*4768–4777* *Nucleic Acids Research, 2010, Vol. 38, No. 14* *Published online 12 April 2010*
*doi:10.1093/nar/gkq232*

# Comprehensive reanalysis of transcription factor knockout expression data in *Saccharomyces cerevisiae* reveals many new targets

**Jüri Reimand[1,2,*], Juan M. Vaquerizas[1], Annabel E. Todd[1], Jaak Vilo[2] and Nicholas M. Luscombe[1,3,*]**

[1]EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge CB10 1SD, UK, [2]Institute of Computer Science, University of Tartu, Liivi 2, Tartu, Estonia and [3]Genome Biology Unit, EMBL Heidelberg, Meyerhofstrasse 1, Heidelberg D-69117, Germany

Received September 28, 2009; Revised March 11, 2010; Accepted March 21, 2010

## ABSTRACT

**Transcription factor (TF) perturbation experiments give valuable insights into gene regulation. Genome-scale evidence from microarray measurements may be used to identify regulatory interactions between TFs and targets. Recently, Hu and colleagues published a comprehensive study covering 269 TF knockout mutants for the yeast *Saccharomyces cerevisiae*. However, the information that can be extracted from this valuable dataset is limited by the method employed to process the microarray data. Here, we present a reanalysis of the original data using improved statistical techniques freely available from the BioConductor project. We identify over 100 000 differentially expressed genes—nine times the total reported by Hu *et al*. We validate the biological significance of these genes by assessing their functions, the occurrence of upstream TF-binding sites, and the prevalence of protein–protein interactions. The reanalysed dataset outperforms the original across all measures, indicating that we have uncovered a vastly expanded list of relevant targets. In summary, this work presents a high-quality reanalysis that maximizes the information contained in the Hu *et al*. compendium. The dataset is available from ArrayExpress (accession: E-MTAB-109) and it will be invaluable to any scientist interested in the yeast transcriptional regulatory system.**

## INTRODUCTION

High-throughput assays such as microarrays allow users to ask detailed questions about biological relationships between genes. A major use of microarrays has been to measure expression changes in response to perturbations such as deletion and over-expression of genes of interest (1–3). As changes are likely to be triggered by the perturbed gene, such experiments may help reveal its cellular function (4); for example knockouts and partial deletions have been used to identify genes that are essential to survival (5).

Over the past few years, considerable effort has been invested into deciphering the transcriptional regulatory network of the yeast *Saccharomyces cerevisiae*, and several large-scale perturbation datasets of transcriptional regulators are now available. An early review by Svetlov *et al*. (6) compiled over 900 individual biochemical and genetic interactions between 83 transcription factors (TFs) and 494 genes. Using microarrays, Hughes *et al*. (7) published a compendium of 300 experiments including 35 TF knockouts. A newer dataset by Chua *et al*. (8) covered 55 TF mutants.

Most recently Hu *et al*. (9) presented a compendium of 269 TF knockout microarrays. Covering almost all yeast regulators, this is currently the most comprehensive perturbation dataset of TFs for any organism, and it is therefore of great interest in the genome-scale investigation of

*To whom correspondence should be addressed. Tel: +372 737 6137; Fax: +372 737 5468; Email: juri.reimand@ut.ee
Correspondence may also be addressed to Nicholas M. Luscombe. Tel: +44 1223 492 572; Fax: +44 1223 494 468; Email: luscombe@ebi.ac.uk

eukaryotic gene regulation. Having performed these experiments, there is a major challenge to process large and frequently noisy datasets in order to identify differentially expressed genes with confidence. Unfortunately, the authors of the Hu *et al*. study used relatively dated and insensitive approaches for microarray data-processing: as a result the published *P*-values and target-gene ranking are likely to be unreliable. Specific examples include the lack of background and print-tip correction during normalization (10–14), and use of an error model that does not account for systematic experimental biases (7,15–18). Moreover, *P*-values were not corrected for multiple-testing, which is crucial for minimizing false positives (15,16,19–21). In short, the lack of robust data-processing procedures have limited the amount of information that could otherwise be extracted from this substantial body of valuable experimental work, and it has greatly restricted the use of these data in follow-up investigations.

A strong consensus for the best methods for microarray data processing has emerged over the past 5 years. In addition, sensitive analysis techniques have been developed that deliver improved data correction and statistical tests for differential expression (12,18,22). Here, we present a reanalysis of the original raw data published by Hu *et al*. using updated statistical methods that are freely available through the BioConductor software suite (23). We identified 110 487 differentially expressed genes—nearly nine times the total reported by Hu *et al*. The reanalysis recovers 90% of the original dataset, suggesting that we have identified a vastly expanded list of target genes. To validate the biological significance of the dataset, we assessed the enrichment of Gene Ontology (GO) (24), KEGG functional annotations (25) and Reactome pathways (26) for target genes, the occurrence of upstream TF-binding sites, and the prevalence of protein–protein interactions among TFs and target genes. In summary, this work presents a high-quality reanalysis that maximizes the information contained in the Hu *et al*. compendium, and the dataset will be invaluable to any scientist interested in the yeast transcriptional regulatory system. Further, the reanalysis constitutes a prime example of the effect of using up-to-date analysis techniques in maximizing the information obtained from high-throughput generated data.

## METHODS

### Microarray data pre-processing and analysis

Raw microarray data were downloaded from the Longhorn Microarray Database (27). Microarrays were normalized using the VSN package, including print-tip and background correction (12). Array probes that were not annotated as Open Reading Frames (ORFs) in the original dataset were discarded, and duplicate and triplicate array probes were averaged. Differential expression was calculated using a moderated eBayes *t*-test as implemented in the Limma Bioconductor package (18). The resulting *P*-values were FDR-adjusted across the whole microarray dataset to correct for multiple testing (28). An adjusted *P*-value cut-off of 0.05 was used to detect significant differential gene expression. See Supplementary Data for further details.

### Functional enrichment analysis

Functional enrichment was performed using g:Profiler (29) based on Ensembl annotations [(30); release 49]. Gene annotations from low-confidence electronic evidence were removed. For each deletion assay, we computed a log-score by aggregating the log of the *P*-values for each category of GO, Reactome and KEGG with a significant enrichment (Supplementary Data). A global score for our reanalysed dataset and the original analysis was computed as the sum of log-scores for each individual TF.

### TF binding data and analysis

High-confidence DNA–protein interactions derived from ChIP-chip experiments were obtained from (31). Data were filtered further to include only direct interactions as defined in ref. (32). Only matches classified as bound, only in YPD and with a $P < 0.001$ were considered.

Predicted TF binding sites were obtained from refs. (33,34). Erb and van Nimwegen derived a set of 'trusted' position weight matrices (PWMs) for 72 regulatory factors by running the PROCSE and PhyloGibbs algorithms on a set of experimentally derived TF binding sites from SCPD (35) and (31). These PWMs were then used to scan multiple alignments of each intergenic region in *S. cerevisiae* with the orthologous regions of another four *Saccharomyces* species. Predicted binding sites with a posterior probability $> 0.5$ were used in our analysis. MacIsaac *et al*. (34) applied a combination of the conservation-based PhyloCon and Converge algorithms to ChIP-chip data (31), to predict binding sites for 172 TFs (34). Only predictions conserved in more than three species with a $P < 0.001$ were considered for our analysis.

