Combinatorial Analysis of Perturbational Gene Expression Compendia

Steven Maere*[1,2], Patrick Van Dijck[3,4] , Martin Kuiper[1,2]

[1]Department of Plant Systems Biology, VIB, Technologiepark 927, B-9052 Ghent, Belgium
[2]Department of Molecular Genetics, Ghent University, Technologiepark 927, B-9052 Ghent, Belgium
[3]Department of Molecular Microbiology, VIB, Kasteelpark Arenberg 31, B-3001 Leuven, Belgium
[4]Laboratory of Molecular Cell Biology, Katholieke Universiteit Leuven, Kasteelpark Arenberg 31, B-3001 Leuven, Belgium

Email: Steven Maere - steven.maere@psb.ugent.be; Patrick Van Dijck - patrick.vandijck@bio.kuleuven.be; Martin Kuiper - martin.kuiper@psb.ugent.be;

*Corresponding author

Abstract

Background: Large-scale compendia of gene expression profiles under chemical or genetic perturbations constitute an invaluable resource from a systems biology perspective. However, the perturbational nature of such data imposes specific requirements on the methods used to analyze them. In particular, the distance measures used in traditional clustering algorithms have difficulties in detecting one of the most prominent features of perturbational data, namely partial correlations between expression profiles. Biclustering methods on the other hand are specifically designed to capture such partial correlations. However, most biclustering algorithms do not provide measures for pair-wise expression correlation between genes, but rely on emergent properties of groups of genes and conditions (modules) in order to identify statistically significant subpatterns in the data. This reliance complicates the elucidation of less modular regions in the underlying transcriptional network.

Results: We introduce a novel method to extract (partial) expression correlations and transcriptional modules from perturbational gene expression data, based on the use of combinatorial statistics and graph-based clustering. The modules are further characterized by incorporating GO annotation and transcription factor binding information and by learning tentative regulation programs. We have incorporated this methodology in a software tool called ENIGMA. We show that ENIGMA outperforms other methods on both modular and non-modular artificial data. We also applied ENIGMA to the Rosetta compendium of expression profiles for

Saccharomyces cerevisiae. In particular, we were able to discriminate a subset of candidate mating-related genes whose expression appears to be regulated by the transcription factor Tec1, although their promoters lack Tec1 binding sites. We propose that Tec1 in fact mediates antisense expression of these genes through interaction with nearby Ty1 long terminal repeats (LTRs), and that this effect could be functionally relevant for the mating process. We present evidence that at least one Ty1 LTR-associated gene, namely $YLR343W$, causes a mating-related phenotype upon deletion.

Conclusions: It is increasingly recognized that perturbational expression compendia are essential to identify the gene networks underlying cellular function, and efforts to build these for different organisms are currently underway. We show that ENIGMA constitutes a valuable addition to the repertoire of methods to analyze such data.

Background

Over the last decade, the availability of fully sequenced genomes and the development of high-throughput technologies such as DNA microarray-based transcript profiling have fuelled an exponential increase in the volume of functional genomics data. This has led to a renewed interest in the study of molecular biology at the system level [1–3].

The central paradigm in systems theory is that one can learn about a system by perturbing it and measuring the response. This principle also applies to biological systems. Since mRNA levels can nowadays easily be measured on a genome-wide scale, expression profiling has become a first method of choice for assessing the molecular response to experimental perturbation (the molecular phenotype). Considerable efforts are put into creating compendia of expression profiles under genetic, chemical or environmental perturbations [4–6] or in different tissues [5, 7, 8]. Such data compendia basically constitute a series of snapshots of expression states under a variety of conditions, and they contain a wealth of information concerning the underlying transcriptional network structure of an organism. The challenge now is to devise methods to efficiently and reliably extract that information.

Clustering of DNA microarray data allows the inference of functional correlations through what was dubbed the 'guilt-by-association' principle [9]. A classical clustering process generally consists of two steps [10]. First, a matrix of distances between expression profiles is calculated using a distance or

similarity measure, such as Pearson's centered correlation coefficient (PCC). Based on this distance matrix, the actual clustering algorithm, for instance average linkage hierarchical clustering, groups similar profiles together. Traditional measures such as PCC are well suited for analyzing time-series microarray data, but they fall short when applied to perturbational data, because they only capture global tendencies of co- or antiregulation. In compendia of perturbed expression profiles, genes do not necessarily show similar behavior under all experimental conditions: they may be coexpressed under some perturbations, and show uncorrelated or even inversely correlated expression under other perturbations.

This observation stimulated the development of alternative clustering strategies. The process of detecting subsets of genes that exhibit similar expression behavior across a subset of conditions is known as biclustering. Several biclustering strategies exist today, each using its own heuristic approach to tackle this complex problem (see [11] and references therein). Some biclustering methods use a greedy iterative search strategy to uncover biclusters, progressively subdividing, or adding and removing rows and columns from the biclusters obtained in a previous iteration in order to maximize a local score function [12–14]. Others use linear algebra [15] or adopt a graph-theoretic approach to biclustering [16]. Yet other methods identify biclusters by proposing a statistical model and estimating the distribution parameters that minimize a certain model fit criterion [17–20].

Evidently, a wide variety of biclustering algorithms exist, each of them having their own strengths and weaknesses. For example, some of these methods are intrinsically less suited to find overlap between biclusters because they mask previously found biclusters with random noise [12, 20], or because they partition the data [15]. However, a feature that most of the existing biclustering methods share is that they do not explicitly define similarity measures on the global space of expression profiles that are capable of detecting pair-wise correlations between individual genes or conditions. Some methods [13] use local measures instead, i.e. they calculate a standard pair-wise distance measure between two genes/conditions on a subset of features (conditions/genes). However, most algorithms [12, 14–20] avoid pair-wise comparison of genes or conditions altogether, instead focusing on the emergent properties of groups of genes and conditions in order to identify statistically significant subpatterns in the data. Although these methods are perfectly capable of finding biologically relevant biclusters, their inability to compare individual expression profiles can be a disadvantage in some situations. For example, less modular regions of the coexpression network can inherently not be elucidated with biclustering methods. Also, the study of the expression behavior of particular pathways may require a fine-grained coexpression network, where the correlations between individual components are resolved. The functional study of single genes might also

benefit from high-resolution coexpression analysis. For example, Wu et al. [21] demonstrated that, in order to predict the function of an uncharacterized gene, simply analyzing the functional profile of the top-10 correlated genes is more efficient than analyzing the output of any of the clustering algorithms they tested. Traditional similarity measures, such as PCC, meet this resolution requirement, but they are less suited to uncover partial correlations in expression.

The aims of this study are: (i) to develop a combinatorial statistic that can extract partial expression correlations between individual genes from perturbational expression data, (ii) to devise a method to cluster the resulting network of significant correlations into overlapping expression modules, (iii) to test the methodology on artificial expression data and compare its performance to other methods, (iv) to identify expression modules from perturbational microarray data for budding yeast [4] and analyze the properties of the resulting network, a.o. the extent of overlap and crosstalk between the modules, and (v) to assess the potential of our method to generate testable hypotheses by analyzing a few modules in more detail.

Results and Discussion

ENIGMA

Here, we introduce a novel method, called ENIGMA (Expression Network Inference and Global Module Analysis), to analyze perturbational expression data. ENIGMA uses a novel combinatorial statistic to correlate pairs of perturbed expression profiles (see Methods). Briefly, gene expression profiles are discretized into three categories (upregulated, downregulated, unchanged) based on p-values for differential expression. For each pair of profiles, we then assess the probability that the observed overlap of upregulated and downregulated fields is generated by chance. The resulting correlation p-values are corrected for multiple testing and translated to edges in a coexpression network, which is then clustered into (potentially overlapping) expression modules using a using a graph clustering procedure that identifies densely connected components in the network. The graph clustering procedure depends on two parameters that control the size and density of individual clusters (ν_1), and the overlap between clusters (ν_2). These parameters are optimized using a simulated annealing strategy (see Methods). ENIGMA then determines relevant condition sets for each module and screens the modules for enrichment of Gene Ontology (GO) [22] categories and transcription factor binding sites. Finally, a regulation program is learned for each module in an attempt to explain the expression behavior of the module's genes as a function of the expression of a limited set of regulators (transcription factors and signal transducers). A global overview of the methodology is given in Figure 1.

Performance of ENIGMA on artificial data and comparison with other methods

To assess the performance of our method and compare ENIGMA to other methods, we performed tests on artificial gene expression data. We generated two types of artificial data, namely expression data containing 20 overlapping biclusters (modular data) and expression data containing 500 partial expression correlations but no biclusters (non-modular data, see Methods). In both cases, we built 10 expression datasets of 1000 genes by 100 experiments. For each dataset, we tested the performance of ENIGMA on two levels by assessing the overlap between the artificial input correlation network and (i) the network of significant correlations obtained in the first step of the ENIGMA algorithm (before clustering, referred to as first-stage ENIGMA); (ii) the module network inferred by ENIGMA (after clustering). The performance of the combinatorial statistic built into ENIGMA was compared with that of two other similarity measures, namely PCC and the χ^2-statistic. We also compared ENIGMA with two established biclustering methods, namely SAMBA [16] and ISA [14, 23] (see Methods).

