Fungal Genetics and Biology 47 (2010) 539–550

Contents lists available at ScienceDirect

Fungal Genetics and Biology

journal homepage: www.elsevier.com/locate/yfgbi

# Identification of modules in *Aspergillus niger* by gene co-expression network analysis

Robert A. van den Berg [b,c,*,1], Machtelt Braaksma [b,1], Douwe van der Veen [a,1], Mariët J. van der Werf [b], Peter J. Punt [b], John van der Oost [a], Leo H. de Graaff [a]

[a] *Laboratory of Microbiology, Fungal Genomics Group, Wageningen University and Research Centre, Dreijenplein 10, 6703 HB Wageningen, The Netherlands*
[b] *TNO Quality of Life, P.O. Box 360, 3700 AJ Zeist, The Netherlands*
[c] *SymBioSys, Katholieke Universiteit Leuven, Tiensestraat 102, 3000 Leuven, Belgium*

ARTICLE INFO

*Article history:*
Received 13 October 2009
Accepted 13 March 2010
Available online 27 March 2010

*Keywords:*
*Aspergillus niger*
Gene co-expression network
Co-expression
Transcriptional analysis
RNA profiling
Transcription factor binding site

ABSTRACT

The fungus *Aspergillus niger* has been studied in considerable detail with respect to various industrial applications. Although its central metabolic pathways are established relatively well, the mechanisms that control the adaptation of its metabolism are understood rather poorly. In this study, clustering of co-expressed genes has been performed on the basis of DNA microarray data sets from two experimental approaches. In one approach, low amounts of inducer caused a relatively mild perturbation, while in the other approach the imposed environmental conditions including carbon source starvation caused severe perturbed stress. A set of conserved genes was used to construct gene co-expression networks for both the individual and combined data sets. Comparative analysis revealed the existence of modules, some of which are present in all three networks. In addition, experimental condition-specific modules were identified. Module-derived consensus expression profiles enabled the integration of all protein-coding *A. niger* genes to the co-expression analysis, including hypothetical and poorly conserved genes. Conserved sequence motifs were detected in the upstream region of genes that cluster in some modules, e.g., the binding site for the amino acid metabolism-related transcription factor CpcA as well as for the fatty acid metabolism-related transcription factors, FarA and FarB. Moreover, not previously described putative transcription factor binding sites were discovered for two modules: the motif 5′-CGACAA is overrepresented in the module containing genes encoding cytosolic ribosomal proteins, while the motif 5′-GGCCGCG is overrepresented in genes related to 'gene expression', such as RNA helicases and translation initiation factors.

© 2010 Elsevier Inc. All rights reserved.

## 1. Introduction

Genome-wide gene expression levels as generated by DNA microarray technology give insight into the behavior of individual genes at the cellular level. Such expression levels can be considered as a reflection of the physiological state of an organism and can be used to reveal details of metabolic regulation. The first fungal microarray studies were reported for the model organism *Saccharomyces cerevisiae* (DeRisi et al., 1997; Lashkari et al., 1997), followed by application of this technology for over 20 filamentous fungi including *Aspergillus niger* (Breakspear and Momany, 2007). *A. niger* is of industrial importance as it is the major production organism for citric acid world-wide (Magnuson and Lasure, 2004) and an efficient producer of both homologous and heterologous proteins (Pel et al., 2007; Punt et al., 2002). So far, *A. niger* transcriptome studies have been used to characterize polysaccharide-degrading enzyme systems (Andersen et al., 2008; Jørgensen et al., 2009; Martens-Uzunova and Schaap, 2008; van der Veen et al., 2009; Yuan et al., 2008a,b), to describe spatial colony development (Levin et al., 2007), and to study the *A. niger* response towards reductive stress (Guillemette et al., 2007) or cell wall damage (Meyer et al., 2007).

Most DNA microarray studies, including all above-mentioned *A. niger* studies, focus on differential gene expression between few experimental conditions. However, solely an observation of fold changes of expression of individual genes does not explain how biological processes work together to achieve the cell's objectives. Additional information regarding the activation and co-operation of biological processes can be obtained by comparing gene expression profiles over a range of conditions. For example, genes that encode subunits of a protein complex may have a consistently similar change of expression levels over many conditions. The similar expression of two or more genes over a range of conditions is referred to hereafter as gene co-expression.

Featherstone and Broadie deduced from a *S. cerevisiae* gene co-expression study (Featherstone and Broadie, 2002) that specific

* Corresponding author. Fax: +32 16 325 993.
*E-mail address:* robert.vandenberg@psy.kuleuven.be (R.A. van den Berg).
[1] These authors contributed equally to this study.

1087-1845/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.fgb.2010.03.005

sets of genes interact extensively at the level of gene expression, and thus can be described in terms of an interconnected network. Such gene co-expression networks can provide a large-scale, global view of the transcriptional response of an organism.

A comparison of gene co-expression networks constructed from DNA microarray data of the evolutionary distinct organisms *S. cerevisiae*, *Homo sapiens*, *Escherichia coli*, *Arabidopsis thaliana*, *Caenorhabditis elegans*, and *Drosophila melanogaster* indicated that these networks share common structural, or topological, properties (Bergmann et al., 2004). For example, the observed gene co-expression networks consist of co-expressed groups of genes, termed clusters or modules, which are associated with the same cellular function. However, while modules of genes involved in similar cellular functions were identified in all species analyzed (e.g., "glycolysis", "proteasome"), this study also indicated that the higher-order relations between modules differ significantly between the organisms (Bergmann et al., 2004). For example, the average gene expression profiles for genes in the "secreted protein" and "proteasome" modules correlate positively in yeast and *A. thaliana*, negatively in *D. melanogaster*, and do not appear to correlate significantly in *H. sapiens*.

Before gene co-expression networks can be generated and analyzed, two problems need to be solved. First, genomes often encode thousands of proteins. Construction of a co-expression network of this amount of genes will in most cases results in a network that is difficult to interpret due the number of genes and their many connections. Second, the physiological role of a large proportion of proteins is unknown, or at best poorly understood (Hughes et al., 2004). This lack of understanding further hampers the interpretation of any network generated. Different strategies to circumvent these problems can be employed in gene co-expression network analyses. The first strategy is to limit the analysis of co-expression networks to descriptive parameters only, such as the number of connections that a gene has with other genes (connectivity), e.g., see Jordan et al. (2008), van Noort et al. (2004). While this approach provides an understanding of the network at a higher abstraction level, such knowledge cannot be converted easily into understanding the actual underlying biological processes. The second strategy is to investigate only a subset of genes that is relevant for a certain research interest, e.g., see Bergmann et al. (2004), Lee et al. (2004), Neretti et al. (2007). For example, Bergmann and coworkers selected genes participating in eight well-defined *S. cerevisiae* biological processes, and examined co-expression of their orthologous genes in five other organisms (Bergmann et al., 2004). In a variation of this strategy, co-expression between all genes is calculated but the analysis is focused on a part of the network that is of particular biological interest, e.g., a certain oncogene (Basso et al., 2005). These approaches, however, are biased towards already known biological processes. In addition, the selected biological processes might operate distinctly in organisms other than *S. cerevisiae*. To reduce bias of using only known biological processes, yet another approach limits the analysis to genes conserved in different species, for instance, in *S. cerevisiae*, *D. melanogaster* and *C. elegans* (Daub and Sonnhammer, 2008).

In this study, we extend the latter approach to the analysis of gene co-expression networks. For the construction of a gene co-expression network of *A. niger*, a subset of genes is selected that is based on evolutionary conservation of the proteins they encode among 19 different fungal species. Even when no defined function can be assigned to such proteins, their evolutionary conservation suggests a biological role. From expression data of these conserved protein-encoding genes, a gene co-expression network is generated. Subsequently, the topology of this network is used to extend the analysis to less conserved genes excluded from the initial analysis. This approach is followed for the analysis of two *A. niger* DNA microarray data sets cultivated under distinct experimental conditions.

## 2. Materials and methods

### 2.1. Culturing

Mildly perturbed conditions: *A. niger* 872.11 (*ΔargB pyrA6 prtF28 goxC17 cspA1*) is derived from CBS 120.49. All media were based on Pontecorvo's minimal medium (pMM) (Pontecorvo et al., 1953), contained 100 mM sorbitol as carbon source and were supplemented with uridine and arginine. Glass 2.5-l fermentors (Applikon) with 2.2 l of pMM were kept at a constant temperature of 30 ± 0.5 °C while fermentor headplates were kept at 8 °C. A total of $1.0 \times 10^6$ of spores per ml were added to a fermentor. During germination, each fermentor was aerated through the headspace (50 l/h) and stirred at 300 rpm. When dissolved oxygen tension levels dropped below 60% for over 5 min, the stirrer speed was set to 750 rpm and aeration was switched to sparger inlet. In one experiment, fermentors were induced with either 0.1 mM sorbitol or D-xylose at $T = 14$ h as previously described (van der Veen et al., 2009). In a second experiment, fermentors were induced with 1 mM of various oils at $T = 14$ h and samples were taken before induction and up to 2 h after induction (Table 1).

