNA chips).

To minimize errors in module identification due the inherent noise in the expression data, rigorous criteria of error risk assessment had to be applied, limiting our analysis to the subset of complexes with at least five components. These numbered 113, representing about one half (46.5%) of the 243 complexes in the MIPS catalogue.

Table 2. Correspondence between the transcriptional modules identified here and those identified on the basis of *cis*-regulatory sequence motifs[19]

MIPS complexes	No. of prot.	Prot. in module^{Exp}	Prot. in module^{Pat}	Prot. in common	Selected conditions	Sig^{max} patterns	Coverage module^{Exp}	Coverage module^{Pat}
Nucleosomal-protein-complex	8	8	8	8	135	4.4	1.00	1.00
19-22 S-regulator	18	17	17	16	101	8.72	0.94	0.94
26 S-proteasome	36	32	30	29	146	15.76	0.89	0.83
Cytochrome-*c*-oxidase	8	8	7	7	37	1.44	1.00	0.88
20 S-proteasome	15	15	13	13	87	5.73	1.00	0.87
Replication-fork-complexes	30	16	17	14	47	13.4	0.53	0.57
Nuclear-splicing-complexes-spliceosome	66	9	4	0	63	1.67	0.14	0.06
Replication-complexes	49	22	20	16	88	15.64	0.45	0.41
Cytoplasmic-translation-elongation	9	7	7	6	12	2.39	0.78	0.78
Cytoplasmic-ribosomal-small-subunit	57	56	45	44	304	8.89	0.98	0.79
RNA-polymerase-II-holoenzyme	35	7	4	0	22	1.82	0.20	0.11
Cytochrome-*bc*1-complex	9	9	6	6	51	1.16	1.00	0.67
Cytoplasmic-ribosomes	138	134	94	91	385	19.47	0.97	0.68
F0-F1-ATP-synthase	14	14	9	9	83	3.91	0.93	0.67
Cytoplasmic-ribosomal-large-subunit	81	76	57	52	334	11.32	0.94	0.70
Microtubules	32	11	5	1	5	1.48	0.34	0.16
Respiration-chain-complexes	36	33	21	19	211	8.49	0.89	0.59
Replication-complex	19	11	8	5	25	1.71	0.58	0.42
H+-transporting-ATPase-vacuolar	13	13	6	6	38	2.44	0.87	0.40
RNA-polymerase-III	13	13	6	6	104	1.51	1.00	0.46
eIF3	7	7	3	3	9	1.3	1.00	0.43
rRNA-processing-complexes	18	13	4	3	45	1.82	0.72	0.22
Spindle-pole-body	32	22	4	3	35	2.97	0.69	0.13
Cytoplasmic-translation-initiation	27	22	5	5	170	6.49	0.81	0.19
Tim22p-complex	3	3	2	1	9	1.99	1.00	0.66

Column 1 lists the name of complex entry in MIPS. Column 2 lists the number of components (proteins) in the complex, and columns 3 and 4 list the number of components in the modules assigned for this complex by the discriminant analysis based, respectively, on the expression profiles (the present study) and on the *cis*-regulatory sequence motifs.[19] Column 5 lists the number of components in common between the two modules; column 6 lists the number of experimental conditions in which the complex is either up or down-regulated relative to all other genes (as selected by the *t*-test in this study); column 7 lists the maximum significance score computed for the shared sequence patterns identified in the up-stream regions of the genes coding to complex components.[19] The two right-most columns list the coverage of the two types of modules: those identified on the basis of the expression analysis here and those identified on the basis of the regulatory sequence patterns, respectively. Coverage is computed as the fraction of the components in the complex assigned to the module.

This not withstanding, our focus on the condition-dependent coherent transcriptional response of protein complexes provides new insights into the regulation of these complexes.

We find that the 113 MIPS complexes can be subdivided into three main categories: those whose components are coherently expressed under many different conditions, complexes coherently expressed under a few conditions only, and those coherently expressed under none of the considered conditions. Complexes of the first category include the cytoplasmic ribosome, the proteasome and the respiration chain complexes. These complexes have previously been termed stable or permanent by several authors.[14,20] But our analysis of the expression levels revealed that these complexes are more often down than up-regulated under the selected conditions, indicating that they are dynamically regulated at the transcriptional level under many different conditions. Such dynamic regulation is also observed for most complexes from the second category, but the number of conditions under which it occurs is often significantly smaller.

