Analysis of Genome-Scale Data

Analysis of Genome-Scale Data
Patrick Kemmeren
Utrecht, Universiteit Utrecht, Faculteit Geneeskunde
Proefschrift Universiteit Utrecht, met een samenvatting in het Nederlands

ISBN 90-393-4099-4
Cover Design Patrick Kemmeren
Reproduction Ponsen & Looijen B.V.

Analysis of Genome-Scale Data

Analyse van genoomschaal data

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht
op gezag van de Rector Magnificus, Prof. Dr. W.H. Gispen,
ingevolge het besluit van het College voor Promoties
in het openbaar te verdedigen op
dinsdag 6 december 2005 des middags te 14.30 uur

door

Patrick Petrus Cornelis Wilhelmus Kemmeren
Geboren op 9 september 1976 te Waalwijk

Promoter Prof. Dr. P.C. van der Vliet
Copromoter Dr. F.C.P. Holstege

Department of Physiological Chemistry
Division of Biomedical Genetics
University Medical Center Utrecht
Utrecht University, the Netherlands

The work presented in this thesis was financially supported by the UMC Utrecht “Genvlag” program and by grants from the Netherlands Organization for Scientific Research (NWO), Horizon and BioRange. The work was carried out within the Utrecht Graduate School for Developmental Biology (OOB).

Financial support for the publication of this thesis by Dell B.V., the J.E. Jurriaanse Stichting and the University Medical Center Utrecht is gratefully acknowledged.

Anyone who has never made a mistake has never tried anything new
Albert Einstein

Contents

Chapter 1	General Introduction	9
Chapter 2	OffBase - an Integrated Microarray Data Analysis Platform	31
Chapter 3	Monitoring Global Messenger RNA Changes in Externally Controlled Microarray Experiments	45
Chapter 4	Expression Profiler: Next Generation - an Online Platform for Analysis of Microarray Data	57
Chapter 5	Expression Profiler: Next Generation - Discerning Regulatory Mechanisms behind Regulatory Processes	67
Chapter 6	Protein Interaction Verification and Functional Annotation by Integrated Analysis of Genome-Scale Data	73
Chapter 7	Predicting Gene Function through Systematic Analysis and Quality Assessment of High-Throughput Data	89
Chapter 8	General Discussion	103
Appendix	Color Figures	111
	Summary	129
	Samenvatting	133
	Dankwoord	137
	Curriculum Vitae	140
	Publications	141

General Introduction

Chapter I

Part of this introduction has been published in

Biochemichal Society Transactions 2003 Dec;31(Pt 6):1484-7

General Introduction

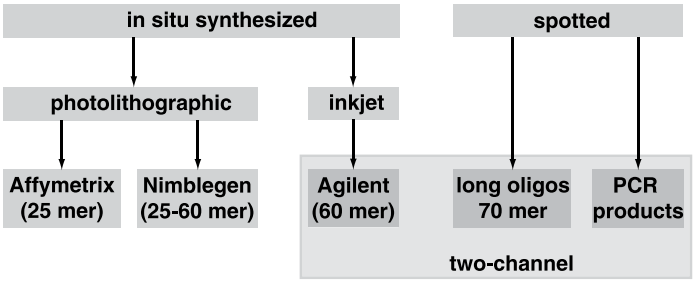
The availability of whole genome sequences has given rise to the parallel development of other high-throughput approaches such as determining mRNA expression level changes, gene-deletion phenotypes, chromosomal location of DNA binding proteins, cellular location of proteins, protein-protein interactions, biological activity using protein arrays, synthetic lethality with combinations of mutants and RNA interference phenotypes. Such functional genomic datasets can be used in a variety of ways. For example, microarray gene expression data has been applied to predict the function of genes using mutant expression-profiling "phenotypes", to identify possible drug targets by exposing cells to certain pharmacological agents, to improve cancer diagnosis and prognosis and to decipher complex biological systems such as transcriptional regulatory networks. Many challenges are encountered when analyzing such data for biological or biomedical purposes. These include data management, data preprocessing, data normalization, differences in data quality within and between datasets, integration of different types of genome-scale data, prioritization of hypotheses and improvement of existing annotation. This chapter reviews these challenges, their implications on biological and biomedical research and introduces the goals of the work described in this thesis.

Microarrays for Expression Profiling

The central dogma in biology describes the flow of information from DNA to RNA and finally to the translation of proteins, which in turn form complex biological systems. Based on this assumption, many scientists investigate protein function at the RNA level. DNA microarrays are frequently used to investigate the mRNA expression levels of many genes in parallel, often referred to as "expression profiling". A DNA microarray is a collection of DNA samples attached to a solid surface, such as a nylon membrane, glass or silicon chip.

Several distinct DNA microarray technologies for expression profiling are in use at the moment (Figure 1). DNA arrays can be produced through in situ synthesis using photolithographic methods (Affymetrix; Nimblegen)[1-4] or inkjet technology (Agilent)[5-7]. Alternatively, prefabricated DNA molecules, such as PCR products (e.g. cDNA) or long oligos, can be printed on glass slides or nylon membranes using robots[8,9]. Each technology has specific advantages and disadvantages. For instance, the photolithographic technique applied by Affymetrix uses 25-mer oligos. To obtain reliable mRNA expression levels multiple probes and mismatch probes per gene are present on the array. This poses the challenge of summarizing data from this probe set into a single measure that estimates the mRNA expression level. As a consequence, different low-level analysis methods exist that can have a large impact on the final outcome of the results[10-17]. A disadvantage of both the photolithographic technique used by Affymetrix[3] and the spotted arrays[8,9], is the relatively inefficient or expensive creation of new array designs, as this requires either a new set of photolithographic masks or manufacturing of a complete collection of new DNA molecules. Greater flexibility is provided through the technologies used by both Nimblegen and Agilent that use a maskless light-directed oligonucleotide synthesis[4] or inkjet printing technology[5,6] respectively. An advantage of all long oligonucleotide platforms is that they provide greater specificity than cDNA arrays, thus having the potential of distinguishing single-nucleotide polymorphisms or splice variants[18-20]. For two-channel platforms (Agilent; spotted arrays) additional biases may be introduced related to the two different dyes used in this setup and

Figure 1 | Different Microarray Technology Platforms
Schematic overview of different technology platforms used for microarray expression profiling. Differentiation can be made in terms of the manufacturing process: in situ, spotted, photolithographic and inkjet technology; and how samples are compared, e.g. one channel versus two channel comparisons.



need to be accounted for during subsequent filtering and normalization processes[21,22].

These technological platforms have been used in a variety of ways. The most prominent applications of expression profiling will be reviewed later in this chapter. Examples include analysis of gene function using co-expression[23], analysis of gene deletion mutants[24], identification of possible drug targets by exposing cells to certain pharmacological agents[24], improvement of cancer diagnosis and prognosis[25-30] and deciphering of complex biological systems such as the transcriptional regulatory network[31-33]. There is in fact hardly any area within the life sciences where expression profiling is not applied. Many challenges are encountered when analyzing microarray data that will be reviewed in more detail below. The focus will be on two-channel microarray data analysis, but most items illustrated here are also applicable to single-channel microarray data analysis.

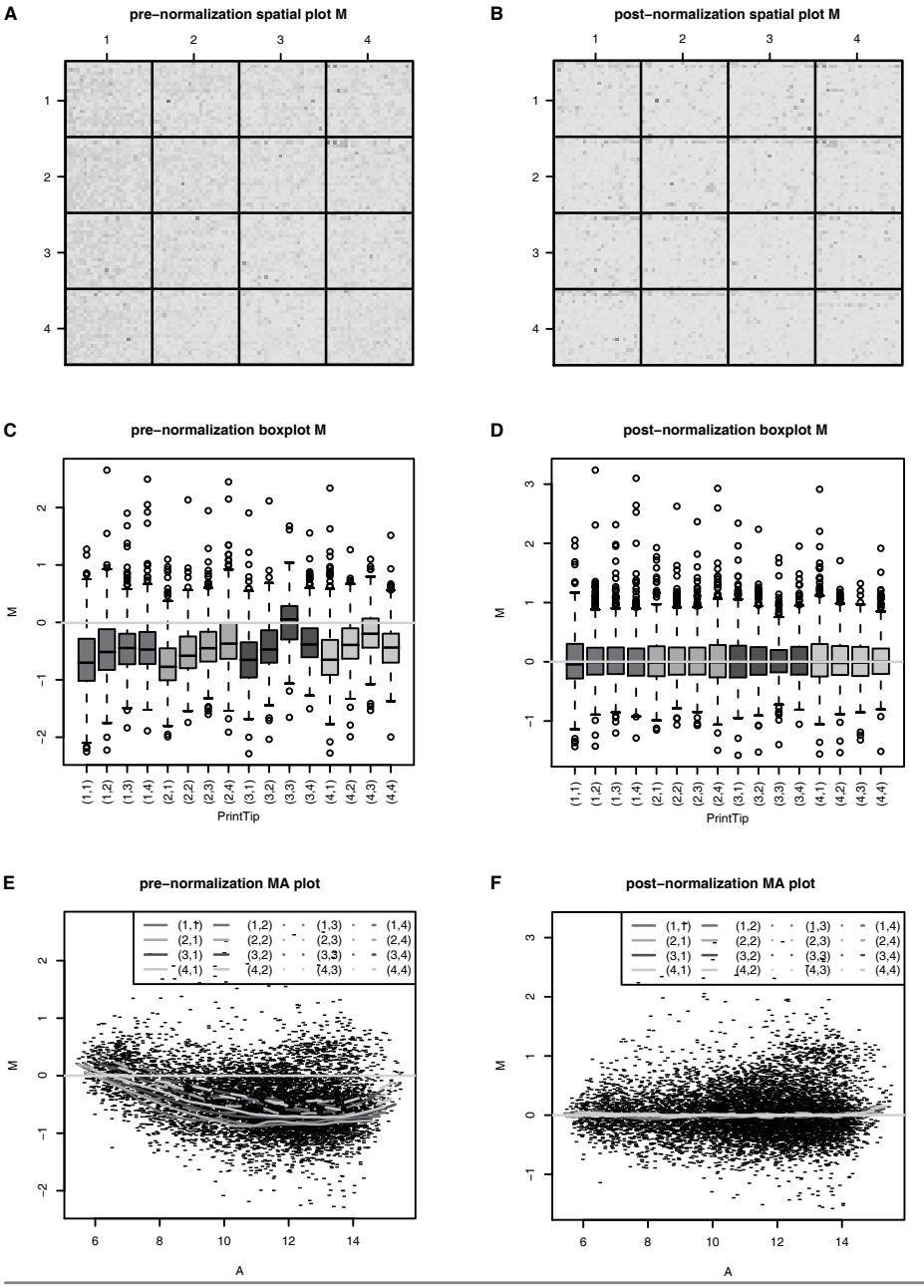
Microarray Data Preprocessing

Experiments using DNA microarrays for expression profiling generally consist of many small steps that need to be tightly controlled in order to get reliable and reproducible results. These include RNA extraction, cDNA/cRNA generation, fluorescent labeling, washing, hybridization and scanning of the microarrays. After initial sample processing and scanning, again many different steps are involved in microarray data preprocessing, including image analysis, background subtraction, quality control, filtering and normalization. Various methods exist for image analysis and background subtraction that can have a large impact on the results but fall outside the scope of this chapter (for reviews see refs 34-37). The following sections will illustrate some of the issues involved in quality control and normalization.

Quality Control

Quality control is an aspect of particular importance throughout a microarray experiment and needs to be applied rigorously at many places to minimize variances introduced by experimental procedures. Most of these quality control steps can be easily included in the experimental procedures even before any data preprocessing has occurred. One example includes monitoring of RNA degradation. RNA degradation has been shown to greatly affect gene expression levels, causing unwanted bias if samples show different levels of RNA degradation[38]. Using capillary electrophoresis, RNA degradation can be detected through an altered 28S/18S rRNA ratio[39,40] and these samples can be excluded from further downstream processing. Another example is monitoring of labeling quality using either a spectrophotometer or a capillary electrophoresis-based methodology. An optimal

Figure 2 | Diagnostic Plots for Quality Control and Normalization
Several diagnostic plots are shown from poor quality data obtained from swirl zebra fish embryos[42]. Data has been normalized with print-tip lowess using a window-size of f = 0.4. Depicted are spatial plots of $\log_2$ ratio M, before (**A**) and after normalization (**B**), coordinates of each subgrid are shown along the *x* and *y*-axis. Box plots before (**C**) and after normalization (**D**) are shown. Each box represents one subgrid on the array. MA plots of pre (**E**) and post (**D**) normalization data are shown. Lowess lines are fitted through the data points from each subgrid. See also Appendix; Color Figure 1.2.



labeling incorporation and sufficiently large product length have been shown to be vital for successful hybridizations and monitoring of labeling quality therefore provides a valuable quality control for microarray experiments[41].

At the data preprocessing level, a number of quality control steps can be incorporated. First, there are a couple of diagnostic plots that can identify certain artifacts as depicted in Figure 2, where poor quality data is used as an example[42,43]. In 2D spatial plots, different shades of colors are used to represent a certain property for each spot on the array positioned according to their physical location. The property depicted in these spatial plots can be the $\log_2$ ratio M ($M = \log_2 R/G$), the background intensity levels, the signal-to-noise ratios or any other spot quality measure. That way, any spatial biases in the values of a certain property due to for instance print-tip or cover-slip effects, can be readily detected (Figure 2A, B)[42]. Another diagnostic plot often used to detect print-tip effects is a box plot. A box plot, or box-and-whisker plot, is a simple graphical summary of the distribution of a certain variable and consists of the median, the upper and lower quartiles, the range, and individual extreme values. When used for diagnostic purposes in microarray data analysis, each box typically represents one subgrid on the array (Figure 2C, D). A final diagnostic plot frequently used is an MA plot (Figure 2E, F), which depicts the intensity log ratio M ($M = \log_2 R/G$) versus the mean log intensity A ($A = \log_2\sqrt{(RG)}$). These MA plots are often used to reveal spot artifacts and for normalization purposes[22].

A second quality control step that can be taken is the implementation of quality metrics. Differentiation should be made between spot quality metrics and array or hybridization quality metrics. Spot quality metrics can be based on either different statistics of individual spots, such as the mean-median correlation, fluorescent intensity, target area, background flatness and signal intensity consistency[44-46], or the repeatability of repeated reporter sequences[21,47]. These spot quality metrics can then be used to eliminate poor quality spots or down-weight lower quality spots in further downstream processes such as normalization[48,49]. Array or hybridization quality metrics are predominantly derived from a combination of the aforementioned spot quality metrics and allow for a qualitative assessment of individual arrays or hybridizations[43,45,50]. How to interpret these quality metrics is still open to debate. Although these quality metrics provide an unbiased view on the quality of individual spots or arrays, it is still unclear where to place thresholds to distinguish good quality data from bad quality data[43,45,50]. Furthermore, the quality of hybridization largely depends on the biological sample, array batch and protocols used and should therefore be assessed in that context. Until more evidence is found where to put the exact thresholds, these array quality metrics can therefore be best used to rank individual hybridizations within an experiment and eliminate the most extreme outliers.

Normalization

Normalization of two-channel microarray data can be subdivided into two very distinct processes. The first normalization process called within-array normalization is aimed at removing any systematic biases introduced due to differences in labeling efficiency, scanner settings and hybridization efficiency between the two channels[21,22,51]. The second normalization process called between-array normalization is aimed at removing any non-biological variation introduced due to differences in array batches, sample processing and processing of the arrays[21,22,52]. However, depending on the goal of the experiment, this second normalization process might not be required, or is implicitly incorporated in further downstream analyses, for instance, when using a mean centered distance measure for clustering analysis.

Within-array normalization requires the selection of an appropriate algorithm and selection of the features on the array that are used by the algorithm. Ideally, these features should not differ between the two samples. Most scientists have dismissed the idea of using housekeeping genes for normalization purposes, as it is impossible to find genes that maintain a constant mRNA expression level across many diverse conditions[53]. Instead, mRNA expression levels of all genes are used as normalization features. The underlying

assumption with this "all-genes" approach is that relatively few mRNA transcript levels change between the two conditions, or if they do occur, these changes are balanced. When using this approach, all changes are calculated relative to the majority of genes and global shifts in mRNA levels are concealed. However, several studies have shown that global shifts in mRNA levels do occur[32,54,55] and therefore need to be accounted for in the normalization procedure. One approach is to use external controls as normalization features. These external controls that consist of a set of exogenous genes, added equivalent to the amount of total RNA, can also be used to correct for spatial, intensity dependent variation[22] if deposited in sufficient numbers on each subgrid[55].

Most algorithms used at the moment for within-array normalization, such as lowess[22], vsn[56] or scale normalization[22] have the ability to correct for local, intensity dependent effects (Figure 2A, B, C, D). However, as stated before, the ability to deal with these effects depends largely on the number of normalization features present on each subgrid on the array. Another important bias introduced in two-channel microarray experiments are dye artifacts. The most pronounced one of these is commonly being referred to as the "banana effect" (Figure 2E; colored lines). The influence of this dye artifact, caused by a different behavior of one of the channels, most notably in the lower intensity region, can be reduced by applying an appropriate normalization procedure (Figure 2F; colored lines) such as lowess[22] or vsn[56].

Microarray Data Analysis

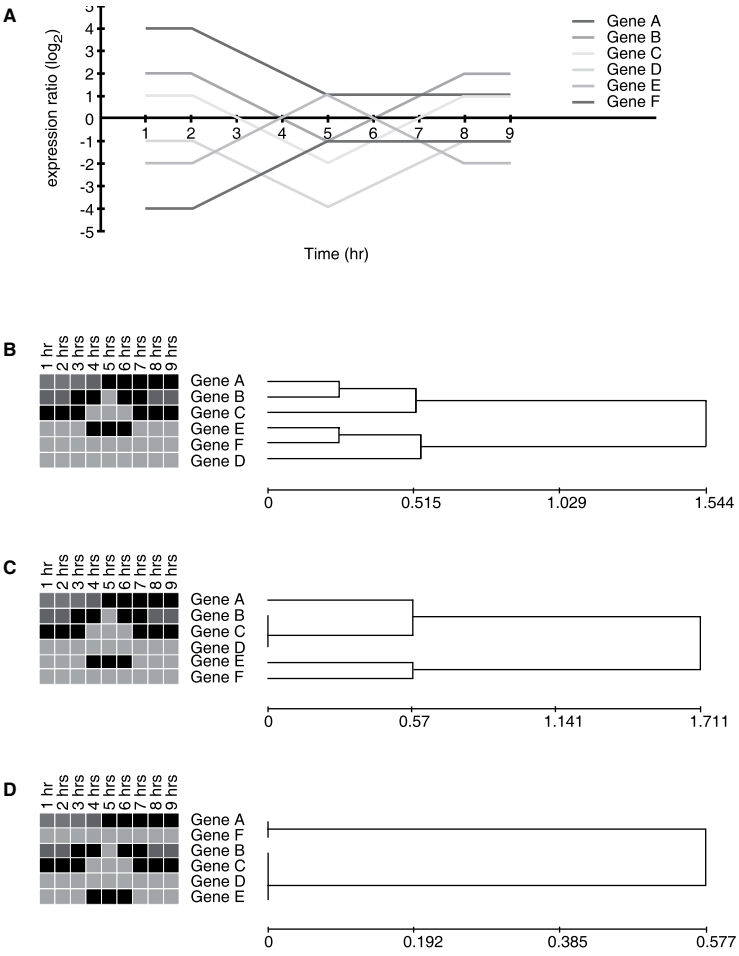
The next stage of analysis, after initial data preprocessing, quality control and normalization, involves an in depth study of the microarray data at hand. Several statistical and computational methods are available for this purpose. Some of these and their implications will be examined in the following sections.