Binding sites from these data sources were then mapped into gene promoters. If the centre of a binding site was located between −1000 and +100 bp from the transcription start site of a given gene, it was said to be located in the gene's promoter region. Similar results were obtained for shorter upstream promoter regions (−600 to +100 bp; data not shown). Our dataset of mapped binding sites covered 142 knockout TFs. Enrichments of binding sites in gene promoters for both direct and indirect interactions were calculated using a cumulative hypergeometric test (Supplementary Data).

### Protein–protein interaction analysis

Protein–protein interactions were obtained from refs (36,37). The enrichment of differentially expressed genes for interacting TFs was assessed using a cumulative hypergeometric test. To test enrichments of TFs targeting protein complexes, we constructed protein–protein interaction modules for each TF target. We then compared the number of protein–protein interactions among TF targets, and between TF targets and non-targets using a cumulative hypergeometric test. In both cases, *P*-values were corrected for multiple testing using FDR (Supplementary Data).

## RESULTS

### The reanalysis detects nine times more differentially expressed genes

The compendium consisted of 588 two-colour cDNA microarray hybridizations for 269 mutants (9). The original analysis used a modification of the error model developed by Hughes *et al*. (7). Briefly, systematic errors in gene expression measurements were estimated from 10 control experiments in which the co-hybridized samples were taken from identical RNA preparations. The resulting log-ratio values provided a model for the error distribution. For each mutant experiment, genes displaying changes in expression values beyond the error distribution were identified. The level of expression change was quantified using a statistic $X$, calculated from the mean log ratios of replicate samples using the minimum-variance weighted average method (i.e., genes with larger variance than observed in the control data were assigned proportionately larger standard errors). The significance of the ratios was computed from the $X$-scores, and a threshold of $P < 0.001$ was applied to define the set of differentially expressed genes. This resulted in a dataset of 12 284 differentially expressed genes across 266 mutant strains (the original analysis detected no differentially expressed target genes for ARG82, YDR026C and YJL206C).

To reanalyse the dataset, we downloaded the unprocessed numerical text files from the Longhorn Database (27). For each array, we applied background correction and print-tip normalisation using the VSN package available from the BioConductor software project (12,23). We then extracted expression values for 6253 protein-coding genes presented on the arrays. Most genes were represented by single probes; for the 360 genes represented by multiple probes we averaged expression measurements across all replicates. Probes corresponding to non-protein-coding regions of the genome were excluded from further analysis.

We identified differentially expressed genes using the Limma eBayes package (18) distributed through the BioConductor project. The experimental design generally consisted of duplicate hybridizations for mutant and wild-type strains against a reference RNA sample and growth level control arrays (Supplementary Data). Therefore to obtain differential expression measurements between the mutant and wild-type, we integrated data from the two sets of comparisons (i.e. mutant versus reference RNA; and reference RNA versus wild-type). To correct for biases such as batch effects, we incorporated the control arrays into the error model. Variations to this experimental design were handled appropriately on a case-by-case basis (Supplementary Data). Finally, we applied a compendium-wide FDR correction to adjust for multiple testing (28), and used a $P$-value threshold of 0.05 to define differentially expressed genes.

Our reanalysis returned a list of 110 487 differentially expressed genes across 269 mutants—almost nine times as many targets—and we recovered 90% of genes presented by the original analysis (Figure 1A; Supplementary Table S1). The difference in number of target genes between our analysis and the original one is due to the increased sensitivity of our approach and not to the usage of a different threshold, as most additional targets in our reanalysis are not obtained by simply adjusting the $P$-value cut-off of the original analysis (Supplementary Figure S1). Moreover, there is a strong correlation ($r = 0.83$; Pearson correlation) between the two datasets for numbers of genes affected by each TF knockout (Figure 1A). As previously observed for regulatory interactions (38), there is a non-uniform distribution of target gene numbers: there is a large number of TFs with relatively few targets, and some very influential TFs that affect more than a third of the yeast genome (Figure 1B; Supplementary Figure S2). Among the top 20 mutants with the greatest impact are chromatin-based regulators including the SWI/SNF remodelling complex (SNF2, 5, 6 and SWI3), individual histone modifiers (SPT10, SIN3, SIR2) as well as global DNA-binding TFs (HAP2, RAP1 and MCM1) (Supplementary Figure S2). Interestingly, there is little correlation between wild-type expression level of a TF (calculated as wild-type $A$-value) and the number of genes it affects ($r = 0.13$; $P = 0.027$; Pearson correlation), indicating that factors with low expression can play a significant gene regulatory role. This is additionally supported by the observation that TFs with no significant depletion in the corresponding knockout mutant often affect a large number of target genes (Figure 1B; Supplementary Figure S2). The 10% of target genes reported in the initial study but not detected by our approach are likely to include both genes with marginal $P$-values in our study, as well as artefacts from the initial study resulting from the different normalization procedures.

Together, these observations suggest that our reanalysis identifies a vastly expanded repertoire of potential regulatory targets compared with the original dataset. In order to assess the biological significance of our results, we examined the gene lists by integrating several different sources of evidence.

### Deleted TFs are down-regulated in mutant strains

First we checked the expression levels of the TFs themselves. Intuitively we expect the TF under consideration to have lower expression in the mutant strain compared with the wild type strain. Our analysis confirms this for 155 TFs compared with just 88 for the Hu *et al*. analysis (Figure 2).

Of the remaining 114 TFs, 78 display a negative fold change in the mutant strain albeit at statistically non-significant levels. They tend to be expressed at much lower levels than other TFs ($P < 10^{-18}$; Wilcoxon test), indicating that low-level expression changes are harder to detect with current microarray technology. Among these regulators are several that affect many genes, emphasising that even small adjustments to TF expression levels can have a dramatic effect on target genes.

For the 36 TFs for which we do not observe a negative fold change, we suggest that changes in their expression levels are too subtle to detect given the experimental noise. Surprisingly, both analyses show that the cell cycle regulator MCM1 has significantly elevated expression in the

mutant strain. As deletion of this gene is lethal, it was placed under the control of an inducible promoter; therefore we suspect that there may have been a fault in the construct rather than in the microarray experiment.

### Reanalysed dataset displays greater functional enrichment

Next we examined the functional annotations of the differentially expressed genes. As most TFs are considered to regulate distinct cellular processes, their target genes should be associated with a coherent set of molecular and biological functions.

