

File ID 110288  
Filename Chapter 5 FunKey : prediction of yeast gene functions using gene  
expression and fitness profile data

---

SOURCE (OR PART OF THE FOLLOWING SOURCE):

Type Dissertation  
Title Dissection of transcriptional regulation networks and prediction of gene  
functions in *Saccharomyces cerevisiae*  
Author A. Boorsma  
Faculty Faculty of Science  
Year 2008  
Pages 136  
ISBN 9078675303

FULL BIBLIOGRAPHIC DETAILS:

<http://dare.uva.nl/record/271379>

---

*Copyright*

*It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or  
copyright holder(s), other than for strictly personal, individual use.*

---

DNA microarray T-profiler mRNA  
unpaired t-test FUNKEY post-genomic  
motif prediction of function YNL155W  
gene groups **Chapter 5** Bonferonni  
Gene Ontology consensus  
*Saccharomyces cerevisiae* T-base  
webapplication Fitness profiles Uncharacterized



**FunKey: Prediction of yeast gene functions using gene expression and fitness profile data**

André Boorsma, Anna Zakrzewska, Klaas J. Hellingwerf,  
Frans M. Klis and Gertien J. Smits

Manuscript in preparation



**Abstract**

Since the completion of the *Saccharomyces cerevisiae* genome sequence about 75% of the genes have been annotated and characterized. 25% remains uncharacterized, and many of the annotated genes are only poorly characterized. Therefore, there is a great need for methods that can generate specific hypotheses about the putative function of these genes. Here we present a novel approach that enables the prediction of the function of genes based on the correlation of the behavior of genes and functional groups in a large collection of gene expression and fitness experiments. We applied T-profiler, a method that scores changes in the average expression level of predefined groups of genes on multiple transcriptome and global fitness datasets. We show functional association is accompanied by correlated gene-gene groups behavior. Based on the analysis of a test-set we were able to make reliable predictions for 64% of all genes, which could be improved to 78% if a strict correlation cutoff was applied. In addition our method allows for the improvement of gene annotation, which can be shown by phenotypic analysis. Our method is implemented in a web application named FunKey, and was used to predict functions for all uncharacterized genes.

## Introduction

The number of completely sequenced organisms is growing every day. According to recent information available on [www.genomesonline.org](http://www.genomesonline.org) over 1800 genomes have been sequenced. A great challenge lies in making sense of the huge amounts of sequence information produced by these projects. Especially the physiological role of individual genes cannot be inferred from sequence information only. Since the unraveling of the first eukaryotic sequence of *Saccharomyces cerevisiae* [5], 4459 genes have obtained a functional annotation (process and function) according to Gene Ontology [100]. Despite the fact that *S. cerevisiae* is one of the most studied eukaryotic organisms, the function of almost 25% of the genes remains enigmatic. In addition, the function of many annotated genes is still poorly understood [6]. Clearly, there is a need for large-scale methods that can generate predictions for the function of uncharacterized and poorly characterized genes. Such predictions can then serve as a guideline for researchers to investigate these genes at the lab bench and test proposals for their functions.

In the post-genomic era a variety of functional genomics methods have been developed that can help elucidate the functional properties of genes. For the yeast *Saccharomyces cerevisiae*, gene expression profiling [21, 22] and fitness profiling [17] are widely used. In recent years the amount of data from such experiments has rapidly increased [8, 64, 92, 152]. Gene expression profiles provide three different types of information. First, expression profiles can be used to infer information about how gene expression is controlled [43, 88]. The combination of expression profiles and information obtained from global transcription factor binding assays (ChIP-chip) [61, 62] has been especially useful to unravel transcriptional regulatory networks [162, 180]. Second, functionally related genes that respond in concert can give information about the physiological and functional changes in a cell [181]. Thirdly, gene expression profiles may indicate the function of uncharacterized genes, for example, when their expression is strongly correlated with the expression of functionally related genes [182]. This has been the basis of various methods predicting function based on gene expression [72, 183-185].

Fitness profiling experiments examines the fitness of all yeast deletion mutants that grow competitively in the presence of an inhibitor [8, 17]. Each of the deletion mutants is uniquely identified by a molecular 'bar-code'. These bar-codes are used to identify the amount of each deletion mutant on a high density oligo array. Similar experiments are also performed by plating assays [63] and parallel analysis [186] using the deletion collection. Information obtained from fitness profiling experiments is about the relative fitness of a deletion mutant under a particular condition. In analogy to gene expression profiles, functionally related genes that show a reduced or a better fitness give information about the physiological and functional responses to a stress or a compound treatment. [65]. In a similar way, the relative fitness of genes with a known function that correlate with uncharacterized genes might give insight into the function of these genes. A number of reports show that there is no clear correlation between gene expression and fitness profiles under similar stresses [8, 187]. This might imply that information obtained from fitness profiling experiments is complementary to information derived from gene expression profiles.

Recently we introduced T-profiler [83], a method that uses the unpaired t-test to score changes in the average activity of predefined groups of genes. This method can use gene sets as defined by Gene Ontology [100] and MIPS to calculate the significance of co-expression of these gene groups. T-profiler has the advantage that no arbitrary cutoffs on the level of gene expression have to be made. Furthermore the transformation to t-values comprises an internal normalization allowing the comparison between heterogeneous gene expression data sets. It also allows for a comparison between gene expression and fitness profile data [65]. We applied T-profiler to a library of 936 gene expression profiles and 159 fitness profiling experiments from various sources, and calculated the t-values for each gene group used. Next, we used the obtained t-values to perform correlation analysis with all individual genes, over all experiments. For the characterized genes, functional association is reflected in high correlation. We therefore tested whether correlation alone serves as an accurate functional prediction. We assessed the reliability of our predictions by testing a set of well-characterized genes under strict conditions. Depending on the correlation coefficient cut-off used we were able to make reliable predictions for 78% of the genes using combined gene expression and fitness profiles. Furthermore, analysis of gene groups suggests novel functions. The data were implemented in a gene function prediction web-tool named FunKey. Using strict criteria FunKey reliably predicted a function for 169 dubious and uncharacterized genes.

## Materials & Methods

### T-profiler.

For a given gene group G, the t-value is given by the following formula:

$$t_G = \frac{\mu_G - \mu_{G'}}{s \sqrt{\frac{1}{N_G} + \frac{1}{N_{G'}}}} \quad \text{where} \quad s = \sqrt{\frac{(N_G - 1) * s_G^2 + (N_{G'} - 1) * s_{G'}^2}{N_G + N_{G'} - 2}}$$

Here  $\mu_G$  is the mean expression log-ratio of the  $N_G$  genes in gene group G;  $\mu_{G'}$  is the mean expression log-ratio of the remaining  $N_{G'}$  genes; and  $s$  is the pooled standard deviation, as obtained from the estimated variances for groups G and G'. The associated two-tailed p-value can be calculated from  $t$  using the t-distribution with  $N-2$  degrees of freedom. We accounted for multiple testing by computing an E-value equal to the p-value multiplied by the number of gene groups (Bonferonni correction). All t-values of groups with an E-value of 0.05 or smaller are considered to be significant. To reduce the influence of outliers, which may result in false positives or false negatives, we discard the highest and lowest expression value in each gene group. This method is similar to the jack-knife procedure [103].