Differential Gene Expression

The purpose of many microarray experiments is to identify genes regulated due to alterations in biological conditions. A classical approach to identify these genes is to take a fixed threshold cutoff, for instance a two-fold up- or downregulation, and select all genes that meet this criterion[57,58]. However, since many systematic and biological sources of variation occur, this approach is inefficient[59]. Only some of these sources of variation can be removed by normalization and it is therefore better to use a statistic based on replicate experiments to select differentially expressed genes. Many sophisticated methods have been proposed for this purpose[59-69] (for reviews of statistical methods for selecting differentially expressed genes see refs 70, 71). Replication of microarray experiments is crucial to obtain an estimate of the variation of gene expression levels. Two types of replication can be distinguished; biological and technical replication. Technical replication provides estimates about variation introduced due to differences in experimental procedures. In addition, biological replicates provide estimates about variation as a result of differences in biological samples and are therefore preferred.

In the most straightforward case two conditions are compared and a simple pair wise comparison can be performed using a student's *t*-test[62,65,67], ANOVA[61,63,68], linear model[69] or Bayesian approach[62,63,67]. However, most of these approaches, such as ANOVA, linear models or Bayesian methods can be extended to include multiple comparisons. Examples of these include time course experiments, where conditions correspond to different time points, and factorial designs, where the influence of multiple factors and their possible interactions is investigated. When selecting differentially expressed genes, a significance cutoff has to be placed to select only those genes that are significant. This can be problematic as an acceptable balance has to be found between the false positives (Type I error) and false negatives (Type II error). Furthermore, since not just one gene is tested, but tens of thousands, a multiple hypothesis testing problem is created. For example, an

Figure 3 | Diverse Distance Measures reveal Different Data Characteristics
Line graph of a hypothetical dataset, consisting of six genes and nine timepoints (**A**). Hierarchical average linkage clustering results using an uncentered correlation distance (**B**), a centered correlation distance (**C**) and an absolute centered correlation distance (**D**). See also Appendix; Color Figure 1.3.



error rate of 5% might be acceptable in an individual test, but if 1000 of such tests are performed, there will probably be 50 false positive results reported, even if not a single gene is differentially expressed. Different approaches can be taken to control the false positive rate. The generalized family-wise error rate (FWER) minimizes the probability of finding at least one false positive[72]. Another approach is to use the false discovery rate (FDR) to control the number of false positives[65,73], which is defined as the proportion of falsely discovered differentially expressed genes. The FDR is less conservative than the FWER. The FWER is more appropriate for analyses in which a single false positive is unacceptable, such as when comparing various drug treatments with a control. On the contrary, the FDR is more applicable in screens for candidate genes, in which a certain proportion of false positives among the discovered genes is acceptable.

Exploratory Data Analysis

Exploratory data analysis or unsupervised data analysis is often used to find groups of genes with similar expression patterns. Two classes of unsupervised algorithms can be distinguished. The first class, dimensionality reduction algorithms, such as principal component analysis (PCA)[74,75], singular value decomposition (SVD)[76,77], multidimensional scaling (MDS)[78,79] and correspondence analysis (COA)[80,81], tries to visualize important data characteristics by ignoring redundant dimensions, i.e. ignoring certain genes or experiments. The second class consists of different clustering algorithms, such as hierarchical clustering[23], K-means clustering[82], self organizing maps (SOMs)[83] and biclustering[84-86] that aim to partition the data into groups with similar expression patterns (for detailed reviews about unsupervised data analyses see refs 51, 87, 88).

It is advisable to use different exploratory data analyses or parameters, as these might elucidate different properties of the microarray data as illustrated in Figure 3. When using a centered correlation measure, genes with similar profiles, but different expression levels (Figure 3A, C; gene B, C and D) cluster together. However, an uncentered correlation measure will favor genes with similar expression levels (Figure 3A, B; gene A and B; gene E and F) over genes with similar profiles. When searching for not just the correlated genes, but also anti-correlated genes, for instance genes being suppressed or activated by the same transcription factor, a more appropriate distance measure would be a squared or absolute centered correlation measure, as this will group these genes together (Figure 3A, D; gene A and F; gene B, C, D and E). The same principles hold for the choice of different clustering or dimensionality reduction algorithms. Different types of unsupervised data analyses are therefore mostly used for generating insights into the different aspects of the microarray data at hand and generating new hypotheses, but additional experimental verification and interpretation of the biological context is required as the results obtained from these analyses might not always be biologically meaningful.

Prediction

Another important area within the microarray data analysis field is class prediction or supervised data analysis. The purpose of these supervised analyses is to train a classifier, based on expression profiles of a certain number of training samples assigned to different groups, after which the classifier can be used to assign new samples to the correct group. Applications of these types of analyses have so far been mainly used to predict the clinical outcome of different disease types[25-27,30,89-92] and show a great promise to improve clinical diagnostic tools[93]. Many different algorithms have been used so far, including regression based algorithms[89,90], classification trees[91], probabilistic models[94], discriminant analysis[95], nearest-neighbor classifiers and support vector machines[96]. However, most of these share three distinct steps: dimension reduction, modeling and generalization. The first step, dimension reduction aims to select only a subset of genes likely to be predictive, thus reducing the complexity of the underlying data. The modeling step will then more accurately define the relationship between the predictors (genes) and the corresponding group (disease state, cell type, etc). These models will then need to be generalized, so that the appropriate group can be assigned to new samples. This last step is also the most difficult one. Since in most cases only a relatively few number of samples are used for training purposes, a high risk of over-fitting of the data exists. This basically results in the classifier being highly effective for the training samples, but not for new samples. To avoid over-fitting, validation of the classifier is required. This can be performed using an independent validation dataset[97]. Alternatively, cross-validation can be used to validate the classifier. With this approach, the data is split in K portions, and the classifier is trained K times, setting aside a different portion each time for validation. The average classification rate resulting from this cross-validation procedure thus provides an unbiased estimate of the correct classification rate. General consensus these days is to use a 10-fold out approach, whereby the data is split in two portions; 90% of the original data is used for training and 10% of the original data is set aside for validation purposes. Reproducibility of

the results should then be further tested on a second independent validation set[98].

Biological Context

Interesting genes found in the previously described analyses do not act on their own. They interact and cooperate with other genes and proteins to form complex biological systems. It is therefore crucial, not only for a higher understanding of these biological systems, but also for a better understanding of the function of these genes to assess these genes in the correct biological context.

Several systematic approaches are available to relate genes to the relevant biological context. The first approach uses the Gene Ontology (GO)[99]. This is an international effort to assign functional annotations to genes through manual curation of the scientific literature. Many tools exist that can automatically assign significantly overrepresented GO categories using a statistical approach[100-103]. Careful consideration has to be placed however on how these programs use the evidence codes provided by the GO consortium, as these provide essential information on how these annotations were provided and therefore how reliable they are. A second approach attempts to extract relevant information from the biomedical literature through a process called text mining or natural language processing[104-106]. These potentially provide a much richer resource than current GO annotations. However, current biomedical literature databases such as PubMed only contain abstracts and not complete articles, therefore limiting this approach to these abstracts. A third approach, often used in concurrence with cluster analysis, is promoter analysis[82,107-109]. If several genes belong to the same cluster, i.e. are co-expressed, these genes might be co-regulated through a common transcription factor. Searching for overrepresented patterns in the upstream region of these genes might elucidate more about the biological process and regulation of these genes. A fourth approach aspires to map significantly expressed genes against pathway databases such as KEGG[110] or rebuild existing metabolic or regulatory pathways using text mining[111]. That way, any significantly altered metabolic or regulatory pathways should become apparent.

Applications of Expression Profiling

DNA microarrays for expression profiling have been applied in many different areas of scientific research and have begun to have a large impact on biological and biomedical research. Frequently used and recently emerging promising applications include screening for genes responsive to certain perturbations, creation and usage of gene expression compendia, improvement of clinical diagnostic tools and reconstruction of regulatory networks.

Probably the most well known application of DNA microarrays is the screening for activated or suppressed genes in response to a genetic mutation, an altered physiological condition or exposure to a chemical reagent. Examples include exposure to different environmental stresses[112], variation in ploidy[113] and exposure of human fibroblasts to serum[114]. Through these perturbations, direct associations can be revealed between groups or classes of genes and the resulting physiological response or phenotype. Gene functions for uncharacterized genes can then be postulated on the basis of the altered response or phenotype. As most of the experiments compare the gene expression levels before and after the perturbation, no distinction can be made between direct and indirect effects of the perturbation on gene expression levels. As this distinction is important to expose key regulatory genes, that in turn affect other genes, time-series are now more commonly used in these types of experiments. Furthermore, higher resolution time-series also allow us to detect mRNA changes at a more granular level, thus providing more detailed insights into the regulatory processes behind these changes[115].

Another type of application is the use of a compendium of gene expression profiles[24,116]. A compendium is built up of many diverse reference profiles, created by

introducing perturbations, such as gene mutations, in a wide range of biological pathways, exposing samples to pharmaceutical or chemical agents, or profiling different disease states or stages. Two goals can be achieved using such compendia. First, many different perturbations and physiological circumstances exist in this database of reference profiles. As a consequence, patterns will emerge within the gene expression levels that allow genes to be associated with certain biological processes or pathways on the basis of similarity between (parts of) their expression profiles[24,84]. Second, new perturbations, for instance a gene deletion of an uncharacterized gene, can be added to the compendium and compared with the existing expression profiles. Resemblance with existing profiles, associated with certain cellular processes, can shed light on the function and involvement of such a gene in these processes.

Expression profiles are also being used to improve clinical diagnostic tools[25-30]. With many diseases, in particular cancer, current clinical diagnostic tools, such as histological examination, do not predict the prognosis of a particular patient well enough. Attempts have therefore been made to improve the accuracy of these clinical tools through expression profiling of early onset samples of particular diseases. Promising results have already been made in some areas of disease classification, most notably in the distinction between different subtypes of leukemia[25] and large B-cell lymphomas[26], and differentiation between metastatic and non-metastatic breast tumors[28] and head and neck squamous cell carcinomas[29,30]. Preliminary results obtained from these studies show that microarray expression profiles can be used to more accurately predict the outcome of a certain disease. However, since these studies have been performed on a limited number of samples, the validity of the approach still has to be tested thoroughly in a clinical setting. The first clinical trials are therefore already underway to test whether microarray expression profiles can be used as a standardized diagnostic tool alongside currently existing diagnostic tools[93].

A last application illustrated here is the reconstruction of regulatory networks. Important progress has been made in this area, using microarray expression profiling. Initial studies have focused on biological systems and environments such as the cell cycle[33], transcription machinery[32], sporulation[117], pheromone treatment[118] and unfolded protein responses[119], laying the foundations in microarray data analysis. At that time, relatively simple techniques were used for the analysis of the data at hand, such as screening for differentially expressed genes, cluster analysis and upstream promoter analysis. However, as our understanding of both the biological systems and how to model these using computational algorithms developed, more sophisticated analyses where performed. Initial models focused on the topology of the underlying regulatory networks from biological systems such as the cell cycle[82], galactose metabolism[120] and transcriptional regulators[109]. Again, as our understanding of these systems and models is improving, the focus is already shifting towards modeling the dynamics of these networks using more advanced types of computational methods[121]. However, since most data sources only provide snapshots at certain time points and conditions, evidence for these types of networks is still scarce.

Other Types of Genome-Scale Data

With the advent of whole-genome sequences, many other high-throughput approaches, besides DNA microarrays, have been developed accordingly. Each of these approaches provides different pieces of the puzzle, that, once assembled may provide us with a complete picture of the underlying biological system. Many different types of data have already been used to explore gene function on a large scale, including protein-protein interactions[122-125], phenotypes[126-131], subcellular location[132,133], genetic interactions[134-136] and chromosomal location data[109,137-140]. Most of these have initially been used for *S. cerevisiae*, but are gradually being implemented in other organisms as well.

Large-scale protein interaction mapping studies have so far been performed using

two different approaches. The first approach uses the yeast two-hybrid assay[122-125], initially setup in *E. coli*[122] and *S. cerevisiae*[123,125], but gradually being used in higher organisms as well[124]. The second approach is based on affinity purification of tagged proteins, followed by mass spectrometry analysis of the associated proteins[141,142]. Although little is known about the possible pitfalls of the second approach, the potential artifacts inherent to the yeast two-hybrid assay indicates that each such reported interaction should be considered putative[123].

Another commonly used large-scale method to uncover gene function is phenotype screening. Systematic approaches used so far include gene deletions[126,127,130] and RNA interference (RNAi)[128,129,131], whereby a gene is either knocked-out or knocked-down and subsequently screened for a number of known phenotypes. Although very powerful approaches, possible drawbacks are that many genes might not show obvious phenotypes when inactivated or might be missed because the appropriate phenotype was not screened. Furthermore, many gene deletions result in loss of viability, excluding them from further phenotype screening.

Different cellular compartments provide specific functions within living organisms. Determining the location of proteins might therefore provide new clues to their functions. Large-scale studies have been performed to expose the subcellular location of proteins using tagged fluorescent protein fusion proteins under native conditions[133], or through a plasmid-based overexpression system[132]. Both approaches have their own specific advantages and disadvantages. While the overexpression system might introduce biases due to saturation of intracellular transport mechanisms, protein expression levels under native conditions might result in insufficient protein yields to be visualized by fluorescent microscopy. Despite these drawbacks, both techniques have uncovered the subcellular location of many proteins thus far unknown.

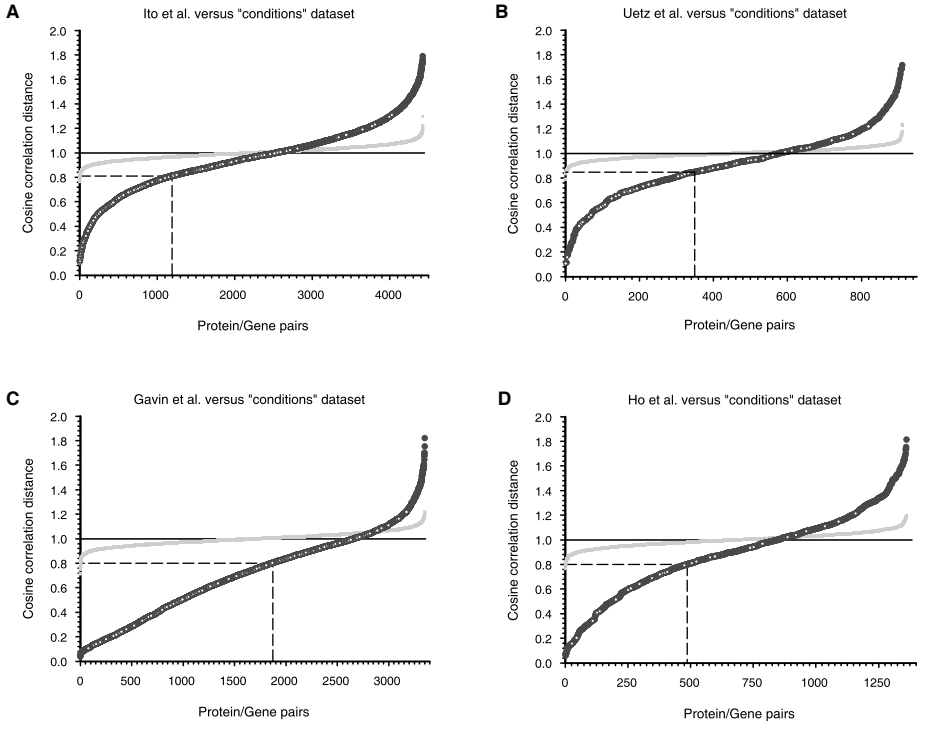
Synthetic genetic interactions can be used to expose redundant functions in proteins. Two genes show a synthetic lethal interaction if the combination of two mutations is lethal, whereas either separately is not[143,144]. These synthetic lethal interactions may occur when two genes operate in the same biochemical pathway, or function in two parallel pathways that can compensate or buffer each others defects[145]. Screening for synthetic lethality might therefore provide additional insights into the function of certain genes or proteins. Large-scale synthetic interaction screens or synthetic genetic array (SGA) analysis, have been performed to discover potential gene function on a genome-scale[134-136]. Although providing additional evidence for possible gene function, results obtained from these studies should be considered as circumstantial evidence, as introduction of two non-fatal injuries can result in fatality, but this does not necessarily mean that these injuries are related.

Regulation of biological processes and systems is often facilitated through transcriptional activators or repressors that bind to certain sequence elements in the upstream region of a group of functionally related genes. Technologies have been developed that allow the monitoring of these DNA binding proteins on a genome-scale[109,137-140]. Genome-wide location analysis, or chip-on-chip analysis, uses chromatin immunoprecipitation (chip) to detect any proteins bound to genomic DNA under certain physiological conditions, thus providing new insights into the function of co-regulated genes. Since many of the underlying technologies used for genome-wide location analysis are similar to regular microarray expression profiling, the same challenges and pitfalls apply to this approach for regulatory circuitry discovery.

Integrating Genome-Scale Data

Whole-genome sequencing projects form a tremendous resource for biological discovery[146]. Besides genome annotation[147], an important challenge arising from the availability of genomes is the elucidation of the function of thousands of newly discovered genes. For instance, the *S. cerevisiae* genome has been available since 1996; however at present

Figure 4 | Comparison of Protein Interaction Datasets
Data from protein interaction datasets of ito et al.[125] (**A**), Uetz et al.[123] (**B**), Gavin et al.[141] (**C**) and Ho et al.[142] (**D**) were analyzed for mRNA co-expression within the "conditions" dataset using the correlation distance (dark grey lines). This "conditions" dataset consisted of 326 expression profiles from various conditions, such as different stress responses, cell cycle, sporulation, pheromone treatment and unfolded protein degradation. The degree of co-expression is plotted on the *y*-axis. A value of 0 corresponds to complete mRNA co-expression across all the expression data. A value of 2 corresponds to perfect anti-correlation under all conditions. The protein pairs are ranked according to the degree of co-expression and the rank number is plotted on the *x*-axis. The previously known interactions within the interaction datasets are plotted as open circles in order to demonstrate the marked skewing towards a higher degree of co-expression for these interactions of higher confidence. The light grey lines are the result of an identical analysis using a randomized expression dataset. The dashed lines indicate the significance threshold for catching 0.1% of the interactions with the randomized expression data.



about half of the genes have been annotated as unknown according to the Gene Ontology (GO) consortium that provides gene annotations for most organisms[99]. High-throughput data could be used to provide improved functional annotation. However, several issues need to be addressed first. First of all, data quality within and between datasets varies significantly. The number of false positives needs to be addressed and dealt with in order to obtain more reliable data. Secondly, hypotheses about gene function arising from high-throughput data need to be prioritized, so that the more likely hypotheses can be given precedence over others.

One approach to deal with these issues is to integrate genome-wide datasets to obtain the better-quality data. The aim in this case is to reduce the overall error rate through combination of independent datasets. Until now two main approaches have been used to integrate these datasets. The first approach combines all possible relationships, obtained from various sources such as phylogenetic profiles, mRNA co-expression and experimental data. This results in huge interaction networks, with high levels of false positives. From these networks, subsets are then used for predicting functional relationships based on the

supported data type[148]. The second approach reduces the number of possible relationships by looking at the significant overlap between two or more different types of data[148-155], e.g. mRNA co-expression and protein-protein interaction.