Transcriptional modules, as defined by the discriminant analysis, were detected in the majority of the complexes amenable to this analysis (the 51 complexes with at least five components coherently expressed under at least five experimental conditions). In about half of these complexes (26 in total), the entire complex behaves as a dynamically regulated transcriptional module. In about a third of the examined complexes (15 in total) a sizable fraction (0.5–0.9) of the components make up such modules. Twenty additional complexes with five or more components were found to be coherently expressed, but under fewer than five experimental

conditions. In 12 of those (60%), which were mainly the smaller complexes (less than ten components) highly correlated expression profiles of their components were observed over the selected conditions.

Detailed analysis of the results, carried out for many of these complexes (see also Supplementary Data) revealed that the identified transcriptional modules often correspond to one or more known sub-complexes that play specialized functional roles. In addition, when these modules encompassed most of the components of a complex, the remaining few unassigned components had more frequently an unknown function, or a less well-defined function than the assigned components.

An important finding of our analysis is that in instances where groups of proteins (genes) are shared between several complexes, these groups are nearly always part of transcriptional modules identified in these complexes. This was illustrated in detail for the RNA polymerase I, II and III complexes, but was also found to occur in nearly all other cases where larger complexes contained sizable sub-complexes (see section II of Supplementary Data). It is noteworthy that in these instances the different groups of complexes, which share components, carry out related functions. In all these cases the transcriptional modules are larger than the groups of shared components, and their composition may vary according to the particular complex. Since each complex is also characterised by a set of selected conditions under which its components are coherently expressed, our findings suggest that transcriptional regulation of protein complexes that carry out related functions may be achieved *via* the so-called Multiple Input Motifs, where different transcription factors themselves expressed under different conditions, target a partially overlapping set of genes[23] (see Figure S9 in Supplementary Data). As in the case of the RNA polymerases, several of these genes can furthermore be homologs, indicating a common origin.

These observations taken together leave us pondering on the underlying biological reasons for which certain complexes might be tightly up or down-regulated in a coherent fashion, whereas others are not. A recent study[15] focused on the temporal pattern of mRNA expression for components of protein complexes active during the cell cycle, using Spellman's data. It monitored the peaking in expression levels of the corresponding genes as a function of the cell cycle phases, which showed that some of the genes were transcribed as required for the just in time assembly of the complexes whose function was needed. Specific functional requirements most certainly also dictate the level of coherence in the condition-specific transcriptional response observed here for the MIPS complexes. But identifying these requirements is not straightforward, especially if one considers that with the exception of Spellman's cell cycle experiment, the remaining expression data used here were measured in non-synchronized, inhomogeneous cell populations. We must therefore conclude that the coherent response detected here for complexes such as the ribosome, proteasome, nucleosomal protein complexes, and the like, reflects the common regulatory response of the mixed cell population in the considered experimental conditions. Indeed, given that the majority of the experimental conditions in our dataset are stressful to the cell, a strong common transcriptional response in house-keeping protein complexes might be expected. Based on the same reasoning, other complexes that carry out functions more closely linked to the cell cycle may not behave as transcriptional modules in our analysis, because any coherent transcriptional response exhibited by these complexes in subsets of the cell population would be averaged out. To avoid such averaging out effects the experiments need to be redesigned to measure the expression levels in cell populations that are as uniform as possible.

Another intriguing possibility might be that the differences in the transcriptional response between complexes and between components thereof are related to the life-times of these components, or possibly to those of the corresponding mRNA molecules, with complexes composed of short-lived proteins (or mRNA) requiring tighter up and down-transcriptional co-regulation. A preliminary analysis of the published decay times of mRNAs[24] corresponding to components of the MIPS complexes analysed here, indicates indeed that complexes found here to behave as tightly co-regulated transcriptional modules (the ribosome, proteasome, RNAP- I and –III, etc.), tend to display lower averages and dispersion in their mRNA half-life times, than complexes with more loosely co-regulated components. Further work is however needed to confirm these conclusions.

Lastly, it should be mentioned that the complex-centric view taken in this study makes it difficult to identify transcriptional modules whose components map into different complexes in a time or condition-dependent fashion. We tackled this issue in part by examining how shared proteins mapped into the transcriptional modules that we identified in individual complexes or sub-complexes. But addressing it systematically would require identifying transcriptional modules under different sets of experimental conditions independently of complexes, and then mapping these modules back into the complexes. Procedures for performing such identification have been proposed[25] and their application to the analysis of protein complexes is currently in progress in our laboratory.