Strongly perturbed conditions: *A. niger* N402 (*cspA1*) (van Hartingsveldt et al., 1987) is derived from CBS 120.49. All media were based on Bennett's minimal medium (bMM) (Bennett and Lasure, 1991). Both shake flask and fermentor cultures were grown in bMM medium at a constant temperature of 30 ± 0.5 °C and with differing combinations of carbon source, nitrogen source and concentration, and pH of the medium (Table 1) (Braaksma et al., 2009). Fermentor inoculum was pre-cultured in baffled 500 ml Erlenmeyer flasks containing 100 ml bMM (pH 6.5) supplemented with the carbon source and nitrogen source concentrations corresponding to fermentor conditions. These flasks were inoculated with $10^6$ spores per liter and incubated in a rotary shaker at 125 rpm until approximately half of the available carbon source was consumed. Cultivations were carried out in 6.6-l fermentors (New Brunswick Scientific) with 5.0 l of bMM. The fermentors were inoculated with 4% (weight/vol) pre-culture. To prevent foaming, 1% (vol/vol) Struktol J-673 antifoam was added to the medium and additional antifoam was added during cultivation when necessary. Each fermentor was sparged with 75 l/h of air with the stirrer speed set at 400 rpm at the start of the cultivation. When dissolved oxygen tension levels dropped below 20%, the stirrer speed was automatically increased to maintain oxygen tension at 20% or until the maximum of 1000 rpm was reached. The pH was controlled by automatic addition of 8 M KOH or 1.5 M $H3PO_4$.

### 2.2. RNA isolation

Culture samples from mildly perturbed conditions were filtrated and biomass was snap-frozen into liquid nitrogen and stored at −80 °C. Culture samples from strongly perturbed conditions were quenched immediately in methanol at −45 °C as described previously (Pieterse et al., 2006) and centrifuged at −20 °C to remove supernatant. Biomass was frozen into liquid nitrogen and stored at −80 °C. A Trizol–chloroform extraction preceded total RNA extraction with RNeasy mini columns (Qiagen) according to the manufacturer's protocol for yeast. Concentration of total RNA was determined by spectrophotometry. RNA integrity was assessed on an Experion system (Biorad) for samples from mildly perturbed conditions by visual inspection of the electropherograms. Graphs depicting RNA integrity categories were used as visual aids (Schroeder et al., 2006). Electropherograms approximating an RNA integrity number of 8 or lower or with a 28S/18S ratio below 1.8 were discarded. For samples from strongly per-

**Table 1**
Fermentation conditions for DNA microarray samples used in this study.

| Sample name | pH | Carbon source | Nitrogen source | Initial concentration of nitrogen source (mM) | Inducer compound (mM) | Sampling time (h) | Growth phase[a] |
|---|---|---|---|---|---|---|---|
| *Mildly perturbed conditions* | | | | | | | |
| 29[b] | 3.5 | Sorbitol | $NaNO_3$ | 70.5 | D-Xylose (0.1 mM) | 14 | E |
| 44[b] | 3.5 | Sorbitol | $NaNO_3$ | 70.5 | D-Xylose (0.1 mM) | 14 | E |
| 52[c] | 3.5 | Sorbitol | $NaNO_3$ | 70.5 | Sorbitol (0.1 mM) | 14 | E |
| 76[b] | 3.5 | Sorbitol | $NaNO_3$ | 70.5 | D-Xylose (0.1 mM) | 14 | E |
| 86-1[b] | 3.5 | Sorbitol | $NaNO_3$ | 70.5 | D-Xylose (0.1 mM) | 14 | E |
| 96[c] | 3.5 | Sorbitol | $NaNO_3$ | 70.5 | Sorbitol (0.1 mM) | 14 | E |
| Triton-0[c] | 3.5 | Sorbitol | $NaNO_3$ | 70.5 | – | 14 | E |
| Triton-0.5 | 3.5 | Sorbitol | $NaNO_3$ | 70.5 | 0.002% Triton-X-100 | 14.5 | E |
| Triton-1 | 3.5 | Sorbitol | $NaNO_3$ | 70.5 | 0.002% Triton-X-100 | 15 | E |
| Triton-2 | 3.5 | Sorbitol | $NaNO_3$ | 70.5 | 0.002% Triton-X-100 | 16 | E |
| Olive-0[c] | 3.5 | Sorbitol | $NaNO_3$ | 70.5 | – | 14 | E |
| Olive-0.5 | 3.5 | Sorbitol | $NaNO_3$ | 70.5 | Olive oil (1 mM) | 14.5 | E |
| Olive-1 | 3.5 | Sorbitol | $NaNO_3$ | 70.5 | Olive oil (1 mM) | 15 | E |
| Olive-2 | 3.5 | Sorbitol | $NaNO_3$ | 70.5 | Olive oil (1 mM) | 16 | E |
| DGDG-0[c] | 3.5 | Sorbitol | $NaNO_3$ | 70.5 | – | 14 | E |
| DGDG-0.5 | 3.5 | Sorbitol | $NaNO_3$ | 70.5 | DGDG[d] oil (1 mM) | 14.5 | E |
| DGDG-1 | 3.5 | Sorbitol | $NaNO_3$ | 70.5 | DGDG[d] oil (1 mM) | 15 | E |
| DGDG-2 | 3.5 | Sorbitol | $NaNO_3$ | 70.5 | DGDG[d] oil (1 mM) | 16 | E |
| Wheat-0[c] | 3.5 | Sorbitol | $NaNO_3$ | 70.5 | – | 14 | E |
| Wheat-0.5 | 3.5 | Sorbitol | $NaNO_3$ | 70.5 | Wheat oil (1 mM) | 14.5 | E |
| Wheat-1 | 3.5 | Sorbitol | $NaNO_3$ | 70.5 | Wheat oil (1 mM) | 15 | E |
| Wheat-2 | 3.5 | Sorbitol | $NaNO_3$ | 70.5 | Wheat oil (1 mM) | 16 | E |
| *Strongly perturbed conditions* | | | | | | | |
| 4G $4NO_3$-1 | 4 | Glucose | $NaNO_3$ | 282.4 | – | 66 | LS |
| 4G $4NO_3$-2 | 4 | Glucose | $NaNO_3$ | 282.4 | – | 96 | LS |
| 4G $8NO_3$ | 4 | Glucose | $NaNO_3$ | 564.8 | – | 53 | LE |
| 4G $4NH_4$ | 4 | Glucose | $NH_4Cl$ | 282.4 | – | 57 | LS |
| 4G $8NH_4$-1a[e] | 4 | Glucose | $NH_4Cl$ | 564.8 | – | 36 | LE |
| 4G $8NH_4$-2a[e] | 4 | Glucose | $NH_4Cl$ | 564.8 | – | 36 | LE |
| 4G $8NH_4$-1b[f] | 4 | Glucose | $NH_4Cl$ | 564.8 | – | 60 | LS |
| 4G $8NH_4$-2b[f] | 4 | Glucose | $NH_4Cl$ | 564.8 | – | 60 | LS |
| 4X $4NO_3$ | 4 | Xylose | $NaNO_3$ | 282.4 | – | 66 | LS |
| 4X $8NO_3$ | 4 | Xylose | $NaNO_3$ | 564.8 | – | 91 | LS |
| 4X $4NH_4$ | 4 | Xylose | $NH_4Cl$ | 282.4 | – | 60 | LS |
| 4X $8NH_4$ | 4 | Xylose | $NH_4Cl$ | 564.8 | – | 66 | LS |
| 5G $4NO_3$ | 5 | Glucose | $NaNO_3$ | 282.4 | – | 48 | S |
| 5G $8NO_3$ | 5 | Glucose | $NaNO_3$ | 564.8 | – | 49 | LE |
| 5G $4NH_4$ | 5 | Glucose | $NH_4Cl$ | 282.4 | – | 35.25 | LE |
| 5G $8NH_4$ | 5 | Glucose | $NH_4Cl$ | 564.8 | – | 35 | LE |
| 5X $4NO_3$ | 5 | Xylose | $NaNO_3$ | 282.4 | – | 93.5 | S |
| 5X $8NO_3$ | 5 | Xylose | $NaNO_3$ | 564.8 | – | 112 | S |
| 5X $4NH_4$ | 5 | Xylose | $NH_4Cl$ | 282.4 | – | 41 | LE |
| 5X $8NH_4$ | 5 | Xylose | $NH_4Cl$ | 564.8 | – | 47.5 | LE |

[a]LS, late stationary growth, over 10 h of carbon depletion; S, stationary phase, carbon source will become depleted within 1 h; LE, late exponential growth phase; E, exponential growth phase.
[b,c]These DNA microarray samples are independent biological replicates, i.e., all [b] or [c] labeled samples are grown in different fermentor vessels but with identical media composition and are sampled at identical time point, even though they are part of different experiments and grown at different dates.
[d]DGDG, digalactoside–diglyceride.
[e,f]These DNA microarray samples are technical replicates. The thus labeled microarray samples are derived from one fermentor sample for which the extracted RNA was processed further in duplicate.

turbed conditions, RNA integrity was assessed on agarose gel, by its A260/A280 ratio, and on an Agilent 2100 Bioanalyzer.

### 2.3. Microarray processing

cDNA and cRNA synthesis and labeling, and array hybridization were performed following the Affymetrix users' manual (Affymetrix, 2004) using the One-cycle Target Labeling and Control Reagents Kit to synthesize 15 μg of cRNA from 5 μg of total RNA as template material for mildly perturbed conditions samples. For strongly perturbed conditions samples, the Bioarray High Yield RNA Transcript Labeling kit (Enzo) was used to synthesize at least 30 μg of cRNA from 10 μg of total RNA as template material. Fifteen micro gram of fragmented and labeled cRNA was hybridized to custom-made *A. niger* arrays at 45 °C for 16 h. Washing and staining was done using the Hybridization, Wash and Stain Kit (Affymetrix) using a GeneChip FS-450 Fluidics station and an Agilent G2500A Gene Array scanner. Scanned images were converted into .CEL files using MicroArray Suite software (Affymetrix).

### 2.4. Microarray data accession number

Raw and RMA-normalized array data were deposited at the NCBI Gene Expression Omnibus database (Edgar et al., 2002) under series entries GSE11405 and GSE14285 for the mildly perturbed conditions and under series entry GSE17329 for the strongly perturbed conditions.