Unlike for real data (see further), we used $\log_2$ ratio thresholds to discretize the expression values for the artificial datasets, since the generation of meaningful artificial p-values proved to merit further study in its own right. We used a $\log_2$ ratio threshold of 1 for upregulation and -1 for downregulation, corresponding to the means of the distributions used to generate the bicluster profiles (see Methods). In other words, half of the datapoints in the biclusters are presumed not to be significantly over- or underexpressed. Therefore, the performance results for ENIGMA obtained here are likely underestimates.

On modular data, ENIGMA outperforms all other methods tested (see Figure 2A and Tables S1 and S2). The rather poor performance of the ISA algorithm may seem somewhat surprising. Prelić et al [24], using the same implementation of ISA but other methods to generate artificial data and to assess biclustering performance, previously established that the performance of ISA decreases with increasing overlap between biclusters. Our results seem to confirm that ISA is not the optimal method in case there is substantial overlap between modules.

SAMBA, however, performs relatively well on modular data. On two out of 10 datasets, the performance of SAMBA was higher than that of ENIGMA , on another 2 out of 10, the performance of SAMBA and ENIGMA were comparable (see Tables S1 and S2). On the remaining 6 datasets, ENIGMA outperformed SAMBA. In all cases, the performance of SAMBA and ENIGMA was higher than the optimal PCC performance (corresponding to a PCC threshold of 0.20-0.30 depending on the dataset), except for 2 datasets on which the performance of SAMBA was comparable to the PCC optimum. The relatively good performance of PCC may give the impression that complicated methods such as ENIGMA or SAMBA are

in fact not really needed. However, the performance of PCC critically depends on the choice of the PCC threshold, and determining the optimal PCC threshold on real data is far from trivial. ENIGMA has the advantage of having an easily tunable threshold parameter: the False Discovery Rate (FDR) level (see Methods). To illustrate this, we plotted the performance curve of first-stage ENIGMA , for different p-value thresholds (before FDR correction), on Figure 2A and 2B. For all artificial datasets, the performance of first-stage ENIGMA at FDR=0.05 (medium gray dot) is close to the optimum of this curve, indicating that FDR control at a reasonable level gives near-optimal performance, although it may be a bit too conservative for modular data (first-stage ENIGMA performance is consistently to the left of the optimum for modular data).

Compared to SAMBA and ISA, ENIGMA has the additional advantage that it recovered the correct number of modules (20) in 8 out of 10 cases (one run with 19 modules and one run with 21 modules), whereas ISA and SAMBA consistently predicted more modules than there actually were (ISA: 27 ± 6 modules and SAMBA: 53 ± 6 modules).

Not surprisingly, the performance of the (bi)clustering methods SAMBA, ISA and ENIGMA on non-modular artificial data is very low (see Figure 2B and Tables S3 and S4). Among the pair-wise methods, first-stage ENIGMA invariably featured the highest performance, indicating that our combinatorial statistic detects partial correlations between expression profiles more efficiently than PCC and χ^2. An attractive feature of ENIGMA is that it finds few modules in the non-modular data (3 ± 1 modules containing on average 5 genes each, precision of clustering result = 0.27), in contrast to ISA and SAMBA, which recover 80 ± 5 modules (containing on average 27 genes) and 127 ± 2 modules (containing on average 16 genes), respectively.

We also constructed 10 artificial datasets that show a mixture of modular and non-modular characteristics, by simply concatenating one of the modular and one of the non-modular artificial datasets, resulting in expression datasets of 2000 genes on 100 conditions with 20 implanted biclusters and 500 implanted partial expression correlations. We tested the capacity of ISA, SAMBA and ENIGMA to retrieve only the truly modular part (the biclusters) from those datasets. Relative to the corresponding modular datasets, a drop in performance is to be expected since the mixed datasets contain more genes, and thus more potential interactions, while the amount of biclusters and their size remain the same (the biclusters are basically diluted). Indeed, all methods showed a drop in F-measure. ISA exhibited the sharpest F-measure drop, on average 0.097 ± 0.081. SAMBA and ENIGMA showed a more moderate drop of the F-measure, 0.035 ± 0.058 for SAMBA and 0.038 ± 0.014 for ENIGMA . As can be seen from the standard deviations, the behavior of

ISA and SAMBA upon addition of non-modular data is rather erratic, with drops in performance ranging from 0.21 (ISA) and 0.14 (SAMBA) to -0.06 for ISA and SAMBA (i.e. greater performance). In contrast, ENIGMA exhibited more predictable behavior, with performance changes in the range 0.01-0.05.
In summary, ENIGMA clearly outperforms other methods on modular artificial data, in terms of recall and F-measure. In addition, the correlation graph produced in the first stage of ENIGMA efficiently captures partial correlations between individual genes from both the modular and non-modular artificial datasets. ENIGMA consistently recovered the right number (± 1) of modules from the modular datasets, and discovered very few modules in the non-modular datasets, in contrast to ISA and SAMBA. On composite datasets with a modular and non-modular part, ENIGMA consistently exhibits a moderate drop in performance, while ISA and SAMBA exhibit a severe resp. moderate drop in performance on average, but with large differences between individual datasets. The characteristics of ENIGMA are particularly useful for analyzing datasets that show a mix between modular and less modular regions. This is the case for real datasets, which typically contain a limited number of perturbation experiments that target a few specific processes. These processes (modules) can be expected to be rather well resolved in terms of their coexpression relationships, whereas other processes will probably give rise to more fragmented (less modular) regions in the network. Moreover, even for well-resolved processes, the true extent of transcriptional modularity remains enigmatic.

Analysis of the Rosetta gene expression compendium

We applied our methodology on 280 perturbed expression profiles from the Rosetta expression compendium for *S. cerevisiae* [4]. Using a differential expression p-value threshold of 0.01 in the discretization step and an FDR threshold of 0.05, ENIGMA identified a network of 100,762 significant positive correlations involving 2,855 genes. The clustering parameters (ν_1,ν_2) were optimized by simulated annealing (SA) as described in the Methods. To assess the efficiency of the SA procedure, we performed an exhaustive screen of the parameter space in order to locate the global maximum of the F-measure criterion (see Figure S1). The SA procedure found back the optimal clustering parameters (ν_1,ν_2) = (0.30, 0.55) with 100% efficiency. Using these parameters, ENIGMA discovered 205 modules in the Rosetta dataset (see supporting data for module details and figures).
According to the GO overrepresentation analysis results, 104 out of 205 modules have a moderate to high degree of functional coherence. Additionally, we screened each module for enrichment of ChIP-determined targets for 102 transcription factors (TFs) [25]. We found that 54 modules are enriched in targets of one or

more TFs. 35 modules show enrichment of both GO Biological Process categories and TF binding sites. Together, 60% of the modules show enrichment of GO categories and/or TF binding sites, indicating that our method is capable of identifying biologically relevant expression modules.

Next, we investigated the global topological characteristics of the coexpression network obtained from the Rosetta dataset. Since many cellular functions are carried out in a highly modular manner [26], most cellular networks, including protein interaction networks, metabolic networks and gene expression networks, are modular in nature [27–32]. On the other hand, many cellular networks, including coexpression networks, have been claimed to exhibit a node degree (k) distribution of the power-law type, $P(k) \sim k^{-\gamma}$, indicative of scale-free properties [32–34]. The coexistence of modularity and a scale-free degree distribution can be explained by assuming a hierarchical modular network organization [28, 32, 34]. According to this view, the network consists of a hierarchy of nested topological modules of increasing size and decreasing coherence. In other words, small coherent modules combine to form larger and less coherent modules in a hierarchical fashion. At reasonable levels of module resolution, the modules consist mainly of sparsely connected but highly clustered nodes (low k, high C). The modules are linked together through a small number of highly connected nodes with a low clustering coefficient (high k, low C), often referred to as hubs. In the case of coexpression networks, these hubs would represent genes that are linked to different expression modules depending on the experimental conditions.

A few papers [35, 36] have cast doubt on the ubiquity of power-law degree distributions in biological networks, claiming that some of the supposed power-laws actually turn out to be closer to exponentials when rigorously analyzed. Indeed, the degree distribution of the ENIGMA network appears to be exponentially distributed (Figure 3A), at least for lower k. For higher k, the picture is different. Relative to the distribution obtained for lower degrees, the most highly connected nodes (hubs) seem to be underconnected. This observation is exactly the opposite of what would be expected for a power-law ('fat-tailed') degree distribution (i.e. highly connected nodes should be overconnected with respect to the exponential distribution), indicating that the coexpression hubs are not nearly as central in the network as would be the case in a scale-free network. However, from the plots of the clustering coefficient C versus the degree k (Figure 3B), it is apparent that the highly connected nodes still possess hub-like characteristics: they generally have a lower clustering coefficient and are assigned to multiple modules. Thus, highly connected nodes act more as local hubs that hold together a few modules.

The hubs in the ENIGMA network, by virtue of their polytomous expression behavior, could be good candidates for key metabolic or regulatory functions. In order to assess whether particular functional

classes of genes are more likely to belong to multiple clusters, we functionally profiled the hubs (i.e. genes belonging to 2 or more modules) with BiNGO [37]. None of the GO Biological Process categories was found enriched among hubs relative to all genes in the ENIGMA network (at FDR = 0.05). Although no functional trends were found, individual hubs might still represent genes that function at the interface of several processes (see further).