**Gene Ontology (GO)-based gene groups.** GO-based gene groups contain the genes associated with a specific GO category as well as all of its child categories (SGD, January 2007). Only Gene Ontology groups with at least 7 members were used for calculation. This approach resulted in a reduction of the original 3836 GO-based groups to 1346 GO-based gene groups, which were used for T-profiler analysis. Significantly scoring GO-based gene groups directly indicate which functions or cellular processes have changed as a result of altered gene expression. We only used categories that showed a significant t-value in at least 5 experiments for our correlation analysis. This excludes gene-groups that showed no coherent regulation. This resulted in 918 GO categories that were used for the correlation analysis.

**MIPS –based gene groups.** MIPS-based gene groups contain genes assigned by the MIPS organization (<ftp://ftpmips.gsf.de/yeast/catalogues>, 15-02-2005). The groups can be divided into three major categories based on function, protein complex and localization. We only used the MIPS function gene groups with at least 7 members for our analysis. Significantly scoring MIPS-based gene groups directly indicate which functions have changed as a result of altered gene expression.

**Motif-based gene groups.** Motif-based groups are defined as groups of genes with a match to a particular consensus motif within 600 base pairs upstream of the ORF [104], allowing no overlap between neighboring ORFs. The consensus motifs used in T-profiler [83] are derived from three different sources. First, motifs were extracted from the SCPD database (<http://cgsigma.cshl.org/jian/>). Additionally, motifs were found by comparing the genome sequence of highly related yeast species [66]. Finally, motifs discovered in various microarray experiments by the REDUCE algorithm [88] were added. Most of these motifs are similar or identical to motifs described in the literature. In total, 153 motif groups have been included in T-profiler calculations.

**Gene groups based on transcription factor binding data.** We used the transcription factor binding (TFB) data, obtained by Harbison *et al.* [62] using ChIP-chip analysis, as input in T-profiler. This data set contains ChIP-chip results of 203 transcription factors from experiments per-

formed in rich medium (YPD). For 84 of these transcription factors, their binding to promoter regions was also measured in at least 1 of 12 other environmental conditions. A gene was considered to be part of a TFB group if the p-value reported by the authors was smaller than 0.001. In addition, TFB groups were required to have at least 7 gene members. This resulted in 252 TFB groups that were used for T-profiler analysis.

#### Data libraries

Our expression library of transcription profiles contains data of 936 hybridization experiments carried out with *S. cerevisiae* from 19 publications (Table 1). This expression library contains data from different DNA-array platforms such as Genefilter, Affymetrix, and spotted slides, and includes experiments with gene deletion strains, synchronized cells for cell cycle analysis, sporulating cells, and cells subjected to various physical and chemical perturbations.

**Table 1 Description of the datasets used for FunKey**

Author	Description	platform
Tai [189]	Anearobic N-C-P-S chemostats	gene expression
Boer [190]	C-S-P-N chemostat limitation	gene expression
Yoshimoto [46]	Calcineurin	gene expression
Daran-Lapujade [191]	Carbon-limited chemostats	gene expression
Spellman [58]	Cell Cycle	gene expression
Lagorce [120]	Cell wall mutants	gene expression
Boorsma [51]	Cell wall perturbants	gene expression
Sahara [192]	Cold shift	gene expression
Murata [193]	Compounds and stress	gene expression
Gasch [194]	DNA damage	gene expression
Gasch [40]	Environmental stress	gene expression
Bro [195]	Lithium response	gene expression
Harris [196]	Map kinase	gene expression
Fleming [197]	Proteasome inhibitor	gene expression
Devaux [198]	regulation by PDR1	gene expression
Hughes [59]	gene deletions and compounds	gene expression
Chu [199]	Sporulation	gene expression
McCammon [200]	TCA cycle mutants	gene expression
Mnaimneh [165]	Titrateable promoter alleles	gene expression
Zakrzewska [65]	Chitosan	fitness
Dudley [201]	Compound & Conditions	fitness
Brown [63]	Compounds	fitness
Wu [202]	DNA - damage	fitness
Parsons [64]	Compounds	fitness
Warringer [186]	Compounds	fitness
Birrel [203]	DNA-damage	fitness
Giaever [8]	Conditions	fitness

The expression library has been analyzed using T-profiler and the data have been uploaded to a database named T-base, which can be found at <http://www.science.uva.nl/~boorsma/T-base-all>. The library of fitness profiles contains data of 159 experiments that were carried using *Saccharomyces cerevisiae*. The data is extracted from 8 different publications. The majority of

the data is derived from the Dudley study, which also includes the  $\log_2$  ratio data from the studies of Giaever, Wu and Birrel. Parsons and Zakrzewska provided  $\log_2$  ratios of their fitness data. The study of Brown performed plate assays of the deletion mutant collection and scanned the intensities and size of all individual colonies. We used  $\log_2$  ratios of the intensities of the treated colonies of each individual mutant and the intensities of the untreated (YPD grown) colonies from the same mutant as a measure of fitness.

### **Correlation analysis and functional predictions**

First we normalized each expression or fitness profile by calculating the z-score, which express the distance of the expression or fitness of a gene towards the mean in units of standard deviation.