For each TF knockout, we used the g:Profiler web-tool (29) to identify the GO, KEGG and Reactome pathway annotations that are over-represented in the target gene list (Supplementary Table S2). Figure 3 illustrates the top 50 enriched GO functional categories among target genes. We calculated a score measuring the enrichment of functional annotations by summing the absolute logarithms of all *P*-values below 0.05 (Supplementary Data). Across all TF knockouts, our reanalysis has a higher score than the original analysis (log-score = 36 230 compared with log-score = 18 519). Comparing individual TFs, the gene lists from our reanalysis score equal or higher in 214 out of 269 cases. Note that the greater functional enrichment in our dataset is not due to increased numbers of target genes, as we apply a hypergeometric test to compare the functional annotation in the test set against a randomly picked sample of similar size. Thus the results indicate that our additional targets are biologically meaningful and not noise.

Our dataset recovers 95% of enriched functional categories for the Hu *et al.* analysis, and in fact improves the significance of the enrichment (in terms of number of genes with a particular annotation) in 85% of these cases. Moreover, functional categories that we observe are generally in good agreement with previous

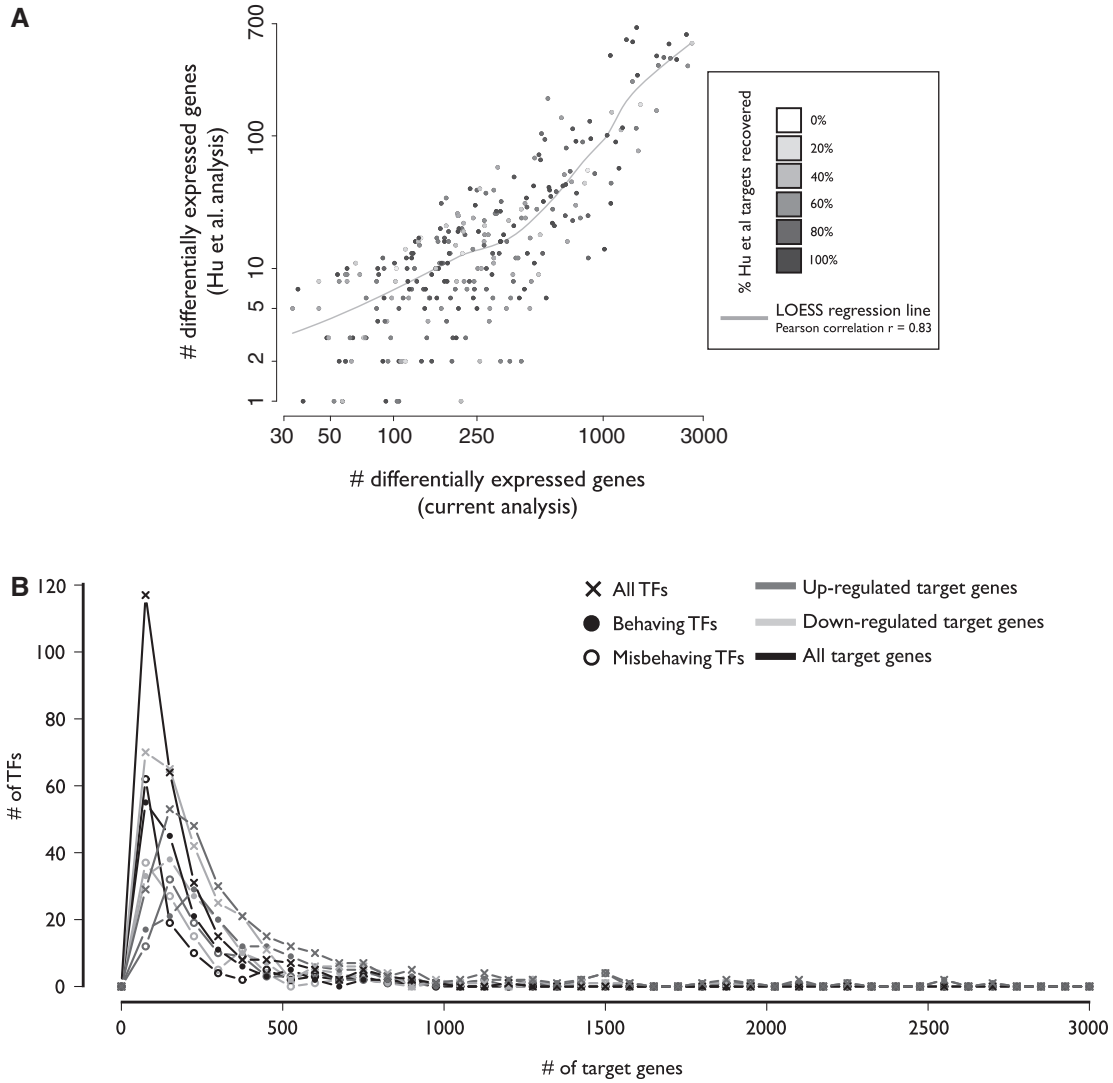


**Figure 1.** Reprocessed dataset displays good agreement with original analysis. (**A**) Scatter plot displaying the correlation in the numbers of target genes between the reanalysed and original datasets. The data points, one for each knockout experiment, are colour-coded according to the proportion of the original gene list that is recovered. (**B**) Non-uniform distribution of TF targets for up-regulated (dark grey), down-regulated (light grey) and all differentially expressed genes (black). Data for all TFs are labelled with crosses. Filled circles denote regulators that are down-regulated in their deletion experiment (behaving TFs), while empty circles denote regulators with no significant down-regulation in their deletion.