Mating modules

In order to assess the potential of ENIGMA to generate testable functional predictions, we performed initial wet-lab validation experiments for some hypotheses, focusing on the yeast mating system. Since the Rosetta compendium contains expression data on at least 20 mating-related perturbations, we expected that the mating system would be well resolved in the coexpression network. Several mating-related modules were uncovered in the Rosetta coexpression network (notably modules 28,77,115 and 171, see supplementary material). A considerable fraction of the genes in these modules contain bona fide binding places for one or several of pheromone response and/or invasive growth-related transcription factors Yhr084w (Ste12), Ypl049c (Dig1), Ymr043w (Mcm1) and Ybr083w (Tec1) (data from [25]).

Module 28 is most strongly related with mating (p = 3.09E-27 for conjugation, GO:0000746). Most of the genes in module 28 are indeed annotated to mating-related GO categories (see Figure 1). The regulation program for this module contains two regulators. The top regulator is the transcription factor *YHR084W* (*STE12*), one of the main regulators of the mating response. The second regulator, *YER155C* (*BEM2*), is a Rho GTPase activating protein involved in the control of cytoskeleton organization and cellular morphogenesis. Since *BEM2* is transcriptionally regulated by Ste12, and since the expression patterns in the leafs 2 and 3 are not very different, *BEM2* is probably less influential as a regulator. In general, one can expect that the regulators farther down the regulation tree are progressively less relevant, as the data splits that they try to explain become less important (i.e. the expression profiles at either side of the split are progressively less divergent).

Module 77 exhibits a more complicated substructure, with six major patterns (1-6) in the condition dimension and six in the gene dimension (a-f, see Figure 4). The top regulator is *YBR083W* (*TEC1*), a transcription factor involved in the regulation of haploid invasive and diploid pseudohyphal growth. The mating and filamentous growth signaling pathways share many of the same components, and *TEC1* is believed to mediate an invasive growth response upon low levels of pheromone signaling [38,39]. Module 77 is also enriched in *TEC1* binding sites, especially in the leafs b, c, e and f. Leafs b, e and f show the

strongest response under condition set 1, which suggests that the regulator *YBR083W* is more relevant for these subsets of genes than for the others. Most of the known mating-related genes of module 77 reside in the gene leafs e and f. Several genes in these leafs overlap with the mating module 28 which was discussed earlier. In contrast, most of the genes in the leafs b and c overlap with module 12, which is mainly enriched in genes involved in cell wall biogenesis (p = 4.68E-08). Compared to the genes in leaf c, the genes in leaf b show a distinctive subpattern in condition leaf 2, which mainly contains perturbations that affect the cell cycle, DNA maintenance and DNA repair. The discriminating regulator for leaf 2 is *YMR199W* (*CLN1*), which codes for a G1 cyclin that activates Cdc28 kinase to promote the G1 to S phase transition. Interestingly, the genes that in leaf b distinguish themselves from the ones in leaf c (except for *YLR332W*) by the presence of TF binding sites for Yer111c (Swi4) and Ylr182w (Swi6), which together form the SBF complex that regulates transcription at the G1/S transition [40]. Both Swi4 and Swi6 are potential substrates of Cdc28, supporting the candidacy of *CLN1* as a regulator for leaf b [41]. Most of the genes in leafs b and c also contain binding sites for the mating-related transcription factors Ste12 and Dig1, and 4 out of 10 contain binding sites for Tec1, justifying their presence in a mating-related module. Together, these data suggest that the set of genes in leaf b functions at the interface of cell wall biogenesis, the G1/S transition and mating/filamentous growth. Such a link makes sense since upon activation of the pheromone signaling pathway, the yeast cell cycle is arrested in G1 and extensive cell wall rearrangements take place [42].

The third regulator, *YFL026W* (*STE2*), is a receptor for the mating α-factor pheromone. Based on the expression profiles in leafs 3 versus 4-5, *YFL026W* seems most relevant as a regulator for the mating-related genes in leaf f and to a lesser extent a, d and e, and less relevant for the cell-wall biosynthesis related genes in leafs b and c. The fourth regulator, *YHR005C* (*GPA1*) is the α subunit of the heterotrimeric G protein that couples to pheromone receptors. Overall, the bottom regulator *YHR005C* again seems less influential (as for module 28), except maybe for the mating-related genes in leaf f. Together with the genes in leafs b and f, most genes in leaf e are strongly repressed under the perturbations in condition leaf 1. Unlike leafs b and f, only a few genes in leaf e (*YCL027W* and *YOL105C*) feature bona fide Tec1 binding sites. However, the expression of the other genes in leaf e (with the exception of *YLR256W*) is specifically and strongly downregulated upon haploid *TEC1* deletion (arrow on Figure 4), suggesting that these genes are somehow transcriptionally regulated by Tec1. Further investigation made apparent that several of these genes are flanked by or overlapping with an antisense Ty1 retrotransposon long terminal repeat (LTR) on the 3' side (*YLR343W*, *YLR334C*, *YOL106W*) or the 5' side (*YOL104C*).

The presence of these Ty elements is highly relevant, since *TEC1* was originally described as a gene required together with *STE12* for full Ty1 expression [43, 44]. Interestingly, three of these genes (*YLR343W*, *YLR334C* and *YOL104C*) were found to be directly or indirectly associated with *TEC1* in a previous study in which the Rosetta compendium was analyzed using a Bayesian network framework [45]. The downregulation of *YOL104C* upon haploid *TEC1* deletion can be directly explained by the presence of a 5' Ty1 LTR in antisense direction (Ty1 LTRs have been found to drive expression in an orientation-independent manner [44]). For *YLR343W*, *YLR334C* and *YOL106W*, the situation is different given the 3' location of the flanking Ty1 LTRs. Tec1 and Ste12 activation of these Ty1 elements could cause the production of antisense transcripts of these loci. Since the probes spotted on the microarray used by Hughes et al [4] contained both strands of the gene sequences, such antisense transcripts could be responsible for the observed coexpression of these genes with *TEC1*.

We set out to investigate whether the hypothetical Ty1-mediated expression of antisense transcripts of these genes is functionally relevant for the mating process. Only two genes in leaf e (*YIL117C* and *YCL027W*) are known to be involved in mating. Neither of them is flanked by a Ty1 LTR. One gene overlapping with an antisense Ty1 LTR, *YOL106W*, was previously reported to elicit a mating-related phenotype upon deletion [46]. We performed mating experiments for the two other 3' Ty1-associated genes in leaf e, namely *YLR334C* (overlapping antisense Ty1 LTR) and *YLR343W* (non-overlapping antisense Ty1 LTR), in addition to a wild type (WT) strain and *sst2*Δ, a mutant that is supersensitive to mating factor-induced G1-arrest.

In the halo assay, the strain deleted for *YLR343W* exhibited an interesting phenotype, characterized by extensive colony formation inside the halo (see Figure 5), which indicates that deletion of *YLR343W* somehow facilitates the recovery from α-factor induced growth arrest. In the mating and growth assays, we did not observe any effect of *YLR343W* deletion on the mating ability (see Table S5, Table S6 and Figure S2). *YLR343W* (*GAS2*) is homologous to *GAS1*, which encodes a 1,3-β-glucanosyltransferase required for cell wall assembly. In a recent study, *GAS2* was found to be involved in spore wall assembly [47]. Ectopic expression of *GAS2* under control of the *GAS1* promoter was found to complement the *gas1*Δ phenotype only partially, and only at pH = 6.5 [47]. Furthermore, whereas Gas1 localizes to the cell wall, Gas2 is found in the cytoplasm [48]. It is therefore unlikely that *YLR343W* directly functions in regular cell wall assembly or maintenance. However, antisense transcripts of *YLR343W*, produced under control of Tec1, might interfere with the expression of its homolog *GAS1* and hence indirectly with the formation and maintenance of the cell wall. An altered cell wall morphology might influence the localization of a.o. Bar1

(Sst1), an aspartyl protease secreted into the periplasmic space of *MAT***a** cells that cleaves and inactivates α-factor, allowing cells to recover from α-factor-induced cell cycle arrest [49]. The localization of Bar1 might in turn influence the efficiency with which it inactivates α-factor, which could explain the observed *ylr343w*Δ phenotype. Obviously, this is only a hypotheses, and further experimentation is needed to unravel how these functional data are linked to the observed mating phenotype. This is however outside the scope of the present study.

The *ylr334c*Δ deletion strain yielded halos indistinguishable from the WT strain, but exhibited mildly reduced mating ability. The effect was more pronounced after 4 hours than after 24 hours, indicating that deletion of *YLR334C* primarily leads to a retardation of the mating response (see Table S5, Table S6 and Figure S2). However, given that the observed reduction in mating ability is not really dramatic, the extent of the involvement of *YLR334C* in the mating process remains uncertain.