From these studies two things have become apparent. The first is that there is hardly any overlap between the datasets used. For instance when looking at protein interaction data from four different studies[123,125,141,142], the total number of interactions amounts to 10121, of which only 251 are shared between at least two out of four datasets. When searching for possible explanations, a distinction has to be made between looking for overlaps between similar types of data and different types of data. When looking for overlap between similar types of data, one possibility is that even though these datasets are called "genome-wide", they still do not cover the complete interaction space and need to be expanded further. Another possibility is the high level of false positives present in the high-throughput datasets. Combining two datasets, both with high levels of false positives, will lead to relatively little overlap. When comparing datasets of different data types a third issue might play a role. Taking the intersection of different types of data automatically assumes that valid relationships should be present in all of the datasets, which is not necessarily true. For instance when combining mRNA co-expression data and protein interaction data, one assumes that protein-protein interaction also means mRNA co-expression, which might not be appropriate under all conditions, as regulation can also occur at the post-transcriptional or translational level.

The second thing that has become apparent is that to more efficiently use these datasets for functional annotation and hypothesis generation, the data quality needs to be assessed. Through comparison of high-throughput interaction data with mRNA co-expression, it has become obvious that there are large differences in data quality between the different datasets (Figure 4). Indicated in this figure are the protein interactions from Ito et al.[125] (Figure 4A), Uetz et al.[123] (Figure 4B), Gavin et al.[141] (Figure 4C) and Ho et al.[142] (Figure 4D), ranked according to mRNA co-expression in various conditions[33,112,117,119] (dark grey lines), ranging from 0 (correlation) to 2 (anti-correlation). Comparison with previously established interactions and random data results in a quality indicator for each individual dataset. Using this quality indicator, one can see that the data from Gavin et al.[141] behave much more like the previously established interactions, whereas the two-hybrid studies contain a lot more false positives. These differences are probably due to the overexpression system used for the two-hybrid approaches and the approach used by Ho et al.[142] The overexpression system is more sensitive and thus better capable of picking up more transient interactions; however, the consequence of this increased sensitivity is that the number of false positives also increases.

The same graphs (Figure 4) can be used to determine which interactions are less likely. Interactions on the right side of the graph show a high degree of anti-correlation. These are most likely false positives as anti-correlation between mRNA levels and protein interaction is much harder to reconcile than mRNA co-expression and protein interaction. This is supported by the fact that 47% of the novel interactions show anti-correlation, whereas only 6% of previously established interactions show the same degree of anti-correlation. This way one can exclude either the most likely false positives (right side of the graph), or include the most likely true positives (left side of the graph) for further downstream analysis.

Another advantage of integrating different types of data is that more insight is gained about the particular relationship. Since in this case mRNA co-expression is added to the protein interaction data, the specific conditions under which mRNA co-expression is observed are also known. Looking at the specific conditions where mRNA co-expression is observed might result in increased knowledge about the potential role of both genes. For example the mRNA co-expression patterns for the interacting partners NTH1 and YLR270W show that both genes are co-expressed during heat shock and nutrient stresses (Figure 5). This could also facilitate follow-up experimentation. When using the same data for functional annotation, only the most likely true positives

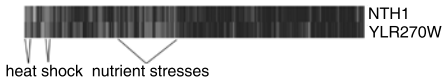


Figure 5 | Expression Profiles of NTH1 and YLR270W Obtained from the "conditions" Dataset
The "conditions" dataset consisted of 326 expression profiles from various conditions, such as different stress responses, cell cycle, sporulation, pheromone treatment and unfolded protein degradation. Red indicates upregulation, green indicates downregulation. mRNA expression levels of both genes are up-regulated during heat shock and other forms of stress. See also Appendix; Color Figure 1.5.

should be included. Therefore a strict cutoff needs to be set. Using a randomized expression matrix (light grey lines), a cutoff can be determined so that 0.1% of the random data is expected to occur in the real dataset. All interactions above this cutoff are of high confidence ($P < 0.003$). Interactions with one characterized partner and above this cutoff can thus be more reliably used for functional annotation.

Predictions of functional annotation can in this way be used to improve the existing annotation of initiatives like GO[99]. However, for improved usefulness the annotations themselves should also be annotated such that information about how the annotation was found is not lost. This is a very important aspect, as this allows us to differentiate between hand-curated annotations, based on single experiments, and automatically predicted annotations. If this type of information is lost, two extremes could happen. When annotations are added without mentioning the source, chances are that predictions are based upon predictions, upon predictions and so forth, leading to less trustworthy annotations. On the other hand, if annotations obtained from automated predictions are completely ignored then valuable information is lost which might be useful to guide future research.

Estimations of data quality have so far been mainly based on protein interaction data. However, the findings from these estimations also have implications for other types of high-throughput data like cellular location, phenotype, metabolic pathways, chromosomal location, proteomic data, etc. When using these high-throughput datasets for further downstream analysis or functional annotation, data quality should be assessed and incorporated into the methodology in order to be able to use these types of data more efficiently. Besides this, each data type has its own properties that need to be taken into account. For instance location data are categorical data, while mRNA co-expression covers the complete range of possible values. These properties need to be incorporated, so that the different data types can be compared and integrated. Some promising results have already been made that take these different aspects into account when integrating different types of data[153,155].

When expanding this approach with different types of data, one has to realize however, that requiring a relationship to be present in all data types, results in highly confident, yet very few functional relationships[155]. Using the intersection of all these datasets will no longer be sufficient and more sophisticated methods have to be applied. For example, the assignment of a data quality estimator to each individual dataset or data point, which then successively can be applied to rank the relationships based on the sum of these quality estimators[153,155].

Until now most of these approaches have been performed in *S. cerevisiae*; however, as more data are becoming available for higher organisms, similar issues will arise there as well. How useful these approaches are remains to be seen. Other issues also play a role, such as scalability, complexity and standardization. Approaches using higher eukaryotes, such as human cell lines, face enormous challenges in areas of poorly annotated, large genomes (order of magnitude larger than *S. cerevisiae*), non-standardized gene names and more complex interactions with the environment. One advantage however is that, because these approaches have been used before in other organisms, valuable information about how to apply these methods is readily available. One can even take the existing data,

available from other species, and integrate these into the analysis, using orthologues or any other way of mapping information from one species to another.

Outline of this Thesis

The goal of the work presented in this thesis is to develop new and improve existing methodologies for analyzing genome-scale datasets, so that the overall efficiency, accuracy, precision and biological significance of functional genomic data increases. In chapter 2, a general framework is presented that facilitates many aspects of microarray data analysis, including array production, array annotation, quality control, sample annotation and creation of standardized workflows. Chapter 3 proposes a data normalization method to monitor global mRNA changes and consequences thereof on further downstream microarray data analyses. In chapter 4, a tool is presented that provides a versatile and extensible, web-based platform for microarray gene expression analysis. In chapter 5, this tool is further extended so that regulatory mechanisms behind biological processes can be revealed. Chapter 6 presents a method that uses microarray expression datasets to assess the quality of high-throughput protein interaction datasets and how to improve functional annotations through the integration of both of these data types. In chapter 7, a new approach is presented for integrating different types of high-throughput datasets, thus improving overall accuracy and precision when used for predicting functional annotations.

References

1. Lockhart, D.J. et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol* **14**, 1675-80 (1996).
2. Marton, M.J. et al. Drug target validation and identification of secondary drug target effects using DNA microarrays. *Nat Med* **4**, 1293-301 (1998).
3. Lipshutz, R.J., Fodor, S.P., Gingeras, T.R. & Lockhart, D.J. High density synthetic oligonucleotide arrays. *Nat Genet* **21**, 20-4 (1999).
4. Nuwaysir, E.F. et al. Gene expression analysis using oligonucleotide arrays produced by maskless photolithography. *Genome Res* **12**, 1749-55 (2002).
5. Blanchard, A.P., Kaiser, R.J. & Hood, L.E. High-density oligonucleotide arrays. *Biosens. Bioelectron* **6/7**, 687-690 (1996).
6. Hughes, T.R. et al. Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat Biotechnol* **19**, 342-7 (2001).
7. Kronick, M.N. Creation of the whole human genome microarray. *Expert Rev Proteomics* **1**, 19-28 (2004).
8. Schena, M., Shalon, D., Davis, R.W. & Brown, P.O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467-70 (1995).
9. DeRisi, J. et al. Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat Genet* **14**, 457-60 (1996).
10. Schadt, E.E., Li, C., Ellis, B. & Wong, W.H. Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data. *J Cell Biochem Suppl* **Suppl 37**, 120-5 (2001).
11. Hill, A.A. et al. Evaluation of normalization procedures for oligonucleotide array data based on spiked cRNA controls. *Genome Biol* **2**, RESEARCH0055 (2001).
12. Li, C. & Wong, W.H. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc Natl Acad Sci U S A* **98**, 31-6 (2001).
13. Hoffmann, R., Seidl, T. & Dugas, M. Profound effect of normalization on detection of differentially expressed genes in oligonucleotide microarray data analysis. *Genome Biol* **3**, RESEARCH0033 (2002).
14. Lazaridis, E.N., Sinibaldi, D., Bloom, G., Mane, S. & Jove, R. A simple method to improve probe set estimates from oligonucleotide arrays. *Math Biosci* **176**, 53-8 (2002).
15. Zhou, Y. & Abagyan, R. Match-only integral distribution (MOID) algorithm for high-density oligonucleotide array analysis. *BMC Bioinformatics* **3**, 3 (2002).
16. Irizarry, R.A. et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249-64 (2003).
17. Li, C. & Wong, W.H. DNA-Chip Analyzer (dChip). in *The Analysis of Gene Expression Data: methods and software* (eds. Parmigiani, G., Garrett, E.S., Irizarry, R.A. & Zeger, S.L.) 120-141 (Springer-Verlag, New York, 2003).
18. Guo, Z., Guilfoyle, R.A., Thiel, A.J., Wang, R. & Smith, L.M. Direct fluorescence analysis of genetic polymorphisms by hybridization with oligonucleotide arrays on glass supports. *Nucleic Acids Res* **22**, 5456-65 (1994).
19. Kothapalli, R., Yoder, S.J., Mane, S. & Loughran, T.P., Jr. Microarray results: how accurate are they? *BMC Bioinformatics* **3**, 22 (2002).
20. Jurata, L.W. et al. Comparison of microarray-based mRNA profiling technologies for identification of psychiatric disease and drug signatures. *J Neurosci Methods* **138**, 173-88 (2004).
21. Tseng, G.C., Oh, M.K., Rohlin, L., Liao, J.C. & Wong, W.H. Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Res* **29**, 2549-57 (2001).
22. Yang, Y.H. et al. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res* **30**, e15 (2002).
23. Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* **95**, 14863-8 (1998).
24. Hughes, T.R. et al. Functional discovery via a compendium of expression profiles. *Cell* **102**, 109-26 (2000).
25. Golub, T.R. et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531-7 (1999).
26. Alizadeh, A.A. et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503-11 (2000).
27. Perou, C.M. et al. Molecular portraits of human breast tumours. *Nature* **406**, 747-52 (2000).
28. van 't Veer, L.J. et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530-6 (2002).
29. Chung, C.H. et al. Molecular classification of head and neck squamous cell carcinomas using patterns of gene expression. *Cancer Cell* **5**, 489-500 (2004).

30. Roepman, P. et al. An expression profile for diagnosis of lymph node metastases from primary head and neck squamous cell carcinomas. *Nat Genet* **37**, 182-6 (2005).
31. Cho, R.J. et al. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol Cell* **2**, 65-73 (1998).
32. Holstege, F.C. et al. Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* **95**, 717-28 (1998).
33. Spellman, P.T. et al. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell* **9**, 3273-97 (1998).
34. Brown, C.S., Goodwin, P.C. & Sorger, P.K. Image metrics in the statistical analysis of DNA microarray data. *Proc Natl Acad Sci U S A* **98**, 8944-9 (2001).
35. Yang, Y.H., Buckley, M.J. & Speed, T.P. Analysis of cDNA microarray images. *Brief Bioinform* **2**, 341-9 (2001).
36. Jain, A.N. et al. Fully automatic quantification of microarray image data. *Genome Res* **12**, 325-32 (2002).
37. Leung, Y.F. & Cavalieri, D. Fundamentals of cDNA microarray data analysis. *Trends Genet* **19**, 649-59 (2003).
38. Auer, H. et al. Chipping away at the chip bias: RNA degradation in microarray analysis. *Nat Genet* **35**, 292-3 (2003).
39. van de Goor, T.A. The principle and promise of labchip technology. *PharmaGenomics* **3**, 16-18 (2003).
40. Carter, D.E., Robinson, J.F., Allister, E.M., Huff, M.W. & Hegele, R.A. Quality assessment of microarray experiments. *Clin Biochem* **38**, 639-42 (2005).
41. Lage, J.M. et al. Microgel assessment of nucleic acid integrity and labeling quality in microarray experiments. *Biotechniques* **32**, 312-4 (2002).
42. Dudoit, S. & Yang, J.Y.H. Bioconductor R Packages for Exploratory Analysis and Normalization of cDNA Microarray Data. in *The Analysis of Gene Expression Data: methods and software* (eds. Parmigiani, G., Garrett, E.S., Irizarry, R.A. & Zeger, S.L.) 73-101 (Springer-Verlag, New York, 2003).
43. Buness, A., Huber, W., Steiner, K., Sultmann, H. & Poustka, A. arrayMagic: two-colour cDNA microarray quality control and preprocessing. *Bioinformatics* **21**, 554-6 (2005).
44. Wang, X., Ghosh, S. & Guo, S.W. Quantitative quality control in microarray image processing and data acquisition. *Nucleic Acids Res* **29**, E75-5 (2001).
45. Chen, Y. et al. Ratio statistics of gene expression levels and applications to microarray data analysis. *Bioinformatics* **18**, 1207-15 (2002).
46. Tran, P.H. et al. Microarray optimizations: increasing spot accuracy and automated identification of true microarray signals. *Nucleic Acids Res* **30**, e54 (2002).
47. Jenssen, T.K. et al. Analysis of repeatability in spotted cDNA microarrays. *Nucleic Acids Res* **30**, 3235-44 (2002).
48. Smyth, G.K. & Speed, T. Normalization of cDNA microarray data. *Methods* **31**, 265-73 (2003).
49. Wang, X., Hessner, M.J., Wu, Y., Pati, N. & Ghosh, S. Quantitative quality control in microarray experiments and the application in data filtering, normalization and false positive rate prediction. *Bioinformatics* **19**, 1341-7 (2003).
50. Sauer, U., Preininger, C. & Hany-Schmatzberger, R. Quick and simple: quality control of microarray data. *Bioinformatics* **21**, 1572-8 (2005).
51. Quackenbush, J. Computational analysis of microarray data. *Nat Rev Genet* **2**, 418-27 (2001).
52. Bolstad, B.M., Irizarry, R.A., Astrand, M. & Speed, T.P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185-93 (2003).
53. Lee, P.D., Sladek, R., Greenwood, C.M. & Hudson, T.J. Control genes and variability: absence of ubiquitous reference transcripts in diverse mammalian expression studies. *Genome Res* **12**, 292-7 (2002).
54. Wang, Y. et al. Precision and functional specificity in mRNA decay. *Proc Natl Acad Sci U S A* **99**, 5860-5 (2002).
55. van de Peppel, J. et al. Monitoring global messenger RNA changes in externally controlled microarray experiments. *EMBO Rep* **4**, 387-93 (2003).
56. Huber, W., von Heydebreck, A., Sultmann, H., Poustka, A. & Vingron, M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* **18 Suppl 1**, S96-104 (2002).
57. Schena, M. et al. Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proc Natl Acad Sci U S A* **93**, 10614-9 (1996).
58. DeRisi, J.L., Iyer, V.R. & Brown, P.O. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**, 680-6 (1997).
59. Chen, Y., Dougherty, E.R. & Bittner, M.L. Ratio-based decisions and the quantitative analysis of cDNA microarray images. *J Biomed Opt* **2**, 364-367 (1997).
60. Ideker, T., Thorsson, V., Siegel, A.F. & Hood, L.E. Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data. *J Comput Biol* **7**, 805-17 (2000).
61. Kerr, M.K., Martin, M. & Churchill, G.A. Analysis of variance for gene expression microarray data. *J Comput*

Biol **7**, 819-37 (2000).
62. Baldi, P. & Long, A.D. A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. *Bioinformatics* **17**, 509-19 (2001).
63. Long, A.D. et al. Improved statistical inference from DNA microarray data using analysis of variance and a Bayesian statistical framework. Analysis of global gene expression in Escherichia coli K12. *J Biol Chem* **276**, 19937-44 (2001).
64. Newton, M.A., Kendziorski, C.M., Richmond, C.S., Blattner, F.R. & Tsui, K.W. On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *J Comput Biol* **8**, 37-52 (2001).
65. Tusher, V.G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A* **98**, 5116-21 (2001).
66. Wu, T.D. Analysing gene expression data from DNA microarrays to identify candidate genes. *J Pathol* **195**, 53-65 (2001).
67. Lonnstedt, I. & Speed, T. Replicated Microarray Data. *Stat Sinica* **12**, 31-46 (2002).
68. Wu, H., Kerr, M.K., Cui, X. & Churchill, G.A. MAANOVA: A Software Package for the Analysis of Spotted cDNA Microarray Experiments. in *The analysis of gene expression data: methods and software* (eds. Parmigiani, G., Garett, E.S., Irizarry, R.A. & Zeger, S.L.) (Springer, New York, 2002).
69. smyth, G.K. Limma: linear models for microarray data. in *Bioinformatics and Computational Biology Solutions using R and BioConductor* (eds. Gentleman, R., Carey, V., Dudoit, S., Irizarry, R. & Huber, W.) (Springer, New York, 2005).
70. Pan, W. A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics* **18**, 546-54 (2002).
71. Cui, X. & Churchill, G.A. Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol* **4**, 210 (2003).
72. Dudoit, S., Yang, Y.H., Speed, T.P. & Callow, M.J. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Stat Sinica* **12**, 111-139 (2002).
73. Reiner, A., Yekutieli, D. & Benjamini, Y. Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics* **19**, 368-75 (2003).
74. Raychaudhuri, S., Stuart, J.M. & Altman, R.B. Principal components analysis to summarize microarray experiments: Application to sporulation time series. *Pac Symp Biocomput*, 455-466 (2000).
75. Yeung, K.Y. & Ruzzo, W.L. Principal component analysis for clustering gene expression data. *Bioinformatics* **17**, 763-74 (2001).
76. Alter, O., Brown, P.O. & Botstein, D. Singular value decomposition for genome-wide expression data processing and modeling. *Proc Natl Acad Sci U S A* **97**, 10101-6 (2000).
77. Holter, N.S. et al. Fundamental patterns underlying gene expression profiles: simplicity from complexity. *Proc Natl Acad Sci U S A* **97**, 8409-14 (2000).
78. Khan, J. et al. Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays. *Cancer Res* **58**, 5009-13 (1998).
79. Bittner, M. et al. Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature* **406**, 536-40 (2000).
80. Fellenberg, K. et al. Correspondence analysis applied to microarray data. *Proc Natl Acad Sci U S A* **98**, 10781-6 (2001).
81. Culhane, A.C., Perriere, G., Considine, E.C., Cotter, T.G. & Higgins, D.G. Between-group analysis of microarray data. *Bioinformatics* **18**, 1600-8 (2002).
82. Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J. & Church, G.M. Systematic determination of genetic network architecture. *Nat Genet* **22**, 281-5 (1999).
83. Tamayo, P. et al. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci U S A* **96**, 2907-12 (1999).
84. Ihmels, J. et al. Revealing modular organization in the yeast transcriptional network. *Nat Genet* **31**, 370-7 (2002).
85. Lazzeroni, L. & Owen, A.B. Plaid models for gene expression data. *Stat Sinica* **12**, 61-86 (2002).
86. Getz, G., Gal, H., Kela, I., Notterman, D.A. & Domany, E. Coupled two-way clustering analysis of breast cancer and colon cancer gene expression data. *Bioinformatics* **19**, 1079-89 (2003).
87. Sherlock, G. Analysis of large-scale gene expression data. *Brief Bioinform* **2**, 350-62 (2001).
88. Valafar, F. Pattern recognition techniques in microarray data analysis: a survey. *Ann N Y Acad Sci* **980**, 41-64 (2002).
89. West, M. et al. Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc Natl Acad Sci U S A* **98**, 11462-7 (2001).
90. Tibshirani, R., Hastie, T., Narasimhan, B. & Chu, G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci U S A* **99**, 6567-72 (2002).