### 2.5. Data preprocessing

DNA microarrays were normalized using Affymetrix' MicroArray Suite Software version 5 (MAS5) with the target value set at 100 (Affymetrix, 2001). MAS5 was preferred over another often-used normalization strategy, Robust Multichip Average, or

RMA (Irizarry et al., 2003), as MAS5 normalization is calculated over each individual array alone, thus excluding a potential influence of normalization to the correlation structure of the whole data set. Some probe sets have a signal above background in only few of the experimental conditions examined. Since their limited number of observations hampers the calculation of reliable correlations, for each of the three data sets, probe sets flagged "absent" in more than 80% of the microarrays per data set were discarded from that specific data set. Under mildly perturbed conditions, 7955 probe sets (55%) were discarded while under strongly perturbed conditions 5084 probe sets (35%) were discarded. Probe set values were normalized per microarray by dividing each probe set value by the mean signal over the whole microarray. Signals that are close to the detection limit are more influenced by random noise signal and thus yield varying values on different arrays. This variation hampers the calculation of reliable correlations as well and therefore the mean "absent" call value divided by two was taken as uniform "lowest in the data set" value. The remaining signals flagged as "absent" (those not included in the removed probe sets) as well as all other probe set signals with a value below this lowest uniform value were replaced by this uniform value. Probe sets were not filtered for a certain fold change threshold as the magnitude of fold change is not necessarily a measure for biological relevance (van den Berg et al., 2006). In addition, the correlation analysis of the data has its own selection criterion, namely the $\rho$ threshold value. No artifacts or outliers in the signals distribution for the microarrays within the data sets were observed, and the per-array signal distributions were similar.

### 2.6. Correlation analysis

The correlations between genes were determined by the Spearman correlation coefficient $\rho$. The correlation coefficient ranges from 0 (no correlation) till either 1 (full positive correlation between expression levels) or $-1$ (full negative correlation between expression levels, i.e., perfect antagonists). The Spearman $\rho$ is a non-parametric correlation measure based on the rank of the expression values instead of the detected values, and is robust against outliers and mild non-linear behavior (Zar, 1996). The Spearman correlation measure was recently shown to be slightly more robust in the analysis of gene co-expression compared to other correlation methods including the Pearson correlation, Euclidian distance, and the mutual information measures (Daub and Sonnhammer, 2008). The $p$-value for the Spearman correlation for the mildly perturbed conditions data set, which is the smallest data set, was $4.12 \times 10^{-6}$ for the lowest cut-off value for $\rho$ ($\rho = 0.85$).

Correlation networks were drawn in Cytoscape (Shannon et al., 2003). Initial networks were constructed using the "spring embedded" layout function, and individual genes within the resulting networks were manually re-positioned for improved interpretation (Fig. 1). Manual arrangement was based on the $\rho$ values associated with each gene pair, by the sign of $\rho$, and by the number of connections per gene that were visible at a certain $\rho$ threshold value. In this iterative process, while switching back and forth between $\rho$ threshold values, only genes visible at a certain $\rho$ threshold value were relocated. The length of the connecting lines does not represent the degree of correlation between the two connected genes. The term "module" is used for a group of genes that has a core of interconnected genes at high $\rho$ threshold values, and to which group genes appear to attach preferentially upon lowering of the $\rho$ threshold. A coloring scheme was deduced from the combined data network, where at $\rho$ 0.95, eight modules can be identified. These modules were labeled A–H, and genes within each module were assigned a color to assist localization of these genes within the networks.

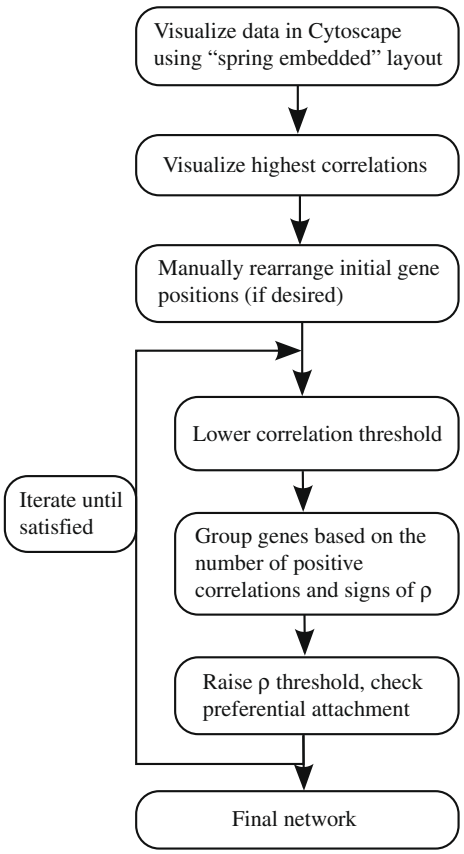


**Fig. 1.** Procedure for construction of gene co-expression networks. Schematic representation of the process to construct gene co-expression networks.

### 2.7. Validation

The influence of individual microarrays on the correlation analysis was evaluated by a "leave $N$ samples out" validation. This validation procedure tests whether the strongest co-expressed gene pairs also remain the strongest co-expressed gene pairs in case DNA microarray samples are removed from the complete set of DNA microarrays. The following procedure was used: (i) random selection of two (for each single data set) or five (for the combined data sets) microarrays and removal from the data set; (ii) calculation of new correlation coefficients; (iii) continuation until all microarrays were excluded whilst ensuring that previously removed arrays were not removed again; (iv) repetition of this procedure for 20 times; (v) calculation of the mean correlation coefficient per gene pair; (vi) selection of gene pairs matching the 2.5 and 97.5 percentiles of the mean correlation coefficient; (vii) comparison of the selected genes with the genes present in the correlation networks. For the mildly perturbed data set, 80% of its strongest correlating gene pairs fell within the 2.5th and 97.5th percentiles in this validation procedure. All of the actually observed strongest correlating gene pairs for both the strongly perturbed data set and the combined data set belonged to the 5% strongest correlating gene pairs of the validation.

### 2.8. Consensus expression profile analysis

The "combined data set" network was used for consensus expression profile analysis. Genes with three or more connections with other genes within a module were selected. Their expression profiles were converted in a rank order analogous to the rank order used for the Spearman correlation. Following (Horvath and Dong,

2008), the consensus expression profile was defined as the profile obtained from the first principal component score vector of a principal component analysis (PCA) (Jackson, 1991; Joliffe, 2002) of the converted expression profiles. Subsequently, the correlation between the obtained consensus expression profile and the expression profile of all measured genes was calculated.

All calculations were performed on a Pentium 4 personal computer with 1 GB internal memory using Matlab (The Mathworks), the Statistics Toolbox (The Mathworks), and homemade scripts.

### 2.9. Promoter analysis

Promoter analysis was done in GeneSpring, version 7.2 (Agilent) using the "find potential regulatory sequences" tool. The promoter region from 10 to 800 bases upstream of a gene was searched for oligonucleotides ranging from 5 to 10 bases, with at maximum one single point discrepancy allowed, and correcting for local nucleotide density. The likelihood of random occurrence of identified sequences was compared relative to the upstream region of all 14,165 genes in the *A. niger* genome.

### 2.10. KEGG pathway analysis

For the combined data set network, all genes present within each module A–H at $\rho$ 0.90 were exported to a tab-delimited file and imported into the KegArray program (Wheelock et al., 2009). The "PathwayMap" tool was applied to extract *A. niger* genes linked to a KEGG pathway from the KEGG database (Kanehisa and Goto, 2000) by using the "Ang" organism abbreviation.

### 2.11. Gene ontology

The genomes of the fungi *Aspergillus niger*, *Aspergillus fumigatus*, *Aspergillus nidulans*, *Penicillium chrysogenum*, *Neurospora crassa*, *Magnaporthe grisea*, *Stagonospora nodorum*, *Ustilago maydis*, and *Trichoderma reesei*, and the yeasts *Ashbya gossypii*, *Candida albicans*, *Candida neoformans*, *Debaromyces hansenii*, *Giberella zeae*, *Kluyveromyces lactis*, *Phanerochaete chrysosporium*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, and *Yarrowia lipolytica* were used to construct an in-house built database of orthologous protein sequences (S. Basmagi and P. Schaap, unpublished data). Protein sequences were placed into an orthology cluster based on bi-directional first-hit BLAST alignment of protein sequences of other species. Conserved proteins were defined as having an ortholog in at minimum 15 of 19 species; the absence of a gene in a species while present in over 15 other genomes is due mostly to mis-annotation or incorrect intron-predictions in our experience. A total of 2749 genes fulfilled this criterion. All 455 genes that have a *S. cerevisiae* ortholog but did not meet the criterion were added because of the extensive body of knowledge that is available for genes of this model organism. On average, these latter 455 genes have an ortholog in 11 species. Gene ontology terms, available from *S. cerevisiae* orthologous genes per module, were browsed at the Saccharomyces Genome Database website (Hong et al., 2008).

## 3. Results

Gene co-expression networks based on a subset of genes have been generated and some of their generic properties will be described. Two series of *A. niger* DNA microarray data sets were used, that were analyzed as separate data sets as well as in combination. Groups of co-expressed genes, termed modules, were observed within the visualized networks. Next, the biological properties of these modules were analyzed using the combined data set network as our reference network. Subsequently, the module structure found for the combined data set was compared with the co-expression networks based on the two individual data sets. We conclude our results by extending our network analysis from a subset of genes to all genes for which a probe set is available on the *A. niger* DNA microarray.

### 3.1. Construction of gene co-expression networks

It is expected that gene co-expression networks will be influenced by the experimental setups of the microarray data sets used to generate these networks. Therefore, a total of 42 microarrays that originated from *A. niger* strains grown in batch fermentation under two different experimental setups were used (Table 1). The microarrays used in this study were obtained from different experimental perspectives (e.g., investigate D-xylose metabolism (van der Veen et al., 2009) or lipid metabolism (van der Veen, 2009), or extracellular protease activity (Braaksma et al., 2009)), but were selected on the basis of covering diversity, especially in relation to cell culture perturbations. We expect that these perturbations will have a larger impact on the physiology than the differences between the closely related strains used. It was decided to not include further *A. niger* microarray data that are available in public repositories, as at the time of this study only shake flask-cultivated experiments were deposited. Shake flask cultivation generally introduces more culture heterogeneity as pH and the transfer of oxygen, nutrients, and heat are not controlled (van der Veen et al., 2009).