In summary, we believe that the occurrence of Ty1 LTRs on the 3' side of candidate mating-related genes associated with *TEC1* is not coincidental, but that they mediate Tec1-driven expression of the corresponding antisense transcripts. Furthermore, given the fact that deletion of several Ty1 LTR associated genes gives rise to a mating-related phenotype (*YLR343W*: this study; *YLR334C* (mild effect): this study, *YOL106W*: [46]), Ty1-mediated antisense expression of these genes (possibly causing silencing of these or other genes) might be functionally relevant for the mating process.

Conclusions

We have developed a novel method, called ENIGMA, to analyze perturbational microarray data. One of the major innovations of our methodology is the use of a combinatorial statistic that is capable of detecting partial correlations between expression profiles. In this respect, our method can be considered similar in purpose to biclustering methods, although ENIGMA assesses expression correlation between individual genes rather than expression coherence in a group of genes under a group of conditions. Our method produces both a detailed network of significant pair-wise expression correlations and a high-level representation of the modularity in the expression network.

Tests on artificial data have shown that ENIGMA outperforms other methods. On modular data, ENIGMA retrieved the correct number of modules with high efficiency. Few modules were found in non-modular data, as expected, but the correlation graphs obtained in the first stage of the ENIGMA algorithm reconstructed the non-modular input networks more accurately than other methods. Together, these properties render ENIGMA very well suited for the analysis of expression datasets that show a mixture of

modular and non-modular characteristics, which is invariably the case for real datasets.

In our re-analysis of the Rosetta compendium for *S. cerevisiae* [4], we have shown that ENIGMA captures biologically meaningful partial correlations in expression, and we uncovered a considerable overlap between expression modules, indicative of extensive crosstalk between processes at the transcriptional level. We analyzed a few mating-related modules in more detail and demonstrated the potential of ENIGMA to generate testable predictions.

Although numerous approaches have already been used to mine the Rosetta compendium [4, 21, 34, 45, 46], ENIGMA allowed us to get novel perspectives on the data. This illustrates that no single approach can extract all the information hidden in large compendium datasets. The elucidation of the regulatory networks governing the many different aspects of cellular function will therefore not only require the integration of different types of data, but also the integrated use of several complementary methods to analyze these data. We believe that ENIGMA constitutes a valuable addition to the existing repertoire of analysis methods.

Methods

Simulated and real expression data

We simulated two types of expression data, namely expression data containing overlapping biclusters (modular data) and expression data containing partial expression correlations but no biclusters (non-modular data). In both cases, we built 10 expression datasets of 1000 genes by 100 experiments (in $\log_2$ ratio format). For each dataset, artificial background expression data were randomly sampled from a normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 0.16$. For the modular datasets, we implanted 20 artificial biclusters in this background, each encompassing between 1-5% of all genes and 10-50% of all conditions. Bicluster sizes, member genes and conditions are chosen at random, with the restriction that at most 30% of the genes and 50% of the conditions overlap between any 2 biclusters (percentages relative to the smallest of the 2 biclusters). Except for a noise component (see further), all genes in a bicluster share the same expression profile over the bicluster conditions. However, a bicluster can be partially overwritten by other biclusters. The bicluster profiles are sampled from a bimodal distribution consisting of 2 normal modes with means μ_1 = -1 (for down-regulated expression) and $\mu_2 = 1$ (for up-regulated expression) and variances $\sigma_{1,2} = 0.49$. The expression profiles of individual genes in a bicluster are noisified by adding normally distributed noise (mean $\mu_n = 0$ and standard deviation $\sigma_n = 0.2\|x\|$ with $\|x\|$ the amplitude of the log ratio expression of the gene in the given condition). The variances and bicluster size and overlap

parameters are chosen so that the overall distribution of simulated log ratio expression values mimicks the distribution of log ratio expression values in the Rosetta compendium [4], up to a scale factor (see Figure S3).

For the non-modular datasets, we simulated 500 partial expression correlations between individual genes, by implanting pairs of correlated expression profiles (encompassing 10-50% of all conditions) in the background. The expression profiles are constructed as in the modular case. The resulting expression value distribution again mimicks the Rosetta distribution (see Figure S3).

For the analyses on real data, we used the Rosetta compendium of expression profiles, representing data on 300 different experimental perturbations in *S. cerevisiae* [4]. Experiments on 20 strains that were marked as aneuploid in the original dataset were left out, because they can give rise to artificial expression correlations between genes on the aneuploid chromosomes. The data were downloaded in prenormalized and preprocessed form. We used the mean $\log_{10}$ values of the expression ratios (perturbation vs. control) and the corresponding p-values for differential expression.

Combinatorial distribution

Consider the expression profiles of two genes A and B under N perturbations. Each gene is represented by a profile of N fields (see Figure 1). A gene is considered to be upregulated in a given perturbation experiment if the $\log_2$ expression ratio (perturbation vs. control) is ≥ 1 (artificial data) or if the p-value for differential expression is significant ($p < 0.01$) and $\log_2$ ratio > 0 (real data). These fields are labeled blue. Experiments in which the gene is downregulated ($\log_2$ ratio ≤ -1 for artificial data or $\log_2$ ratio < 0 and $p < 0.01$ for real data) are similarly labeled yellow, and the remaining fields are labeled black. In order to compare the profiles of the two genes A and B, let us assume that profiles A and B have a_x and b_x blue fields respectively, as well as a_y and b_y yellow fields, and that they have x blue and y yellow fields in common. We want to assess whether this overlap is statistically significant. If the response of the genes A and B to the perturbations were uncorrelated (null hypothesis), the blue and yellow fields would be independently distributed on both profiles. Under this hypothesis (if we randomly distribute a_x blue and a_y yellow fields on profile A, and b_x blue and b_y yellow fields on profile B) the probability that the profiles overlap on exactly x blue and y yellow positions is given by the following recursive formula:

$$P(x,y) = \frac{\binom{a_x}{x}\binom{a_y}{y}\binom{N-x-y}{b_x-x}\binom{N-b_x-y}{b_y-y}}{\binom{N}{b_x}\binom{N-b_x}{b_y}} - \sum_{\substack{x'=x \\ (x',y')\neq(x,y)}}^{\min(a_x,b_x)} \sum_{y'=y}^{\min(a_y,b_y)} \binom{x'}{x}\binom{y'}{y} P(x',y') \qquad (1)$$

The probability of observing an overlap of at least x blue and y yellow fields by chance is then expressed by the cumulative distribution:

$$P_c(x,y) = \sum_{x'=x}^{\min(a_x,b_x)} \sum_{y'=y}^{\min(a_y,b_y)} P(x',y') \tag{2}$$

Equation 1 can be understood by assuming that profile A is given, and that we randomly distribute b_x blue and b_y yellow positions on profile B. The denominator of the first term then represents the total number of possible profiles B. The numerator represents the combinations in which x blue and y yellow matching positions are selected, and the residual positions are chosen at random. However, in this manner, a number of combinations are selected while having more than exactly x blue and/or y yellow matching positions. Moreover, combinations with $x' > x$ blue and/or $y' > y$ yellow matching positions are counted $\mathrm{C}(x',x).\mathrm{C}(y',y)$ times, hence the second term (see supporting methods).

Although the probabilistic question we aim to answer can be formulated in terms of contingency tables, the hypothesis tested by our statistic is fundamentally different from that tested by standard contingency table analysis methods such as the chi-square test, Fisher's exact test or mutual information (MI). For example, situations in which the blue (upregulated) fields in profile A are perfectly mapped onto the black fields (up nor down) in profile B would yield significant chi-square and MI values, whereas they would not yield a significant p-value with our scoring scheme. Indeed, our statistic only considers mappings of up- and down-regulation of the expression of a gene to up- or down-regulation of another gene to be meaningful for assessing coregulation, a premise which is motivated by the perturbational nature of the data we aim to analyze. Black fields are considered non-informative from the perspective of coregulation.

Multiple testing correction on coexpression p-values

In our probabilistic framework, each comparison of two profiles can be considered an individual test. For N genes, $N(N-1)/2$ tests are performed to fish for correlated expression partners. The obtained p-values have to be adjusted in order to control the type I error rate. The p-values are corrected for multiple testing with the Benjamini & Hochberg procedure, which controls the False Discovery Rate (FDR) [50].

Graph-based clustering

The set of significant partial expression correlations at FDR = 0.05 is translated to a network, with nodes and edges representing genes and significant correlations, respectively. We identify expression modules from this correlation network using a graph-based clustering technique similar to MCODE [51]. To identify

potential module seeds, all nodes v are weighted based on the density of the highest k-core of the node neighborhood (denoted as the k_{max}-core of v). A k-core of a graph is a maximal subgraph in which each node has at least degree k. Analogous to Bader and Hogue [51], the core-clustering coefficient C_{core} of v is defined as the density of the k_{max}-core of v, and the weight of v as the product of the core-clustering coefficient of v and k_{max}.

The k_{max}-core of the node with the highest weight is taken as the first module seed. This module seed then grows by accreting nodes on which it exerts a pull above a certain threshold ν_2. The pull of a module with seed S on a node v outside the module is defined as the number of nodes in the neighborhood of v that belong to S, divided by the size of S. The next module is then initiated by taking the k_{max}-core of the node with the highest weight in the remaining graph. An additional constraint is set by requiring that the overlap between the new seed S and any existing module C does not exceed $\nu_1 . \min(N_S, N_C)$, with N_S and N_C the size of the seed and the module, respectively. While the threshold ν_2 controls the size and density of individual modules, ν_1 controls the spacing or overlap between modules. Modules are named after the gene that defined its seed.