91. Zhang, H. & Yu, C.Y. Tree-based analysis of microarray data for classifying breast cancer. *Front Biosci* **7**, c63-7 (2002).
92. van 't Veer, L.J. et al. Expression profiling predicts outcome in breast cancer. *Breast Cancer Res* **5**, 57-8 (2003).
93. Garber, K. Genomic medicine. Gene expression tests foretell breast cancer's future. *Science* **303**, 1754-5 (2004).
94. Garrett, E.S. & Parmigiani, G. POE: Statistical Methods for Qualitative Analysis of Gene Expression. in *The Analysis of Gene Expression Data: methods and software* (eds. Parmigiani, G., Garrett, E.S., Irizarry, R.A. & Zeger, S.L.) 362-387 (Springer-Verlag, New York, 2003).
95. Li, W. & Yang, Y. How many genes are needed for a discriminant microarray data analysis? in *Methods for Microarray Data* (eds. Lin, S.M. & Johnson, K.F.) 137-150 (Kluwer Academic, Dordrecht, 2002).
96. Brown, M.P. et al. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc Natl Acad Sci U S A* **97**, 262-7 (2000).
97. Dudoit, S., Fridlyand, J. & Speed, T.P. Comparison of discrimination methods for the classification of tumors using gene expression data. *JASA* **97**, 77-87 (2002).
98. Ransohoff, D.F. Rules of evidence for cancer molecular-marker discovery and validation. *Nat Rev Cancer* **4**, 309-14 (2004).
99. Harris, M.A. et al. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* **32**, D258-61 (2004).
100. Khatri, P., Draghici, S., Ostermeier, G.C. & Krawetz, S.A. Profiling gene expression using onto-express. *Genomics* **79**, 266-70 (2002).
101. Robinson, M.D., Grigull, J., Mohammad, N. & Hughes, T.R. FunSpec: a web-based cluster interpreter for yeast. *BMC Bioinformatics* **3**, 35 (2002).
102. Doniger, S.W. et al. MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome Biol* **4**, R7 (2003).
103. Al-Shahrour, F., Diaz-Uriarte, R. & Dopazo, J. FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics* **20**, 578-80 (2004).
104. Blaschke, C., Oliveros, J.C. & Valencia, A. Mining functional information associated with expression arrays. *Funct Integr Genomics* **1**, 256-68 (2001).
105. Jenssen, T.K., Laegreid, A., Komorowski, J. & Hovig, E. A literature network of human genes for high-throughput analysis of gene expression. *Nat Genet* **28**, 21-8 (2001).
106. Raychaudhuri, S. & Altman, R.B. A literature-based method for assessing the functional coherence of a gene group. *Bioinformatics* **19**, 396-401 (2003).
107. Pilpel, Y., Sudarsanam, P. & Church, G.M. Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat Genet* **29**, 153-9 (2001).
108. Shen-Orr, S.S., Milo, R., Mangan, S. & Alon, U. Network motifs in the transcriptional regulation network of Escherichia coli. *Nat Genet* **31**, 64-8 (2002).
109. Lee, T.I. et al. Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science* **298**, 799-804 (2002).
110. Grosu, P., Townsend, J.P., Hartl, D.L. & Cavalieri, D. Pathway Processor: a tool for integrating whole-genome expression results into metabolic networks. *Genome Res* **12**, 1121-6 (2002).
111. Lappe, M. & Holm, L. Algorithms for protein interaction networks. *Biochem Soc Trans* **33**, 530-4 (2005).
112. Gasch, A.P. et al. Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell* **11**, 4241-57 (2000).
113. Galitski, T., Saldanha, A.J., Styles, C.A., Lander, E.S. & Fink, G.R. Ploidy regulation of gene expression. *Science* **285**, 251-4 (1999).
114. Iyer, V.R. et al. The transcriptional program in the response of human fibroblasts to serum. *Science* **283**, 83-7 (1999).
115. Radonjic, M. et al. Genome-wide analyses reveal RNA polymerase II located upstream of genes poised for rapid response upon S. cerevisiae stationary phase exit. *Mol Cell* **18**, 171-83 (2005).
116. Mnaimneh, S. et al. Exploration of essential gene functions via titratable promoter alleles. *Cell* **118**, 31-44 (2004).
117. Chu, S. et al. The transcriptional program of sporulation in budding yeast. *Science* **282**, 699-705 (1998).
118. Roberts, C.J. et al. Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. *Science* **287**, 873-80 (2000).
119. Travers, K.J. et al. Functional and genomic analyses reveal an essential coordination between the unfolded protein response and ER-associated degradation. *Cell* **101**, 249-58 (2000).
120. Ideker, T. et al. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* **292**, 929-34 (2001).
121. Hasty, J., McMillen, D., Isaacs, F. & Collins, J.J. Computational studies of gene regulatory networks: in numero

molecular biology. *Nat Rev Genet* **2**, 268-79 (2001).
122. Bartel, P.L., Roecklein, J.A., SenGupta, D. & Fields, S. A protein linkage map of Escherichia coli bacteriophage T7. *Nat Genet* **12**, 72-7 (1996).
123. Uetz, P. et al. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature* **403**, 623-7 (2000).
124. Walhout, A.J. et al. Protein interaction mapping in C. elegans using proteins involved in vulval development. *Science* **287**, 116-22 (2000).
125. Ito, T. et al. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A* **98**, 4569-74 (2001).
126. Ross-Macdonald, P. et al. Large-scale analysis of the yeast genome by transposon tagging and gene disruption. *Nature* **402**, 413-8 (1999).
127. Winzeler, E.A. et al. Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. *Science* **285**, 901-6 (1999).
128. Fraser, A.G. et al. Functional genomic analysis of C. elegans chromosome I by systematic RNA interference. *Nature* **408**, 325-30 (2000).
129. Maeda, I., Kohara, Y., Yamamoto, M. & Sugimoto, A. Large-scale analysis of gene function in Caenorhabditis elegans by high-throughput RNAi. *Curr Biol* **11**, 171-6 (2001).
130. Giaever, G. et al. Functional profiling of the Saccharomyces cerevisiae genome. *Nature* **418**, 387-91 (2002).
131. Kamath, R.S. et al. Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. *Nature* **421**, 231-7 (2003).
132. Kumar, A. et al. Subcellular localization of the yeast proteome. *Genes Dev* **16**, 707-19 (2002).
133. Huh, W.K. et al. Global analysis of protein localization in budding yeast. *Nature* **425**, 686-91 (2003).
134. Tong, A.H. et al. Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* **294**, 2364-8 (2001).
135. Jorgensen, P. et al. High-resolution genetic mapping with ordered arrays of Saccharomyces cerevisiae deletion mutants. *Genetics* **162**, 1091-9 (2002).
136. Tong, A.H. et al. Global mapping of the yeast genetic interaction network. *Science* **303**, 808-13 (2004).
137. Ren, B. et al. Genome-wide location and function of DNA binding proteins. *Science* **290**, 2306-9 (2000).
138. Iyer, V.R. et al. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* **409**, 533-8 (2001).
139. Lieb, J.D., Liu, X., Botstein, D. & Brown, P.O. Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. *Nat Genet* **28**, 327-34 (2001).
140. Harbison, C.T. et al. Transcriptional regulatory code of a eukaryotic genome. *Nature* **431**, 99-104 (2004).
141. Gavin, A.C. et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141-7 (2002).
142. Ho, Y. et al. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature* **415**, 180-3 (2002).
143. Novick, P., Osmond, B.C. & Botstein, D. Suppressors of yeast actin mutations. *Genetics* **121**, 659-74 (1989).
144. Guarente, L. Synthetic enhancement in gene interaction: a genetic tool come of age. *Trends Genet* **9**, 362-6 (1993).
145. Hartman, J.L.t., Garvik, B. & Hartwell, L. Principles for the buffering of genetic variation. *Science* **291**, 1001-4 (2001).
146. Grunenfelder, B. & Winzeler, E.A. Treasures and traps in genome-wide data sets: case examples from yeast. *Nat Rev Genet* **3**, 653-61 (2002).
147. Rust, A.G., Mongin, E. & Birney, E. Genome annotation techniques: new approaches and challenges. *Drug Discov Today* **7**, S70-6 (2002).
148. Marcotte, E.M., Pellegrini, M., Thompson, M.J., Yeates, T.O. & Eisenberg, D. A combined algorithm for genome-wide prediction of protein function. *Nature* **402**, 83-6 (1999).
149. Ge, H., Liu, Z., Church, G.M. & Vidal, M. Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. *Nat Genet* **29**, 482-6 (2001).
150. Jansen, R., Greenbaum, D. & Gerstein, M. Relating whole-genome expression data with protein-protein interactions. *Genome Res* **12**, 37-46 (2002).
151. Kemmeren, P. et al. Protein interaction verification and functional annotation by integrated analysis of genome-scale data. *Mol Cell* **9**, 1133-43 (2002).
152. Schlitt, T. et al. From gene networks to gene function. *Genome Res* **13**, 2568-76 (2003).
153. Lee, I., Date, S.V., Adai, A.T. & Marcotte, E.M. A probabilistic functional network of yeast genes. *Science* **306**, 1555-8 (2004).
154. Parsons, A.B. et al. Integration of chemical-genetic and genetic interaction data links bioactive compounds to cellular target pathways. *Nat Biotechnol* **22**, 62-9 (2004).
155. Kemmeren, P., Kockelkorn, T.T., Bijma, T., Donders, R. & Holstege, F.C. Predicting gene function through

systematic analysis and quality assessment of high-throughput data. *Bioinformatics* **21**, 1644-52 (2005).

OffBase - an Integrated Microarray Data Analysis Platform

Chapter II

OffBase – an Integrated Microarray Data Analysis Platform

Abstract
Microarrays are used throughout many different areas of biological research. Managing data obtained from microarrays is still a challenge. To facilitate this process, microarray data analysis platforms are required that can function as a central storage and management system for all microarray related data. Special emphasis should be placed on rigorous quality control and annotation, as this will increase overall data integrity, reproducibility and utility. Here we present a framework that provides many additional features on top of currently existing systems. Most notably are rigorous quality control, MIAME (Minimal Information About a Microarray Experiment) compliant array and sample annotation and support for standardized workflows.

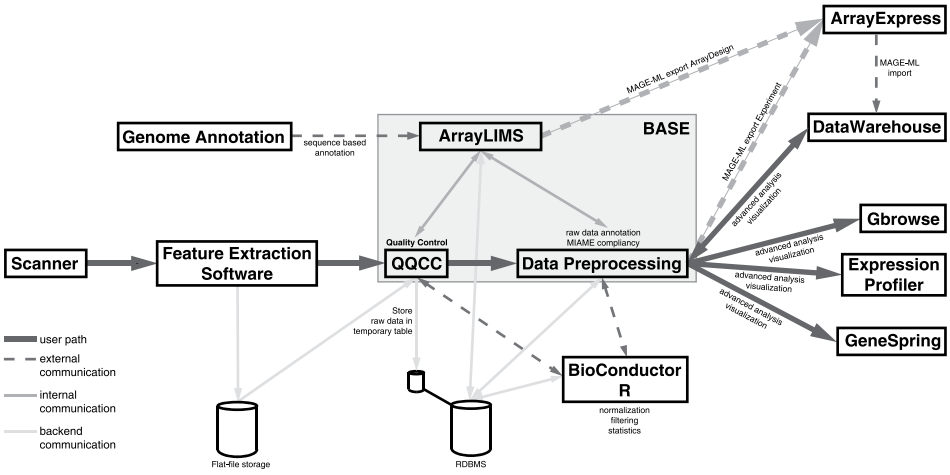
Introduction

Microarrays are used throughout the life sciences, including studying transcription regulation, drug development, pathological conditions, tumor profiling, etc[1,2]. Many genes can be studied in parallel using microarrays, typically ranging from thousands to tens of thousands. However, managing data obtained from microarrays is still a challenge. Different aspects of managing this type of data, such as array production, array annotation, quality control and sample annotation, have to be taken care of even before any data analysis can occur. Although different attempts have been made in the past, both commercially[3-5] and academic[6-9], still no single platform exists that covers all these aspects as well as providing a starting point for many different analysis routines. The best-developed, publicly available system at the moment seems to be BASE (BioArray Software Environment)[6]. BASE provides a way to store and manage microarray related data in a structured database management system, linking the raw microarray data to the slides and array designs used. Furthermore, additional information about the samples, its origins and how they were treated can be recorded in BASE. Within BASE different analyses routines can be started and results linked back to the original data. However, it does not incorporate rigorous quality control, enforcement of proper sample annotation and standardized workflows. Therefore a system is required that can function as a central storage and management system for all microarray related data, incorporating rigorous quality control, annotation and workflow management.

To ensure that the data can be interpreted, reanalyzed and compared with other data, the international Microarray Gene Expression Database Society (MGED)[10] has established standards to exchange and publish microarray data. The minimal amount of information that should be reported to ensure the interpretability of the experimental results is described in MIAME (Minimal Information About a Microarray Experiment)[11]. In addition, many commonly used terms are encapsulated inside the MGED ontology[12]. An ontology describes the relationships between different concepts using controlled vocabularies. Through these controlled vocabularies sharing, reusage and overall integrity of data are better ensured. Another standard the MGED society has developed is the data exchange format MAGE-ML (MicroArray Gene Expression Markup Language)[13]. MAGE-ML will facilitate interoperability between different software packages, data submission to public microarray data repositories such as GEO (Gene Expression Omnibus)[14] and ArrayExpress[15] and data sharing within large consortia or collaborating laboratories.

To incorporate rigorous quality control, array and sample annotation and workflow management, we developed a system called OffBase that provides an integrated framework for storing, analyzing and sharing microarray data. OffBase has strong roots in the open-

Figure 1 | Microarray Data Analysis Pipeline
A schematic overview of the microarray data analysis pipeline is shown. The main user path is depicted with black arrows, starting with the scanning of slides. See also Appendix; Color Figure 2.1.



source community through the integrated use of currently existing open source projects such as BASE[6], R[16], BioConductor[17] and MAGEstk[18]. OffBase is a MIAME-supportive database and analysis platform, easily accessible through a web interface.

Results

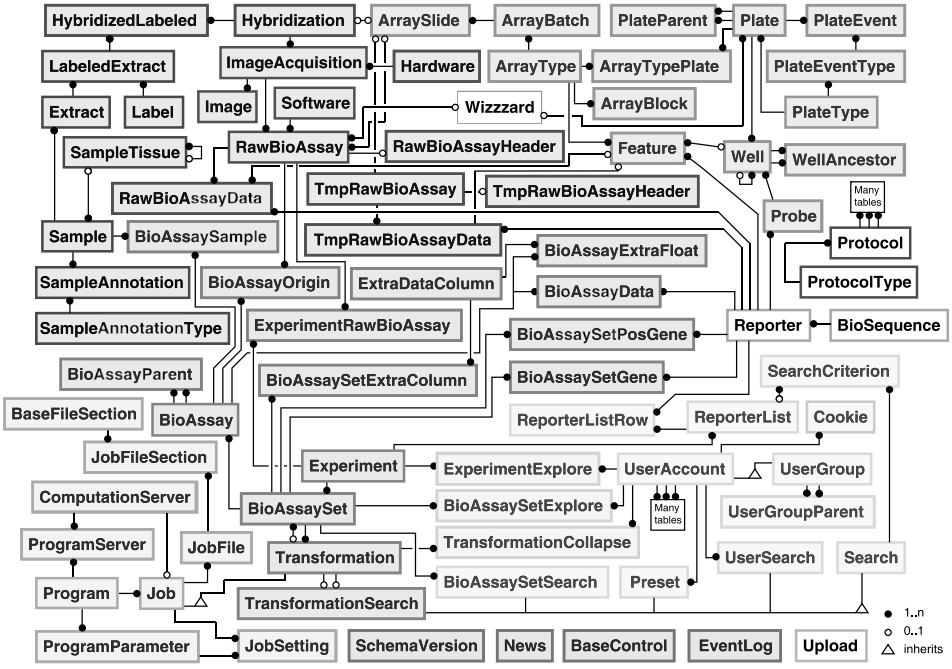
Microarray Data Analysis Framework

To enable an intuitive, flexible and efficient microarray data analysis environment, a general framework is setup as depicted in Figure 1. A typical user path is indicated with blue arrows and starts with scanning of microarray images. Extraction of signal and background intensities is performed through feature extraction software such as ImaGene[25] or GenePix[26]. The resulting files are read from disk, stored inside a temporary table within the database and subjected to rigorous quality checks through a program called QQCC (Quick Quality Check for Chips). Once the slides have been run through QQCC, the data can be selected for further annotation and analysis. The ArrayDesigns, QQCC and MIAME compliant annotation are all integrated within BASE[6], an open source microarray database system. Various filtering, normalization and statistical analysis routines are available within BASE, more advanced analyses can be performed using either Expression Profiler NG[27], GeneSpring[3] or GBrowse[28]. In the future a more generalized data warehouse will also be available for comparison with publicly available microarray data. Before publication, data adhering to the MIAME standards[11] should be submitted to a public repository. In this setup ArrayExpress[15] is used as the public repository to which data will be submitted through MAGE-ML[13].

BASE

Within the microarray data analysis framework, BASE has been chosen as the central component, connecting the different functional modules with each other. To ensure that BASE is fully MIAME supportive the database schema has been designed in accordance with MIAME standards (Figure 2). The tables depicted in this figure can be subdivided into different functional categories; tables dealing with arrays (orange), hybridizations (purple), samples (brown), data analysis (pink), jobs (green) and users (blue). The database back-

Figure 2 | BASE Database Schema
A schematic overview of the full BASE database schema is depicted, adopted from the original version[36]. The database schema can be subdivided in different categories; tables dealing with arrays (orange), hybridizations (purple), samples (brown), data analysis (pink), jobs (green), reporters (green border), protocols (blue border) and users (blue). See also Appendix; Color Figure 2.2.



end is run in PostgreSQL[22], an open-source relational database management system. The php[20] generated web front-end provides a unified, platform independent interface that integrates the most commonly used components and analyses routines and provides export capabilities to widely used third party external programs.

ArrayLIMS

The first component that is an integral part of BASE is the arrayLIMS section. Array designs can be based on either commercially available arrays or in-house developed ones. Information about the array designs, array batches and array slides is stored inside the database. The sequences of the oligos or cDNAs deposited on the arrays are used to (re)annotate the array designs through genome databases such as ensembl[29] and SGD[30]. This way important information about the genes, such as chromosomal position or possible cross-hybridization, is readily available throughout the entire system. Furthermore, future updates of gene annotations can easily be incorporated into the system.