Twenty microarrays were obtained from fungal cells growing exponentially with 100 mM sorbitol as primary carbon source, to which either 0.1 mM sorbitol or D-xylose or 1 mM vegetable oils were applied. Under these growth conditions, the cells do not experience any nutrient limitation and grow at maximum growth rate. We reasoned that the applied pulses would provoke only a minor disturbance in global gene expression levels, and hence labeled these cultivation conditions "mildly perturbed". In contrast, the other 22 microarrays were obtained from fungal cells growing in much more perturbed conditions at the time of sampling (e.g., carbon source deprivation). These cells were expected to yield more drastic changes in gene expression levels, and therefore these conditions were labeled "strongly perturbed". From a biological point of view, the differing experimental conditions are expected to yield both condition-specific gene expression (e.g., induction with D-xylose leads to increased expression of the xylan–metabolic system) as well as expression of genes involved in general metabolic processes required for both conditions (e.g., growth in Minimal Medium broth requires de novo amino acid biosynthesis under both conditions).

Construction of a co-expression network using the data of the over 14 thousand predicted *A. niger* genes will result in a network that is difficult to interpret due to the many resulting gene–gene interactions. Therefore, a subset of genes was selected according to their evolutionary conservation among fungal species, and their signal value. The evolutionary conserved subset consisted of only those protein-encoding genes for which an ortholog is present in 15 or more of the 19 fungal species analyzed, or for which a *S. cerevisiae* ortholog is identified (see Section 2). Even in case no clear biological function has been assigned to such protein, its evolutionary conservation suggested a functional role. In addition, a present signal for a gene in at least 20% of the arrays ensures that enough relevant data points are available to calculate an expression profile for that gene. The selected gene list comprised 2773 genes.

The similarity in expression of two genes was expressed in the correlation coefficient $\rho$ and was calculated for all pair-wise combinations of the 2773 genes for each of the three data sets. The $\rho$ distribution detailed the strength of pair-wise correlations and gave an impression on the nature of the three gene co-expression

networks (Fig. 2). For the mildly perturbed conditions data set, the $\rho$ values were centered around zero (Fig. 2, left), which suggests that most gene pairs in this data set were weakly co-expressed with only relatively few genes being strongly co-expressed. In contrast, the histogram for the strongly perturbed conditions data showed a much broader base. This broader base translated into a tendency of gene pairs to be more strongly correlated or anti-correlated (i.e., two genes that have antagonistic expression patterns) (Fig. 2, middle). The histogram of the combined data set resembled the histogram of the strongly perturbed conditions in shape, although less strongly correlating $\rho$ values were observed for this network (Fig. 2, right). The different $\rho$ value distributions suggested that co-expression was different between the data sets.

Three gene co-expression networks were constructed using the calculated $\rho$ values. In these networks, a connecting line was drawn between each pair of genes for which their expression profile correlated stronger than by the set $\rho$ threshold. The networks were visualized at different $\rho$ threshold values. When the $\rho$ threshold value was lowered, more connecting gene pairs appeared in the network. The network based on the combined data set at different $\rho$ threshold values is visualized in Fig. 3 (panels A–D), while for the mildly and strongly perturbed conditions-derived networks only the lowest $\rho$ threshold value with a meaningful clustering was visualized in panels E and F of Fig. 3 (full networks are accessible in Supplementary material file 1). Upon lowering the $\rho$ threshold, additional connecting gene pairs preferred attachment to genes already present, instead of being randomly placed within the network (Fig. 3, panels A–D). This preferential attachment of new genes to genes already in the network is a common observation for biological networks (Almaas, 2007; Barabási and Albert, 1999; Barabási and Oltvai, 2004). A result of preferential attachment was the presence of a small number of genes that correlated strongly with many other genes within a network, while many genes only correlated strongly with few other genes. The distribution of the number of correlations per gene, or the gene connectivity, is given in Fig. 4. For the three networks described here, the gene connectivity distribution could be described by a power-law distribution with a connectivity exponent $\gamma$ around 1.2 (Fig. 3). Similar values were found for other gene co-expression networks: a 4077-genes network of *S. cerevisiae* had $\gamma$ around 1.0 (van Noort et al., 2004), whereas this value ranged between 1.1 and 1.8 for gene co-expression networks for six distinct organisms (Bergmann et al., 2004).

### 3.2. Modules relate to biological functions

As the combined data network was derived from expression data obtained from both mildly and strongly perturbed conditions, we expected that co-expressed genes within this network are less prone to condition-specific peculiarities. Therefore, the combined data network was analyzed with respect to biological processes. Eight modules were observed in the combined data set network. These were colored and labeled A–H (Fig. 3, panel D). As genes with similar expression level profiles often encode proteins that are involved in a similar biological process (Walker et al., 1999; Wolfe et al., 2005), we searched for indications for biological processes that were overrepresented in the modules using the *S. cerevisiae* Gene Ontology vocabulary for genes in these modules that had a *S. cerevisiae* ortholog and the annotation of these genes. In addition, we examined whether genes within each module were assigned to metabolic pathways using the KEGG database. Lastly, we examined the upstream regions of genes within these modules for conserved upstream elements that hint to co-regulation by a common transcription factor. Indeed, when using the Gene Ontology annotations, an overrepresentation of similar ontology terms was found for genes that were present in some modules (Fig. 5; Supplementary material file 2). Also, conserved sequences were identified for genes of some modules.

Module A contains an overrepresentation of genes predicted to encode proteins involved in amino acid metabolic processes, including amino acid biosynthetic enzymes and tRNA-ligases. For 62 of 137 genes (45%), the conserved sequence 5′-TGA-(C/G)-TCA was identified ($p$-value $4.6 \times 10^{-15}$ and $6.7 \times 10^{-14}$, respectively) which is a known binding site of the DNA-binding protein CpcA (Wanke et al., 1997). This transcription factor is a global regulator in *A. niger*. Upon amino acid starvation, CpcA co-ordinates a transcriptional response by derepressing transcription of many genes encoding enzymes involved in amino acid biosynthetic pathways, as well as enzymes involved in nucleotide biosynthesis.

Module B consists of genes encoding proteins involved in fatty acid metabolism or peroxisome organization. Genes encoding the peroxins Pex6, Pex10, and Pex11, as well as the FoxA bifunctional enzyme that catalyzes the second and third step of fatty acid β-oxidation are in this module. We identified the conserved sequence 5′-CCTCGG or its reverse complement sequence within the upstream region of 16 of 20 genes of this module ($p$-value $2.6 \times 10^{-5}$). This sequence has been shown to be present upstream of a large number of genes predicted to encode proteins involved in fatty acid metabolism and peroxisome proliferation in filamentous fungi (Hynes et al., 2006). Hynes and co-workers showed experimentally that two transcription factors involved in fatty acid utilization, FarA and FarB, bind to this sequence in *A. nidulans* (Hynes et al., 2006).

Module C contains mostly cytosolic ribosomal protein-encoding genes. In the upstream region of 47% of the 88 genes within this module, the conserved sequence 5′-CGACAA was identified, while the core sequence 5′-CGAC was found upstream 80% of the genes. The probability of observing these upstream sequences for these

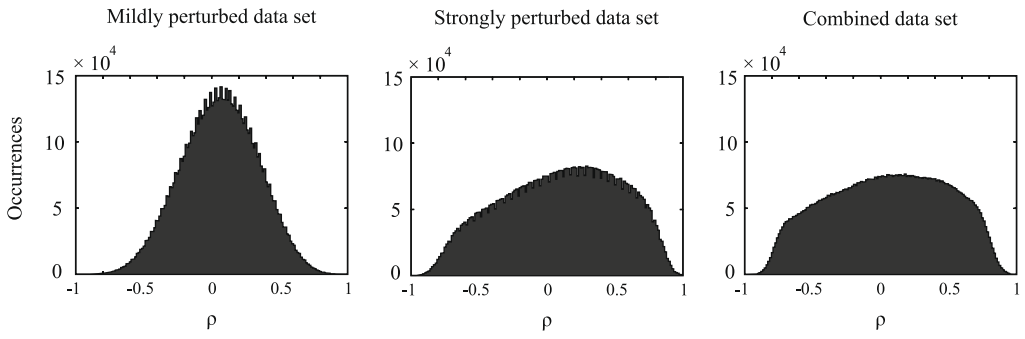


**Fig. 2.** Correlation coefficient distribution per data set. Histogram of the values of $\rho$ as calculated for all possible gene pair combinations in for each data set. The distribution of $\rho$ for all gene pairs possible is visualized for the subset of 2773 genes by dividing the range of $\rho$ values in equally spaced bins (e.g., one bin would range from 0 to 0.1, the next from 0.1 to 0.2, and so on), followed by counting the number of occurrences for a range of $\rho$ values per bin.