$$z = \frac{x - \mu}{\sigma}$$

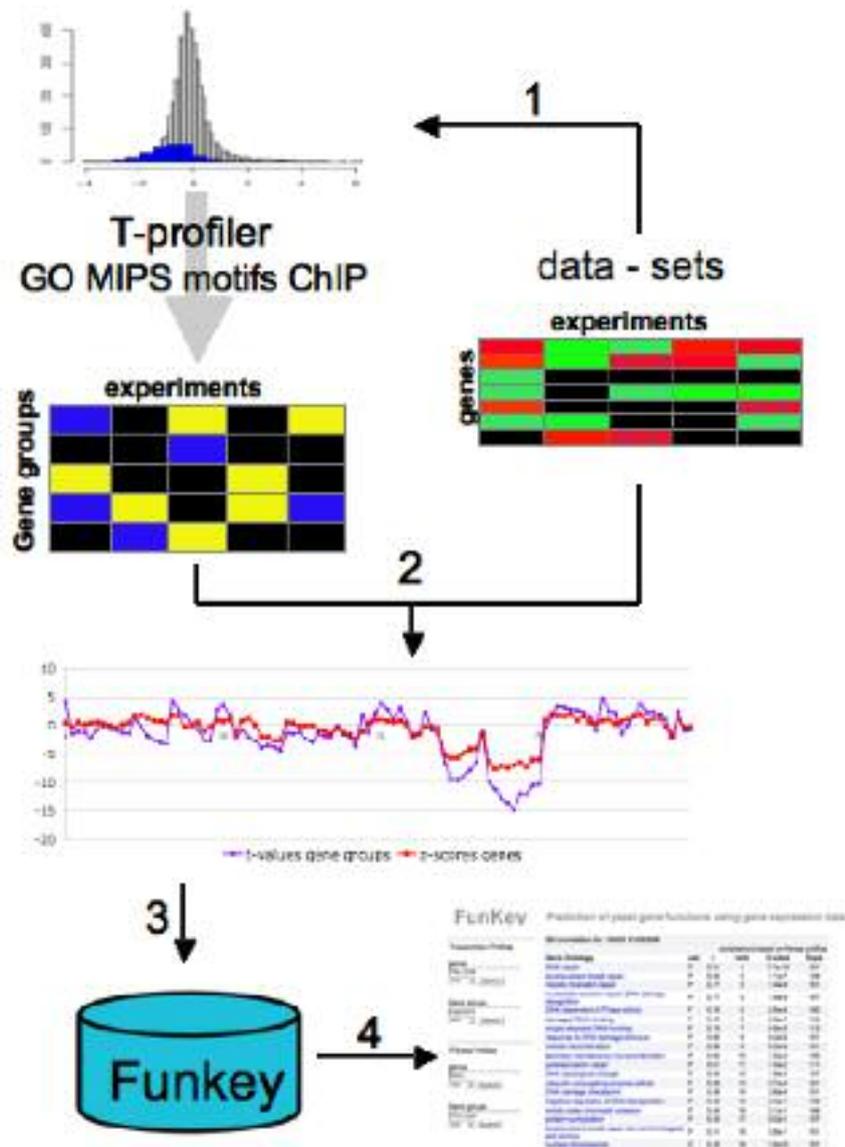
where  $x$  = the log-ratio ,  $\mu$  = population mean and  $\sigma$  = the standard deviation. To quantify the extent to which an individual gene follows the behavior of a given gene group, we computed the Pearson correlation  $r$  between the mRNA expression z-scores of the gene and the t-score of the gene group across all hybridizations in our expression library.

## Results

### Correlation analysis between gene and pathway modulation applied on gene expression and fitness profiles

Recently we introduced T-profiler, a tool that analyses gene expression profiles by applying a t-test to score the changes in the average gene expression of predefined groups of genes. We also demonstrated that the same approach could be applied to data extracted from global fitness profiling experiments [65]. Since the t-statistic is invariant to scaling of the gene expression or fitness profile it allows for comparison of data from heterogeneous data-sources (Boorsma *et al.*, submitted; Chapter 4). In the whole T-profiler procedure no arbitrary cutoffs for the expression of individual genes are used. Instead all genes and all experiments are used, minimizing data loss. To analyze the relationship between the expression or relative fitness of individual genes to all gene-groups we applied T-profiler to a library of expression and fitness profiles that covers 936 mRNA hybridizations and 159 fitness profiling experiments (**Materials & Methods**). We calculated the Pearson correlation  $r$  over all experiments, between either the expression of individual genes or the relative fitness of individual mutants represented by their z-scores within each experiment, (**Materials & Methods**) with the t-values of all gene-groups (**Figure 1**). This resulted in a correlation profile for each gene in which a high  $r$ -value indicates that the z-values of a gene behaved like the t-values of a gene-group over many conditions.

As an example of such an analysis we took the well-characterized gene *GRX2*. In table 2 the 5 GO-terms showing the highest correlation in behavior over all experiments are shown. According to the Saccharomyces Genome Database (SGD) [204] *GRX2* is functionally annotated as “glutathione peroxidase activity”, “glutathione transferase activity”, and “thiol-disulfide exchange intermediate activity” in the category molecular function. The GO-group of “glutathione peroxidase activity” has 6 members while the minimum for T-profiler analysis is 7. This group is therefore not analyzed, but the parent group “peroxidase activity” is the group with the highest correlation to *GRX2* ( $r = 0.76$ ). The 2<sup>nd</sup> annotated group has an  $r$  of 0.69, and is 9<sup>th</sup> in the list. The group “thiol-disulfide exchange intermediate activity” has significant t-values in fewer than 5 experiments, and is therefore not included in the analysis, but again the parent group (disulfide oxidoreductase activity) has an  $r$  of 0.65 and is ranked 10<sup>th</sup> in the list. The 2 annotated GO-ontologies for biological process (“regulation of cell redox homeostasis” and “response to oxidative stress”) are ranked 4<sup>th</sup> and 7<sup>th</sup> in the correlation profile, with  $r$ -values of 0.75 and 0.71, respectively.



**Figure 1. Schematic overview of the generation of correlation profiles.** A library of 936 gene expression and 159 fitness profiles is analyzed by T-profiler using GO and MIPS gene groups. In addition the gene expression profiles were also analyzed using Motif and ChIP gene groups (only using gene expression data) (1). The obtained t-values for each gene group (blue line) over all conditions are compared by correlation analysis with the z-score (based on expression or on relative fitness) of individual genes (red line) (2). The correlation values for each individual gene between all gene groups (correlation profiles) are stored in a database, (3) which can be queried to generate functional predictions ([www.science.uva.nl/~boorsma/funkey](http://www.science.uva.nl/~boorsma/funkey)) (4).