Modifications to the public instance of BASE have been made. First, information about the features (physical spots) on the array is split up in a reporter and biosequence part. The reporter table (Figure 2; table with green border) contains detailed information about the oligos or cDNAs used on the array such as description, length, sequence and reporter group. In contrast the biosequence table (Figure 2; table with green border) contains detailed information about the gene or intergenic region the oligo or cDNA was designed for, such as species, systematic name, gene symbol, chromosomal position and database accession numbers. Second, additional administrative features are added

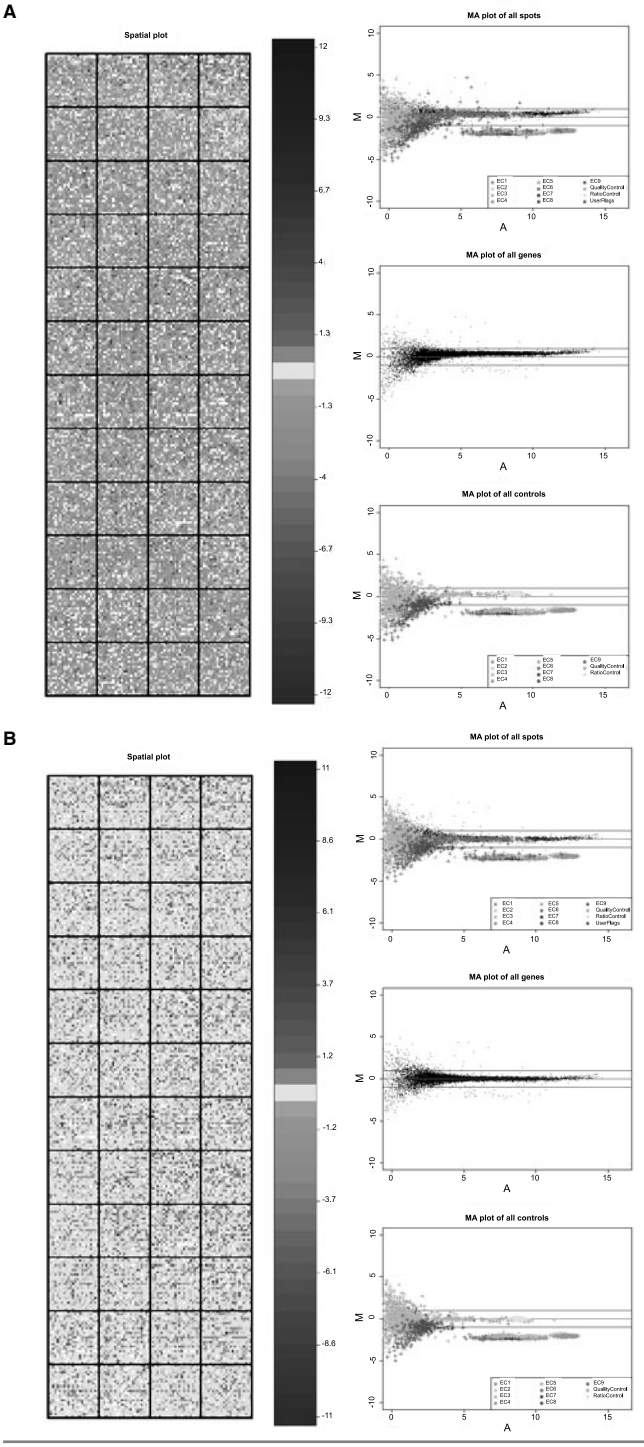
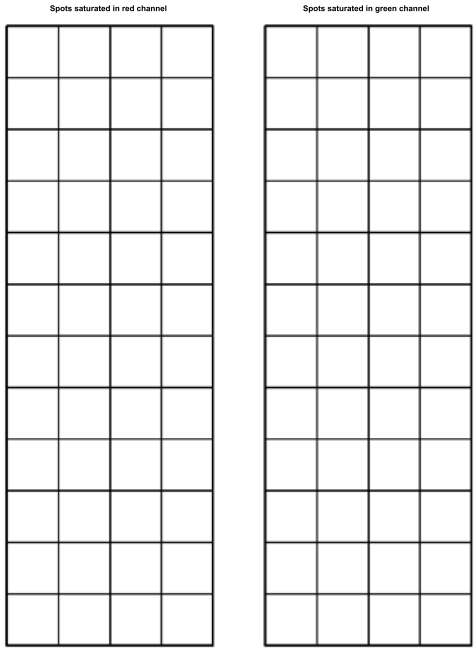
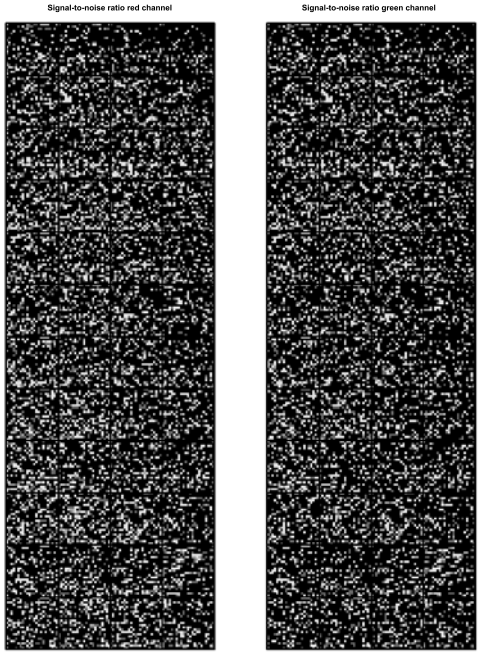


Figure 3 | QQCC Quality Report
The quality report depicted in this figure was obtained from a human self-self hybridization with the external normalization controls spiked in at a four fold difference between the two samples. (**A**) Displayed here are the ratios across the slide (spatial plot, left) for the raw data. Three MA plots of the raw data (ratio vs. intensities) are shown on the right: the first plot shows all spots with the controls and user flags highlighted, the second plot shows only the genes and the third plot shows only the controls. (**B**) Same as A, data shown is obtained after lowess normalization on genes. (**C**) Displayed here are the saturated spots (average background corrected signal intensity above 45000). Red and green dots represent saturated spots in the red and green channel respectively. In this case no saturated spots are detected. (**D**) Displayed here is a spatial plot with the raw data binned according to the number of standard deviations above background (SNR): SNR < 2 (black), SNR < 3 (red), SNR < 4 (orange) and SNR >= 4 (yellow). (**E**) Displayed here are the normalized ratios of all genes on their respective chromosome location. (**F**) Displayed here are the percentage of good, marginal and bad spots based on the quality metrics obtained from Chen et al.[24]. See also Appendix; Color Figure 2.3.

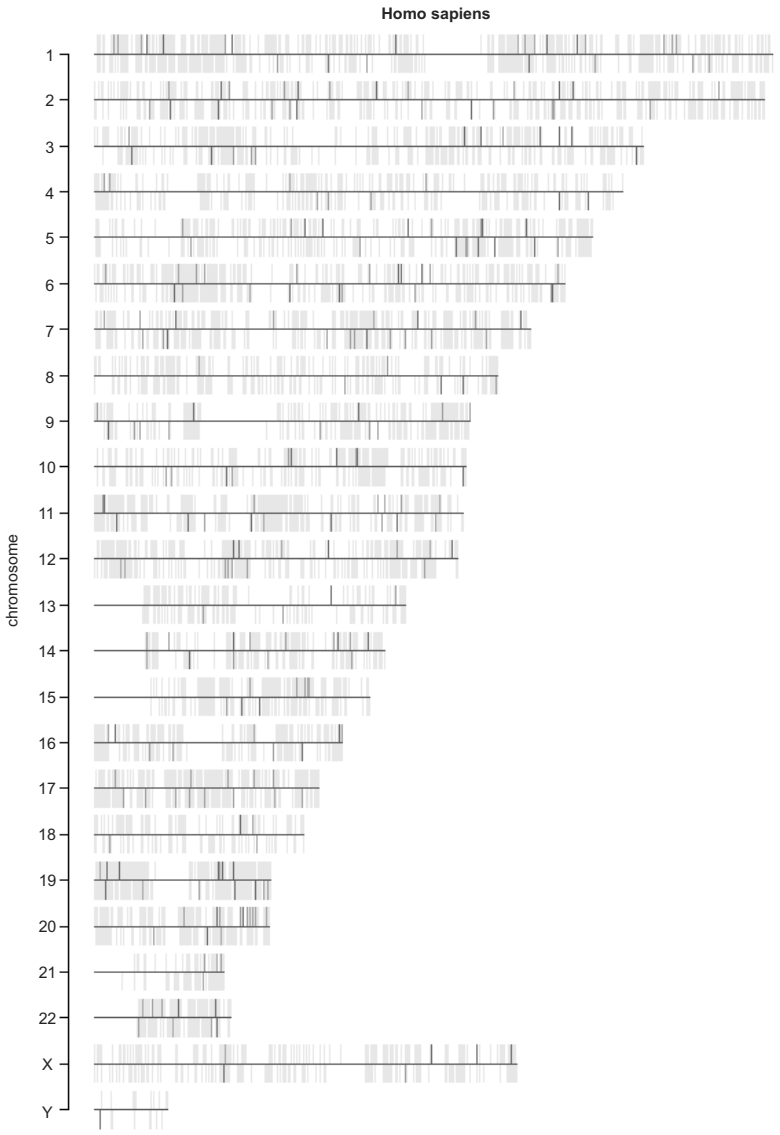
C


Spots saturated in red channel
Spots saturated in green channel

D


Signal-to-noise ratio red channel
Signal-to-noise ratio green channel

E



F

% good spots	94.8
% marginal spots	0.6
% bad spots	4.6
% spots poor Signal-to-Noise ratio	0.5
% spots poor background flatness	1.1
% spots poor Signal Consistency	0.2
% Black Holes	2.8
Banana Factor	1.13

to the arrayLIMS. For instance, slides can be reserved and checked out for certain users, enabling automatic linking of future data uploads to users based on the slide barcode. Third, additional MIAME required fields are added to ensure that array designs are stored in accordance with MIAME compliance. Single array designs can then be submitted to ArrayExpress through MAGE-ML, so that future publications and experiments within BASE can use the accession numbers associated with these array designs.

QQCC

Another component that has been integrated in BASE is QQCC (Quick Quality Check for Chips), allowing for rigorous quality control of single hybridizations. The major part of QQCC is written in R[16], an open-source statistical programming environment and available through the QQCC package within the BioConductor suite[17] of microarray related analyses routines. QQCC allows a user to assess the quality of a hybridization based on a single quality report created as a PDF file. This quality report consists of two parts.

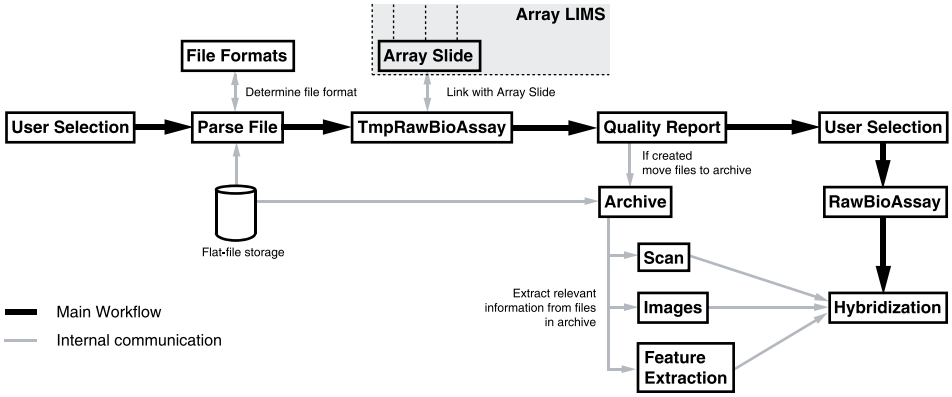
First, a number of figures are drawn for visual inspection of the hybridization at hand. The first two figures show a spatial plot and three MA plots ($\log_2$ ratio vs. $\log_2$ intensity) of the raw and normalized data respectively (Figure 3A, B). The spatial plot shows the ratios according to their physical location on the array, thus showing any biases introduced due to the physical position. The three different MA plots show a scatter plot of all features, all genes and all controls on the array respectively. As both raw and normalized data are shown, an assessment can be made whether the spatial biases in the raw data are corrected after normalization. Another figure shown is a signal-to-noise ratio (SNR) plot for both channels (Figure 3C). In this figure the SNR is divided into four different categories and displayed according to the physical position on the array. This way, areas with a relatively poor signal are easily distinguished and can be excluded from further downstream analysis. The fourth figure displays the saturated features according to their position on the array (Figure 3D). In this particular case no saturated spots are present on the array. The final figure shows the ratios of the different genes according to their chromosomal position (Figure 3E). As such, chromosomal rearrangements can easily be detected and accounted for. Furthermore, parts of the different chromosomes not represented on the array also become evident. In this particular case it is clear that no reporters are present for genes at the start of chromosome 13, 14 and 15 (Figure 3E).

The second part of the quality report shows an overview about the percentage of good, marginal and bad features, based on the signal-to-noise ratio, background flatness, black holes and signal consistency (see Experimental Procedures)[24]. Based on these four different quality metrics a quality value is assigned to each individual feature that can be combined into one quality value per array, allowing for an unbiased assessment of the data quality.

Data Integrity and MIAME Annotation

Ensuring data integrity and correct MIAME compliant annotation is an important task, which is taken care of by BASE in different ways. First of all, many of the steps that have to be taken have been automated (Figure 4). The workflow depicted in this figure, shows what happens when a new hybridization is added to the system. The first step is to read data from the result files generated by the feature extraction software. The different file formats are automatically detected based on the file headers and file formats stored in the database. If no appropriate match is found, a warning is given and the correct file format can be added, after which the process continues. The result file is subsequently uploaded into a temporary table (TmpRawBioAssay) (Figure 2) inside BASE and automatically linked to the correct array slide by extracting the barcode from the filename. QQCC is then used to generate the quality report. If created successfully, the result files and images are moved to a storage area for archival purposes. Files can then be selected for further analysis. These are moved to a permanent table (RawBioAssay) (Figure 2) inside BASE and relevant information about the feature extraction, scanning and images is extracted

Figure 4 | Standardization and Quality Control of Hybridizations
Workflow representation of hybridization data uploads to BASE. The main workflow or user path is highlighted with thick black arrows and starts with selecting the files for upload. Communication between the different components within BASE is depicted with thin grey arrows.



from the files on disk and added to BASE. Using these workflows, users are forced to complete certain steps (e.g. every result file should have a quality report) before they can continue with further downstream analysis.

Another way that BASE ensures data integrity is the use of the MGED ontology[12]. A separate section has been added to BASE that integrates the MGED ontology within BASE. Annotation of samples and experiments is facilitated in many places through the use of drop-down menus, obtained directly from the MGED ontology structure.

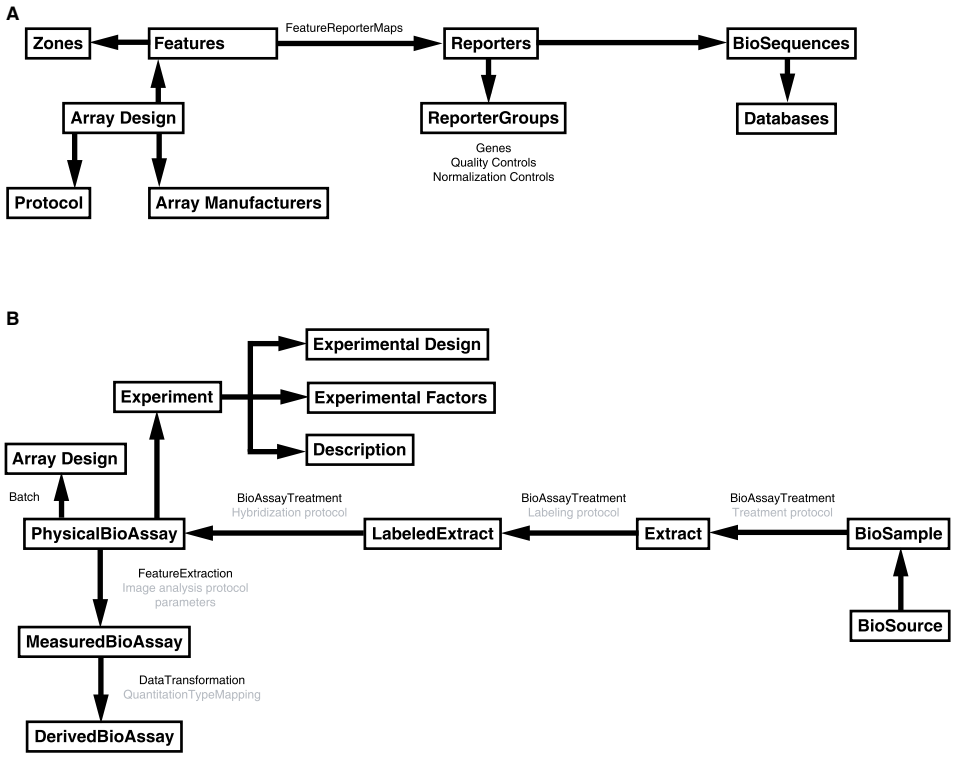
BioConductor Connectivity

BioConductor[17] is a collection of commonly used analysis routines for microarray data analysis in R. As many of the filtering, normalization and statistical analyses are run inside BioConductor, a direct link between the packages within BioConductor and BASE has been developed. Given a certain hybridization and user id, the BioConductor packages can obtain data directly from BASE using three different functions added to a commonly used BioConductor package for microarray data analysis (marray package). The functions "readDB.Base.marrayLayout", "readDB.Base.marrayRaw", "readDB.Base.Gnames" and "readDB.Base.samples" obtain the array design layout, raw data, gene annotation and sample annotation respectively. Once the data is obtained and a marrayRaw object is created using these routines, further downstream analyses can be performed such as normalization using lowess[31] or vsn[32] or statistical analyses using SAM[33], maanova[34] or limma[35].

MAGE-ML Export

Once the data has been annotated according to the MIAME guidelines, the data can be exported to MAGE-ML. Using the MAGEstk[18] Java API, a mapping has been created between all relevant information inside BASE and the MAGE objects necessary for MAGE-ML export. The MAGE-ML contains detailed information about the array designs (Figure 5A) and experiments (Figure 5B). After MAGE-ML export, the MAGE-ML files are uploaded to an ftp site at ArrayExpress and validated against the dtd (DocumentTypeDefinition) and database. When the MAGE-ML has been validated, the data is loaded into ArrayExpress. An accession number is assigned and reported back to BASE.

Figure 5 | MAGE-ML Representation of Array Designs and Experiments
(**A**) Simplified representation of MAGE-ML structure for an array design. Arrows indicate the parent-child relationships between the different objects. (**B**) Simplified representation of MAGE-ML structure for an experiment. Black arrows indicate the parent-child relationships between the different objects.



Discussion

Here we present OffBase, an integrated framework for microarray data storage, retrieval and analysis. Through the integration of solid and widely supported open-source projects, such as BASE[6] and BioConductor[17], a solid foundation has been created upon which many different analyses routes can be explored and developed (Figure 1). Furthermore, the data integrity is ensured through additional rigorous quality control of individual hybridizations and subsequent MIAME compliant annotation[11] through the use of the MGED ontology[12]. Publication of high-quality, well-annotated data is further facilitated by direct submission of array designs and experiments to the public microarray repository ArrayExpress[15].

A particular point of interest, reflected throughout the entire framework, is rigorous quality control. This already starts at the early stages of the experiment, through careful setup of the experimental design, number of replicates, mRNA quality control and monitoring of labeling efficiency. Within the framework, QQCC plays an integral part in ensuring that a reliable assessment of the data quality can be made. Currently, this still relies heavily on the expertise of the individual user when interpreting the different figures represented within the quality report (Figure 3). However, an unbiased quality value is calculated for each hybridization based on four independent quality metrics[24]. At first this quality value can be used to rank the different hybridizations within an experiment, so

that the most extreme outliers can be detected and removed form further analysis. As the number of hybridizations will increase, a knowledge base will arise, that can be used to determine cutoffs, separating poor quality hybridizations from good quality hybridizations, based on the species, array design or array batch. That way more automated, unbiased assessments can be made about the data quality. However, the quality of hybridization largely depends on the biological sample, array batch and protocols used and should therefore be assessed in that context.