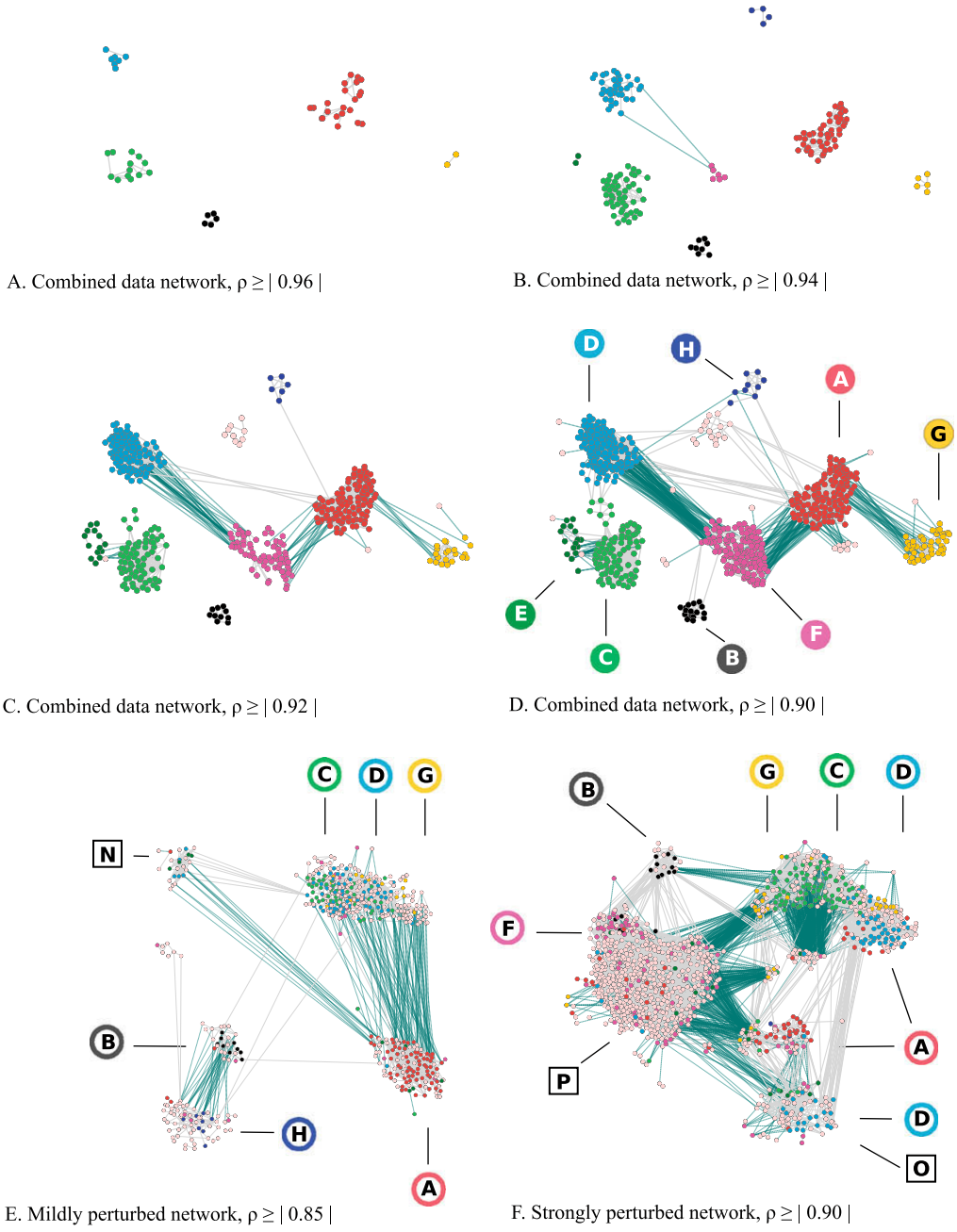


**Fig. 3.** Gene co-expression networks. Panels A–D: the gene co-expression network constructed from all 42 microarrays for four threshold settings of $\rho$ as indicated in each lower left corner. Circles represent genes, while lines represent a $\rho$ value above the set threshold. Positive $\rho$ values are shown as solid gray lines while negative $\rho$ values are represented as green lines. The networks constructed from mildly perturbed conditions data set (panel E) and strongly perturbed conditions data set (panel F) are given at their lowest $\rho$ threshold value only. Coloring is based on the modules identified in the combined data sets network (panel D), with module labels indicated in solid colored circles. This coloring is superimposed on the networks, and groups of genes with identical color are indicated by open colored circles (panels E and F). Boxed letters indicate modules that are not present in the combined data sets network.

genes by chance alone is very low ($p$-value $4 \times 10^{-6}$ and $2 \times 10^{-4}$, respectively). This sequence does not resemble any of the known binding sites associated with ribosomal proteins in *S. cerevisiae* or *S. pombe* (Tanay et al., 2005). The presence of such conserved sequence hints to the existence of a yet unidentified DNA-binding transcription factor that is involved in the regulation of genes encoding cytosolic ribosomal proteins in a fungal system.

In the D-labeled module, genes categorized by the generic Gene Ontology term "gene expression" are overrepresented. For example, this module contains genes that encode putative RNA helicases, spliceosome assembly proteins, and 16 putative translation initiation factors. We identified the sequence 5′-GGCCGCG for 111 of 152 genes ($p$-value $8 \times 10^{-4}$). This upstream element is located 400 base pairs or more away from the gene's start site for 60% of these 111 genes. Also for this module, the presence of a specific conserved upstream sequence suggested regulation by a yet unidentified DNA-binding transcription factor involved in the regulation of genes whose products appear involved in "gene expression" processes. Next to the above-mentioned sequence motif, we identified an overrepresentation of pyrimidine-rich sequences upstream for 80% of the genes of module D. However, it should be noted that CT-rich regions are relatively common in upstream re-

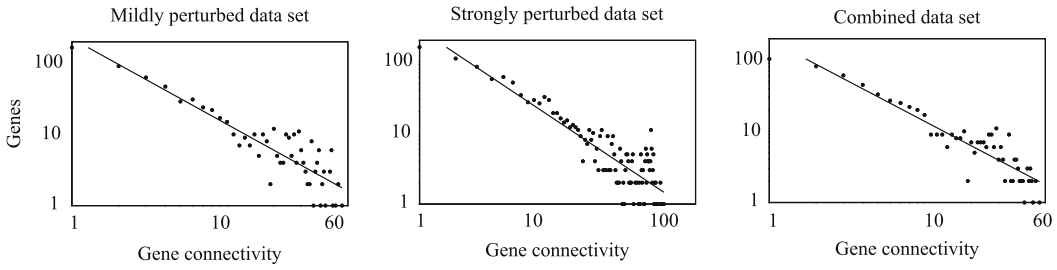


**Fig. 4.** Gene connectivity distribution per data set. For each gene within the networks at a $\rho$ threshold of 0.90, the number of gene pairs it partners with (horizontal axis) is plotted against the number of genes with identical number of gene pairs (vertical axis). The fitted line is for a power-law distribution, $P(k) \sim k^{-\gamma}$, which describes the probability that a gene has $k$ gene pairs. For all networks, $\gamma$ is around 1.2.

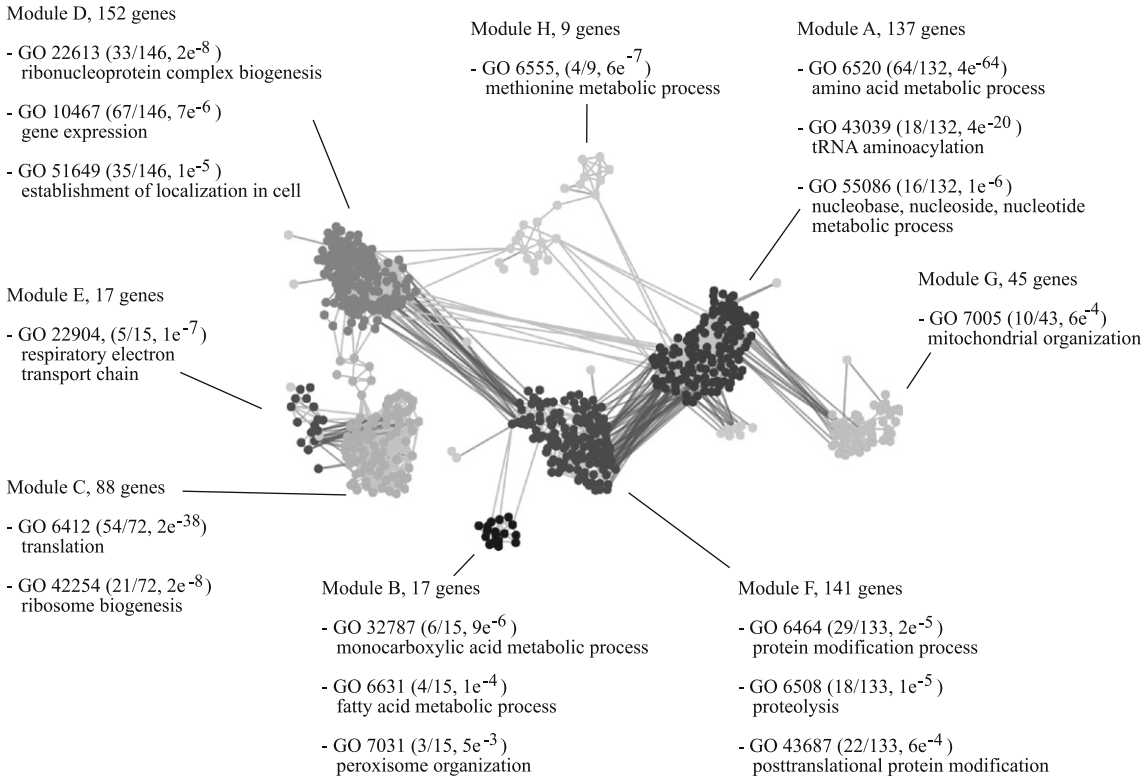


**Fig. 5.** Assignment of biological functions to modules. For the combined data sets network at a $\rho$ threshold of 0.90 (as shown in Fig. 3, panel D), enriched biological processes are indicated within modules using the Gene Ontology terms of genes with a *S. cerevisiae* ortholog. The number of *A. niger* genes per module is indicated after the module code. The GO-number points to the observed Gene Ontology process. Between brackets, the number of genes with that annotated Gene Ontology process relative to the total number of genes queried is given, followed by a *p*-value that gives the likelihood that the identified GO process is found by chance alone. Genes that encode conserved proteins but have no *S. cerevisiae* ortholog make up the difference in the total number of *A. niger* genes per module and the number of genes queried for Gene Ontology enrichment.

gions of filamentous fungi and that they are mostly related to the position of the transcription start site (Punt and van den Hondel, 1992).