Materials and Methods

Data on multi-protein complexes

Data manually annotated from the literature on 243 protein complexes in the yeast *S. cerevisiae* are retrieved

from the catalogue of complexes in CYGD-MIPS.[4] These data include complexes known to form a single physical entity under certain experimental conditions, as well as larger assemblies composed of several complexes whose formation is thought to be interdependent. The complete list of analysed complexes can be obtained from our website§.

Microarray data and their treatment

Gene expression levels for the yeast *S. cerevisiae* are obtained from publicly available microarray data measured under a total of 549 different experimental conditions. They include data measured for 77 conditions during the yeast cell cycle,[7] for 173 different stress, drug and carbon source conditions[6] and for 279 mutants and 20 drug interaction experiments.[16]

The data in the original Tables, which represent the logarithms of the expression ratio of individual genes under specific experimental conditions, are standardized to *Z*-scores. In order to avoid bias from outliers, the population mean is computed as the median of the sample, and the standard deviation as an expression depending on the inter-quartile range (*IQR*) as follows:

$$z_{ij} = \frac{x_{ij} - \hat{\mathrm{m}}_j}{\hat{\sigma}_j} = \frac{x_{ij} - \tilde{\mathrm{m}}_j}{\frac{IQR_j}{IQR_{\mathrm{norm}}}} \qquad (1)$$

where x_{ij} is the logarithm of the expression ratio (original data) of gene i in condition (column) j, $\hat{m}_j$ and $\hat{\sigma}_j$ are the estimates of the mean and standard deviation, respectively, $\hat{m}_j$ is the median of the column j, IQR_j the inter-quartile range of column j and $IQR_{\mathrm{norm}} = 1.34898$ is the inter-quartile range of the standard normal distribution (with a mean of 0 and a standard deviation of 1).

Selecting experimental conditions under which complexes show a significant transcriptional response

Welsh's *t*-test is used to select the experimental conditions for which the mean *Z*-score of the genes coding for components of a given complex is significantly different from the mean *Z*-score of all other yeast genes. Selection on the basis of a *t*-test, which applies criteria of statistical significance is less prone to influence from noise in the experimental data than simply relying on the raw *Z*-score values.

The *t*-test was carried out on all complexes , yielding a *P*-value per complex and experimental condition. This value was corrected for multi-testing by computing an *E*-value = 549 *P*-value, following Bonferoni's rule, with 549 being the total number of conditions for which the *t*-test was performed. An *E*-value threshold (≤0.05) was then used to select for each protein complex the experimental conditions in which components of the considered complex are up or down-regulated relative to all other yeast genes.

To validate this procedure we applied it to three datasets, the MIPS complexes, normally distributed random values, and expression levels of random selection of gene sets of the same size as the corresponding complexes. Analysis of the results revealed that the rate of false positives was as expected for the background model (<5%) only for complexes containing at least five components, coherently expressed under at least two experimental conditions selected by the *t*-test (see Supplementary Data for details). Our subsequent analysis was therefore restricted to those complexes (numbering 71 in total).

Clustering complexes and groups of experimental conditions

The analysis described in this section was performed solely for the purpose of generating a pictorial illustration of the grouping of complexes and conditions selected by the *t*-test (Figure 2). To this end the considered experimental conditions were assembled into groups. The experimental conditions analysed by Gasch[6] were grouped as described by the author into a total of 21 groups with between 3–22 conditions in each group. For Spellman's data[7] conditions were grouped according to the different phases of the cell cycle, yielding a total of six groups with 4–17 conditions per group. Thus, overall, 248 experimental conditions were grouped into 27 groups for this analysis. DNA chips from mutant strains[8] were not considered here (but used elsewhere throughout this work), because they could not be readily grouped . Details of the grouping are provided in Table S2 of the Supplementary Data.

The analysis was performed on a total of 57 complexes containing five or more components, identified as displaying coherent expression in two or more of the 248 considered conditions.

Complete linkage hierarchical clustering analysis was applied in two dimensions, to identify: (1) groups of protein complexes that display similar behaviour with regards to the expression pattern of their components under the considered condition groups, and (2) groups of conditions that affect similarly the expression of the considered complexes. The clustering was performed with the R-package‖ using as metric the Euclidian distance between the vectors representing the fraction of the conditions in each condition group that passed the *t*-test for a given complex.