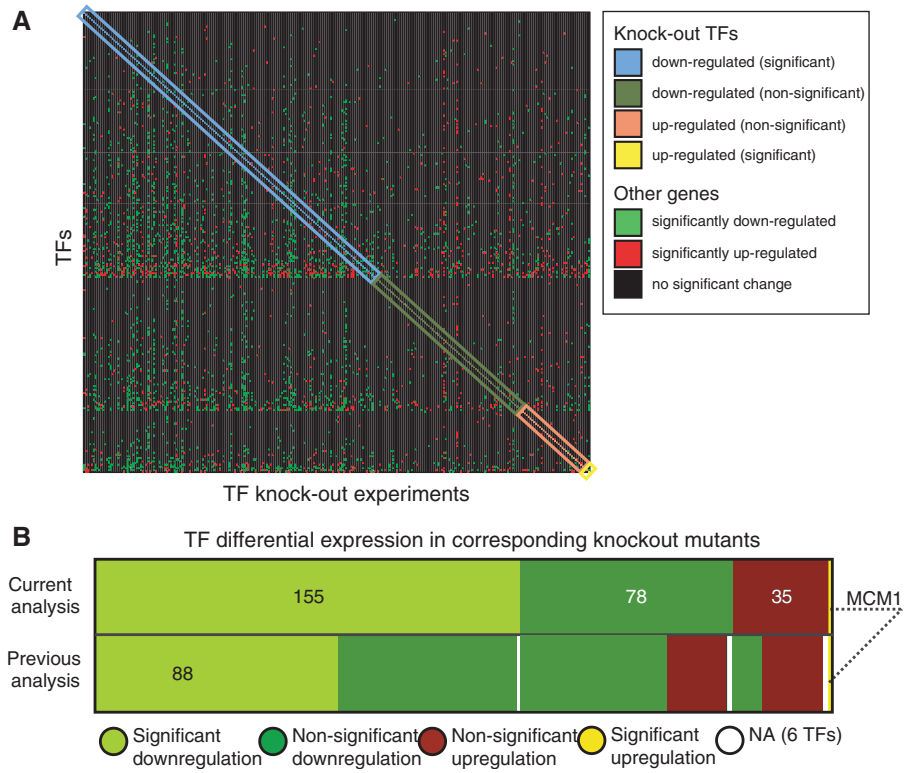


**Figure 2.** Most TFs are significantly depleted in the corresponding knockout mutants. (**A**) Heat-map of TF expression levels in 269 knockout mutants. Intersecting cells are coloured according to changes in expression values relative to wild-type: significantly up-regulated (red), significantly down-regulated (green) and no change (black). Cells on the diagonal indicate the expression levels of the TFs in their own knockout mutants. Blue cells on the diagonal represent 155 TFs that display lower expression in their own knockout. In the original analysis there were only 88 such TFs. Dark green cells on the diagonal show 78 TFs with a negative, although non-significant, fold-change in the deletion assay. Orange cells on the diagonal represent 35 TFs with positive fold-change (non-statistically significant). Finally, a single yellow cell on the diagonal represents MCM1, which is significantly up-regulated. (**B**) Comparison of TF differential expression in the current (top row) and original analysis (bottom row). Cells are coloured according to the differential expression of each TF in its own knockout mutant. Light green cells represent TFs whose expression is statistically significantly depleted (155/88 TFs in current/original analysis). Dark green cells show TFs with a negative, although non-significant, fold-change in the deletion assay (78/131 TFs in current/original analysis). Dark red cells represent TFs with positive non-significant fold-changes (35/43 TFs in current/original analysis). White cells represent six TFs that had undetermined (NA) fold change values in the original analysis. Finally, a single yellow cell represents MCM1, which is significantly up-regulated in both analyses.

knowledge. For example, ISW1 is a component of several chromatin remodelling complexes; its deletion causes up-regulation of 188 genes involved in eukaryotic translation initiation ($P < 10^{-6}$) and components of the ribosome ($P < 10^{-11}$) (39,40). Additionally, we identify 247 down-regulated genes related to several metabolic processes including glycolysis ($P < 10^{-4}$), alcohol biosynthesis ($P < 10^{-3}$) and fungal-type cell wall ($P < 10^{-6}$), functions that were previously not obviously associated with this TF. In contrast, the original analysis reported just 13 differentially expressed genes for ISW1, which do not show any functional enrichment.

In another example, FKH2 is best known for its role as a cell cycle TF, which regulates genes during the G2/M phase; it is also involved in transcriptional elongation and chromatin silencing (41–43). The knockout causes down-regulation of 343 genes with strong representation in functions relating to ribosomal biogenesis and assembly ($P < 10^{-32}$), and up-regulation of 64 related genes that encode membrane-associated proteins ($P < 10^{-5}$). The previous analysis describes just two differentially expressed genes. Here, it is possible that the mutant does not display as great an effect as expected owing to the back-up provided by the homologue FKH1; thus these results highlight the importance of detecting small expression changes using sensitive methods.

## Reanalysed dataset shows better overlap with TF-binding site data

In the original analysis, Hu *et al.* reported a surprisingly low overlap between their list of differentially expressed genes and publicly available ChIP-chip datasets. They suggested several explanations, ranging from data quality issues to the fact that perturbation experiments can reveal secondary regulatory interactions that are absent from ChIP-chip experiments.

To re-examine the overlap between the knockout and TF-binding datasets, we considered information from large-scale ChIP-chip (31) and motif-finding studies (33,34) (see 'Methods' section). By incorporating the results published in a recent study by Zhu *et al.*, we filtered the ChIP-chip data to focus on high-confidence direct DNA–protein interactions (31,32). We mapped all observed and potential TF-binding sites reported in the