Optimization of clustering parameters

No clear-cut criterion exists to score the clustering performance as a function of the parameters ν_1 and ν_2. Standard internal criteria for partitional clustering performance, such as the silhouette width or Dunn's index [52, 53], do not apply for clustering strategies in which clusters are allowed to overlap.

If we consider true/false positives (tp resp. fp) to be correlations between genes (edges) inferred by the clustering that are present/absent in the original network, and false negatives (fn) as edges present in the original network that are not inferred by the clustering, we can define the precision $P = tp/(tp + fp)$ and the recall $R = tp/(tp + fn)$ of the clustering result. An edge $A - B$ that is inferred multiple times from the clustering, because it belongs to the intersection of multiple (say x) modules, is counted as 1 tp and $x - 1$ fp. This is equivalent to drawing x edges between the genes A and B in the correlation graph inferred from the clustering. Since there is only one edge in the original graph, the $x - 1$ remaining edges can be considered superfluous (or false). The strategy to penalize overpredicted edges has the intuitively pleasing property of not affecting the recall, but lowering the precision of the clustering result when the amount of edges 'explained' by multiple modules increases. We use the F-measure, i.e. the harmonic mean of recall (R) and precision (P), $F = 2PR/(P + R)$, as a measure for the quality of the clustering.

The parameters ν_1 and ν_2 are then optimized by simulated annealing (SA) [54], using F as the

optimization criterion. In practice, ENIGMA performs 3 runs of SA, starting from randomly chosen (ν_1,ν_2). The convergence of the solutions of the three runs can be used as a check on the adequacy of the SA parameter choice. We used a two-stage SA procedure, in the first stage, a rough SA search of the clustering parameter space is performed in order to identify the most interesting parameter region (SA settings: begin temperature = 0.1, end temperature = 0.001, cooling rate = 0.99, step size = 0.05). In the second stage, a finer SA search is performed starting from the optimum obtained in the first stage (SA settings: begin temperature = 0.01, end temperature = 0.0001, cooling rate = 0.995, step size = 0.01). At the end of each stage, an additional gradient descent is performed towards the nearest optimum of F.

Calculating condition sets

For each gene module, we determine a condition set by selecting those conditions that show overrepresentation of significant up- or downregulated genes in the module relative to the whole microarray (hypergeometric test, FDR = 0.05). Thus, for a given module, the condition set contains the experimental conditions that elicit a significant and specific response in the module (as compared to the overall response) and, by consequence, have been most influential in shaping the module. The resulting 'bicluster' does not necessarily have a uniform expression pattern over all genes, but may show subpatterns for some genes under certain conditions, possibly indicating involvement in other expression modules. Although conditions that show differential patterning within one module might appear to be irrelevant for the module as a whole, they are important for at least part of the module and may provide insight into inter-module connections or further substructure within the module.

Learning regulation programs

The purpose of a regulation program is to explain the expression behavior of a module as a function of the expression of a limited number of potential regulators [19, 55]. To this end, the expression data subset defined by each module is clustered in both dimensions by average linkage hierarchical clustering, using the cosine correlation coefficient as a similarity measure ($\cos(\theta) = \mathbf{x}.\mathbf{y}/(\|\mathbf{x}\|.\|\mathbf{y}\|)$ for expression vectors $\mathbf{x}$ and $\mathbf{y}$). The average linkage trees are cut at $\cos(\theta) = 0.65$. Regulators are assigned to the internal nodes of the resulting condition tree (regulation tree) on the basis of a conditional entropy criterion proposed in [55]. Given the propensity of condition tree splits involving small-size leafs to generate spurious regulatory predictions, leafs encompassing < 3 conditions are left out of the tree for the purpose of learning the regulation program.

At a given internal node T_α of the regulation tree, the experiment set E_α is partitioned into two distinct sets E_{α_1} and E_{α_2} according to the tree structure. On the other hand, given a regulator r and split value s, E_α can be partitioned in the sets

$$R_1 = \{m \in E_\alpha \colon x_{r,m} \leq s\}$$
$$R_2 = \{m \in E_\alpha \colon x_{r,m} > s\},$$

where $x_{r,m}$ is the expression value of regulator r in experiment m.
The uncertainty in the partitioning E given knowledge (through the data) of the partitioning R is given by the conditional entropy [55, 56]

$$H(E \mid R) = p_1 h(q_1) + p_2 h(q_2), \tag{3}$$

where $p_i = \frac{|R_i|}{|E_\alpha|}$, h is the binary entropy function

$$h(q) = -q \log(q) - (1-q) \log(1-q),$$

and q_i are the conditional probabilities

$$q_i = \frac{|E_{\alpha_1} \cap R_i|}{|R_i|},\ i = 1, 2.$$

To each internal node of a regulation tree, we assign the regulator – split value pair that minimizes the conditional entropy (3). The conditional entropy is equal to 0 only when the regulator – split value pair 'explains' the split in the regulation tree exactly. Only transcription factors and signal transducers are considered potential regulators (list taken from Segal et al. [19]). Transcription factors of which the binding sites are enriched in the module are prioritized when their entropy is comparable to that of the minimum-entropy regulator (entropy difference < 0.001).

Integration with ChIP data and GO profiling of modules

We obtained data on genome-wide binding and phylogenetically conserved motifs for 102 TFs from Harbison et al [25]. Genes that were bound with $p < 0.005$ were considered true targets. For each TF, we determined whether its targets were significantly enriched in any expression module (hypergeometric test, FDR = 0.05). We used BiNGO [37] to determine significantly enriched GO terms for each expression module (hypergeometric test, FDR = 0.05).

Software

The methodology outlined above is implemented in a command-line Java application called ENIGMA , which stands for 'Expression Network Inference and Global Module Analysis'. ENIGMA is open-source and freely available (under GNU General Public License) from [57].

Performance criterion and comparison with other methods

In order to quantify the performance of ENIGMA on artificial data, we calculated the F-measure of the correlation networks generated by ENIGMA (before and after clustering, see main text) and other methods with respect to the input correlation networks. The first-stage ENIGMA and χ^2 networks were constructed by translating correlations with FDR-corrected $p < 0.05$ to edges in the graph. The performance of PCC was measured for different thresholds (for each threshold t, gene pairs with PCC $< t$ define an edge in the network). Output networks for SAMBA, ISA and ENIGMA are obtained by converting biclusters/modules to fully connected network components. The F-measure is calculated as described above for the optimization of the graph-clustering parameters, except that multiply inferred edges are not penalized, in order to compare the different methods on equal footing (SAMBA, ISA and ENIGMA give different numbers of clusters and different amounts of overlap between clusters, while PCC, χ^2 and first-stage ENIGMA , i.e. before clustering, do not generate clusters or multiply inferred edges). The performance of the methods is only assessed in the gene dimension, since the cluster structure in the condition dimension is not resolved for the PCC, χ^2 and first-stage ENIGMA networks. We used the version of SAMBA [16] incorporated in the EXPANDER 3.0 package [58], and the implementation of ISA [23] available as part of the biclustering tool BicAT [59], both with default parameter settings.

Mating experiments

Yeast strains were grown overnight in YPD [yeast extract (1%), peptone (2%) and glucose (2%)] and diluted to an $OD_{600} = 0.5$ in fresh YPD. 500 μl of each strain ($MAT\mathbf{a}$) was mixed with 500 μl of the wild type strain ($MAT\boldsymbol{\alpha}$). The cells were shaken with 180 rpm at 30 °C. At time points 0h, 4h and 24h, 100 μl samples were serially diluted and plated on medium lacking either methionine ($MAT\boldsymbol{\alpha}$), lysine ($MAT\mathbf{a}$) or methionine and lysine (diploids).

Halo assay

A halo assay to measure response to and recovery from pheromone-induced growth arrest was performed as follows. Yeast cells (*MAT***a**) were grown overnight and diluted to OD_{600}=1. 500 μl was plated on YPD plates (1.5% agar in YPD). When the plates were dry, 2 μl of the α mating factor (= 1000 pmol) was spotted. The cells were allowed to grow for 48 hrs before the plates were scanned.

Growth assay

Yeast strains (*MAT***a**) were incubated with the wild type strain (*MAT*$\boldsymbol{\alpha}$) for 4 hours as described above and diluted to OD_{600}= 0.1. The length of the lag phase and the maximum growth rate of yeast strains in SDglu without lysine and methionine were monitored automatically by OD_{600} measurements with a BioscreenC apparatus (Labsystems). The parameters were as follows: 300 μl of culture in each well, 30 s of shaking each 3 min (medium intensity), and OD_{600} measurement every hour. Readings are saturated at OD_{600}s above 1.5.

Authors contributions

S.M. designed the study, analyzed and interpreted the data, and wrote the paper. P.V.D. performed experiments and analyzed data, and M.K. designed the study and supervised the project.