Overrepresented Gene Ontology processes were also found for modules E–H (Fig. 5; Supplementary material file 2). However, no conserved upstream sequences were found. Module E pertains to energy metabolism related processes like "electron transport chain", "oxidative phosphorylation", and "ATP synthesis coupled electron transport". Module F contains the following overrepresented processes: "cellular catabolic process", "proteolysis", and "protein modification process". Module G is related to "organelle organization" and "mitochondrion organization". Module H pertains to processes related to different amino acid metabolic processes.

For each module, we assessed whether their genes could be related to metabolic pathways (Supplementary material file 2). We observed a good agreement between *A. niger* genes within a module that can be linked to biological pathways, and the observed overrepresentation of Gene Ontology processes. For example, module A, which contains an overrepresentation of genes encoding proteins involved in amino acid metabolic processes, contains 71 genes that encode proteins of amino acid related biochemical pathways (Supplementary material file 2). For the 16-gene containing module E, which has overrepresenting Gene Ontology terms related to energy metabolism, the four genes that encode proteins that relate to KEGG biochemical pathways are within the "oxidative phosphorylation" pathway.

### 3.3. Modular structure is retained in the two other networks

Seventy-five percent of the genes that were present in the combined data sets network were found in at least one other network

(Fig. 6). The localization of these genes within each network was examined by coloring of each module identified in the combined data sets network (Fig. 3, panel D), and superimposition of these colors to both other networks (Fig. 3, panels E and F).

Both the mildly perturbed and strongly perturbed conditions networks appeared less structured compared to the combined data sets network, but modules can be recognized nevertheless.

The modules labeled C, D, and G, were relatively well separated in the combined data sets network while these modules overlapped or were closely connected in the two other networks. In the combined data sets network, these modules were enriched for genes encoding proteins involved in "ribosome biogenesis" and "translation", "gene expression" and "ribonucleoprotein complex biogenesis", and "mitochondrial organization" respectively. In the mildly perturbed conditions network, 323 genes are in the module that contains many of the C, D, and G-colored genes; 138 of these genes (43%) were present as well in the combined data sets network. A Gene Ontology terms search on all 323 genes yielded similar GO terms for this C–D–G module (Supplementary material file 3). Likewise, in the strongly perturbed conditions network, the 284 genes that included many of the genes of the C, D, and G modules in the combined data sets network yielded the same GO terms (Supplementary material file 3).

Modules A and B found in the network based on the combined data set are present in the mildly and strongly perturbed networks as well. Other genes are also associated to these modules, and these genes have the same GO terms associated to them as in the combined data set network (Supplementary material file 3), namely "cellular amino acid biosynthetic process" and related GO terms for module A and "fatty acid β-oxidation" and "carboxylic acid metabolic process" for module B.

In addition, network-specific modules appeared that were not visible in the combined data sets network. The mildly perturbed conditions network gave one such module, labeled N (Fig. 3, panel E). The N module consisted of 24 genes, of which half are related to the respiratory electron transport chain. Indeed, this module contains subunits of the ubiquinol–cytochrome C oxidase complex. The expression of these genes could be a specific adaptation to the exponential growth phase these cells were in at the time of sampling.

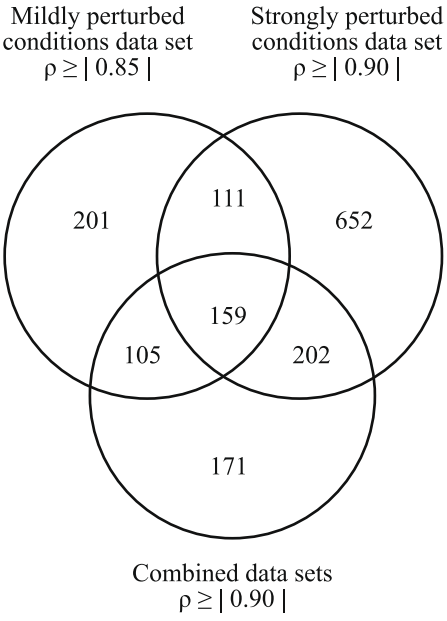


**Fig. 6.** Overlap of genes between networks. Venn-diagram showing the overlap of genes present in any of the three networks analyzed.

For the strongly perturbed network, two specific modules were observed and were labeled O and P in Fig. 3, panel F. The 84 genes present in module O were enriched for "metabolic processes", to which term 62 of 84 genes are assigned. Twenty-four percent of the 84 genes were involved in the generation of precursor metabolites and energy.

The second module P (Fig. 3, panel F) was large and contained half of the genes present in the strongly perturbed data set. No biological processes were overrepresented for this module although the consensus expression profile (see below) seemed to correlate strongly to the presence or absence of a carbon source in the medium. No overrepresented motifs were detected in the upstream region of genes in this module. Ten percent of the genes located in this module were also located in module F in the combined data sets network. These observations supported our choice to use the combined data set as a basis as indeed the individual data sets seemed to contain condition-specific modules.

### *3.4. Extending the subset of genes by means of a consensus expression profile*

The thus far constructed gene co-expression networks were based on a subset of 2773 genes that encode evolutionary conserved proteins. However, these genes made up only 43% of the total of 6416 *A. niger* protein-encoding genes that were evaluated as present in more than 20% of the arrays on the combined data set microarrays. Genes that did not take part in our initial selection were examined using the modules identified in the network. Genes within a module have similar gene expression profiles, and this similarity was used to calculate a consensus expression profile (Horvath and Dong, 2008) for each module in the combined data set. The correlation between each module's consensus expression profile and all 6416 genes' expression profiles was calculated and expressed as an associated consensus expression profile correlation coefficient $\rho_{cons}$ for each gene. Associated $\rho_{cons}$ values for all modules are given in Supplementary material file 4. Here, the results of this approach are exemplified by description of module B. This module contained 19 genes at $\rho$ of 0.90, with most genes being related to peroxisome proliferation and fatty acid metabolism. As genes in this module are relatively well characterized, interpretation of the resulting data and analysis of this proof-of-concept is made easier.

Table 2 presents the genes with a $\rho_{cons}$ to the module B consensus expression profile of 0.90 or higher. Half of the 20 genes that correlate most strongly with the module B consensus expression profile did fall outside our initial selection criteria (Table 2). Most of these genes could be associated with fatty acid metabolic activity or peroxisome functioning based on inspection of their gene annotation. Interestingly, the ortholog of the *A. nidulans* FarA fatty acid-related transcription factor, encoded by gene An14g00920, also correlated strongly with the module B consensus expression profile. A motif sequence for the FarA and FarB transcription factors could be identified in the upstream region of all but three genes listed in Table 2, including the FarA-encoding gene itself.

Similar results were obtained when the consensus expression profiles derived from the other modules were analyzed (Supplementary material file 4). For example, module F in the combined data sets network had an overrepresentation of "protein modification process" and related Gene Ontology terms. These Gene Ontology terms are also overrepresented when the genes with a *S. cerevisiae* ortholog that have an expression profile $\rho_{cons}$ of 0.80 or more were analyzed. For instance, the Gene Ontology term "protein modification process" is found for 22% of these genes ($p$-value $4.4 \times 10^{-11}$). In addition, of the 86 genes with an expression profile similar to the consensus expression profile by over $\rho_{cons}$ 0.90, 40