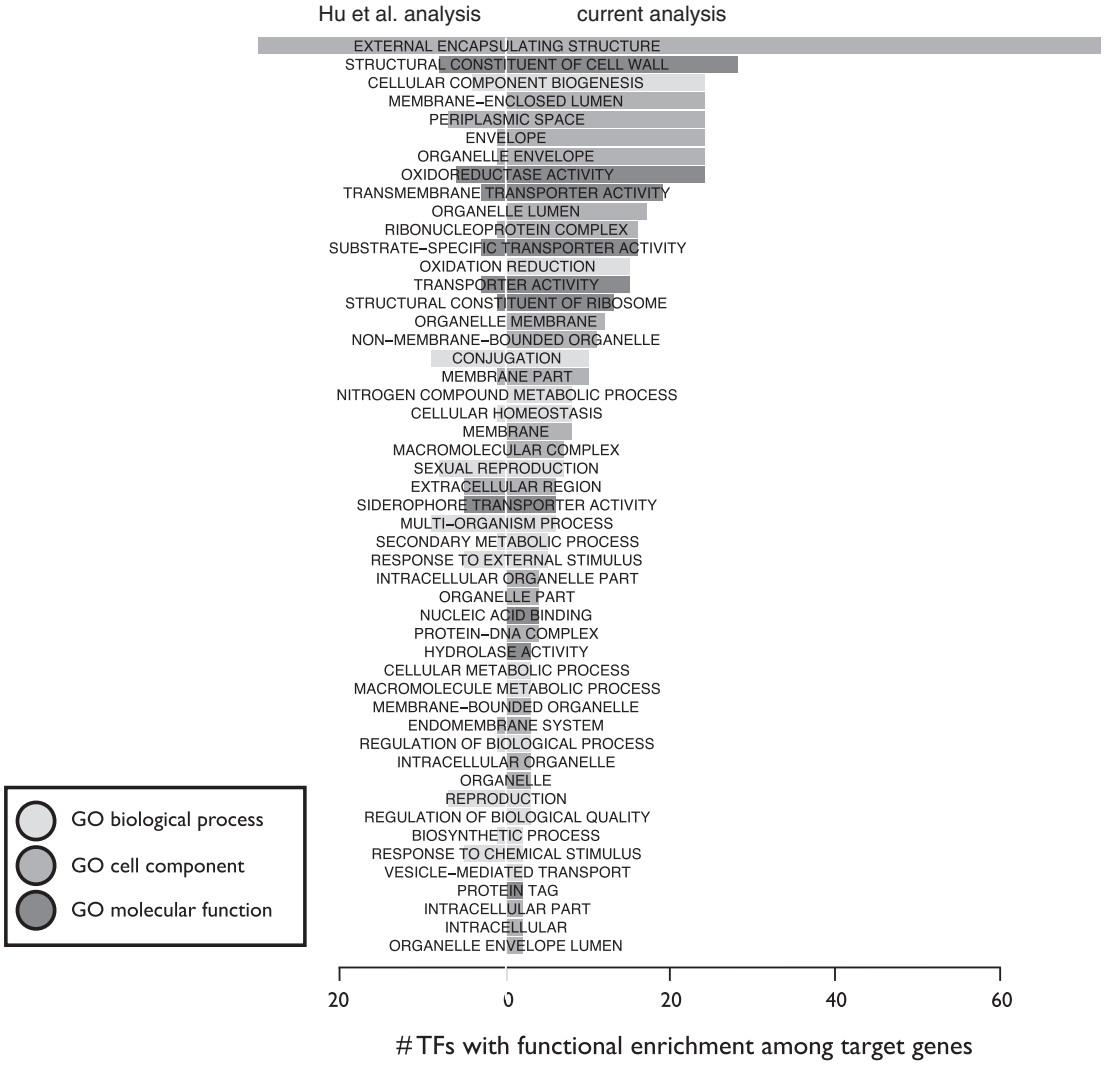


**Figure 3.** Top 50 enriched functional categories among target genes. List of the top 50 enriched Gene Ontology functional categories among target genes as determined by g:Profiler (see 'Methods' section). Categories are ranked based on the number of TF knockouts in which they are found significantly over-represented. The current analysis (right-side) identifies many more functional enrichments than the original one (left-side).

motif-finding studies to a recent version of the yeast genome (*Saccharomyces Genome Database*, 1 December 2008) and considered only those that are located within promoter regions (defined as: from −1000 to +100 bp of the transcription start site). In total, we collected binding sites for 142 TFs included in our analysis, comprising 5188 ChIP-chip interactions and 17 091 motif predictions.

We calculated the intersection between our list of differentially expressed genes from the TF deletion mutants and targets identified by ChIP-chip or binding-site predictions. Our analysis identified 645 of the ChIP-chip targets (compared with 168 targets identified by the original analysis). The overlap between the ChIP-chip and the deletion datasets is statistically significant in both cases, while our analysis shows a notable improvement ($P < 10^{-71}$ versus $P < 10^{-56}$). Combined analysis with motif predictions shows even stronger agreement; we were able to recapitulate gene expression changes for 2230 binding events or motifs (compared with 585 in the original study) and, for 37 TFs, the target genes are enriched in binding sites for the deleted TF itself ($P < 10^{-193}$) (Figure 4A and C; Supplementary Table S3). A similar level of enrichment was detected for the original analysis for 34 TFs ($P < 10^{-184}$).

We extended this analysis by considering regulatory cascades (44), in which we allowed differentially expressed genes to be secondary targets of the TF under consideration. We detected significant enrichments of binding sites for 129 TFs, compared with 42 TFs for the original dataset (Figure 4B and D, Supplementary Table S3). Altogether, we were able to find gene expression changes associated with 34 232 binding events (compared with 1353 for the original analysis—a 25-fold enrichment). Therefore, our analysis demonstrates that a large proportion (~98%) of the differential expression is likely to be due to secondary-regulatory interactions.

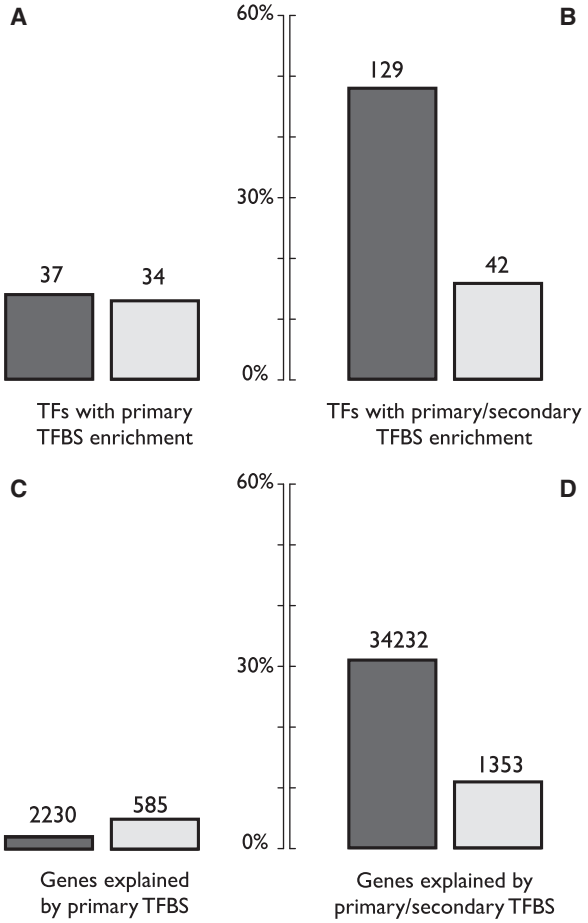


**Figure 4.** Enrichment of TF-binding sites upstream of target genes. (**A**) Proportion of TFs whose target genes are enriched in binding sites for the knockout TF itself. Binding site data are available for 142 TFs (see 'Methods' section). Only primary TF knockout-target gene interactions are considered. (**B**) Proportion of TFs whose target genes are enriched in binding sites for the knockout TF itself, as well as for TFs that are among the set of differentially expressed genes in the TF deletion mutant, i.e. this measure considers both primary regulatory interactions, as well as secondary ones in the regulatory cascade. (**C**) Proportion of differentially expressed genes that can be explained by binding events. Only primary TF knockout-target gene interactions are considered. (**D**) Proportion of differentially expressed genes that can be explained by binding events including regulatory cascades. Results from the current analysis are shown in dark grey; previous analysis results are shown in light grey. Absolute numbers of TFs and target genes are indicated above the bar-plots.

## TFs that physically interact target overlapping sets of genes

The inclusion of protein–protein interaction information provides an additional perspective to the assessment of our dataset. Eukaryotic TFs generally function in a combinatorial manner, and we can identify potential regulatory units by searching for groups of TFs that interact physically with each other (45). For this, we used two datasets: a compilation of protein–protein interactions among stable protein complexes by Collins *et al.* (36) determined using affinity purification/mass spectrometry, and a yeast two-hybrid screen by Yu *et al.* (37), which captures both stable and transient interactions.

Intuitively, we expect TFs that function together to show significant overlap in their target genes. Of the 115 pairs of physically interacting TFs in the dataset, 92 display such an overlap (compared with 49 pairs in the original analysis) (Supplementary Table S4). These include well-known regulatory combinations; for instance HIR2 and HIR3 are subunits of the histone regulatory nucleosome assembly complex that acts as a transcriptional repressor (46,47). The individual knockouts cause up-regulation of genes involved in nuclear assembly and nucleosome functions: of the 550 differentially expressed genes in the HIR2 and HIR3 knockout mutants (240 HIR2; 371 HIR3), 61 are targeted by both HIR2 and HIR3 ($P < 10^{-22}$). The Hu *et al.* dataset, however, shows no overlap among 15 target genes.