Acknowledgements

We thank Yvan Saeys, Thomas Abeel, Yves Van de Peer, Johan Thevelein and Dirk Aeyels for critical comments on the manuscript. Cindy Colombo is acknowledged for her technical assistance and Martine De Cock for help in preparing the manuscript. S.M. is a Postdoctoral Fellow of the Research Foundation Flanders (Belgium).

References

1. Ideker T, Galitski T, Hood L: **A new approach to decoding life: systems biology**. *Annu Rev Genomics Hum Genet* 2001, **2**:343–372.
2. Kitano H: **Systems biology: a brief overview**. *Science* 2002, **295**:1662–1664.
3. Hohmann S: **The Yeast Systems Biology Network: mating communities**. *Curr Opin Biotechnol* 2005, **16**:356–360.
4. Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai H, He YD, Kidd MJ, King AM, Meyer MR, Slade D, Lum PY, Stepaniants SB, Shoemaker DD, Gachotte D, Chakraburtty K, Simon J, Bard M, Friend SH: **Functional discovery via a compendium of expression profiles**. *Cell* 2000, **102**:109–126.

5. Kim SK, Lund J, Kiraly M, Duke K, Jiang M, Stuart JM, Eizinger A, Wylie BN, Davidson GS: **A gene expression map for Caenorhabditis elegans**. *Science* 2001, **293**:2087–2092.

6. Lee Dk, Park JW, Kim YJ, Kim J, Lee Y, Kim J, Kim JS: **Toward a functional annotation of the human genome using artificial transcription factors**. *Genome Res* 2003, **13**:2708–2716.

7. Zhang W, Morris QD, Chang R, Shai O, Bakowski MA, Mitsakakis N, Mohammad N, Robinson MD, Zirngibl R, Somogyi E, Laurin N, Eftekharpour E, Sat E, Grigull J, Pan Q, Peng WT, Krogan N, Greenblatt J, Fehlings M, van der Kooy D, Aubin J, Bruneau BG, Rossant J, Blencowe BJ, Frey BJ, Hughes TR: **The functional landscape of mouse gene expression**. *J Biol* 2004, **3**:21.

8. Schmid M, Davison TS, Henz SR, Pape UJ, Demar M, Vingron M, Schölkopf B, Weigel D, Lohmann JU: **A gene expression map of Arabidopsis thaliana development**. *Nat Genet* 2005, **37**:501–506.

9. Walker MG, Volkmuth W, Sprinzak E, Hodgson D, Klingler T: **Prediction of gene function by genome-scale expression analysis: prostate cancer-associated genes**. *Genome Res* 1999, **9**:1198–1203.

10. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns**. *Proc Natl Acad Sci U S A* 1998, **95**:14863–14868.

11. Madeira SC, Oliveira AL: **Biclustering algorithms for biological data analysis: a survey**. *IEEE/ACM Trans Comput Biol Bioinform* 2004, **1**:24–45.

12. Cheng Y, Church GM: **Biclustering of expression data**. *Proc Int Conf Intell Syst Mol Biol* 2000, **8**:93–103.

13. Getz G, Levine E, Domany E: **Coupled two-way clustering analysis of gene microarray data**. *Proc Natl Acad Sci U S A* 2000, **97**:12079–12084.

14. Ihmels J, Bergmann S, Barkai N: **Defining transcription modules using large-scale gene expression data**. *Bioinformatics* 2004, **20**:1993–2003.

15. Kluger Y, Basri R, Chang JT, Gerstein M: **Spectral biclustering of microarray data: coclustering genes and conditions**. *Genome Res* 2003, **13**:703–716.

16. Tanay A, Sharan R, Shamir R: **Discovering statistically significant biclusters in gene expression data**. *Bioinformatics* 2002, **18 Suppl 1**:136–144.

17. Lazzeroni L, Owen A: **Plaid models for gene expression data**. *Stat Sinica* 2002, **12**:61–86.

18. Segal E, Battle A, Koller D: **Decomposing gene expression into cellular processes**. *Pac Symp Biocomput* 2003, :89–100.

19. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N: **Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data**. *Nat Genet* 2003, **34**:166–176.

20. Sheng Q, Moreau Y, De Moor B: **Biclustering microarray data by Gibbs sampling**. *Bioinformatics* 2003, **19 Suppl 2**:II196–II205.

21. Wu LF, Hughes TR, Davierwala AP, Robinson MD, Stoughton R, Altschuler SJ: **Large-scale prediction of Saccharomyces cerevisiae gene function using overlapping transcriptional clusters**. *Nat Genet* 2002, **31**:255–265.

22. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium**. *Nat Genet* 2000, **25**:25–29.

23. Bergmann S, Ihmels J, Barkai N: **Iterative signature algorithm for the analysis of large-scale gene expression data**. *Phys Rev E Stat Nonlin Soft Matter Phys* 2003, **67**:031902.

24. Prelić A, Bleuler S, Zimmermann P, Wille A, Bühlmann P, Gruissem W, Hennig L, Thiele L, Zitzler E: **A systematic comparison and evaluation of biclustering methods for gene expression data.** *Bioinformatics* 2006, **22**:1122–1129.

25. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J, Jennings EG, Zeitlinger J, Pokholok DK, Kellis M, Rolfe PA, Takusagawa KT, Lander ES, Gifford DK, Fraenkel E, Young RA: **Transcriptional regulatory code of a eukaryotic genome**. *Nature* 2004, **431**:99–104.

26. Hartwell LH, Hopfield JJ, Leibler S, Murray AW: **From molecular to modular cell biology**. *Nature* 1999, **402**:47–52.

27. Ihmels J, Friedlander G, Bergmann S, Sarig O, Ziv Y, Barkai N: **Revealing modular organization in the yeast transcriptional network**. *Nat Genet* 2002, **31**:370–377.

28. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabási AL: **Hierarchical organization of modularity in metabolic networks**. *Science* 2002, **297**:1551–1555.

29. Bar-Joseph Z, Gerber GK, Lee TI, Rinaldi NJ, Yoo JY, Robert F, Gordon DB, Fraenkel E, Jaakkola TS, Young RA, Gifford DK: **Computational discovery of gene modules and regulatory networks**. *Nat Biotechnol* 2003, **21**:1337–1342.

30. Rives AW, Galitski T: **Modular organization of cellular networks**. *Proc Natl Acad Sci U S A* 2003, **100**:1128–1133.

31. Han JDJ, Bertin N, Hao T, Goldberg DS, Berriz GF, Zhang LV, Dupuy D, Walhout AJM, Cusick ME, Roth FP, Vidal M: **Evidence for dynamically organized modularity in the yeast protein-protein interaction network**. *Nature* 2004, **430**:88–93.

32. Barabási AL, Oltvai ZN: **Network biology: understanding the cell's functional organization**. *Nat Rev Genet* 2004, **5**:101–113.

33. Albert R, Barabási AL: **Statistical mechanics of complex networks**. *Rev Mod Phys* 2002, **74**:47–97.

34. Bergmann S, Ihmels J, Barkai N: **Similarities and differences in genome-wide expression data of six organisms**. *PLoS Biol* 2004, **2**:E9.

35. Amaral LA, Scala A, Barthelemy M, Stanley HE: **Classes of small-world networks**. *Proc Natl Acad Sci U S A* 2000, **97**:11149–11152.

36. Tanaka R, Yi TM, Doyle J: **Some protein interaction data do not exhibit power law statistics**. *FEBS Lett* 2005, **579**:5140–5144.

37. Maere S, Heymans K, Kuiper M: **BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks**. *Bioinformatics* 2005, **21**:3448–3449.

38. Erdman S, Snyder M: **A filamentous growth response mediated by the yeast mating pathway**. *Genetics* 2001, **159**:919–928.

39. Bao MZ, Schwartz MA, Cantin GT, Yates JRr, Madhani HD: **Pheromone-dependent destruction of the Tec1 transcription factor is required for MAP kinase signaling specificity in yeast**. *Cell* 2004, **119**:991–1000.

40. Horak CE, Luscombe NM, Qian J, Bertone P, Piccirrillo S, Gerstein M, Snyder M: **Complex transcriptional circuitry at the G1/S transition in Saccharomyces cerevisiae.** *Genes Dev* 2002, **16**:3017–3033.

41. Ubersax JA, Woodbury EL, Quang PN, Paraz M, Blethrow JD, Shah K, Shokat KM, Morgan DO: **Targets of the cyclin-dependent kinase Cdk1.** *Nature* 2003, **425**:859–864.

42. Lesage G, Bussey H: **Cell wall assembly in Saccharomyces cerevisiae.** *Microbiol Mol Biol Rev* 2006, **70**:317–343.

43. Laloux I, Dubois E, Dewerchin M, Jacobs E: **TEC1, a gene involved in the activation of Ty1 and Ty1-mediated gene expression in Saccharomyces cerevisiae: cloning and molecular analysis**. *Mol Cell Biol* 1990, **10**:3541–3550.

44. Laloux I, Jacobs E, Dubois E: **Involvement of SRE element of Ty1 transposon in TEC1-dependent transcriptional activation.** *Nucleic Acids Res* 1994, **22**:999–1005.

45. Pe'er D, Regev A, Elidan G, Friedman N: **Inferring subnetworks from perturbed expression profiles.** *Bioinformatics* 2001, **17 Suppl 1**:S215–S224.