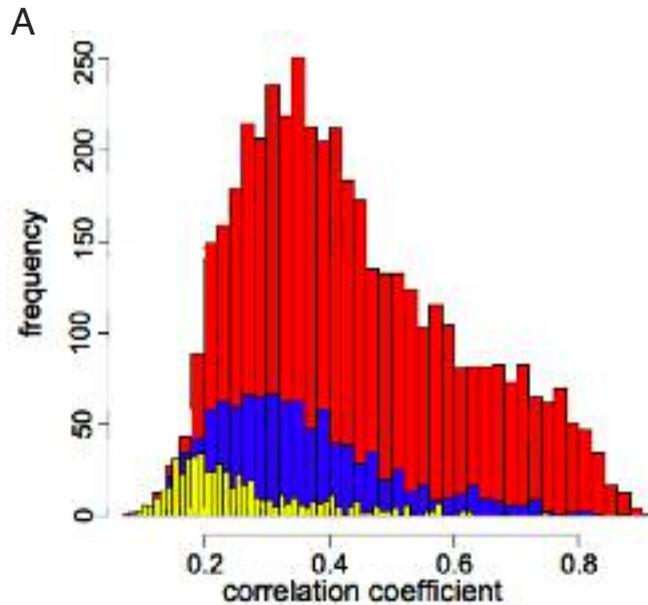
**Table 2. Gene Ontology based correlation profile for the characterized gene *GRX2*.** The five highest and lowest ranked positive and negative gene group correlations, based on Gene Ontologies are shown. The complete prediction profile, based on all experimental conditions, contains values for 918 gene ontology gene groups can be found on [www.science.uva.nl/~boorsma/funkey](http://www.science.uva.nl/~boorsma/funkey). All top 10 GO-categories are closely related to these 5 annotations. For cellular component, both “cytosol” and “mitochondrion” have been annotated. Many mitochondrial gene groups correlate strongly with *GRX2* over all experiments, starting at rank 13 with an *r*-value of 0.6, but cytosolic groups do not appear until the 75<sup>th</sup> position in the list

Gene Ontology	r	E-value	rank
peroxidase activity	0.76	< 1.0e-14	1
oxidoreductase activity, acting on peroxide as acceptor	0.76	< 1.0e-14	2
antioxidant activity	0.75	< 1.0e-14	3
regulation of cell redox homeostasis	0.75	< 1.0e-14	4
cell redox homeostasis	0.75	< 1.0e-14	5
snoRNA binding	-0.55	< 1.0e-14	913
RNA methylation	-0.56	< 1.0e-14	914
tRNA methyltransferase activity	-0.56	< 1.0e-14	915
RNA helicase activity F	-0.57	< 1.0e-14	916
helicase activity	-0.58	< 1.0e-14	917
RNA methyltransferase activity	-0.58	< 1.0e-14	918

### Global analysis of the correlation profiles according to SGD gene classification

For many genes we observed that the functions correlating strongest were amongst or related to the annotations as found in SGD. This is true only if a gene has been annotated, but SGD categorizes genes as ‘verified’ if they have a functional annotation (4471 genes), as uncharacterized if they do not (1039 genes), and as ‘dubious’ genes (560 genes). The latter category contains genes that were initially predicted as open reading frames, but of which sequence comparison with closely related yeast species [66] indicates that they most likely are not. We compared the correlation profiles for the genes of the three classes. Figure 2 shows the distributions of the highest correlation coefficients for each gene of the three gene classes. The distribution of the correlation coefficients of the verified genes peaked around correlation coefficients of 0.7 and 0.35, while the class of uncharacterized genes peaked around 0.6 and 0.3. When we assessed the number of genes above a certain correlation cutoff, a similar picture was seen. The percentage of genes that had a correlation of lower or equal to 0.5 with at least one GO-term for verified genes was 35% and dropped to 21% at a cut-off value of lower or equal to 0.6. For the uncharacterized genes the number was lower; an *r*-value of 0.5 or

higher was found for 16% of the genes, and this number dropped to 8% when a correlation cut-off of 0.6 is used. Finally, the distribution of the dubious genes peaked around a correlation coefficient of 0.2, and the number of correlations for the dubious genes was much lower than for either verified or uncharacterized genes (**Figure 2AB**).



**B**

Genes	SGD	$(r \geq 0.3)$	$(r \geq 0.4)$	$(r \geq 0.5)$	$(r \geq 0.6)$
Verified	4471	3491 ( <b>78%</b> )	2482 ( <b>56%</b> )	1565 ( <b>35%</b> )	954 ( <b>21%</b> )
Uncharacterized	1039	627 ( <b>60%</b> )	353 ( <b>34%</b> )	165 ( <b>16%</b> )	87 ( <b>8%</b> )
Dubious	560	164 ( <b>29%</b> )	90 ( <b>16%</b> )	45 ( <b>8%</b> )	16 ( <b>3%</b> )

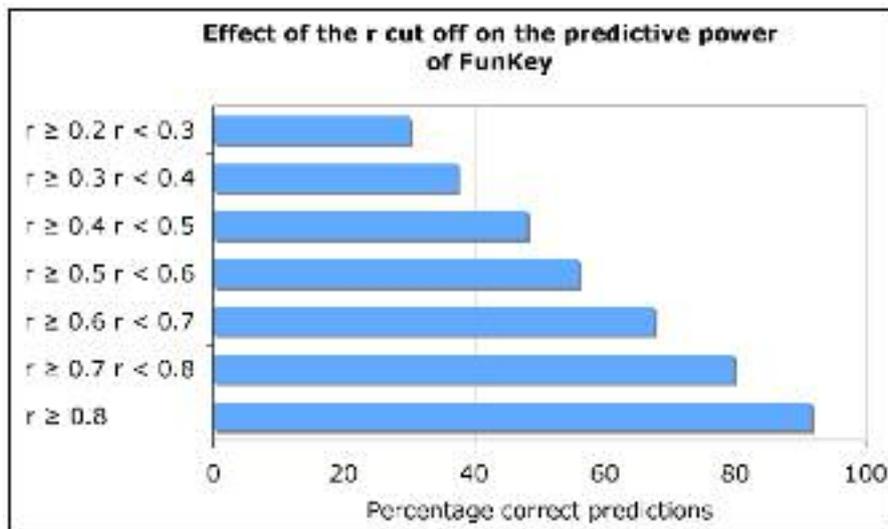
**Figure 2. Distributions of correlation coefficients according to SGD gene classification. (A)** The frequency distribution of the highest correlation coefficient for each gene correlation profile is plotted. In grey, the distributions for the verified genes, dark grey, the distributions for the uncharacterized genes and dubious genes (all according to the *Saccharomyces cerevisiae* Genome Database) are shown in light grey. **(B)** Effect of the Pearson correlation cut-off on the number of functional gene predictions of GO functions. The number of correlations with a certain correlation cutoff is given. In bold is the percentage of correlations for a certain correlation cutoff. Note that the dubious genes still have a substantial number of correlations with a high correlation coefficient.

Since the dubious genes are measured in most of the gene expression and fitness data sets but not expected to behave like functional genes, they can be regarded as background correlation. Concluding, the verified genes on average showed highest correlations with functional groups, reflected both in the distribution of correlations and in the percentage of correlations found above a certain correlation. Still, a considerable number of uncharacterized genes had a high correlation with one or more functional groups.