Similarly, the RTG regulators RTG1 and RTG3 form a complex to activate the retrograde pathway in response to mitochondrial dysfunctions and nutrient starvation (48,49). Again, there is a significant, although small, overlap in the target genes of the two TFs (26 out of 610 genes—144 RTG1, 466 RTG3, $P < 10^{-4}$), whereas there are no overlapping genes among 45 targets in the Hu *et al.* dataset. This demonstrates the potential of our reanalysis in obtaining meaningful biological information. Other examples can be found in the Supplementary Data.

## Differentially expressed genes form protein complexes

Previous studies have reported that TFs tend to co-regulate genes that interact with each other (50). Therefore we used the interaction information from above to test whether we detect similar behaviour in our reanalysed data. Out of 110 487 differentially expressed genes, there are 3846 pair-wise interactions between co-regulated genes, covering 2262 genes in total (36,37). Most TFs (225) target at least one pair of interacting genes, compared with just 39 TFs in the previous analysis.

To check the statistical significance of our observations, we used a simple module construction approach available through the GraphWeb tool (51). For each TF mutant, we defined a set of 'core' connections comprising interactions among the differentially expressed genes and a set of 'neighbourhood' connections that also include interactions between the core and non-differentially expressed genes. For each mutant, we then compared the number of interactions among core and neighbourhood genes, measuring the proportion of interacting genes that are targeted by the same TF.

We find that targets of 154 TFs are enriched for membership to an interaction module (compared with 38 TFs for the original analysis) ($P$-value threshold $<0.05$; Supplementary Table S5). An interesting example consists of 16 TFs—including histone modifiers (SIN3, SPT10, HFI1, CDC73, SDS3, SAS4, SAS5), general TFs (TAF14) and growth- and metabolism-specific TFs (GLN3, UME6, BAS1, SUM1)—that affect modules related to vitamin B6 metabolism, which is essential to successful glycolysis. Deletion of these regulators cause defective growth phenotypes such as reduced fitness in rich medium (all except SAS4, SAS5, SDS3) and altered

glycogen accumulation (SIN3, HFI1, TAF14, GLN3, SUM1) (1,52,53).

## DISCUSSION

The compendium of TF-knockout expression data by Hu *et al*. (9) is an invaluable resource that makes a substantial contribution to our understanding of the transcriptional regulatory system in the yeast *S. cerevisiae*. We performed a comprehensive reanalysis of the raw data to maximize the information that could be gained from these experiments.

We employed standard pre-processing methods that are freely available through the BioConductor software project. These methods improved on the original publication by greatly reducing the noise inherent in microarray experiments, and by applying strict filters such as multiple-testing correction to minimize false positives.

The resulting dataset contained nearly nine times more differentially expressed genes, including 90% of the original gene list. The numerous functional assessments provide strong indications that the additional targets are biologically meaningful. In fact, the reanalysed dataset achieved better results in most of the tests compared with the original data, suggesting that the current method provides greater sensitivity without compromising on the quality of the output. We demonstrated the importance of detecting small but significant changes in gene expression. For some weakly expressed TFs whose deletion nevertheless has a high regulatory impact, their expression changes are too low to detect even when sensitive methods of analysis are applied. Moreover, there is little correlation between wild-type TF expression and the number of genes affected in its knockout.

Adr1 illustrates a perfect example of the power of the reanalysed dataset. During growth on non-fermentable carbon sources, the Adr1 TF activates glucose-repressed genes involved in non-fermentative carbon metabolism, peroxisome biogenesis and beta-oxidation (54–57). In fermentative growth conditions in rich media, Adr1 is inhibited by the phosphatase complex Glc7-Reg1 via an unknown mechanism (58–60) and its targets are not induced. Given the knockout experiments were conducted in rich media and a role for Adr1 in these conditions has not been described previously, it is a surprise to find that the deletion of ADR1 causes up-regulation of genes enriched in respiratory and mitochondrial functions [respiratory electron transport chain ($P < 10^{-8}$), oxidative phosphorylation ($P < 10^{-4}$), mitochondrial respiratory chain ($P < 10^{-8}$)]. These observations hint at a role for Adr1 in glucose conditions either as a direct repressor, or as an activator of a repressor of these genes. A possible role as both an activator and repressor, depending on carbon source availability, is reminiscent of the dual regulators Sko1 and Hap1 which 'switch' regulatory roles in response to osmotic stress and haem concentration, respectively (61,62).

Hal9 provides a second example for which an experimentally testable hypothesis may be gleaned from the reanalysed dataset. The physiological role of this TF is unknown. One hundred and senventy-two genes are differentially expressed upon deletion of HAL9, and down-regulated genes are enriched in several biosynthetic and transport functions, the most significant of which is the GO term branched chain family amino acid biosynthetic process ($P < 2 \times 10^{-3}$). A possible functional link between HAL9 and amino acid sensing and metabolism was made in a previous genetic study that identified HAL9 as a negative regulator of PTR2 expression (63). PTR2 encodes a transporter of di/tripeptides which is transcriptionally induced by extracellular leucine (64) through the SPS plasma membrane amino acid sensor system [reviewed in ref. (65)].

The current study clearly demonstrates the importance of utilizing advanced and sensitive statistical methods in order to benefit fully from microarray experiments, which are often expensive and time-consuming. Moreover, the additional data afforded by our reanalysis represents a useful resource for future studies of gene regulation. For instance, a comparison of the target gene list with additional microarray data may facilitate the identification of new regulators for poorly characterized cellular processes.

The data presented here are publicly available from the ArrayExpress database (E-MTAB-109).

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

## ACKNOWLEDGEMENTS

J.R. acknowledges funding from the Marie Curie Biostar programme, the Tiger University programme of the Estonian Information Technology Foundation, the Artur Lind and Ustus Agur Foundations.

## FUNDING

EMBL and ERDF through EXCS and COBRED LSHB-CT-2007-037730.

*Conflict of interest statement*. None declared.

## REFERENCES

1. Deutschbauer,A.M., Jaramillo,D.F., Proctor,M., Kumm,J., Hillenmeyer,M.E., Davis,R.W., Nislow,C. and Giaever,G. (2005) Mechanisms of haploinsufficiency revealed by genome-wide profiling in yeast. *Genetics*, **169**, 1915–1925.
2. Sopko,R., Huang,D., Preston,N., Chua,G., Papp,B., Kafadar,K., Snyder,M., Oliver,S.G., Cyert,M., Hughes,T.R. *et al*. (2006) Mapping pathways and phenotypes by systematic gene overexpression. *Mol. Cell*, **21**, 319–330.
3. Xie,M.W., Jin,F., Hwang,H., Hwang,S., Anand,V., Duncan,M.C. and Huang,J. (2005) Insights into TOR function and rapamycin response: chemical genomic profiling by using a high-density cell array method. *Proc. Natl Acad. Sci. USA*, **102**, 7215–7220.
4. Dudley,A.M., Janse,D.M., Tanay,A., Shamir,R. and Church,G.M. (2005) A global view of pleiotropy and phenotypically derived gene function in yeast. *Mol. Syst. Biol.*, **1**, 2005.0001.
5. Giaever,G., Chu,A.M., Ni,L., Connelly,C., Riles,L., Véronneau,S., Dow,S., Lucau-Danila,A., Anderson,K., André,B. *et al*. (2002) Functional profiling of the Saccharomyces cerevisiae genome. *Nature*, **418**, 387–391.