46. Tanay A, Sharan R, Kupiec M, Shamir R: **Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data**. *Proc Natl Acad Sci U S A* 2004, **101**:2981–2986.

47. Ragni E, Coluccio A, Rolli E, na JMRP, Colasante G, Arroyo J, Neiman AM, Popolo L: **GAS2 and GAS4, a pair of developmentally regulated genes required for spore wall assembly in Saccharomyces cerevisiae.** *Eukaryot Cell* 2007, **6**:302–316.

48. Huh WK, Falvo JV, Gerke LC, Carroll AS, Howson RW, Weissman JS, O'Shea EK: **Global analysis of protein localization in budding yeast**. *Nature* 2003, **425**:686–691.

49. MacKay VL, Welch SK, Insley MY, Manney TR, Holly J, Saari GC, Parker ML: **The Saccharomyces cerevisiae BAR1 gene encodes an exported protein with homology to pepsin**. *Proc Natl Acad Sci U S A* 1988, **85**:55–59.

50. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing**. *J Roy Stat Soc B* 1995, **57**:289–300.

51. Bader GD, Hogue CWV: **An automated method for finding molecular complexes in large protein interaction networks**. *BMC Bioinformatics* 2003, **4**:2.

52. Halkidi M, Batistakis Y, Vazirgiannis M: **On clustering validation techniques**. *J Intell Inf Syst* 2001, **17**:107–145.

53. Bolshakova N, Azuaje F: **Cluster validation techniques for genome expression data**. *Signal Process* 2003, **83**:825–833.

54. Kirkpatrick S, Gelatt CD, Vecchi MP: **Optimization by simulated annealing**. *Science* 1983, **220**:671–680.

55. Michoel T, Maere S, Bonnet E, Joshi A, Saeys Y, den Bulcke TV, Leemput KV, van Remortel P, Kuiper M, Marchal K, de Peer YV: **Validating module network learning algorithms using simulated data.** *BMC Bioinformatics* 2007, **8 Suppl 2**:S5.

56. Shannon CE: **A mathematical theory of communication**. *The Bell System Technical Journal* 1948, **27**:379 – 423, 623 – 656.

57. **ENIGMA**[http://bioinformatics.psb.ugent.be/ENIGMA/main.htm].

58. Shamir R, Maron-Katz A, Tanay A, Linhart C, Steinfeld I, Sharan R, Shiloh Y, Elkon R: **EXPANDER–an integrative program suite for microarray data analysis**. *BMC Bioinformatics* 2005, **6**:232.

59. Barkow S, Bleuler S, Prelic A, Zimmermann P, Zitzler E: **BicAT: a biclustering analysis toolbox**. *Bioinformatics* 2006, **22**:1282–1283.

Figures

Figure 1 - Global methodology overview

Global overview of the methodology. To the right is a figure of module 28 learned on the Rosetta dataset [4], a module enriched in mating-related genes. See Figure 4 for interpretation guidelines.

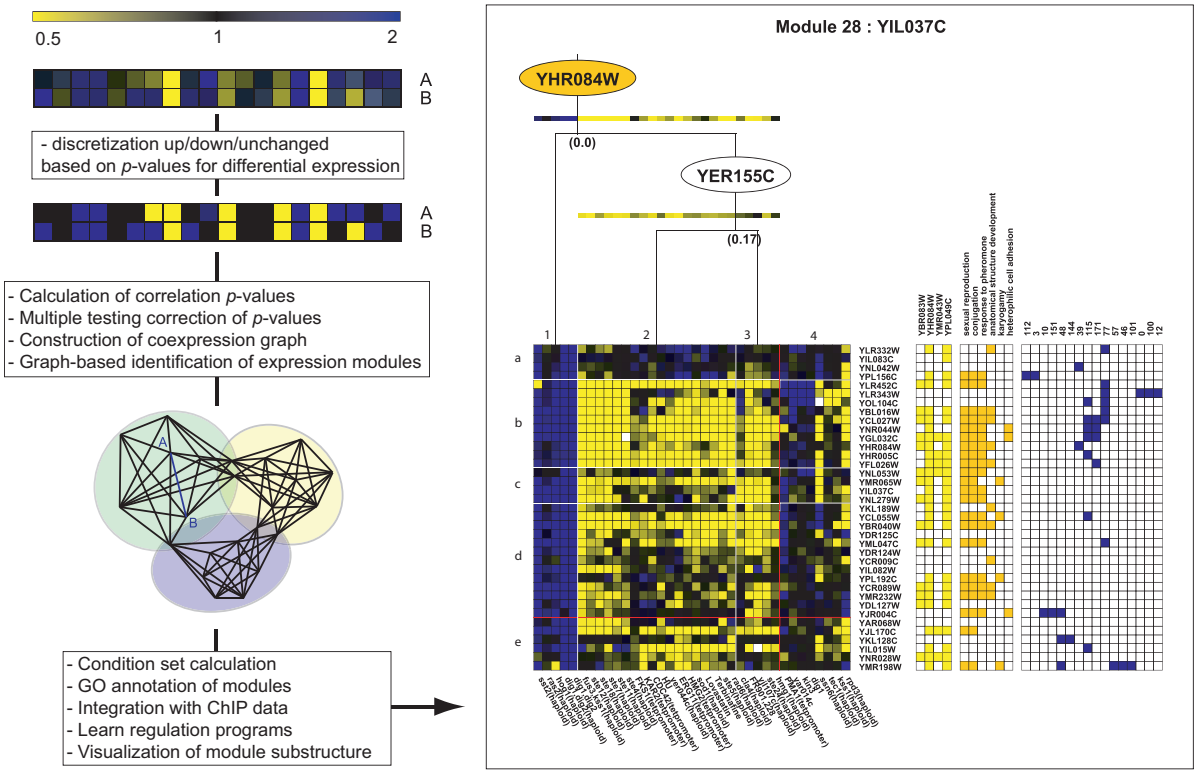

Figure 1: Global overview of the methodology. To the right is a figure of module 28 learned on the Rosetta dataset [4], a module enriched in mating-related genes. See Figure 4 for interpretation guidelines.

Figure 2 - Performance on artificial data

Performance of ENIGMA versus other coexpression measures and biclustering methods on (A) modular and (B) non-modular artificial gene expression data (1000 genes on 100 conditions, (A) 20 overlapping biclusters of varying size and (B) 500 partial expression correlations between individual genes).

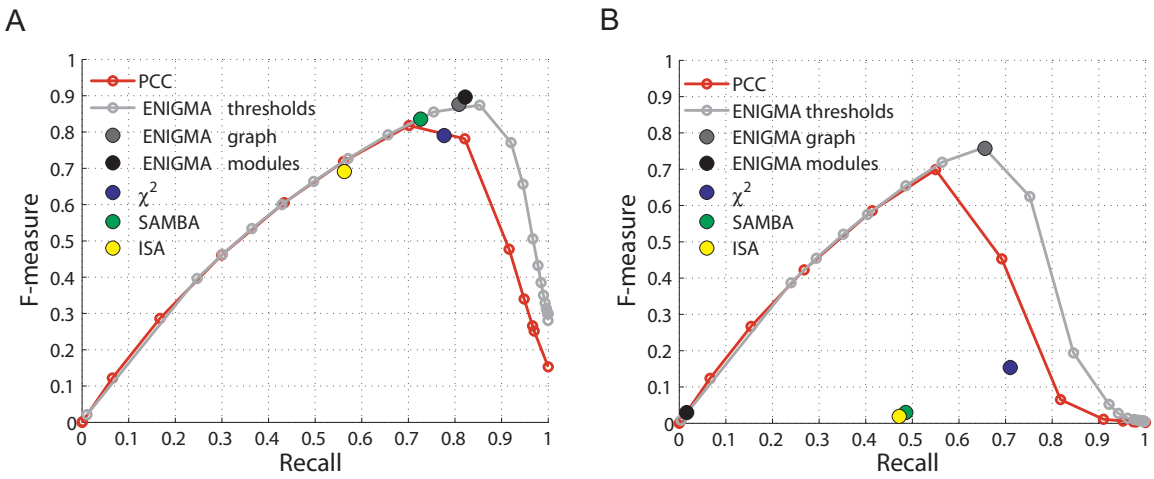

Figure 2: Performance of ENIGMA versus other coexpression measures and biclustering methods on (A) modular and (B) non-modular artificial gene expression data (1000 genes on 100 conditions, (A) 20 overlapping biclusters of varying size and (B) 500 partial expression correlations between individual genes)

Figure 3 - Degree distribution

(A) Semilog rank-degree plot for the ENIGMA network inferred from the Rosetta data [4]. (B) Plot of the clustering coefficient of a node's neighborhood as a function of the node degree k. The data points are colored according to the number of modules to which the corresponding gene is assigned.