### Assessment of correlation as a functional predictor

We tested whether high correlation of genes to gene groups favored the groups to which they have been shown to belong. If this is the case, gene-gene group correlation can be used to assess and predict gene function. Therefore, a set of well-characterized genes that were annotated to at least one group in all three GO classes (molecular function, biological process and cellular component) was selected. In total, 3400 genes fit these criteria and were used in this analysis. For each gene we selected the GO-term with the highest correlation coefficient ( $r$ ) and compared this to the GO-terms to which the gene was assigned according to Gene Ontology (GO). A prediction was considered “correct” only when the highest GO-term matched a GO-term assigned. These criteria are rather strict, since in some cases genes were not annotated to the most highly ranked GO term but of the one ranked later, even if the highest term was highly related. Such a situation was shown in the *GRX2* example above.

**Figure 3** shows the effect of the correlation coefficient cutoff on the quality of the functional predictions. In total the highest GO-term correlations of 50% of the genes (1695 out of 3400) matched one of the assigned GO-terms. If a rank equal or better than 5 was used, 68% of the genes matched. When we analyzed only genes with a correlation of 0.8 or higher, we predicted the assigned GO gene group in 92% of the cases, and even at low correlation many predictions are still correct according to these criteria. Using a correlation cutoff of 0.5, 68% of gene functions were predicted correctly, and this number increased to 78% if functional groups correlating with an  $r$  equal or more than 0.5 but positioned at a rank  $\leq 5$  were included



**Figure 3. Effect of the  $r$ -cut off on the predictive power of the correlation profiles**

We assessed the quality of the correlation profiles by taking a set of well-characterized genes (3400). The correlation value of the highest ranked gene group of each gene was taken. A correct correlation was considered when the gene was actually a true member of this gene group based on Gene Ontology information. **Figure 3** shows the differences in the used  $r$ -cutoffs and the correct gene group correlations. Information of all 3400 genes can be found in the additional information.

Based on the analysis above, we now regarded groups with correlations of 0.5 or higher as specific hypotheses for the function of genes, with a 68% chance of predicting a correct GO-ontology gene-group on top of the list, and a 78% chance of it being amongst the first 5. We used the correlation profiles to generate predictions for all dubious and uncharacterized genes. For 45 dubious assigned genes correlation of 0.5 or higher was found with a GO-term. However, when the chromosomal position of these dubious genes was inspected we found that in about 90% of the cases the genes physically overlapped other (verified) genes. Since the majority of the micro-array experiments present in our expression library were based on spotted PCR products of whole genes, the DNA from such a spot was also able to hybridize to mRNA from the overlapping complementary ORF. Interestingly, we found four dubious genes that showed high correlation but did not overlap with another, characterized gene (**Table 3**). These might represent valid genes that are specific for *Saccharomyces cerevisiae*.

**Table 3. Functional predictions of ‘dubious’ ORFs** Functional predictions of ‘dubious’ ORFs that do not show physical overlap on the chromosome with other ORFs. These ORFs might represent true genes that are specific for *Saccharomyces cerevisiae*.

ORFs	<i>r</i>	GO-term
YGL188C	0.56	cytochrome-c oxidase activity
YAR075W	0.51	amino acid derivative biosynthesis
YMR103C	0.51	structural constituent of cell wall
YOL118C	0.51	urea cycle intermediate metabolism

In total 165 out of 1039 ‘uncharacterized’ genes have a correlation coefficient of 0.5 or higher based on Gene Ontology. **Table 4** shows the functional prediction for these genes. GO-terms include arginine biosynthesis, organellar (mitochondrial) ribosome, lysine metabolism, proteasome complex and histone exchange. The majority of the genes are however associated with GO-terms related to ribosome biogenesis (snoRNA binding, rRNA binding and processing of 27S pre-rRNA), which is in agreements with findings from Hughes *et al.* [6].

Orf	<i>r</i>	GO-id	description
YBR047W	0.69	000051	urea cycle intermediate metabolism
YPL250C	0.67		
YIL165C	0.62		
YJR111C	0.58		
YPL264C	0.55		
YER049W	0.74	000054	ribosome export from nucleus
YNL119W	0.67		
YHL039W	0.65		
YPL207W	0.64		
YPL245W	0.52		
YOR154W	0.53	000055	ribosomal large subunit export from nucleus
YOL007C	0.55	000079	regulation of cyclin-dependent protein kinase activity
YLR049C	0.51		
YHR122W	0.59	000105	histidine biosynthesis
YDR115W	0.76	000313	organellar ribosome
YNL081C	0.66		

YKL137W	0.59	000314	organellar small ribosomal subunit
YPL183W-A	0.50	000315	organellar large ribosomal subunit
YNL305C	0.66	000328	vacuolar lumen (sensu Fungi)
YNL115C	0.63		
YHR138C	0.58		
YOR220W	0.54		
YNL155W	0.72	000502	proteasome complex (sensu Eukaryota)
YBR062C	0.66		
YOR059C	0.52		
YNL313C	0.73	003724	RNA helicase activity
YJL010C	0.82	004004	ATP-dependent RNA helicase activity
YMR315W	0.73	004033	aldo-keto reductase activity
YNL134C	0.71		
YBR053C	0.70		
YBR056W	0.69		
YMR110C	0.66		
YKR011C	0.63		
YDR391C	0.60		
YHR112C	0.54		
YBR204C	0.53		
YGR111W	0.53		
YGL259W	0.60	004190	aspartic-type endopeptidase activity
YHR209W	0.59		
YMR090W	0.71	004364	glutathione transferase activity
YOL083W	0.67		
YOR289W	0.63		
YOL048C	0.51		
YMR251W	0.50		
YLR194C	0.65	005199	structural constituent of cell wall
YIL108W	0.57		
YKR046C	0.57		
YBR071W	0.54		
YOL022C	0.72	005666	DNA-directed RNA polymerase III complex
YOR021C	0.70		
YIL110W	0.69		
YGR173W	0.65		
YIL064W	0.62		
YDL213C	0.52		
YCR095C	0.50	005681	spliceosome complex
YOR287C	0.74	005730	nucleolus
YJR124C	0.62	005736	DNA-directed RNA polymerase I complex
YCR087C-A	0.61		
YEL048C	0.55		
YOL092W	0.55		
YDR493W	0.54	005759	mitochondrial matrix
YCL042W	0.64	005775	vacuolar lumen
YBR269C	0.57		
YBR241C	0.54		
YBR025C	0.50	005829	cytosol
YMR067C	0.59	005839	proteasome core complex (sensu Eukaryota)
YNL200C	0.72	005975	carbohydrate metabolism
YJL161W	0.68		
YJL163C	0.63		
YFR017C	0.54		
YMR291W	0.52		
YER067W	0.70	005977	glycogen metabolism
YLR345W	0.65		
YJR008W	0.62		
YGR243W	0.54		
YOR215C	0.52		