Proper array and sample annotation plays a crucial role in maintaining a high level of accuracy, interpretability and reproducibility of the results. To ensure this, embracement of international annotation standards such as MIAME, tight integration with ontologies such as the MGED ontology and submission of microarray data to public repositories such as ArrayExpress are essential. However, there always is a fine balance between the amount of work that has to be performed when annotating the data and the level of detail that is recorded. Facilitating this annotation process is a very important task that needs to be taken care off by the underlying system. The first steps in that direction have been made here through the integration of the MGED ontology. However, more automated bulk annotation pages that use the information stored within the MGED ontology have to be developed, so that the workload is kept to a minimum, while still storing detailed information about the arrays and samples used throughout experiments.

Support for workflows should also be an essential property of any underlying system for two reasons. First, through the use of workflows, users are forced to take a certain route and can only continue further downstream analyses after certain criteria or steps are satisfied. This way an overall higher level of consistency and interpretability of the data can be obtained. Second, definition of workflows for the most commonly used analysis steps and/or processes will greatly facilitate biological hypotheses generation and verification. At the moment only a limited number of workflows have been implemented within the framework, of which one example has been shown here (Figure 4). However as the different methodologies used for interpreting the biological data will mature over time, more of these workflows should be incorporated into the system to enable efficient and reproducible analysis results.

An important factor determining the success of an experimental setup is the ability to place the data in the correct biological context. For that purpose, providing a way to communicate with external third party programs, pulling in that biological knowledge is essential. These external programs can provide additional information about for instance the regulatory motifs present within the upstream sequences of genes or the biological processes that these genes are involved in, which is not necessarily available within the framework. Currently, communication with external programs can be achieved through export of tab-delimited files, however communication with some of the most commonly used external programs should be more streamlined to increase the usability of these programs.

The framework presented here, provides many additional features on top of currently existing systems. Most notably are rigorous quality control, MIAME compliant array and sample annotation and support for standardized workflows.

Experimental Procedures

Front-end

The front-end consists of a RedHat Linux Advanced Server 3.0[19] (dual CPU, 1.13 GHz, 1 GB ram). The interface consists of php scripts, run through mod_php (version 4.3.2-19ent)[20] inside Apache 2 (version 2.0.46-44.ent)[21].

Database Back-end

The database back-end consists of a RedHat Linux Advanced Server 3.0 (dual CPU, 2.4 GHz, 6 GB ram, 210 GB disk space). The relational database management system (RDBMS) used to manage the database is PostgreSQL (version 7.3.9-2)[22].

Computational Back-end
The computational back-end consists of nine RedHat Linux Advanced Server 3.0 servers (dual CPU, 1.13 GHz, 2 GB ram). All jobs run on the computational back-end are submitted and managed through the PBS Pro queuing system (version 7.0)[23]. All servers are running the same version of php (version 4.3.2-19ent) and R (version 2.0.1)[16].

QQCC
The quality metrics used within parts of the quality report are derived from Chen et al.[24]. The formulas have been taken from that paper and implemented in the R package QQCC. The Signal-to-Noise Ratio (SNR) quality weight is defined by

$$w_{SNR} = \begin{cases} 0, & \frac{R+G}{2\hat{\sigma}_B} \le 3 \\ \frac{R+G}{6\hat{\sigma}_B}, & 3 < \frac{R+G}{2\hat{\sigma}_B} \le 6 \\ 1, & otherwise \end{cases}$$

The background flatness quality weight is defined by $W_B = \min\{W_{BR}, W_{BG}\}$, where

$$w_{BR} = \begin{cases} 0, & BR_k \ge \mu_{BR} + 4\sigma_{BR} \\ \frac{(\mu_{BR} + 6\sigma_{BR}) - BR_k}{3\sigma_{BR}} & \mu_{BR} + 4\sigma_{BR} \le BR_k < \mu_{BR} + 6\sigma_{BR} \\ 1, & BR_k < \mu_{BR} + 4\sigma_{BR} \end{cases}$$

and W_{BG} is defined similarly. The signal consistency quality weight is defined by

$$w_s = \begin{cases} 0, & 1.1 < cv_{\min,k} \\ \frac{cv_{\min,k} - 0.9}{0.2}, & 0.9 < cv_{\min,k} \le 1.1 \\ 1, & cv_{\min,k} \le 0.9 \end{cases}$$

The black hole quality weight is defined by $W_{Bneg} = \min\{W_{BRneg}, W_{BGneg}\}$, where

$$w_{BRneg} = \begin{cases} 0, & I_R - B_R \le 0 \\ 1, & I_R - B_R > 0 \end{cases}$$

and W_{BGneg} is defined similarly.

BioConductor
The following packages inside the BioConductor release v1.2[17] have been used throughout the framework: annotate (version 1.5.1); BioBase (version 1.5.0); geneplotter (version 1.5.2); limma (version 1.8.6); marray (version 1.5.8); Rdbi (version 1.0.4); RdbiPgSQL (version 1.0.9). The marray packages have been modified to incorporate specific functions necessary for the QQCC package.

MAGEstk Java API
The MAGEstk[18] Java API (2003-02-23) is used to export ArrayDesigns, Experiments and Protocols to MAGE-ML[13].

References

1. Lockhart, D.J. & Winzeler, E.A. Genomics, gene expression and DNA arrays. *Nature* **405**, 827-36 (2000).
2. Young, R.A. Biomedical discovery with DNA arrays. *Cell* **102**, 9-15 (2000).
3. *GeneSpring* [http://www.silicongenetics.com/cgi/SiG.cgi/Products/GeneSpring/index.smf].
4. *Rosetta Resolver* [http://www.rosettabio.com/products/resolver/default.htm].
5. *GeneTraffic* [http://www.iobion.com].
6. Saal, L.H. et al. BioArray Software Environment (BASE): a platform for comprehensive management and analysis of microarray data. *Genome Biol* **3**, SOFTWARE0003 (2002).
7. *GeneX* [http://genex.sourceforge.net/].

8. *maxd* [http://bioinf.man.ac.uk/microarray/maxd/].
9. Stanford Microarray Database.
10. *MGED - Microarray Gene Expression Data Society* [http://www.mged.org].
11. Brazma, A. et al. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet* **29**, 365-71 (2001).
12. *Microarray Gene Expression Data (MGED) Society Ontology Working Group (OWG)* [http://mged.sourceforge.net/ontologies/index.php].
13. Spellman, P.T. et al. Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biol* **3**, RESEARCH0046 (2002).
14. Barrett, T. et al. NCBI GEO: mining millions of expression profiles--database and tools. *Nucleic Acids Res* **33**, D562-6 (2005).
15. Parkinson, H. et al. ArrayExpress--a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res* **33**, D553-5 (2005).
16. *R: The R Project for Statistical Computing* [http://www.r-project.org].
17. *BioConductor: Open Source Software for Bioinformatics* [http://www.bioconductor.org].
18. *MGED software - MAGEstk: The MAGE Software Toolkit* [http://www.mged.org/Workgroups/MAGE/magestk.html].
19. *RedHat* [http://www.redhat.com].
20. *PHP: Hypertext preprocessor* [http://www.php.net].
21. *Apache HTTP Server Project* [http://httpd.apache.org].
22. *PostgreSQL* [http://www.postgresql.org].
23. *PBS Professional* [http://www.altair.com/software/pbspro.htm].
24. Chen, Y. et al. Ratio statistics of gene expression levels and applications to microarray data analysis. *Bioinformatics* **18**, 1207-15 (2002).
25. *ImaGene* [http://www.biodiscovery.com/imagene.asp].
26. *GenePix Pro* [http://www.axon.com/GN_GenePixSoftware.html].
27. Kapushesky, M. et al. Expression Profiler: next generation--an online platform for analysis of microarray data. *Nucleic Acids Res* **32**, W465-70 (2004).
28. Stein, L.D. et al. The generic genome browser: a building block for a model organism system database. *Genome Res* **12**, 1599-610 (2002).
29. Hubbard, T. et al. Ensembl 2005. *Nucleic Acids Res* **33**, D447-53 (2005).
30. Christie, K.R. et al. Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from Saccharomyces cerevisiae and related sequences from other organisms. *Nucleic Acids Res* **32**, D311-4 (2004).
31. Yang, Y.H. et al. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res* **30**, e15 (2002).
32. Huber, W., von Heydebreck, A., Sultmann, H., Poustka, A. & Vingron, M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* **18 Suppl 1**, S96-104 (2002).
33. Tusher, V.G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A* **98**, 5116-21 (2001).
34. Wu, H., Kerr, M.K., Cui, X. & Churchill, G.A. MAANOVA: A Software Package for the Analysis of Spotted cDNA Microarray Experiments. in *The analysis of gene expression data: methods and software* (eds. Parmigiani, G., Garett, E.S., Irizarry, R.A. & Zeger, S.L.) (Springer, New York, 2002).
35. smyth, G.K. Limma: linear models for microarray data. in *Bioinformatics and Computational Biology Solutions using R and BioConductor* (eds. Gentleman, R., Carey, V., Dudoit, S., Irizarry, R. & Huber, W.) (Springer, New York, 2005).
36. *BASE database schema* [http://base.thep.lu.se/documentation/development/baseschema.ps].

Monitoring Global Messenger RNA Changes in Externally Controlled Microarray Experiments

Chapter III

Patrick Kemmeren, Jeroen van de Peppel, Harm van Bakel, Marijana Radonjic, Dik van Leenen & Frank C.P. Holstege

Embo Reports 2003 Apr; 4(4): 387-93

Monitoring Global Messenger RNA Changes in Externally Controlled Microarray Experiments

Patrick Kemmeren[1,3], Jeroen van de Peppel[1,3], Harm van Bakel[2], Marijana Radonjic[1], Dik van Leenen[1] & Frank C.P.Holstege[1]

[1]Department of Physiological Chemistry, Division of Biomedical Genetics, UMC Utrecht, the Netherlands; [2]DMG Section Research, Division of Biomedical Genetics, UMC Utrecht, the Netherlands; [3]These authors contributed equally to this work.

Abstract

Expression profiling is a universal tool, with a range of applications that benefit from the accurate determination of differential gene expression. To allow normalization using endogenous transcript levels, current microarray analyses assume that relatively few transcripts vary, or that any changes that occur are balanced. When normalization using endogenous genes is carried out, changes in expression levels are calculated relative to the behavior of most of the transcripts. This does not reflect absolute changes if global shifts in messenger RNA populations occur. Using external RNA controls, we have set up microarray experiments to monitor global changes. The levels of most mRNAs were found to change during yeast stationary phase and human heat shock when external controls were included. Even small global changes had a significant effect on the number of genes reported as being differentially expressed. This suggests that global mRNA changes occur more frequently than is assumed at present, and shows that monitoring such effects may be important for the accurate determination of changes in gene expression.

Introduction

Microarray expression profiling is a universal tool, with a range of applications that benefit from the accurate determination of differential gene expression[1,2]. The detection of changes in messenger RNA expression requires normalization between samples. This counters non-biological variation, such as differences in labeled material, local array differences, dye-specific biases and so on[3-5].

Normalization methods consist of an algorithm and the features on the array to which the algorithm is applied. Ideally, such features should have identical signals in all the samples under investigation. Most researchers carrying out microarray analyses have dismissed the idea of using invariant house-keeping genes that are stably expressed across a wide range of experimental conditions[6]. With a few exceptions[5,7], most microarray experiments make use of the expression levels of all genes as normalization features. The assumption underlying this “all-genes” approach is that relatively few transcript levels vary between samples, or that any changes that occur are balanced. Normalization using the expression levels of endogenous genes means that changes are calculated relative to the majority of transcripts. These relative changes do not reflect the absolute changes at the cellular level that occur if global shifts in mRNA populations take place.

The aim of this study was to set up microarray experiments, incorporating external controls, to monitor the effects of inactivating components of the generally required transcription machinery, such as RNA polymerase II[8,9]. The purpose of external normalization controls is to derive a set of signals for which the final outcome is known to be equal among samples. This can be achieved by the addition of equal amounts of control RNA molecules to samples before processing. Here, we present results from experiments

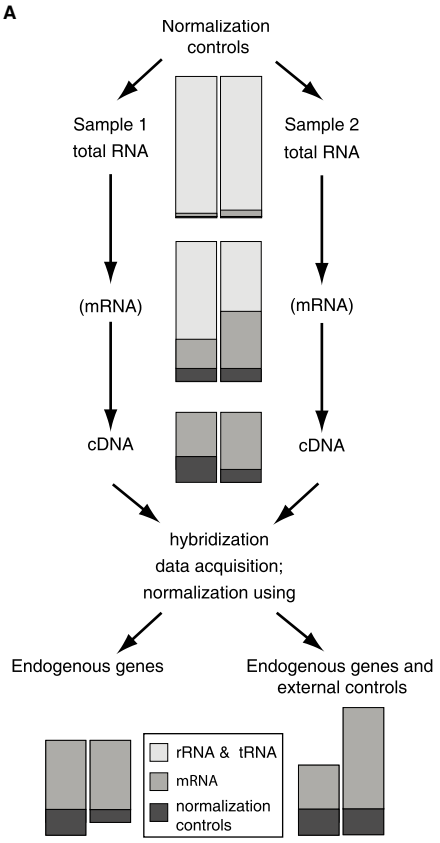


Figure 1 | Normalization Using an External Control (**A**) An example where sample 2 total RNA contains more messenger RNA than sample 1 total RNA. Another possibility is that samples contain different compositions of mRNA. Such differences are not detected when normalization is carried out using expression levels of endogenous genes (bottom left). For normalization using an external control (bottom right), controls are added equivalent to the amount of total RNA. Processing can be varied to exclude mRNA enrichment or to include amplification. (**B**) External control concentrations were varied over three orders of magnitude. The amounts of each *in vitro* transcribed control RNA added as a single mix to total RNA are shown. rRNA, ribosomal RNA; tRNA, transfer RNA.

in which we observed global changes occurring under conditions that, in the past, have been studied without the use of external controls. As well as showing how global effects can be monitored, the results suggest that global mRNA changes occur more frequently than is presently assumed by researchers carrying out microarray analyses. This has important implications for the interpretation of such experiments.

Results

External Normalization Controls

A generalized scheme for carrying out microarray analyses using external controls for normalization is shown in Figure 1A. In this study, total RNA was spiked with a mixture of nine control RNAs. The concentrations of the control RNAs were varied over three orders of magnitude to cover a range of mRNA expression levels (Figure 1B). Oligonucleotide probes (of 70 nucleotides in length) representing each control were spotted at least twice onto the microarray subgrids to generate sufficient data points (960 in total; 20 per subgrid). This allows local, expression-dependent or intensity-dependent normalization[5]. The microarrays also incorporated other controls in addition to the gene probes (6,371 for *Saccharomyces cerevisiae*, each spotted twice, and 16,735 for the human arrays).

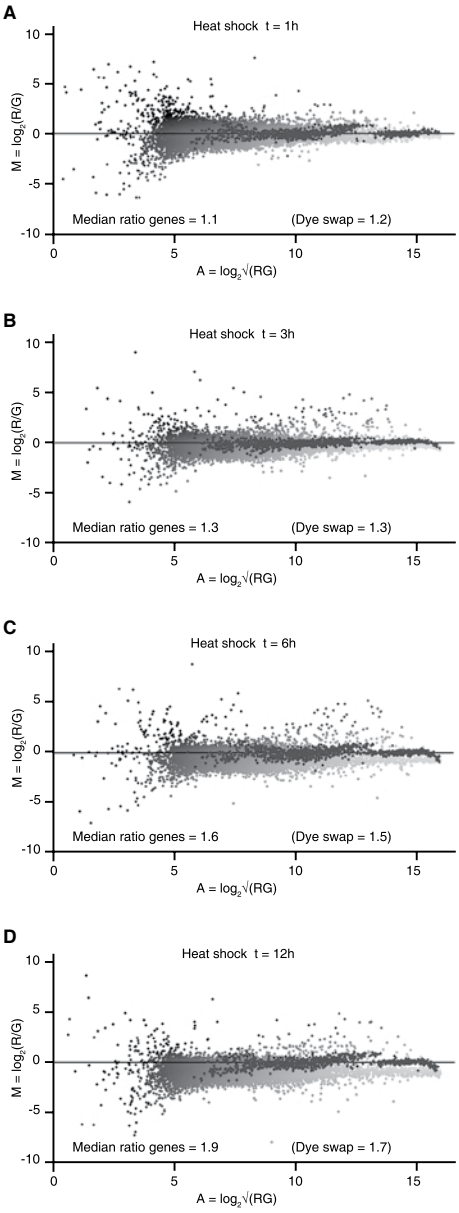
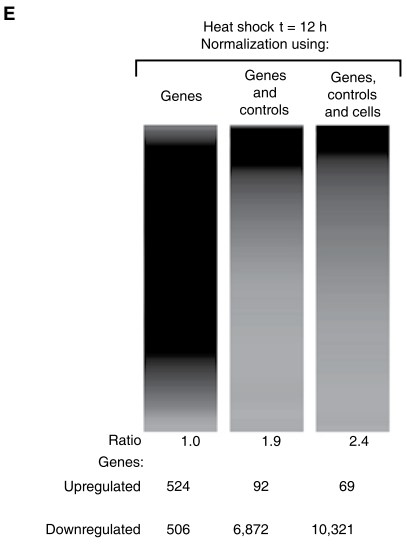


Figure 2 | Global Messenger RNA Changes during Mammalian Heat Shock

Human umbilical vein endothelial cells were heat-shocked and RNA was isolated at the timepoints shown (**A–D**). The result of hybridization of each sample (R) against that of the non-heatshocked reference cells (G) is shown. Each graph is an MA scatterplot (where $M = \log_2(R/G)$ and $A = \log_2\sqrt{(RG)}$)[5]. The *y*-axis shows the $\log_2$ ratio (mean spot intensity minus mean local area background) after normalization using genes and controls (see Experimental Procedures). The values plotted on the *x*-axes are derived from the intensities of both channels. The median change for all genes is indicated, as well as the result of a corresponding dye-swap experiment. (**E**) Comparison of different normalization strategies. From left to right are the results of three strategies applied to the 12-h timepoint. Each column is made up of 16,735 coloured lines, stacked vertically. Each line represents the change for a single gene.Upregulated genes are shown in red and downregulated genes are shown in green. Numbers below the bars indicate the median change of all genes after normalization (as a ratio) and how many genes are reported as being upregulated or downregulated if an arbitrary twofold cut-off is applied. The first column shows the result of Lowess normalization per subgrid using endogenous genes (see Experimental Procedures). The second column shows the result of incorporating external controls into the normalization strategy by carrying out normalization to equivalent amounts of total RNA (normalization using genes and controls; see Experimental Procedures). This leads to a markedly altered perception of the changes that have taken place, with many more transcripts reported as being downregulated. The third column shows the result when the slight drop in the total RNA content of the cells is taken into account. See also Appendix; Color Figure 3.2.