**Table 2**
Genes strongly correlating with consensus expression profile of module B.

| Rank | *A. niger* probe ID (corresponding gene ID) | $\rho_{cons}$[a] | Description | *S. cerevisiae* ID | Number of orthologous species containing this protein | Gene is present in combined data network? | Upstream motif[b] |
|---|---|---|---|---|---|---|---|
| 1 | An00g06872_at (An13g01920) | 0.982 | Strong similarity to acetyl-CoA C-acetyltransferase precursor – *R. norvegicus* | YPL028W | 17 | Yes | (61), 437 |
| 2[c] | An00g11070_at (An16g07150) | 0.979 | Strong similarity to soluble cytoplasmic fumarate reductase FRDS1 – *S. cerevisiae* | YEL047C | 17 | Yes | (208) |
| 3 | An00g06382_at (An16g05340) | 0.977 | Similarity to trans-2-enoyl-ACP reductase II fabK – *S. pneumoniae* | YJR149W | 15 | Yes | (165), 192 |
| 4[c] | An00g11070_s_at (An16g07150) | 0.970 | Strong similarity to soluble cytoplasmic fumarate reductase FRDS1 – *S. cerevisiae* | YEL047C | 17 | Yes | (208) |
| 5 | An00g11023_at (An01g12960) | 0.970 | Strong similarity to short/branched chain specific acyl-CoA dehydrogenase precursor ACADSB – *H. sapiens* | | 12 | No | 145, (161) |
| 6 | An00g09716_at (An15g01920) | 0.968 | Strong similarity to methylcitrate synthase mcsA – *A. nidulans* | YPR001W | 12 | Yes | (152), 368 |
| 7 | An00g06734_at (An14g00430) | 0.958 | Strong similarity to 3-hydroxybutyryl-CoA dehydrogenase BHBD – *C. acetobutylicum* | | 11 | No | 109 |
| 8 | An00g06380_at (An07g03290) | 0.953 | Similarity to trans-2-enoyl-ACP reductase II fabK – *S. pneumoniae* | | 11 | No | (250) |
| 9 | An00g09575_at (An08g10110) | 0.940 | Strong similarity to lipid transfer protein POX18 – *C. tropicalis* | | 14 | No | (612) |
| 10 | An00g09583_at (An04g03290) | 0.939 | Strong similarity to long-chain acyl-CoA dehydrogenase – *R. norvegicus* | | 8 | No | 784 |
| 11 | An00g06703_at (An12g07630) | 0.938 | Strong similarity to 2-methylisocitrate lyase ICL2 – *S. cerevisiae* | YPR006C | 14 | Yes | 246 |
| 12 | An00g08462_at (An01g09830) | 0.927 | Strong similarity to glutathione S-transferase GTT1 – *S. cerevisiae* | YIR038C | 15 | Yes | 170 |
| 13 | An00g09552_at (An08g07520) | 0.921 | Strong similarity to levodione reductase lvr – *C. aquaticum* | | 11 | No | (259) |
| 14 | An00g05624_at (An02g05230) | 0.921 | Similarity to protein fragment SEQ ID NO: 65270 of patent EP1033405-A2 – *A. thaliana* | | | NO | (298) |
| 15 | An00g09578_at (An12g08270) | 0.919 | Strong similarity to L-lactate 2-monooxygenase LA2M – *M. smegmati* | | 9 | No | 152 |
| 16 | An00g10969_at (An07g00440) | 0.918 | Strong similarity to secretory lipase LIP2 – *C. albicans* | | 9 | No | Not present |
| 17 | An00g11952_at (An14g00990) | 0.915 | Strong similarity to trifunctional protein of the β-oxidation fox-2 – *N. crassa* | YKR009C | 17 | Yes | 226 |
| 18 | An00g12118_at (An07g09190) | 0.906 | Strong similarity to very long-chain fatty acyl-CoA synthetase FAT1 – *S. cerevisiae* | | | No | 318 |
| 19 | An00g13622_at (An02g04350) | 0.906 | Weak similarity to the helix-loop-helix transcription factor Max – *M. musculus* | | 5 | No | Not present |
| 20 | An00g10237_at (An01g03680) | 0.901 | Strong similarity to peroxisomal ABC transporter ALDR – *M. musculus* | YPL147W | 17 | Yes | 315 |
| 21 | An00g07656_at (An14g00920) | 0.900 | Strong similarity to FarA transcription factor – *A. nidulans* | | 11 | No | Not present |

[a] The correlation coefficient of a gene's expression profile with the consensus expression profile constructed from genes present in module B.
[b] Number indicates the position of the upstream motif 5′-CCGAGG relative to the gene's start codon in base pairs; number in brackets indicates the position of the reverse complement motif relative to the start codon.
[c] Gene An16g07150 is represented on the DNA microarray by two probe sets, both of which are in this list.

genes are annotated as "hypothetical protein" (Supplementary material file 4).

## 4. Discussion

Variations in the timing and levels of gene transcription, mRNA translation, and protein maturation have considerable consequences for a cell. For understanding the dynamics of the physiological processes in *Aspergillus niger*, insight into the interactions and combined activity of these processes or events is required, in addition to knowledge of individual components of the cellular system. This study queried transcriptomes obtained from cultures grown under different experimental conditions, with the aim to gain insight into the relations between genes, and, at a higher hierarchical level, into relations between modules. For this, initial analysis of gene co-expression was performed on an evolutionary highly conserved subset of the *A. niger* genes, and the analysis was subsequently extended to the whole genome.

Our approach reveals that the gene co-expression networks consist of modules of co-expressed genes (Fig. 3). Subsequent analysis of the discovered modules provides evidence for their biological relevance: (i) modules are enriched for Gene Ontology terms, (ii) genes within the modules relate to biochemical pathways, and (iii) conserved motifs are present in the upstream region of many genes in several of the modules. Experimentally confirmed upstream sequences corresponding to the DNA-binding sites of the transcription factors CpcA (involved in amino acid related processes) (Wanke et al., 1997) and FarA/FarB (involved in β-oxidation and perosixome biosynthesis) (Hynes et al., 2006) are found in modules A and B, respectively, that have an overrepresentation of related Gene Ontology terms. These findings indicate that our approach is able to infer "true" biological processes.

In addition to observations that can be related to experimentally verified data, our approach yielded novel targets for experimental validation, such as the upstream sequences observed in genes of modules C and D. The sequence 5′-CGACAA in the upstream region of many ribosomal protein-encoding genes in module C appears of special interest, as this sequence does not resemble the conserved upstream sequences found genes encoding ribosomal proteins in the yeasts *S. cerevisiae* and *S. pombe* (Tanay et al., 2005).

Previous gene co-expression network studies used a subset of genes already known (Bergmann et al., 2004; Herrgård et al., 2003) or suspected (Neretti et al., 2007) to be involved in specific biological processes, or discussed network characteristics without zooming into biological details (Jordan et al., 2008; van Noort et al., 2004). In this study, however, an approach similar to the approach of Daub and Sonnhammer (2008) was followed. A subset of genes was selected based on their highly conserved nature among fungal species, without taking into account their role in biological processes. An advantage of this approach is that also protein-encoding genes for which no function is assigned are analyzed, while a potential pitfall of this selection criterion is that evolutionary less well conserved co-expressed genes (e.g., species-specific genes that encode biopolymer-degrading enzymes) will not be examined in this initial selection. Genes within the observed modules are related to essential cellular processes; for instance, ribosomes are required for protein synthesis (Fig. 5, module C), genes must be transcribed (Fig. 5, module D), and amino acids must be synthesized de novo when not supplied in the medium (Fig. 5, Module A).

The advantage of selecting evolutionary conserved genes likely extends to the consensus expression profile analyses. As the modules identified are based on evolutionary conserved sequences, it is likely that the consensus expression profiles of these modules are more robust than consensus expression profiles based on subsets of known genes. The evolutionary conservation suggests that a large time span has past in which the regulation of such a module was tuned, while for organism-specific modules the regulation could be more variable. Thus, using consensus expression profiles based on an evolutionary conserved subset will probably result in more accurate lists of genes with similar expression profiles to the consensus expression profile.

Research towards a better understanding of the higher-order structures that play a role in *A. niger* cellular functioning did only start recently, after high throughput technologies such as DNA microarray platforms became available for this organism. The usefulness of studying these higher-order structures is illustrated in this paper; the networks of evolutionary conserved genes of *A. niger* resulted in the identification of biologically relevant gene co-expression modules. In addition, the use of consensus-profiles extended the analysis to include the full gamut of genes of *A. niger*.

## Acknowledgments

This project was in part carried out within the research programme of the Kluyver Centre for Genomics of Industrial Fermentation (RvdB, MB, PP, and MvdW) which is part of the Netherlands Genomics Initiative/Netherlands Organization for Scientific Research. In addition, RvdB was supported by the Research Fund of Katholieke Universiteit Leuven (EF/05/007 SymBioSys) and by IWT-Flanders (IWT/060045/SBO Bioframe).

The authors would like to thank DSM Food Specialties for access to the DSM *A.niger* Affymetrix DNA microarrays, and for providing the microarrays to MB.

## Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.fgb.2010.03.005.

## References

Affymetrix, 2001. Statistical Algorithms Reference Guide. Technical Report. Affymetrix Inc., Santa Clara, CA.

Affymetrix, 2004. GeneChip Expression Analysis Technical Manual. Affymetrix Inc., Santa Clara, CA.

Almaas, E., 2007. Biological impacts and context of network theory. J. Exp. Biol. 210, 1548–1558.

Andersen, M.R., Vongsangnak, W., Panagiotou, G., Salazar, M.P., Lehmann, L., Nielsen, J., 2008. A trispecies *Aspergillus* microarray: comparative transcriptomics of three *Aspergillus* species. Proc. Natl. Acad. Sci. USA 105, 4387–4392.

Barabási, A., Albert, R., 1999. Emergence of scaling in random networks. Science 286, 509–512.

Barabási, A., Oltvai, Z.N., 2004. Network biology: understanding the cell's functional organization. Nat. Rev. Genet. 5, 101–113.

Basso, K., Margolin, A.A., Stolovitzky, G., Klein, U., Dalla-Favera, R., Califano, A., 2005. Reverse engineering of regulatory networks in human B cells. Nat. Genet. 37, 382–390.

Bennett, J., Lasure, L., 1991. Growth media. In: Bennett, J., Lasure, L. (Eds.), More Gene Manipulations in Fungi. Academic Press, San Diego, CA.

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

Braaksma, M., Smilde, A.K., van der Werf, M.J., Punt, P.J., 2009. The effect of environmental conditions on extracellular protease activity in controlled fermentations of *Aspergillus niger*. Microbiology 155, 3430–3439.

Breakspear, A., Momany, M., 2007. The first fifty microarray studies in filamentous fungi. Microbiology 153, 7–15.

Daub, C.O., Sonnhammer, E.L., 2008. Employing conservation of co-expression to improve functional inference. BMC Syst. Biol. 2, 81.

DeRisi, J.L., Iyer, V.R., Brown, P.O., 1997. Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 278, 680–686.

Edgar, R., Domrachev, M., Lash, A.E., 2002. Gene expression omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 30, 207–210.

Featherstone, D.E., Broadie, K., 2002. Wrestling with pleiotropy: genomic and topological analysis of the yeast gene expression network. Bioessays 24, 267–274.

Guillemette, T., van Peij, N.N.M.E., Goosen, T., Lanthaler, K., Robson, G.D., van den Hondel, C.A.M.J.J., Stam, H., Archer, D.B., 2007. Genomic analysis of the secretion stress response in the enzyme-producing cell factory *Aspergillus niger*. BMC Genomics 8, 158.

Herrgård, M.J., Covert, M.W., Palsson, B.Ø., 2003. Reconciling gene expression data with known genome-scale regulatory network structures. Genome Res. 13, 2423–2434.