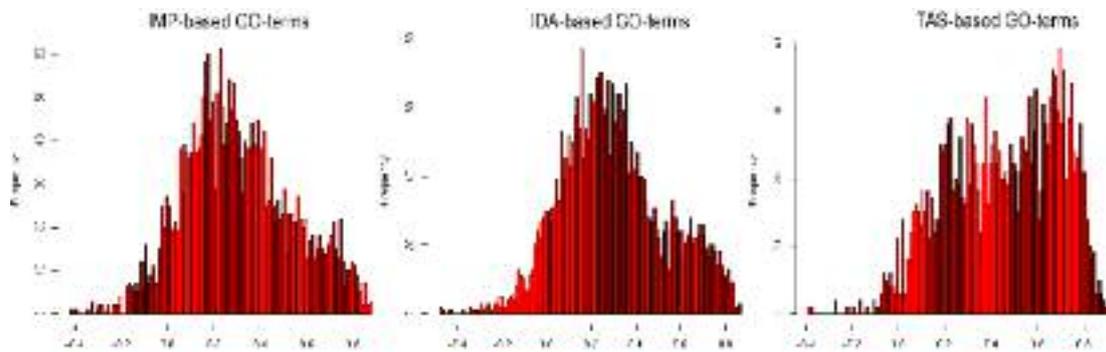
YGR052W	0.51	
YMR196W	0.61	005991 trehalose metabolism
YLR149C	0.73	006112 energy reserve metabolism
YHL021C	0.72	
YPL230W	0.66	
YER079W	0.64	
YGR130C	0.59	
YJL057C	0.53	
YLR177W	0.50	
YJL048C	0.52	006118 electron transport
YLR243W	0.62	006360 transcription from RNA polymerase I promoter
YGR145W	0.87	006364 rRNA processing
YIL127C	0.79	006396 RNA processing
YGR251W	0.64	
YPR169W	0.64	
YLR401C	0.68	006400 tRNA modification
YLR405W	0.52	
YKL033W-A	0.59	006520 amino acid metabolism
YGL117W	0.84	006526 arginine biosynthesis
YHR162W	0.56	
YJL200C	0.74	006553 lysine metabolism
YGR090W	0.64	006611 protein export from nucleus
YLR267W	0.59	006766 vitamin metabolism
YIL056W	0.57	006768 biotin metabolism
YJR154W	0.50	
YLR301W	0.50	006888 ER to Golgi vesicle-mediated transport
YPL166W	0.50	006914 autophagy
YDR196C	0.51	007034 vacuolar transport
YOL125W	0.62	008175 tRNA methyltransferase activity
YLR089C	0.56	008483 transaminase activity
YLR152C	0.51	
YLR218C	0.55	008535 cytochrome c oxidase complex assembly
YCR082W	0.51	008541 proteasome regulatory particle, lid subcomplex (sensu Eukaryota)
YGR110W	0.50	008614 pyridoxine metabolism
YOL107W	0.56	008654 phospholipid biosynthesis
YGR026W	0.50	
YDR330W	0.53	009056 catabolism
YMR321C	0.70	009073 aromatic amino acid family biosynthesis
YLR179C	0.50	009092 homoserine metabolism
YGR250C	0.52	009250 glucan biosynthesis
YDR070C	0.67	009269 response to desiccation
YJL144W	0.67	
YGR043C	0.60	
YNL195C	0.52	
YGR201C	0.50	
YMR181C	0.58	009415 response to water
YPL247C	0.55	009628 response to abiotic stimulus
YHR097C	0.54	
YML131W	0.64	009636 response to toxin
YKL071W	0.52	
YJR085C	0.59	015036 disulfide oxidoreductase activity
YFR042W	0.54	
YDR161W	0.73	016423 tRNA (guanine) methyltransferase activity
YBR271W	0.71	
YMR310C	0.67	
YLR073C	0.66	
YLR063W	0.63	
YBR030W	0.57	
YDR020C	0.55	

YGR079W	0.54	
YLR356W	0.61	016491 oxidoreductase activity
YBR187W	0.50	018193 peptidyl-amino acid modification
YGR149W	0.61	019203 carbohydrate phosphatase activity
YDL027C	0.57	
YPR172W	0.55	019321 pentose metabolism
YGR127W	0.54	
YBR147W	0.66	019794 nonprotein amino acid metabolism
YDR412W	0.79	019843 rRNA binding
YLR003C	0.79	
YGR272C	0.74	
YCR016W	0.76	030489 processing of 27S pre-rRNA
YIL096C	0.75	
YNL022C	0.71	
YKR060W	0.70	
YJR003C	0.69	
YLR287C	0.61	
YLR409C	0.82	030515 snoRNA binding
YDL063C	0.81	
YPL183C	0.74	
YIL091C	0.70	
YJL069C	0.63	
YMR259C	0.61	
YBR238C	0.55	
YLR413W	0.57	030684 preribosome
YML018C	0.53	
YBL054W	0.66	030687 nucleolar preribosome, large subunit precursor
YLR414C	0.53	031505 cell wall organization and biogenesis (sensu Fungi)
YJL097W	0.54	042175 nuclear envelope-endoplasmic reticulum network
YBR231C	0.71	043486 histone exchange
YBR052C	0.63	044248 cellular catabolism

**Table 4. Functional predictions for uncharacterized genes.** Functional predictions for ‘uncharacterized’ genes based on GO-terms ranked by their highest correlation coefficient.

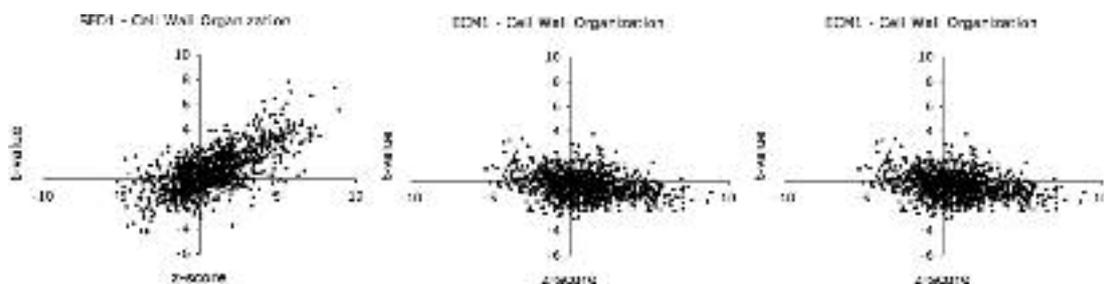
### Analysis of gene group-gene correlations to improve annotation

The previous examples were all based on the prediction of functions of individual genes. It is also possible to assess correlations on the gene group level. In many groups, some gene members showed poor correlation to the group they have been assigned to. This might indicate that these genes have been incorrectly annotated. Conversely, some groups have non-members that correlate closely. Assignment of genes to gene groups is based on different types of evidence. Therefore, we assayed the correlation strengths of genes annotated to functional groups based on three different types of evidence. **Figure 4** shows the frequency distribution of the correlations of all GO-terms for genes with the evidence codes TAS (Traceable Author Statement), IDA (Inferred from Direct Assay) and IMP (Inferred from Mutant Phenotype). Gene annotation based on Traceable Author Statement is generally more solidly evidence based than annotation based on IDA and IMP, which often originated from large-scale analyses. Indeed, genes associated with GO-terms based on TAS correlate much better with their associated terms than those based on IDA and IMP evidence codes (**Figure 4A,B and C**).