Global Changes during Mammalian Heat Shock

One study that has shown global changes in the mRNA population is that of the heat-shock response of primary human umbilical vein endothelial cells (HUVECs). This is similar to many previous microarray studies that have examined cellular responses, and was investigated as part of our program to understand transcriptional regulation. When external controls were used, a change in global mRNA levels during heat shock was shown, as indicated by the gradual separation of the green gene spots away from the blue control spots (Figure 2A-D). This culminated in an almost twofold median drop in mRNA levels on average in the dye-swap experiments (Figure 2D). Typically, these responses have previously been normalized using expression levels of all genes. Approximately equal numbers of genes were interpreted as being up- or downregulated after normalization of the experiment in this way (Figure 2E; left column), which is a consequence of making the assumption that there is no overall change. Such changes in individual transcript levels are relative to the behavior of most of the transcripts. If a global change occurs, as is the case here (Figure 2D), normalization using genes does not correlate with changes at the cellular level. The interpretation of these experiments differs markedly when external controls are used (Figure 2E; middle column). The number of genes reported as being downregulated increased from 506 to 6,872 when an arbitrary twofold threshold was used. These changes are relative to equivalent amounts of total RNA because spiking of the controls was carried out at the total RNA stage. Counting cells before harvesting allows differences in RNA content in the two states to be taken into account (Figure 2E; right column).

Results of Small Global Changes during Serum Starvation

A second example shows the results of subjecting human mammary gland adenocarcinoma (MCF7) cells to serum deprivation for 30 h (Figure 3). Although this resulted in only a small change in global mRNA levels (1.3-fold median drop; Figure 3A), it had a significant effect on the outcome. More than twice as many genes were found to be differentially expressed as when normalization using the all-genes approach was carried out (Figure 3B).

Yeast Stationary Phase

To rule out the possibility that global changes are restricted to mammalian cell cultures, we also studied stationary phase in *S. cerevisiae*. This showed a 1.8-fold median drop in mRNA levels when normalized using external controls (Figure 4A). Significantly, there is a group of more than 1,000 genes that appear to be upregulated when normalized using the all-genes method, although their mRNA levels had actually decreased. This was only revealed after normalization using external controls was carried out (Figure 4C, left and middle columns). The apparent upregulation when the experiment was normalized using the all-genes approach was seen because the amount of downregulation for these genes is less than that of most transcripts. When cellular RNA content was also taken into account, the changes were found to be more extreme (Figure 4C; right column).

Changes Dependent on Expression Levels

An important issue highlighted by the stationary phase experiment is that of an inconsistent distribution of changes across the range of expression levels: that is, along the *x*-axis in Figure 4A, B. This results in the apparently aberrant behavior of the controls seen when normalization using the all-genes approach was carried out (Figure 4B). Highly expressed genes (on the right halves of the graphs in Figure 4) show far more downregulation. One group of highly expressed mRNAs are the ribosomal protein genes[8].The high degree of downregulation of these genes is consistent with the significant decrease in translation that occurs during stationary phase[10]. Unlike approaches using endogenous transcripts, normalization using external controls is not confounded by an uneven distribution of changes, and also contributes to a more accurate determination of mRNA changes in this respect.

Figure 3 | Minor Global Change during Serum Deprivation lead to a Significantly Different Interpretation
(**A**) Human mammary gland adenocarcinoma (MCF7) cells deprived of serum for 30 h (R) compared with the non-deprived culture (G). (**B**) Identical normalization strategies were used, as described in Figure 2. See also Appendix; Color Figure 3.3.

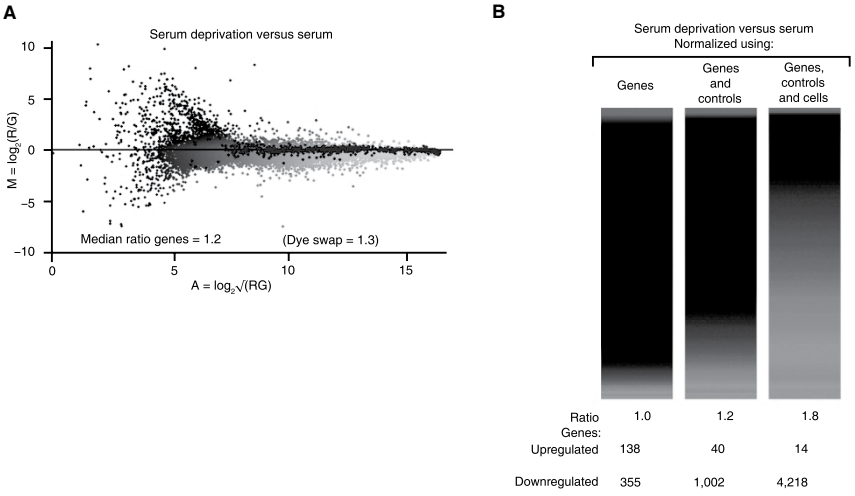
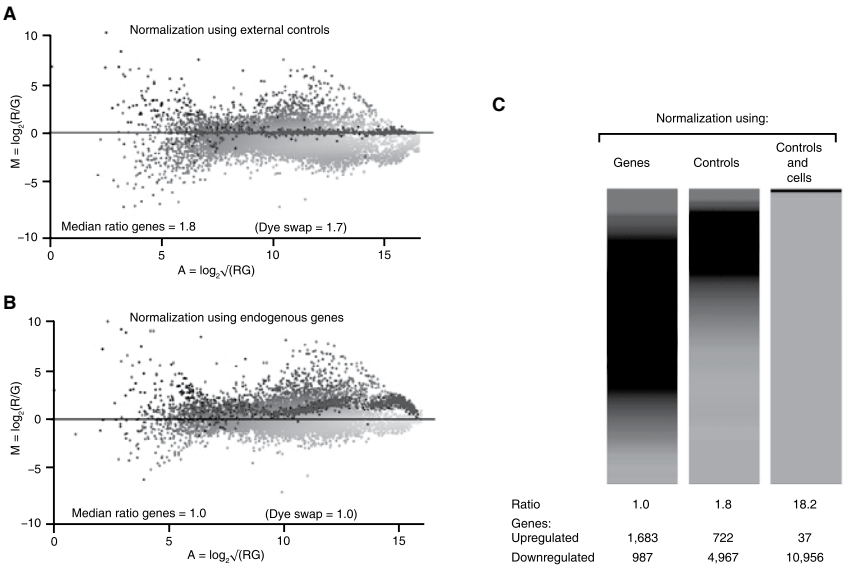


Figure 4 | Yeast Stationary Phase Culture compared with Mid-Log Phase Culture
(**A**) MA scatterplot (where M = $\log_2$(R/G) and A = $\log_2\sqrt{(RG)}$) after Lowess normalization for each subgrid using controls (see Experimental Procedures). (**B**) Lowess normalization for each subgrid using genes (see Experimental Procedures). The aberrant pattern of external control spots (blue) occurred because messenger RNA levels had not changed uniformly across the entire range of expression levels (along the *x*-axis). RNA levels for genes with higher expression levels had dropped to a greater degree than those of genes expressed at lower levels. Due to saturation of the signals in the scanned images, this effect seemingly decreases for a small group of genes expressed at the highest levels, resulting in the curve of the control spots to the far right of the graph. (**C**) Comparison of different normalization strategies. The left and middle columns correspond to the graphs shown in (**B**) and (**A**), respectively. The drop in total RNA yields per cell (see Experimental Procedures) can also be taken into account (right column). Because each probe was spotted twice on the yeast arrays, the numbers reflect how many spots have changed. R, stationary phase culture; G, mid-log phase culture. See also Appendix; Color Figure 3.4.



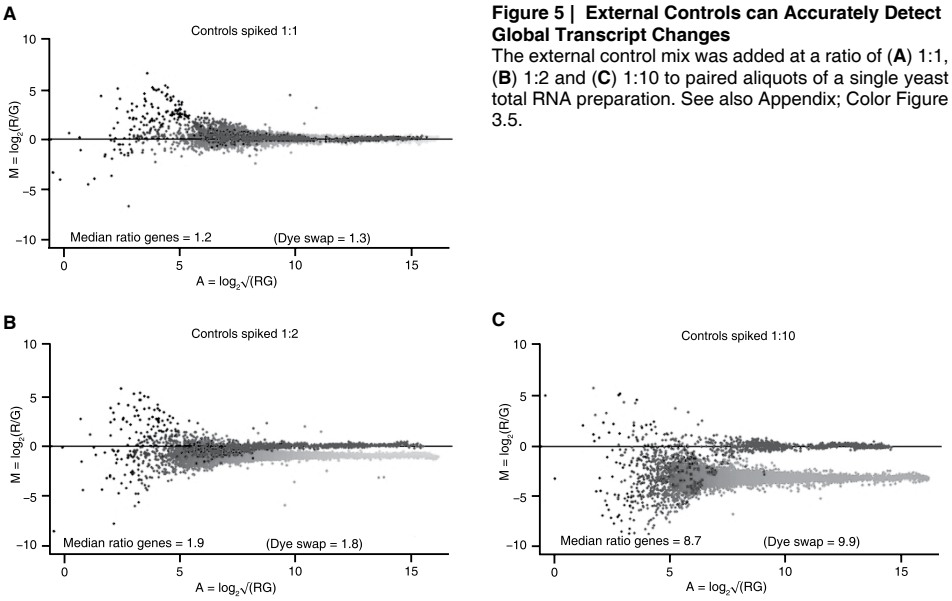


Figure 5 | External Controls can Accurately Detect Global Transcript Changes
The external control mix was added at a ratio of (**A**) 1:1, (**B**) 1:2 and (**C**) 1:10 to paired aliquots of a single yeast total RNA preparation. See also Appendix; Color Figure 3.5.

Quantitative Testing of Reported Changes

How accurate are the changes observed when normalization is carried out using external controls? Carrying out RT–PCR (reverse transcription followed by PCR) on a selected set of mRNAs leads to the question of what reference transcript should be chosen for the normalization of such experiments. The accuracy of the reported global mRNA changes was therefore tested by spiking the control mixtures in different ratios in identical pairs of RNA samples. The ratios chosen were 1:1, 1:2 and 1:10, thereby simulating 1-, 2- and 10-fold changes in the mRNA population relative to the controls. Normalization using external controls allowed the accurate determination of changes in mRNA levels relative to the controls. The gene spots drifted away from the control spots as the spiking ratio was increased to 1:10 (Figure 5A-C). The median amount of change observed in the expression of the genes after normalization was almost identical to the spiking ratios, with values of 1.0-, 1.9- and 9.3-fold, respectively, for the average of each dye swap. This shows that the global changes reported here (up to twofold median changes; Figures 2–4) are likely to be correct.

Discussion

This study focuses on the features used for single-slide normalization and shows how the incorporation of external controls can markedly alter the interpretation of microarray experiments.

External controls have been used previously for normalization in studies examining the artificial inactivation of RNA polymerase II[8,9]. In these cases, global changes were expected. The experiments presented here show that global changes can also occur under more conventional experimental conditions. We have also shown how these controls can be applied to cope with local, intensity-dependent systematic variation[5] by representation in sufficient numbers on each microarray subgrid, and by spiking over a range of levels. As well as being useful for normalization, controls such as these are useful for monitoring sample labeling, optimization of all microarray protocols, and as external controls for the

reported ratios.

Recent papers have discussed the importance of experimental design and normalization algorithm choice[3-5,11-13]. An important improvement has been the adoption of normalization algorithms that take into account local, intensity-dependent systematic variation[5]. Regardless of the algorithm applied, most current analyses rely on assumptions of evenly distributed changes and/or on the absence of global shifts. As well as the wide use of the level of expression of all genes as an invariant feature, alternative normalization features that have been proposed include housekeeping genes, spotted microarray sample pools or spots containing genomic DNA (for overviews, see 5, 12). Normalization using such features does not reveal global, unbalanced changes. The use of a common reference sample, made up of a collection of all the DNAs represented on arrays, has also been proposed[14,15]. This allows better comparisons between slides and also overcomes the problem of obtaining negligible signals for underrepresented mRNAs. However, such approaches do not address the possibility of global, unbalanced changes. It is likely that the use of a combination of external controls and a common pooled, spotted reference sample may be the best way of overcoming several problems simultaneously.

Monitoring global effects is not required, or feasible, for every microarray experiment. The advantage of the use of external controls is that it allows the possibility of detecting such changes in an experimental set-up that is otherwise unchanged. The disadvantage is the requirement for the robust preparation of RNA samples and the reliable quantitation of yields. When RNA yields are too low to be monitored reliably, or when RNA preparations vary qualitatively, the use of external controls for normalization will not be reliable.

The importance of monitoring global and/or unbalanced mRNA changes is dictated by the goals of the experiment. For example, in disease classification studies, determination of the most extreme (relative) markers of a particular state without consideration of the actual changes does not require external control normalization. Neither is it important if the goal is to screen for only the most responsive genes in particular experimental conditions. However, as experimental biology becomes more comprehensive and quantitative due to the availability of whole-genome sequences and to the increased focus on systems biology[16], it will become more of a necessity to monitor with greater precision the actual, rather than the relative, changes in different cellular states. Methods that take into account the possibility of global changes will contribute towards such goals. Examples of studies that will benefit from the use of external controls include comprehensive studies of gene regulation and analyses of drug side-effects, as well as microarray studies incorporating only limited sets of genes.

Experimental Procedures

Accession Numbers and Protocols

For MIAME (minimum information about a microarray experiment)-compliant[17] protocols, datasets in Microarray Gene Expression Markup Language (MAGE-ML)[18] and normalization scripts, see our website (http://www.genomics.med.uu.nl/pub/jvp/ext_controls) or the public microarray database ArrayExpress (http://www.ebi.ac.uk/microarray/ArrayExpress/arrayexpress.html) (Table 1).

External Controls

Constructs containing *Bacillus subtilis* genes (*ycxA*, *yceG*, *ybdO*, *ybbR*, *ybaS*, *ybaF*, *ybaC*, *yacK* and *yabQ*) cloned between the *Xba*I and *Bam*HI sites in pT7T3 (Amersham Pharmacia Biotech) were made, with an additional 30-nucleotide poly(A) sequence between the gene and the *Xho*I site. For making RNA, plasmids were digested with *Xho*I for use in *in vitro* transcription reactions using MEGAscript-T7 (Ambion).

Cell Culture

HUVECs were isolated as described in Jaffe et al.[19]. Cells were cultured in endothelial growth medium (EGM-2) at 37°C in the presence of 5% CO_2. Before heat shock, the medium was removed, preheated EGM-2 was added and cells were incubated at 42.5°C in EGM-2.

MCF7 cells were cultured in DMEM/Ham's F12 medium (1:1) containing 5% fetal calf serum, glutamine (300 mg/ml), penicillin (100 inhibitory units/ml) and streptomycin (100 mg/ml). Cells at 70% confluence were grown for 30 h in phenol-red-free, serumfree medium with 0.2% BSA, transferrin (10 mg/ml) and 30 nM sodium selenite.

Table 1 | ArrayExpress Accession Numbers for Protocols and Datasets

Accession number	Description
A-UMCU-1	UMC Utrecht *Saccharomyces cerevisiae* 16-k array, version 1.1
A-UMCU-2	UMC Utrecht *Homo sapiens* 19-k array, version 1.0
P-UMCU-1	*S. cerevisiae* culture
P-UMCU-2	Human umbilical vein endothelial cell culture
P-UMCU-3	MCF7 cell culture
P-UMCU-4	Total RNA isolation (*S. cerevisiae*)
P-UMCU-5	Total RNA isolation (mammalian cell culture)
P-UMCU-6	Messenger RNA enrichment (*S. cerevisiae*)
P-UMCU-7	Amino-allyl labelling
P-UMCU-8	Microarray production
P-UMCU-9	Hybridization
P-UMCU-10	Scanning protocol
P-UMCU-11	Image analysis
E-UMCU-1	Yeast spiked controls
E-UMCU-2	Human umbilical vein endothelial cell culture heat shock
E-UMCU-3	MCF7 serum deprivation
E-UMCU-4	Yeast stationary phase

S. cerevisiae S288c (*MATa*; *met15*; *ura3*; *his3*Δ1; *leu2*) (Research Genetics) was grown in YEP medium (containing yeast extract and peptone) supplemented with 2% glucose. Cultures for the spiking experiments (Figure 5) were grown to mid-log phase (OD_{600} = 0.5), and for the stationary phase experiment were grown to mid-log phase or to stationary phase (OD_{600} = 10.0; 10-day culture).

RNA Isolation and Labeling
For mammalian cells, total RNA was prepared using Trizol (Gibco BRL) in accordance with the manufacturer's instructions. External controls were added in an appropriately diluted 5-µl mixture to 10 µg of total RNA. Yeast total RNA was prepared using hot phenol. External controls were added, as an appropriately diluted 5-µl mixture, to 500 µg of total RNA, and mRNA was isolated using Oligotex (Qiagen). Complementary DNA synthesis was carried out using 10 µg of mammalian total RNA, or 3 µg of yeast mRNA, in the presence of 2-aminoallyl-dUTP. Samples were purified using Microcon-30 (Millipore) columns and were coupled to Cy3 and Cy5 fluorophores. Before hybridization, free dyes were removed using Chromaspin-30 (Clontech) columns, and the efficiency of cDNA synthesis and dye incorporation was measured using a spectrophotometer (UV1240mini, Shimadzu).

Microarray Hybridization
From each sample, 300 ng cDNA (with a specific activity of 2–4% dye-labeled nucleosides) was hybridized for 16–20 h at 42°C. Slides were scanned in a Scanarray 4000 XL (Perkin Elmer Biosystems). Image analysis was carried out using Imagene 4.0 (Biodiscovery).

Microarray Production
C6-amino-linked oligonucleotides (70 nucleotides in length), the Yeast Genome ArrayReady and the Human Genome ArrayReady Oligo set (version 1.1) were purchased from Qiagen, and were printed on Corning UltraGAPS slides with a MicroGrid II (Apogent Discoveries) using 48-quill pins (Microspot2500; Apogent Discoveries) in 3 x SSC at 50% humidity and at 18°C, and were processed by ultraviolet crosslinking (2,400 millijoules, 10 min) with a Stratalinker2400 (Stratagene).

Normalization
Algorithms were based on Lowess print-tip normalization[5], applied in the statistical package R[20], using the existing packages SMA (http://www.stat.berkeley.edu/users/terry/zarray/Software/smacode.html) and com.braju.sma (http://www.maths.lth.se/help/R/com.braju.sma). Alterations were made for the import of Imagene 4.0 files, flagging of control spots, Lowess line calculation on subsets of spots (either controls or genes) and extrapolation to all spots in the subgrid. This has been incorporated in an R package called genomics.sma.

The first method, normalization using the expression levels of endogenous genes, uses gene spots to calculate the Lowess line for each subgrid and then applies these lines to all spots. The second method, normalization using the expression levels of all genes, and external controls, also shifts all spots linearly according to the median ratio of the control spots. The third method, normalization using external controls, uses the controls to calculate the Lowess line for each subgrid and then applies these lines to all spots for normalization.

The second and third methods gave almost identical results for the experiments shown in Figures 2, 3 and 5. Due to the non-uniform distribution of changes in mRNA levels in the stationary phase experiment, only the third method (normalization using external controls) gave accurate results overall. This would be the method of choice for all experiments. However, there was a significant variation in the amount of the individual oligonucleotides supplied in the collections compared with the control oligonucleotides. This resulted in suboptimal extrapolation of the Lowess line to those gene spots that have extremely low intensity values, as a result of both low amounts of oligonucleotides and low or non-existent levels of mRNA. The variation in the amount of oligonucleotides supplied has been rectified in later versions of the collections.

To take into account differences in RNA yield per cell, cells were counted, and a linear shift was applied to the normalized data using the ratio derived from the total RNA yield per cell from each pair of samples.