Discriminant analysis

Linear discriminant analysis[26] was used to classify the genes involved in a given protein complex according to their standardized expression ratios (*Z*-scores) measured in the experimental conditions selected by the *t*-test. The approach is analogous to that described previously for classifying proteins (genes) in complexes on the basis of shared regulatory motifs,[13] and briefly works as follows. Two gene groups are defined. Group 1 comprises the g genes coding for the components of a given protein complex, in which p conditions have been selected by the *t*-test. Group 2 is a control group of $3g$ genes, selected at random from the yeast genome. A linear discriminant function is built, which optimally separates genes from groups 1 and 2 into their respective groups in the p-dimensional space of the expression ratios, taken as variables. To avoid over fitting, the most discriminant variables are identified in a stepwise fashion.

In order to assign individual genes from a given complex to either group, a leave-one-out procedure is performed, whereby the genes are removed from the complex one at a time, a discriminant function is built each

§ http://ftp.scmbb.ulb.ac.be/pub/nicolas/html_upc_daexpr_05se/mips_synthetic_table.html

‖ http://r-project.org/

time with a different gene removed and used to assign the removed gene to groups one or two.

To account for fluctuation in the results, the entire procedure was repeated 100 times for each complex, using different random selections of genes for group 2, and the probability that a given gene is part of the complex was computed as the mean of the posterior probabilities evaluated in all the trials. Genes with mean posterior probability >0.5 are defined as belonging to the transcriptional module. The posterior probability can be seen as a measure of the reliability of the gene assignment, with higher probabilities representing more reliable assignments. Individual mean probabilities for each gene are provided on the supporting website¶.

Lastly, since the discriminant analysis is performed using the experimental conditions as the parameter space for the classification, this analysis was carried out on all complexes with at least five components for which reliable *t*-test results were obtained for at least five experimental conditions (totalling a subset of 51 complexes out of the 71 with five or more components).

The results of the discriminant analysis were evaluated by two measures, the Coverage=$TP/(TP+FN)$, and the Positive Predictive Value $(PPV)=TP/(TP+FP)$. *TP* is the number of True Positives (genes assigned to a given complex that were originally part of it), *FN* is the number of False Negatives (genes that are part of the original complex but were assigned to the control group) and *FP* is the number of False Positives (genes belonging to the control group that were assigned to the complex).

Correlations between gene expression profiles

The UPC was computed to measure the correlation between the expression profiles of two genes under a set of conditions, as follows:

$$UPC = \sum_{i=1}^{p} \left(\frac{a_i}{\sqrt{\frac{1}{p}\sum_{i=1}^{p} a_i^2}} \right) \left(\frac{b_i}{\sqrt{\frac{1}{p}\sum_{i=1}^{p} b_i^2}} \right) \quad (2)$$

where a_i and b_i are the standardized log expression ratios, for genes *a* and *b* measured in condition *i*, and *p* is the number of conditions in the considered condition group.

The use of the uncentered version was motivated by the wish to keep track of the relative level of expression of each gene above or below the reference expression level considered for each condition so as to enable identification of coherent up and down-regulation.

This and all other analyses were performed with the R package.

Acknowledgements

We thank Ron Kafri (Weizmann Institute, Israel) and Gerald Quon (University of Toronto, Canada) for careful reading of the manuscript and their useful suggestions. We are grateful to Shuye Pu and Jim Vlasblom for the GenePro figures. Support for this work is gratefully acknowledged from the "Action de Recherches Concertées de la Communauté Française de Belgique" and the Hospital for Sick Children, Toronto Canada. S.W. is recipient of the CIHR Establishment grant and D.G. is "chargé de recherches" at the "Fonds National de la Recherche Scientifique", Belgium.

¶http://ftp.scmbb.ulb.ac.be/pub/nicolas/html_upc_daexpr_05se/mips_synthetic_table.html