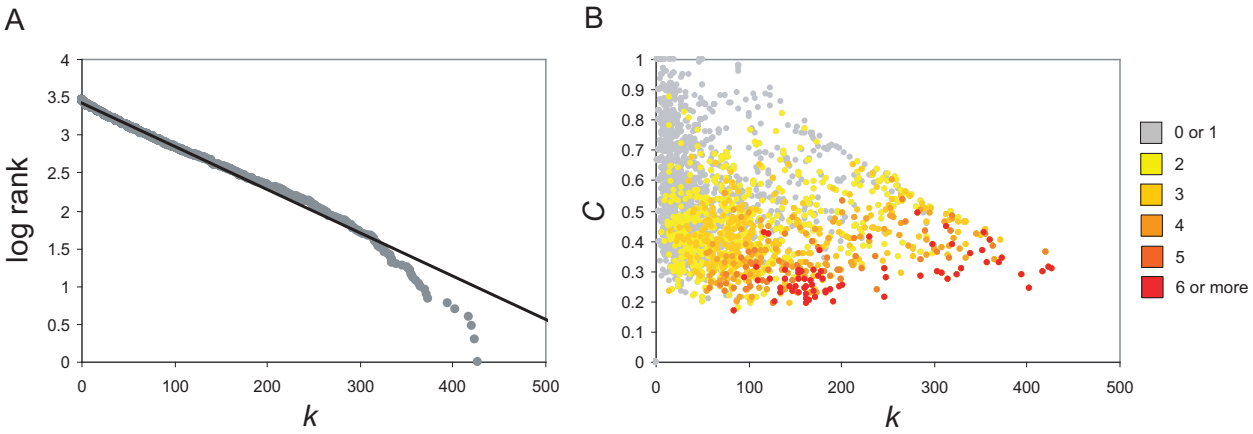

Figure 3: (A) Semilog rank-degree plot for the ENIGMA network inferred from the Rosetta data [4]. (B) Plot of the clustering coefficient of a node's neighborhood as a function of the node degree k. The data points are colored according to the number of modules to which the corresponding gene is assigned.

Figure 4 - Mating module 77

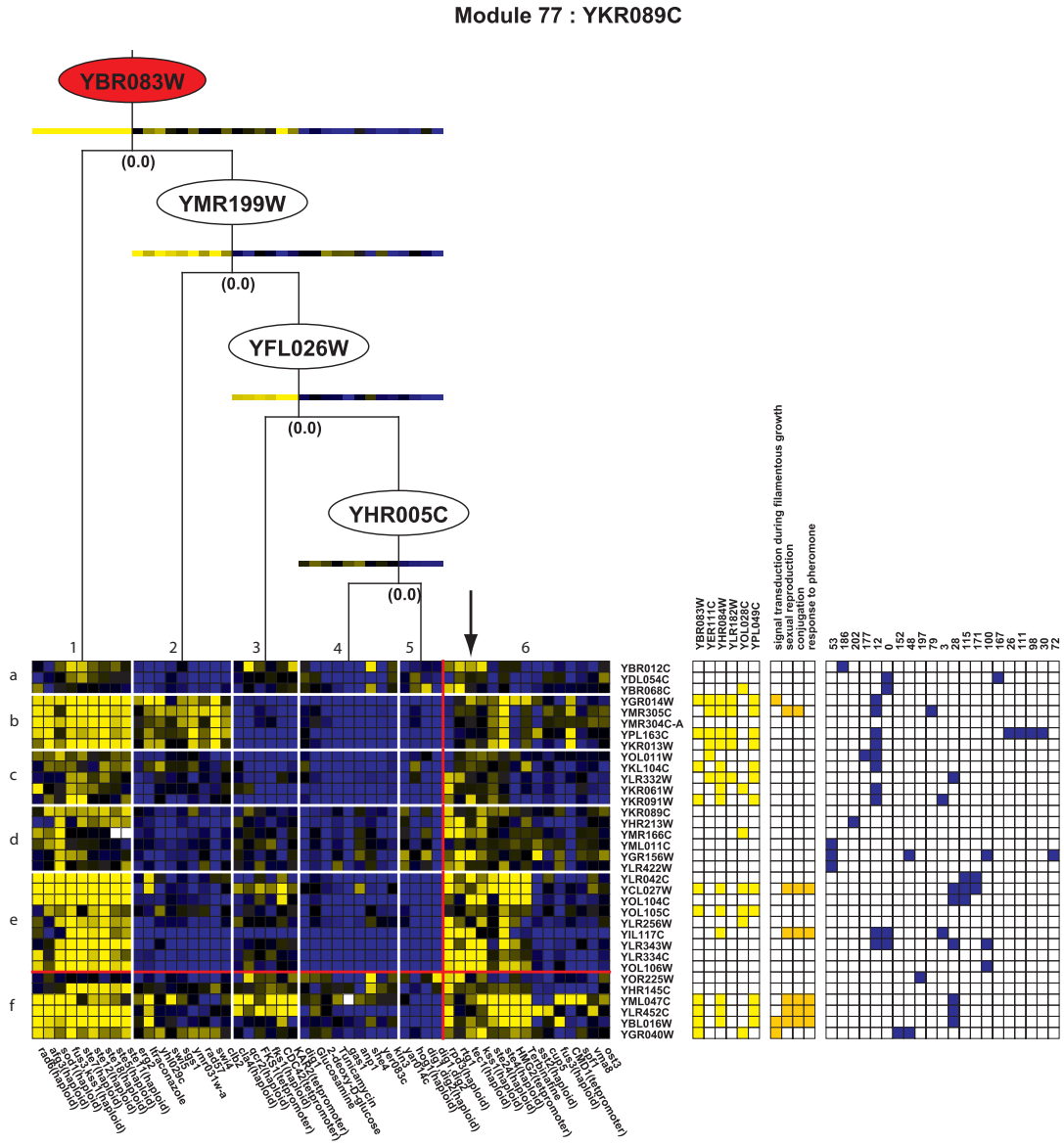

Figure 4: Subset of the expression matrix encompassing the genes and conditions that define module 77, a module enriched in mating-related genes. The colors of individual spots reflect the expression ratio (experiment vs. control, blue = upregulated, yellow = downregulated, white = missing value). The matrix is split in leafs in both dimensions based on average linkage clustering (see Methods). Leafs of size < 3 are grouped in a single leaf beyond the red line (rightmost leaf and bottom leaf). The rightmost leaf is not included in the regulation program (tree on top). Regulators that belong to the module are colored yellow, while regulators of which the binding sites are overrepresented in the module are colored red. In case these two features are combined, the regulator is colored orange. The numbers between parentheses represent the conditional entropy of the regulators. To the right of the expression matrix are other matrices depicting the genes' membership of significant GO categories for the module, the presence of overrepresented binding sites, and membership of other modules, in that order.

Figure 5 - Halo test for α-factor based growth inhibition.

Yeast strains (OD_{600}=1) were plated on YPD plates and 1000 pmol of α-factor was spotted. The pictures are taken after 48 hours of incubation at 30 °C. Strains: A: Wild type BY4741 (*MAT***a** *his*3Δ1 *leu*2Δ0 *met*15Δ0 *ura*3Δ0), B: *sst*2Δ, C: *ylr*343*w*Δ.

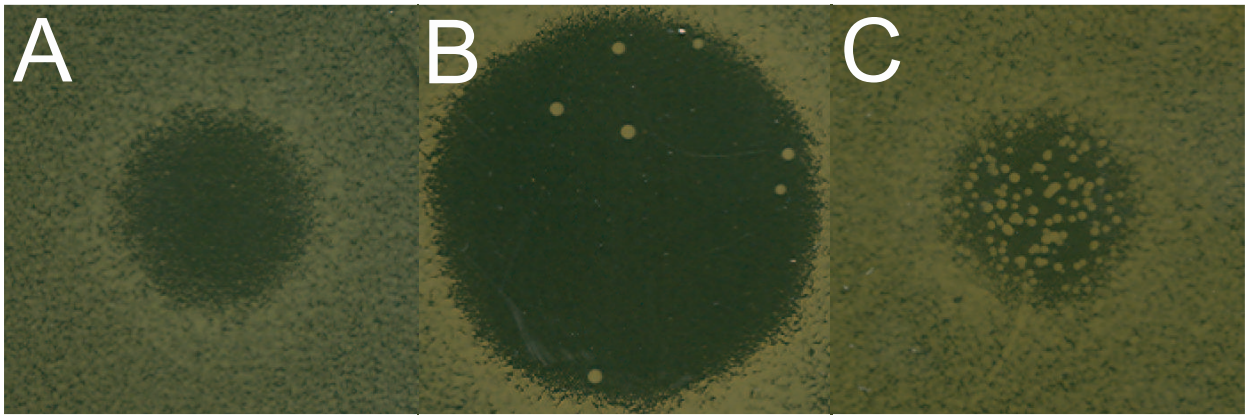

Figure 5: Halo test for α-factor based growth inhibition. Yeast strains (OD_{600}=1) were plated on YPD plates and 1000 pmol of α-factor was spotted. The pictures are taken after 48 hours of incubation at 30 °C. Strains: A: Wild type BY4741 (*MAT***a** *his*3Δ1 *leu*2Δ0 *met*15Δ0 *ura*3Δ0), B: *sst*2Δ, C: *ylr*343*w*Δ.

Additional Files

Supplementary methods and figures

The supplementary pdf file accompanying this article contains the Supplementary Methods, Tables S1-S6 and Figures S1-S3. Additional supplementary material can be downloaded from [57].