**Figure 4. Distribution of correlation coefficients based on the three different GO evidence codes.** Distribution of the correlation coefficients of all GO-terms of all genes with evidence codes A) Inferred from Mutant Phenotype (IMP), B) Inferred from Direct Assay (IDA) and C) Traceable Author Statement (TAS).

This indicates that annotation which is most strongly evidence based also leads to higher correlation, which makes it safe to assume that correct functional annotation should be reflected in correlation. As an example of such a group-centered analysis we describe the analysis of the GO-term ‘cell wall organization and biogenesis’ (GO:007047). This gene group contains 196 gene members correlating with correlation coefficients ranking from 0.59 to -0.30. For example, *SED1*, a gene that codes for a structural GPI-cell wall glycoprotein scores a high correlation ( $r = 0.58$ ) (**Figure 5A**) whereas the gene *ECM1* scored the lowest correlation with the cell wall organization and biogenesis GO gene group ( $r = -0.30$ ) (**Figure 5B**). The evidence that *ECM1* belongs to this GO gene group originates from a transposon mutant screen for hypersensitivity against the cell wall perturbing agent Calcofluor White [205]. In contrast to its low correlation with the ‘cell wall organization and biogenesis’ group, *ECM1* showed the highest positive correlation for the GO term ‘ribosomal large subunit export from nucleus’ ( $r = 0.82$ ) (**Figure 5C**).

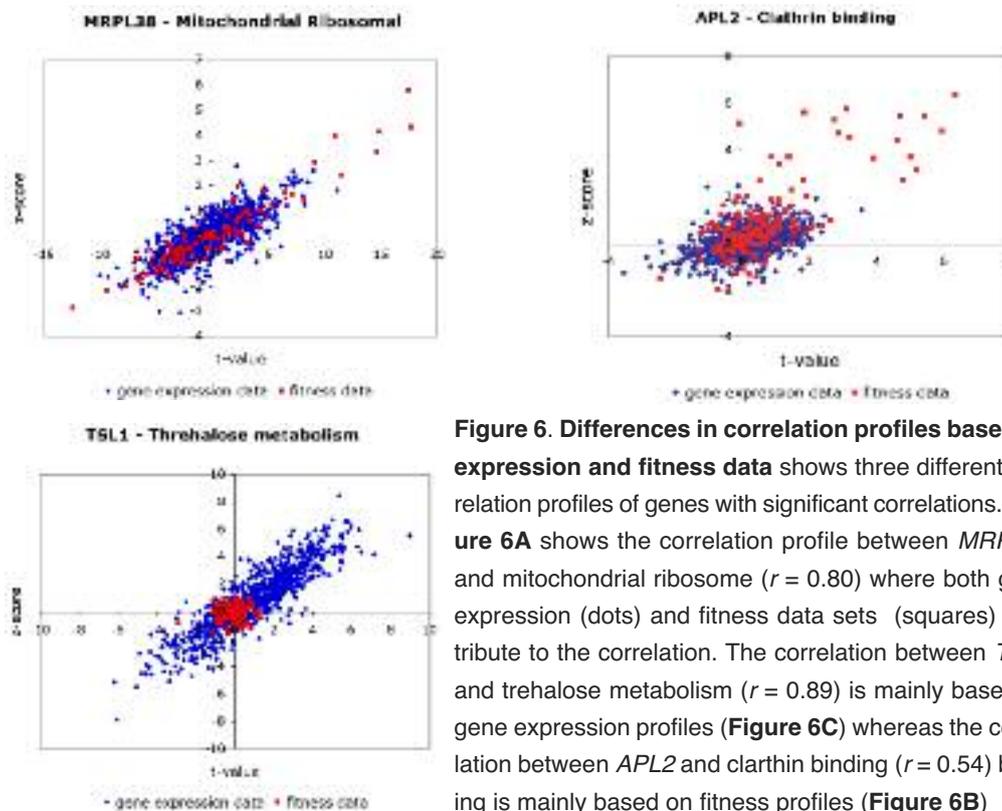


**Figure 5. Scatter plot of correlation profiles of two gene members of the cell wall organization GO gene group (*SED1* and *ECM1*).** Figure 5 shows the correlation profile, visualized by a scatter plot, between the z-scores of genes (based on gene expression of fitness profiles) and the t-values of gene ontology gene groups over 1100 experiments. Figure 5A *SED1* and the GO-term of ‘Cell Wall Organization’,  $r = 0.58$ . Figure 5B; *ECM1* and the GO-term of ‘Cell Wall Organization’,  $r = -0.30$ . Figure 5C; *ECM1* and the GO-term of ‘Ribosomal export from nucleus’,  $r = 0.80$ .

This newly predicted function is strongly supported by the localization of *ECM1* in the nucleoplasm and the nucleolus and genetic interaction with *MTR2*, which is involved in 60S ribosomal protein subunit export [71]. Ten other ECM genes in this GO-group (based on mutant phenotypes) have a negative correlation, and a strong positive correlation with other functional groups and are therefore unlikely to be directly involved in cell wall biogenesis. Additionally, 4 genes (*SRL3*, *CRG1*, *YPS1* and *YIL108w*) not assigned to this GO-group correlate to it with an  $r$  of 0.5 or higher. The example above suggests that our correlation analysis is useful in discrimination between correct and incorrect annotation.

### Expression and fitness profiling data are complementary in identifying gene functions

We used combined gene expression and fitness profile data, but not in all cases do the two independently give the same result. For example, *MRPL38*, a gene coding for a mitochondrial ribosomal protein correlates with the gene group mitochondrial ribosome with an  $r$  of 0.81, to which both fitness and transcription data contribute (figure 6a). A different picture is seen for the gene *TSL1*; its correlation of 0.89 with the GO-term 'Trehalose metabolism' is entirely based on gene expression data sets (**Figure 6b**). A reverse pattern is found for the gene *APL2*; correlation between *APL2* and the GO-term 'Clathrin binding' is mostly based on the fitness data (Figure 6c). The latter two examples suggest that expression profiling studies and fitness profiling studies used in FunKey are complementary in identifying gene functions.