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2006.06.024

References

1. Alberts, B. (1998). The cell as a collection of protein machines: preparing the next generation of molecular biologists. *Cell*, **92**, 291–294.
2. Gavin, A. C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A. *et al.* (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141–147.
3. Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D., Moore, L., Adams, S. L. *et al.* (2002). Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*, **415**, 180–183.
4. Mewes, H. W., Frishman, D., Guldener, U., Mannhaupt, G., Mayer, K. *et al.* (2002). MIPS: a database for genomes and protein sequences. *Nucl. Acids Res.* **30**, 31–34.
5. Cherry, J. M., Adler, C., Ball, C., Chervitz, S. A., Dwight, S. S., Hester, E. T. *et al.* (1998). SGD *Saccharomyces* Genome Databases. *Nucl. Acids Res.* **26**, 73–79.
6. Gasch, A. P., Spellman, P. T., Kao, C. M., Carmel-Harel, O., Eisen, M. B. *et al.* (2000). Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell*, **11**, 4241–4257.
7. Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B. *et al.* (1998). Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.
8. Hughes, J. D., Estep, P. W., Tavazoie, S. & Church, G. M. (2000). Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J. Mol. Biol.* **296**, 1205–1214.
9. Wingender, E., Chex, X., Hehl, R., Karas, H., Liebich, I., Matys, V. *et al.* (2000). TRANSFAC: an integrated system for gene expression regulation. *Nucl. Acids Res.* **28**, 316–319.
10. Zhu, J. & Zhang, M. Q. (1999). SCPD: a promoter database of the yeast *Saccharomyces cerevisiae*. *Bioinformatics*, **15**, 607–611.
11. Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K. *et al.* (2002). Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, **298**, 799–804.
12. Harbison, C. T., Gordon, D. B., Lee, T. I., Rinaldi, N. J., Macisaac, K. D. *et al.* (2004). Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99–104.
13. Simonis, N., Wodak, S. J., Cohen, G. N. & van Helden, J. (2004). Combining pattern discovery and discriminant analysis to predict gene co-regulation. *Bioinformatics*, **20**, 2370–2379.

14. Jansen, R., Greenbaum, D. & Gerstein, M. (2002). Relating whole-genome expression data with protein-protein interactions. *Genome Res.* **12**, 37–46.
15. de Lichtenberg, U., Jensen, L. J., Brunak, S. & Bork, P. (2005). Dynamic complex formation during the yeast cell cycle. *Science*, **307**, 724–727.
16. Hughes, T. R., Marton, M. J., Jones, A. R., Roberts, C. J., Stoughton, R., Armour, C. D. *et al.* (2000). Functional discovery via a compendium of expression profiles. *Cell*, **102**, 109–126.
17. Fruchterman, T. M. J. & Reingold, E. M. (1991). Graph drawing by force-directed placement. *Softw.-Pract. Exp.* **21**, 1129–1164.
18. Krogan, N. J., Peng, W. T., Cagney, G., Robinson, M. D., Haw, R., Zhong, G. *et al.* (2004). High-definition macromolecular composition of yeast RNA-processing complexes. *Mol. Cell.* **13**, 225–239.
19. Simonis, N., van Helden, J., Cohen, G. N. & Wodak, S. J. (2004). Transcriptional regulation of protein complexes in yeast. *Genome Biol.* **5**, R33.
20. Teichmann, S. A. & Babu, M. M. (2002). Conservation of gene co-regulation in prokaryotes and eukaryotes. *Trends Biotechnol.* **20**, 407–410; discussion 410.
21. Han, J. D., Bertin, N., Hao, T., Goldberg, D. S., Berriz, G. F., Zhang, L. V. *et al.* (2004). Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, **430**, 88–93.
22. Duden, R. (2003). ER-to-Golgi transport: COP I and COP II function (Review). *Mol. Membr. Biol.* **20**, 197–207.
23. Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D. & Alon, U. (2002). Network motifs: simple building blocks of complex networks. *Science*, **298**, 824–827.
24. Wang, Y., Liu, C. L., Storey, J. D., Tibshirani, R. J., Herschlag, D. & Brown, P. O. (2002). Precision and functional specificity in mRNA decay. *Proc. Natl Acad. Sci. USA*, **99**, 5860–5865.
25. Ihmels, J., Levy, R. & Barkai, N. (2004). Principles of transcriptional control in the metabolic network of *Saccharomyces cerevisiae*. *Nature Biotechnol.* **22**, 86–92.
26. Huberty, C. J. (1994). *Applied Discriminant Analysis.* John Wiley and Sons, New York.
27. Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D. *et al.* (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504.
28. Vlasblom, J., Wu, S., Pu, S., Superina, M., Liu, G., Orsi, C. *et al.* (in press). Gene Pro: a Cystoscape plug-in for advanced visualization and analysis of interaction network. *Bioinformatics*.

Edited by M. Sternberg

(Received 17 November 2005; received in revised form 14 May 2006; accepted 12 June 2006)
Available online 3 July 2006