6. Svetlov,V.V. and Cooper,T.G. (1995) Review: compilation and characteristics of dedicated transcription factors in Saccharomyces cerevisiae. *Yeast*, **11**, 1439–1484.
7. 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.* (2000) Functional discovery via a compendium of expression profiles. *Cell*, **102**, 109–126.
8. Chua,G., Morris,Q.D., Sopko,R., Robinson,M.D., Ryan,O., Chan,E.T., Frey,B.J., Andrews,B.J., Boone,C. and Hughes,T.R. (2006) Identifying transcription factor functions and targets by phenotypic activation. *Proc. Natl Acad. Sci. USA*, **103**, 12045–12050.
9. Hu,Z., Killion,P.J. and Iyer,V.R. (2007) Genetic reconstruction of a functional transcriptional regulatory network. *Nat. Genet.*, **39**, 683–687.
10. Vaquerizas,J.M., Dopazo,J. and Díaz-Uriarte,R. (2004) DNMAD: web-based diagnosis and normalization for microarray data. *Bioinformatics*, **20**, 3656–3658.
11. Ritchie,M.E., Silver,J., Oshlack,A., Holmes,M., Diyagama,D., Holloway,A. and Smyth,G.K. (2007) A comparison of background correction methods for two-colour microarrays. *Bioinformatics*, **23**, 2700–2707.
12. Huber,W., von Heydebreck,A., Sültmann,H., Poustka,A. and Vingron,M. (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, **18(Suppl. 1)**, S96–S104.
13. Yang,Y.H., Dudoit,S., Luu,P., Lin,D.M., Peng,V., Ngai,J. and Speed,T.P. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.*, **30**, e15.
14. Smyth,G.K. and Speed,T.P. (2003) Normalization of cDNA microarray data. *Methods*, **31**, 265–273.
15. Nadon,R. and Shoemaker,J. (2002) Statistical issues with microarrays: processing and analysis. *Trends Genet.*, **18**, 265–271.
16. Allison,D.B., Cui,X., Page,G.P. and Sabripour,M. (2006) Microarray data analysis: from disarray to consolidation and consensus. *Nat. Rev. Genet.*, **7**, 55–65.
17. Murie,C., Woody,O., Lee,A.Y. and Nadon,R. (2009) Comparison of small n statistical tests of differential expression applied to microarrays. *BMC Bioinformatics*, **10**, 45.
18. Smyth,G.K. (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, **3**.
19. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B.*, **57**, 289–300.
20. Smyth,G.K., Yang,Y.H. and Speed,T.P. (2003) Statistical issues in cDNA microarray data analysis. *Methods Mol. Biol.*, **24**, 111–136.
21. Ge,Y., Dudoit,S. and Speed,T.P. (2003) Resampling-based multiple testing for microarray data analysis. *Test*, **12**, 1–77.
22. McCarthy,D.J. and Smyth,G.K. (2009) Testing significance relative to a fold-change threshold is a TREAT. *Bioinformatics*, **25**, 765–771.
23. Gentleman,R.C., Carey,V.J., Bates,D.M., Bolstad,B., Dettling,M., Dudoit,S., Ellis,B., Gautier,L., Ge,Y., Gentry,J. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.
24. Harris,M.A., Clark,J., Ireland,A., Lomax,J., Ashburner,M., Foulger,R., Eilbeck,K., Lewis,S., Marshall,B., Mungall,C. *et al.* (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, D258–D261.
25. Kanehisa,M., Araki,M., Goto,S., Hattori,M., Hirakawa,M., Itoh,M., Katayama,T., Kawashima,S., Okuda,S., Tokimatsu,T. *et al.* (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.*, **36**, D480–D484.
26. Vastrik,I., D'Eustachio,P., Schmidt,E., Joshi-Tope,G., Gopinath,G., Croft,D., de Bono,B., Gillespie,M., Jassal,B., Lewis,S. *et al.* (2007) Reactome: a knowledge base of biologic pathways and processes. *Genome Biol.*, **8**, R39.
27. Killion,P.J., Sherlock,G. and Iyer,V.R. (2003) The Longhorn Array Database (LAD): an open-source, MIAME compliant implementation of the Stanford Microarray Database (SMD). *BMC Bioinformatics*, **4**, 32.
28. Reiner,A., Yekutieli,D. and Benjamini,Y. (2003) Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, **19**, 368–375.
29. Reimand,J., Kull,M., Peterson,H., Hansen,J. and Vilo,J. (2007) g:Profiler–a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res.*, **35**, W193–W200.
30. Hubbard,T.J., Aken,B.L., Beal,K., Ballester,B., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cunningham,F., Cutts,T. *et al.* (2007) Ensembl 2007. *Nucleic Acids Res.*, **35**, D610–D617.
31. Harbison,C.T., Gordon,D.B., Lee,T.I., Rinaldi,N.J., Macisaac,K.D., Danford,T.W., Hannett,N.M., Tagne,J., Reynolds,D.B., Yoo,J. *et al.* (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99–104.
32. Zhu,C., Byers,K.J., McCord,R.P., Shi,Z., Berger,M.F., Newburger,D.E., Saulrieta,K., Smith,Z., Shah,M.V., Radhakrishnan,M. *et al.* (2009) High-resolution DNA-binding specificity analysis of yeast transcription factors. *Genome Res.*, **19**, 556–566.
33. Erb,I. and Nimwegen,E.V. (2006) Statistical features of yeast's transcriptional regulatory code. *IEEE Proc. ICCSB*, **1**, 111–118.
34. MacIsaac,K.D., Wang,T., Gordon,D.B., Gifford,D.K., Stormo,G.D. and Fraenkel,E. (2006) An improved map of conserved regulatory sites for Saccharomyces cerevisiae. *BMC Bioinformatics*, **7**, 113.
35. Zhu,J. and Zhang,M.Q. (1999) SCPD: a promoter database of the yeast Saccharomyces cerevisiae. *Bioinformatics*, **15**, 607–611.
36. Collins,S.R., Kemmeren,P., Zhao,X., Greenblatt,J.F., Spencer,F., Holstege,F.C.P., Weissman,J.S. and Krogan,N.J. (2007) Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae. *Mol. Cell Proteomics*, **6**, 439–450, 10.1074/mcp.M600381-MCP200.
37. Yu,H., Braun,P., Yildirim,M.A., Lemmens,I., Venkatesan,K., Sahalie,J., Hirozane-Kishikawa,T., Gebreab,F., Li,N., Simonis,N. *et al.* (2008) High-quality binary protein interaction map of the yeast interactome network. *Science*, **322**, 104–110.
38. Guelzim,N., Bottani,S., Bourgine,P. and Képès,F. (2002) Topological and causal structure of the yeast transcriptional regulatory network. *Nat. Genet.*, **31**, 60–63.
39. Morillon,A., Karabetsou,N., O'Sullivan,J., Kent,N., Proudfoot,N. and Mellor,J. (2003) Isw1 chromatin remodeling ATPase coordinates transcription elongation and termination by RNA polymerase II. *Cell*, **115**, 425–435.
40. Tsukiyama,T., Palmer,J., Landel,C.C., Shiloach,J. and Wu,C. (1999) Characterization of the imitation switch subfamily of ATP-dependent chromatin-remodeling factors in Saccharomyces cerevisiae. *Genes Dev.*, **13**, 686–697.
41. Morillon,A., O'Sullivan,J., Azad,A., Proudfoot,N. and Mellor,J. (2003) Regulation of elongating RNA polymerase II by forkhead transcription factors in yeast. *Science*, **300**, 492–495.
42. Hollenhorst,P.C., Bose,M.E., Mielke,M.R., Müller,U. and Fox,C.A. (2000) Forkhead genes in transcriptional silencing, cell morphology and the cell cycle. Overlapping and distinct functions for FKH1 and FKH2 in Saccharomyces cerevisiae. *Genetics*, **154**, 1533–1548.
43. Zhu,G., Spellman,P.T., Volpe,T., Brown,P.O., Botstein,D., Davis,T.N. and Futcher,B. (2000) Two yeast forkhead genes regulate the cell cycle and pseudohyphal growth. *Nature*, **406**, 90–94.
44. Palin,K., Ukkonen,E., Brazma,A. and Vilo,J. (2002) Correlating gene promoters and expression in gene disruption experiments. *Bioinformatics*, **18(Suppl. 2)**, S172–S180.
45. Reményi,A., Schöler,H.R. and Wilmanns,M. (2004) Combinatorial control of gene expression. *Nat. Struct. Mol. Biol.*, **11**, 812–815, 10.1038/nsmb820.
46. Prochasson,P., Florens,L., Swanson,S.K., Washburn,M.P. and Workman,J.L. (2005) The HIR corepressor complex binds to nucleosomes generating a distinct protein/DNA complex resistant to remodeling by SWI/SNF. *Genes Dev.*, **19**, 2534–2539.
47. Spector,M.S., Raff,A., DeSilva,H., Lee,K. and Osley,M.A. (1997) Hir1p and Hir2p function as transcriptional corepressors to regulate histone gene transcription in the Saccharomyces cerevisiae cell cycle. *Mol. Cell. Biol.*, **17**, 545–552.
48. Crespo,J.L., Powers,T., Fowler,B. and Hall,M.N. (2002) The TOR-controlled transcription activators GLN3, RTG1, and