**Figure 6. Differences in correlation profiles based on expression and fitness data** shows three different correlation profiles of genes with significant correlations. **Figure 6A** shows the correlation profile between *MRPL38* and mitochondrial ribosome ( $r = 0.80$ ) where both gene expression (dots) and fitness data sets (squares) contribute to the correlation. The correlation between *TSL1* and trehalose metabolism ( $r = 0.89$ ) is mainly based on gene expression profiles (**Figure 6C**) whereas the correlation between *APL2* and clathrin binding ( $r = 0.54$ ) binding is mainly based on fitness profiles (**Figure 6B**)

If genes have strong correlations in both transcription and fitness data, but correlate to different functions in the different datasets, this indicates dual functionality. An example is the TCA cycle gene *ACO1*. Based on gene expression correlations profiles, the GO-term glutamate biosynthesis and tricarboxylic acid cycle score the highest correlation coefficients ( $r = 0.74$  and  $r = 0.68$ , respectively). If only fitness profiles are used, the GO-term mitochondrial genome maintenance ( $r = 0.58$ ) shows the highest correlation. This is in accordance with a recent finding that describes Aco1p to function in mitochondrial DNA maintenance [206].

### **Gene prediction webtool: FunKey**

We created a web-application named FunKey, which can be used to query the correlation profiles for all *Saccharomyces cerevisiae* genes ([www.science.uva.nl/~boorsma/Funkey](http://www.science.uva.nl/~boorsma/Funkey)). Since Gene Ontology and MIPS describe the physiological role of gene groups they are the most suitable to generate functional predictions. As output we provide the correlation of the query gene to all GO-ontology gene groups. As described in the analysis above, lower correlations have less predictive power. In addition, we applied our method to Motif and Transcription Factor Occupancy (TFO) gene groups. The correlation profiles for the latter gene groups provide information about the transcriptional regulation of a gene and are therefore based on gene expression data only. Besides correlation profiles based on the combined expression and fitness dataset, it is also possible to query FunKey for the two independently. Additionally, it is possible to generate output for gene groups, so that all genes correlating to a gene group can be analyzed. The user can define cutoffs for correlation strength and rank.

## Conclusions & Discussion

There is a need for methods that can, based on the use of high-throughput data, generate useful hypotheses about the function of uncharacterized or poorly characterized genes [6]. Here we present such a method that integrates gene expression and fitness profiles to predict the function of uncharacterized or poorly characterized genes. Unlike existing methods our method first measures the modulation of functionally related genes by means of t-statistics, and next uses this information to predict the function of a gene based on the correlation of the behavior of individual genes to that of gene groups throughout our data library. The first step in the procedure can be regarded as a normalization step where modulation of individual genes is transferred into t-values on the gene groups level. This transformation allows the integration of heterogeneous data sets. The results of our method are available through webservice FunKey that allows for efficient data mining and hypothesis generation.

The predictive power of the correlation profiles was assessed in two ways: First we analyzed the frequency distributions of correlation profiles from verified, uncharacterized and dubious genes. This analysis revealed that the characterized genes in general show higher correlations to at least one gene group than the uncharacterized genes. This may be because gene groups were generated based on knowledge of the genes functioning in them. However, it also shows that uncharacterized genes are not as easily assigned to functional groups. Next we compared the GO annotations of a set of well-characterized genes with the predictions generated by FunKey. This revealed that, when no correlation cutoffs were used, we were able to make correct predictions for 68% of the genes. Using a strict correlation cut-off ( $r > 0.5$ ), predictions were generated for 165 (16%) of the uncharacterized genes. These findings suggest that, in proportion, the functions of uncharacterized genes are harder to predict using gene expression and fitness profile data sets. The method could be improved by adding new gene expression and fitness profiling dataset; therefore GO-terms that are found to be not modulated in the dataset might help to suggest, which experimental conditions should be added. Finally, the group of dubious genes shows very little correlation, and therefore serves as a measure for background correlation with no functional significance. Interestingly, four dubious genes, not overlapping with other genes, showed significant correlation with a GO-term and could be *Saccharomyces cerevisiae* specific genes.

We also show that the origin of gene annotations strongly influences the distribution of the correlation coefficients. For example, genes that are annotated according to the 'Traceable Author Statement' (TAS) perform much better than genes that are annotated according to the 'Inferred from Mutant Phenotype' (IMP) evidence code. This suggests that annotation of genes via TAS is more accurate than IMP annotated genes. An example is shown of genes that belong to the GO-term 'cell wall organization and biogenesis' that show large variation of correlation coefficients. Most of the genes that have poor correlation to this GO-term are annotated with an IMP evidence code and originate from high-throughput assays [205], and do correlate strongly to other GO-ontology categories. We currently are re-evaluating these mutants for their sensitivity to cell wall perturbants. Generally, genes correlating poorly to their annotated gene group,

and correlating strongly to others should be looked into.

We also took the opportunity to compare separate correlation profiles based on gene expression profiles and fitness profiles. Only a portion of the genes (196 genes,  $r \geq 0.5$ , rank  $\leq 10$ ) share predictions of gene function based on gene expression and fitness profiles. Most of these genes function in cytosolic and mitochondrial ribosomal biogenesis, functions which are probably most strongly transcriptionally regulated (our unpublished data). This shows that fitness profiling is complementary to the use of gene expression profiles. Interestingly, we are able to predict two distinct functions for the TCA-cycle gene *Aco1p*, which confirms earlier findings that *Aco1p* has a double function in the TCA-cycle and mitochondrial DNA maintenance [206]. Mining FunKey might reveal more of such multifunctional genes.

Our method uses data of all expression and fitness profiles and makes no prior selection. Huttenhower *et al.* (Huttenhower *et al.*, 2006) suggested that selection based on meta-data about experiments might improve the predictive power for functions whose effects are under-represented in the dataset. For example, there are only 9 gene expression profiles of sporulation conditions. Still, our method is capable of generating reliable functional predictions for genes involved in the sporulation, especially if also the rank is taken in account. Similarly, although no datasets specifically targeting peroxisomal functions are present in our dataset, we still find significant correlations of peroxisomal genes to peroxisomal gene groups. A pre-selection of experiments can also be dangerous, and might lead to accumulation of (positive) errors and therefore generate seemingly solid but false predictions

In summary, we have presented a conceptually simple and transparent method that can be used to generate specific hypothesis of uncharacterized or poorly characterized *Saccharomyces cerevisiae* genes. Our method is scalable and can easily be used to improve the annotation of the genome of *Saccharomyces cerevisiae* and other organisms.