Hong, E.L., Balakrishnan, R., Dong, Q., Christie, K.R., Park, J., Binkley, G., Costanzo, M.C., Dwight, S.S., Engel, S.R., Fisk, D.G., Hirschman, J.E., Hitz, B.C., Krieger, C.J., Livstone, M.S., Miyasato, S.R., Nash, R.S., Oughtred, R., Skrzypek, M.S., Weng, S., Wong, E.D., Zhu, K.K., Dolinski, K., Botstein, D., Cherry, J.M., 2008. Gene Ontology annotations at SGD: new data sources and annotation methods. Nucleic Acids Res. 36, D577–D581.

Horvath, S., Dong, J., 2008. Geometric interpretation of gene coexpression network analysis. PLoS Comput. Biol. 4, e1000117.

Hughes, T.R., Robinson, M.D., Mitsakakis, N., Johnston, M., 2004. The promise of functional genomics: completing the encyclopedia of a cell. Curr. Opin. Microbiol. 7, 546–554.

Hynes, M.J., Murray, S.L., Duncan, A., Khew, G.S., Davis, M.A., 2006. Regulatory genes controlling fatty acid catabolism and peroxisomal functions in the filamentous fungus *Aspergillus nidulans*. Eukaryotic Cell 5, 794.

Irizarry, R., Bolstad, B., Collin, F., Cope, L., Hobbs, B., Speed, T., 2003. Summaries of Affymetrix Genechip probe level data. Nucleic Acids Res. 31, e15.

Jackson, J., 1991. A User's Guide to Principal Components. John Wiley & Sons Inc..

Joliffe, I., 2002. Principal Component Analysis. Springer-Verlag, New York.

Jordan, I.K., Katz, L.S., Denver, D.R., Streelman, J.T., 2008. Natural selection governs local, but not global, evolutionary gene coexpression networks in *Caenorhabditis elegans*. BMC Syst. Biol. 2, 96.

Jørgensen, T.R., Goosen, T., Hondel, C.A.M.J.J.V.D., Ram, A.F.J., Iversen, J.J.L., 2009. Transcriptomic comparison of *Aspergillus niger* growing on two different sugars reveals coordinated regulation of the secretory pathway. BMC Genomics 10, 44.

Kanehisa, M., Goto, S., 2000. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 28, 27–30.

Lashkari, D.A., DeRisi, J.L., McCusker, J.H., Namath, A.F., Gentile, C., Hwang, S.Y., Brown, P.O., Davis, R.W., 1997. Yeast microarrays for genome wide parallel genetic and gene expression analysis. Proc. Natl. Acad. Sci. USA 94, 13057–13062.

Lee, H.K., Hsu, A.K., Sajdak, J., Qin, J., Pavlidis, P., 2004. Coexpression analysis of human genes across many microarray data sets. Genome Res. 14, 1085–1094.

Levin, A.M., de Vries, R.P., Conesa, A., de Bekker, C., Talon, M., Menke, H.H., van Peij, N.N.M.E., Wösten, H.A.B., 2007. Spatial differentiation in the vegetative mycelium of *Aspergillus niger*. Eukaryotic Cell 6, 2311–2322.

Magnuson, J.K., Lasure, L.L., 2004. Organic acid production by filamentous fungi. In: Lange, J., Lange, L. (Eds.), Advances in Fungal Biotechnology for Industry, Agriculture, and Medicine. Kluwer Academic/Plenum Publishers.

Martens-Uzunova, E.S., Schaap, P.J., 2008. An evolutionary conserved D-galacturonic acid metabolic pathway operates across filamentous fungi capable of pectin degradation. Fungal Genet. Biol. 45, 1449–1457.

Meyer, V., Damveld, R.A., Arentshorst, M., Stahl, U., van den Hondel, C.A., Ram, A.F., 2007. Survival in the presence of antifungals: genome-wide expression profiling of *Aspergillus niger* in response to sublethal concentrations of caspofungin and fenpropimorph. J. Biol. Chem. 282, 32935–32948.

Neretti, N., Remondini, D., Tatar, M., Sedivy, J.M., Pierini, M., Mazzatti, D., Powell, J., Franceschi, C., Castellani, G.C., 2007. Correlation analysis reveals the emergence of coherence in the gene expression dynamics following system perturbation. BMC Bioinformatics 8 (Suppl. 1), S16.

Pel, H.J., de Winde, J.H., Archer, D.B., Dyer, P.S., Hofmann, G., Schaap, P.J., Turner, G., de Vries, R.P., Albang, R., Albermann, K., Andersen, M.R., Bendtsen, J.D., Benen, J.A., van den Berg, M., Breestraat, S., Caddick, M.X., Contreras, R., Cornell, M., Coutinho, P.M., Danchin, E.G., Debets, A.J., Dekker, P., van Dijck, P.W., van Dijk, A., Dijkhuizen, L., Driessen, A.J., d'Enfert, C., Geysens, S., Goosen, C., Groot, G.S., de Groot, P.W., Guillemette, T., Henrissat, B., Herweijer, M., van den Hombergh, J.P., van den Hondel, C.A., van der Heijden, R.T., van der Kaaij, R.M., Klis, F.M., Kools, H.J., Kubicek, C.P., van Kuyk, P.A., Lauber, J., Lu, X., van der Maarel, M.J., Meulenberg, R., Menke, H., Mortimer, M.A., Nielsen, J., Oliver, S.G., Olsthoorn, M., Pal, K., van Peij, N.N., Ram, A.F., Rinas, U., Roubos, J.A., Sagt, C.M., Schmoll, M., Sun, J., Ussery, D., Varga, J., Vervecken, W., van de Vondervoort, P.J., Wedler, H., Wosten, H.A., Zeng, A.P., van Ooyen, A.J., Visser, J., Stam, H., 2007. Genome sequencing and analysis of the versatile cell factory *Aspergillus niger* CBS 513.88. Nat. Biotechnol. 25, 221–231.

Pieterse, B., Jellema, R.H., van der Werf, M.J., 2006. Quenching of microbial samples for increased reliability of microarray data. J. Microbiol. Methods 64, 207–216.

Pontecorvo, G., Roper, J.A., Hemmons, L.M., Macdonald, K.D., Bufton, A.W., 1953. The genetics of *Aspergillus nidulans*. Adv. Genet. 5, 141–238.

Punt, P., van den Hondel, C., 1992. Analysis of transcription control sequences in filamentous fungi. In: Stahl, U., Tudzynscki, P. (Eds.), Proc. EMBO Workshop on Molecular Biology of Filamentous Fungi. VCH, Weinheim.

Punt, P.J., van Biezen, N., Conesa, A., Albers, A., Mangnus, J., van den Hondel, C., 2002. Filamentous fungi as cell factories for heterologous protein production. Trends Biotechnol. 20, 200–206.

Schroeder, A., Mueller, O., Stocker, S., Salowsky, R., Leiber, M., Gassmann, M., Lightfoot, S., Menzel, W., Granzow, M., Ragg, T., 2006. The RIN: an RNA integrity number for assigning integrity values to RNA measurements. BMC Mol. Biol. 7, 3.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., Ideker, T., 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 13, 2498–2504.

Tanay, A., Regev, A., Shamir, R., 2005. Conservation and evolvability in regulatory networks: the evolution of ribosomal regulation in yeast. Proc. Natl. Acad. Sci. USA 102, 7203–7208.

van den Berg, R.A., Hoefsloot, H.C.J., Westerhuis, J.A., Smilde, A.K., van der Werf, M.J., 2006. Centering, scaling, and transformations: improving the biological information content of metabolomics data. BMC Genomics 7, 142.

van der Veen, D., 2009. Transcriptional Profiling of *Aspergillus niger*. Ph.D. Thesis. Wageningen University, Wageningen, The Netherlands.

van der Veen, D., Oliveira, J.M., van den Berg, W.A.M., de Graaff, L.H., 2009. Variance components analysis reveals contribution of sample processing to transcript variation. Appl. Environ. Microbiol. 75, 2414–2422.

van Hartingsveldt, W., Mattern, I.E., van Zeijl, C.M., Pouwels, P.H., van den Hondel, C.A., 1987. Development of a homologous transformation system for *Aspergillus niger* based on the *pyrG* gene. Mol. Gen. Genet. 206, 71–75.

van Noort, V., Snel, B., Huynen, M.A., 2004. The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model. EMBO Rep. 5, 280–284.

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

Wanke, C., Eckert, S., Albrecht, G., van Hartingsveldt, W., Punt, P.J., van den Hondel, C.A., Braus, G.H., 1997. The *Aspergillus niger* GCN4 homologue, *cpcA*, is transcriptionally regulated and encodes an unusual leucine zipper. Mol. Microbiol. 23, 23–33.

Wheelock, C.E., Wheelock, A.M., Kawashima, S., Diez, D., Kanehisa, M., van Erk, M., Kleemann, R., Haeggström, J.Z., Goto, S., 2009. Systems biology approaches and pathway tools for investigating cardiovascular disease. Mol. Biosyst. 5, 588–602.

Wolfe, C.J., Kohane, I.S., Butte, A.J., 2005. Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks. BMC Bioinformatics 6, 227.

Yuan, X., Roubos, J.A., van den Hondel, C.A.M.J.J., Ram, A.F.J., 2008a. Identification of InuR, a new $Zn(II)_2Cys_6$ transcriptional activator involved in the regulation of inulinolytic genes in *Aspergillus niger*. Mol. Genet. Genomics 279, 11–26.

Yuan, X., van der Kaaij, R.M., van den Hondel, C.A.M.J.J., Punt, P.J., van der Maarel, M.J.E.C., Dijkhuizen, L., Ram, A.F.J., 2008b. *Aspergillus niger* genome-wide analysis reveals a large number of novel alpha-glucan acting enzymes with unexpected expression profiles. Mol. Genet. Genomics 279, 545–561.

Zar, J., 1996. Biostatistical Analysis. Prentice Hall.