RTG3 are regulated in response to intracellular levels of glutamine. *Proc. Natl Acad. Sci. USA*, **99**, 6784–6789.
49. Jia,Y., Rothermel,B., Thornton,J. and Butow,R.A. (1997) A basic helix-loop-helix-leucine zipper transcription complex in yeast functions in a signaling pathway from mitochondria to the nucleus. *Mol. Cell. Biol*, **17**, 1110–1117.
50. Ge,H., Liu,Z., Church,G.M. and Vidal,M. (2001) Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. *Nat. Genet.*, **29**, 482–486.
51. Reimand,J., Tooming,L., Peterson,H., Adler,P. and Vilo,J. (2008) GraphWeb: mining heterogeneous biological networks for gene modules with functional significance. *Nucleic Acids Res.*, **36**, W452–W459.
52. Powers,R.W., Kaeberlein,M., Caldwell,S.D., Kennedy,B.K. and Fields,S. (2006) Extension of chronological life span in yeast by decreased TOR pathway signaling. *Genes Dev.*, **20**, 174–184.
53. Wilson,W.A., Wang,Z. and Roach,P.J. (2002) Systematic identification of the genes affecting glycogen storage in the yeast Saccharomyces cerevisiae: implication of the vacuole as a determinant of glycogen level. *Mol. Cell Proteomics*, **1**, 232–242.
54. Schüller,H. (2003) Transcriptional control of nonfermentative metabolism in the yeast Saccharomyces cerevisiae. *Curr. Genet.*, **43**, 139–160.
55. Young,E.T., Dombek,K.M., Tachibana,C. and Ideker,T. (2003) Multiple pathways are co-regulated by the protein kinase Snf1 and the transcription factors Adr1 and Cat8. *J. Biol. Chem.*, **278**, 26146–26158.
56. Tachibana,C., Yoo,J.Y., Tagne,J., Kacherovsky,N., Lee,T.I. and Young,E.T. (2005) Combined global localization analysis and transcriptome data identify genes that are directly coregulated by Adr1 and Cat8. *Mol. Cell. Biol.*, **25**, 2138–2146.
57. Simon,M., Adam,G., Rapatz,W., Spevak,W. and Ruis,H. (1991) The Saccharomyces cerevisiae ADR1 gene is a positive regulator of transcription of genes encoding peroxisomal proteins. *Mol. Cell. Biol.*, **11**, 699–704.
58. Dombek,K.M., Camier,S. and Young,E.T. (1993) ADH2 expression is repressed by REG1 independently of mutations that alter the phosphorylation of the yeast transcription factor ADR1. *Mol. Cell. Biol.*, **13**, 4391–4399.
59. Dombek,K.M., Voronkova,V., Raney,A. and Young,E.T. (1999) Functional analysis of the yeast Glc7-binding protein Reg1 identifies a protein phosphatase type 1-binding motif as essential for repression of ADH2 expression. *Mol. Cell. Biol.*, **19**, 6029–6040.
60. Young,E.T., Kacherovsky,N. and Van Riper,K. (2002) Snf1 protein kinase regulates Adr1 binding to chromatin but not transcription activation. *J. Biol. Chem.*, **277**, 38095–38103.
61. Hickman,M.J. and Winston,F. (2007) Heme levels switch the function of Hap1 of Saccharomyces cerevisiae between transcriptional activator and transcriptional repressor. *Mol. Cell. Biol.*, **27**, 7414–7424.
62. Proft,M. and Struhl,K. (2002) Hog1 kinase converts the Sko1-Cyc8-Tup1 repressor complex into an activator that recruits SAGA and SWI/SNF in response to osmotic stress. *Mol. Cell*, **9**, 1307–1317.
63. Cai,H., Kauffman,S., Naider,F. and Becker,J.M. (2006) Genomewide screen reveals a wide regulatory network for di/tripeptide utilization in Saccharomyces cerevisiae. *Genetics*, **172**, 1459–1476.
64. Perry,J.R., Basrai,M.A., Steiner,H.Y., Naider,F. and Becker,J.M. (1994) Isolation and characterization of a Saccharomyces cerevisiae peptide transport gene. *Mol. Cell. Biol.*, **14**, 104–115.
65. Forsberg,H. and Ljungdahl,P.O. (2001) Sensors of extracellular nutrients in Saccharomyces cerevisiae. *Curr. Genet.*, **40**, 91–109.