Acknowledgements

We thank A. Dijk, R. de Roos, and I. Hamelers for cell cultures, K. Duran for technical assistance, E. Holloway, H. Parkinson and U. Sarkans for assistance with MAGE-ML generation and submission to ArrayExpress, C. Wilson and R.A. Young for external control constructs and P.C. van der Vliet, H.Th.M. Timmers, W. van Driel, R. Kerkhoven and C. Wijmenga for advice and discussions. This work was supported by the following grants from the Netherlands Organization for Scientific Research (NWO): 05050205, 90101238, 90101226, 016026009 and 90104219.

References

1. Brown, P.O. & Botstein, D. Exploring the new world of the genome with DNA microarrays. *Nat Genet* **21**, 33-7 (1999).
2. Young, R.A. Biomedical discovery with DNA arrays. *Cell* **102**, 9-15 (2000).
3. Quackenbush, J. Computational analysis of microarray data. *Nat Rev Genet* **2**, 418-27 (2001).
4. Tseng, G.C., Oh, M.K., Rohlin, L., Liao, J.C. & Wong, W.H. Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Res* **29**, 2549-57 (2001).
5. Yang, Y.H. et al. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res* **30**, e15 (2002).
6. Lee, P.D., Sladek, R., Greenwood, C.M. & Hudson, T.J. Control genes and variability: absence of ubiquitous reference transcripts in diverse mammalian expression studies. *Genome Res* **12**, 292-7 (2002).
7. Talaat, A.M. et al. Genomic DNA standards for gene expression profiling in Mycobacterium tuberculosis. *Nucleic Acids Res* **30**, e104 (2002).
8. Holstege, F.C. et al. Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* **95**, 717-28 (1998).
9. Wang, Y. et al. Precision and functional specificity in mRNA decay. *Proc Natl Acad Sci U S A* **99**, 5860-5 (2002).
10. Dickson, L.M. & Brown, A.J. mRNA translation in yeast during entry into stationary phase. *Mol Gen Genet* **259**, 282-93 (1998).
11. Kerr, M.K. & Churchill, G.A. Statistical design and the analysis of gene expression microarray data. *Genet Res* **77**, 123-8 (2001).
12. Kroll, T.C. & Wolfl, S. Ranking: a closer look on globalisation methods for normalisation of gene expression arrays. *Nucleic Acids Res* **30**, e50 (2002).
13. Yang, Y.H. & Speed, T. Design issues for cDNA microarray experiments. *Nat Rev Genet* **3**, 579-88 (2002).
14. Dudley, A.M., Aach, J., Steffen, M.A. & Church, G.M. Measuring absolute expression with microarrays with a calibrated reference sample and an extended signal intensity range. *Proc Natl Acad Sci U S A* **99**, 7554-9 (2002).
15. Sterrenburg, E., Turk, R., Boer, J.M., van Ommen, G.B. & den Dunnen, J.T. A common reference for cDNA microarray hybridizations. *Nucleic Acids Res* **30**, e116 (2002).
16. Ideker, T., Galitski, T. & Hood, L. A new approach to decoding life: systems biology. *Annu Rev Genomics Hum Genet* **2**, 343-72 (2001).
17. Brazma, A. et al. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet* **29**, 365-71 (2001).
18. Spellman, P.T. et al. Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biol* **3**, RESEARCH0046 (2002).
19. Jaffe, E.A., Nachman, R.L., Becker, C.G. & Minick, C.R. Culture of human endothelial cells derived from umbilical veins. Identification by morphologic and immunologic criteria. *J Clin Invest* **52**, 2745-56 (1973).
20. Ihaka, R. & Gentleman, R. R: a language for data analysis and graphics. *J. Comp. Graph. Statist.* **5**, 299-314 (1996).

Expression Profiler: Next Generation - an Online Platform for Analysis of Microarray Data

Chapter IV

Misha Kapushesky, Patrick Kemmeren, Aedín C. Culhane, Steffen Durinck, Jan Ihmels, Christine Körner, Meelis Kull, Aurora Torrente, Ugis Sarkans, Jaak Vilo & Alvis Brazma

Nucleic Acids Research 2004 Jul 1; 32(Web Server Issue): W465-W470

Expression Profiler: Next Generation - an Online Platform for Analysis of Microarray Data

Misha Kapushesky[1], Patrick Kemmeren[1,2], Aedín C. Culhane[3], Steffen Durinck[4], Jan Ihmels[5], Christine Körner[6], Meelis Kull[7,8], Aurora Torrente[1], Ugis Sarkans[1], Jaak Vilo[1,7,8,9] & Alvis Brazma[1]

[1]European Bioinformatics Institute, United Kingdom; [2]Department of Physiologcial Chemistry, Division of Biomedical Genetics, UMC Utrecht, the Netherlands; [3]Conway Institute, University College Dublin, Ireland; [4]ESAT, KULeuven, Belgium; [5]Weizmann Institute of Science, Israel; [6]University of Leipzig, Germany; [7]Egeen International, Estonia; [8]Department of Computer Science, University of Tartu, Estonia; [9]Estonian Biocenter, Estonia

Abstract

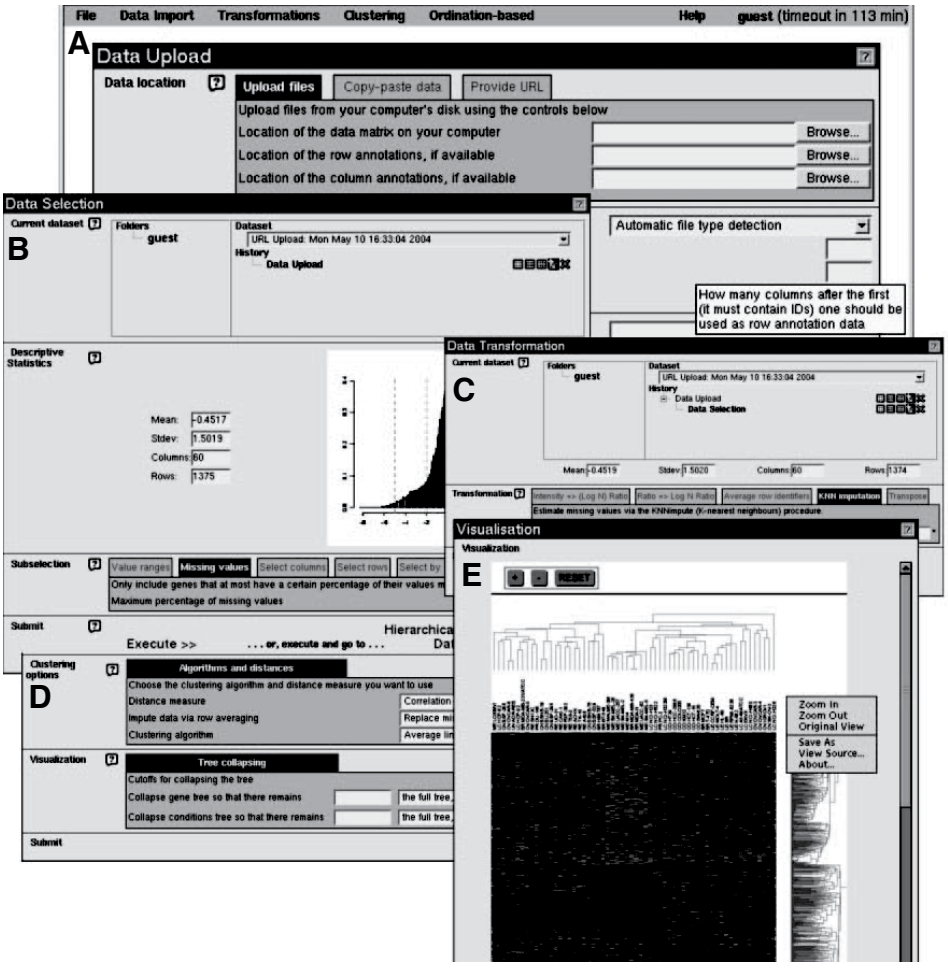
Expression Profiler (EP, http://www.ebi.ac.uk/expressionprofiler) is a web-based platform for microarray gene expression and other functional genomics-related data analysis. The new architecture, Expression Profiler: next generation (EP:NG), modularizes the original design and allows individual analysis-task-related components to be developed by different groups and yet still seamlessly to work together and share the same user interface look and feel. Data analysis components for gene expression data preprocessing, missing value imputation, filtering, clustering methods, visualization, significant gene finding, between group analysis and other statistical components are available from the EBI (European Bioinformatics Institute) web site. The web-based design of Expression Profiler supports data sharing and collaborative analysis in a secure environment. Developed tools are integrated with the microarray gene expression database ArrayExpress and form the exploratory analytical front-end to those data. EP:NG is an open-source project, encouraging broad distribution and further extensions from the scientific community.

Introduction

The complete analysis of microarray gene expression data consists of many steps, starting with image processing, annotation, data normalization and transformation, gene subselection and filtering[1]. The analysis that follows can be one of many available techniques, e.g. clustering or class prediction. While there exist applications that integrate some of these analyses within a single tool (some stand-alone and some web-based), the need for an extensible framework with a unified interface to disparate analysis components persists. Expression Profiler: next generation (EP:NG) provides such an online environment, in which diverse components are brought together under a unified look and feel, so that the user is unaware that these algorithms are developed, maintained and run by different groups. In fact, these data analysis components may be distributed across different systems and physical locations. We have also endeavored to make it easy for algorithm developers to contribute their methods to EP:NG, thus making it a suitable platform to become a public compendium of data analysis methods.

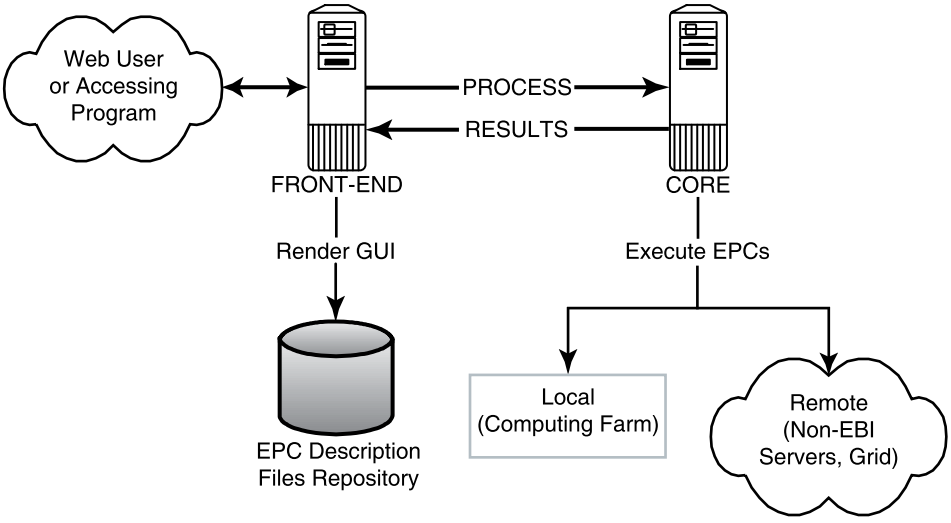
In a typical workflow (Figure 1), the user might either upload gene expression data from an external source via the Data Upload component, or retrieve data from the ArrayExpress public repository at the EBI[2]. The Data Selection component provides a basic statistical overview of the dataset, which can be used to guide the user in subselecting genes relevant for further analysis. The Data Transformation component can impute missing values in the chosen data subset, and perform other data transformations. Following these

Figure 1 | A Typical Analysis Workflow
(**A**) After the user logs in, the data file is uploaded via the Data Upload component. (**B**) A subselection is made via the Data Selection component (the distribution histogram indicates the spread of the data and helps choose the range criteria). (**C**) Data Transformation is used to impute the missing values in the resulting data subset; clicking on the dataset name will return the new matrix with estimated missing values. (**D**) Hierarchical Clustering with tree collapsing is used to get the first overview of the data structure; (**E**) the hierarchical tree is displayed as an SVG image, allowing the user to zoom in on the tree in detail.



optional preprocessing steps, data can be subjected to one or more analyses. The user can explore the overall structure of the data via one of the clustering methods available in the Hierarchical and K-groups Clustering components, and the best number and quality of clusters within the data can be evaluated using a novel algorithm, Clustering Comparison. Alternatively, the user can subject data to a semisupervised method aimed at studying correspondences between groups of samples and genes in the Between Group Analysis component. Users interested in studying a specific gene or a group of genes can use the Similarity Search or the Signature Algorithm components to investigate the structural organization of the data. Thus, EP:NG provides a useful tool for exploratory data analysis. EP:NG also integrates the components of the earlier version of EP[3], which remains

Figure 2 | EP:NG System Overview
All access is realized through the FRONT-END, which uses the Expression Profiler Component (EPC) description files to present a graphical user interface (GUI), and then sends the user's request to the CORE for either local execution on the Linux server farm at the EBI or remote processing on external (non-EBI) computer servers.



available at http://ep.ebi.ac.uk and includes several additional online tools. EP:NG offers a new architecture and interface, using the foundations of the original Expression Profiler, keeping the original fast algorithm implementations and building on the initial idea of a simple to use, flexible online data analysis toolbox.

EP:NG Platform Architecture

The EP:NG platform gives a unified style to all the constituent components, regardless of the origin of their development, their physical location and their manner of execution. EP:NG supports quick and easy integration of third-party tools and algorithms: for instance, several components are implemented via integration with the statistical package R[4]. The software is based on a relational database that stores user account information as well as metadata about the datasets being analyzed. Analysis results and the datasets themselves are stored on the file system in a format that facilitates fast operations on the data. All core algorithms are written in C/C++ or R, while the interfacing between components is done through XML/XSLT transformations. Perl serves as glue between the various parts of the EP:NG platform.

The EP:NG infrastructure is built to support (i) data sharing between groups of collaborators independent of location and format, (ii) execution of algorithms in a pipeline fashion and (iii) the integration of a constantly expanding collection of methods of data analysis. A simple web-service-type interface is exposed for programmatic access to individual components of EP:NG as well as for complex data processing queries. Figure 2 depicts a high-level overview of the entire system. EP:NG is built on the concept of atomic components: each functional software subunit is represented by a separate distinct component, whose inputs and outputs are described separately in a simple format. The system uses these descriptions to present an interface to the user, and runs the component on the back-end computing servers, or sends the data off for remote analysis. This means that in order to integrate a novel method into the EP:NG environment its author needs only

to compose an appropriate XML description file and send it to the site maintainers; the method will then appear seamlessly among the other tools in EP:NG.

Expression Profiler Components (EPCs)

An EP:NG user can log in as guest or can establish a permanent password-protected account. We would recommend the latter as it provides each user with the ability to store and organize data and prior analysis results into nested folders, for further work over multiple sessions. EP:NG user accounts can also belong to one or more user groups, which are managed by an account with group administrator privileges. This facility allows groups of researches to share their data and analysis strategies.

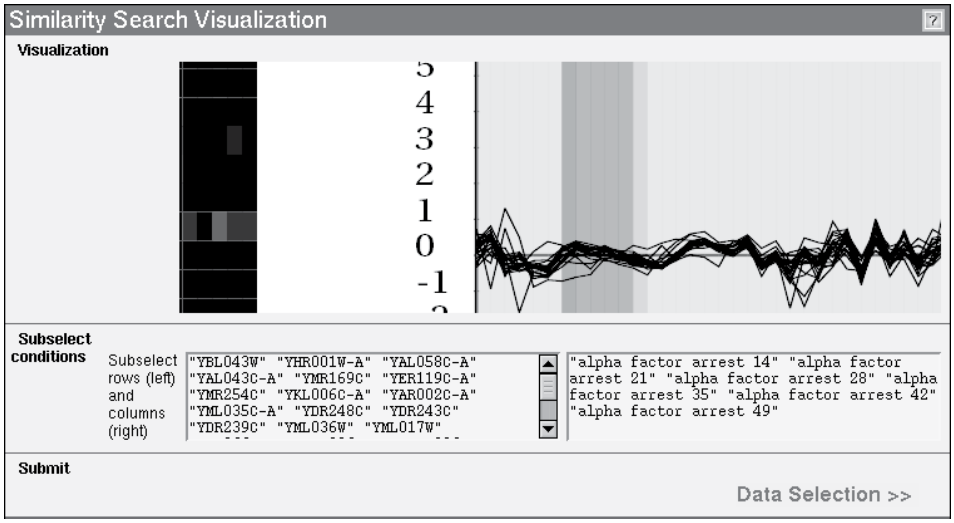
Data Upload
Users can either upload their own data or select a published dataset from the ArrayExpress database. To retrieve data from ArrayExpress, one can explore the database via its online interface and subsequently export an expression matrix by selecting the desired samples and measurements (Assays and Quantitation Types) and then send this to EP:NG. The Data Upload component can accept data in a number of formats including basic delimited files such as those exported by Microsoft Excel. Uploaded expression datasets must be represented as matrices, with rows and columns corresponding to genes and conditions, respectively. In addition to loading the data matrix, the user can provide various metadata for the dataset. Metadata could include a description of the dataset, studied organism, or row and column annotations, which can be either free-text annotations or identifiers from publicly available ontologies, e.g. GeneOntology[5].

Data Selection
Data Selection presents a brief statistical overview of the data (data distribution histogram, mean, standard deviation) and allows the user to select genes and conditions that have particular expression values. For example, one can select only those genes (data rows) that are outside ± 2 standard deviations of the mean in at least 60% of experiments and which contain no more than 80% missing values. Further subselections can be made at any point in the analysis.

Data Transformation
Either before or after subselecting data, various transformation procedures can be applied. These include K-nearest neighbor imputation[6] to fill in missing values (this can be a lengthy procedure especially on large datasets), LOWESS normalization[7] (an integrated third-party component) and conversion of absolute intensity values from a two-channel experiment to log ratios. From here one can proceed to further subselections or apply a suitable analysis component.

Similarity Search
The Similarity Search provides a means of selecting groups of genes related to given ones. The user specifies one or several genes, chooses a similarity measure and receives those genes most closely co-expressed with the selected genes within the dataset. A wide variety of distance measures is available, including the Euclidean metric, Pearson correlation, Manhattan distance, Spearman's ranking and chord distance. From the resulting screen the user can choose a subset of these genes and/or conditions, and then use the Data Selection component to save this subset as a separate dataset for further analysis. Similarity Search Visualization allows a subset of conditions to be subselected under which the displayed genes correlate best (Figure 3).

Figure 3 | Similarity Search Visualization Component
The user can select with the mouse a subset of conditions where the expression line-plot shows correlation by visual inspection. The chosen genes and conditions can be sent to other components for further analysis (e.g. to Data Selection).



Hierarchical and K-groups Clustering

EP:NG provides both hierarchical and partitioning-based clustering methods. As stated above, there are many distance measures available. The clustering algorithms are implemented in C, and results are visualized as publication-quality vector-based SVGs or in a raster format, PNG. Data can be hierarchically clustered on both experiments and conditions simultaneously and one can interactively zoom in on interesting subtrees. The hierarchical clustering tree produced for large datasets can be difficult to comprehend all at once. An attractive feature of EP:NG is that it provides a facility to display a quick high-level overview of the clustered data. Using the 'Tree Collapsing' option, one can display a specified percentage of major branches, with the other ones collapsed into single nodes. There are two K-groups clustering algorithms in EP:NG: K-means and K-medoids. The latter is a variant of another well-known approach to partitioning the data into a specified number of clusters. It differs in that it uses existing objects from the dataset as cluster centers in its calculations. This allows the use of a distance matrix derived from any measure; hence, this component can be applied to more diverse data types, such as sequences.

Clustering Comparison

A problem that arises naturally with all partitioning clustering methods is to find the appropriate K (number of clusters). The Clustering Comparison component implements an algorithm that takes two K-groups clustering results and matches the clusters by membership. By examining the output the user can evaluate the optimal (according to some criteria) number of clusters in the dataset.

Signature Algorithm

Signature Algorithm is the R implementation of an algorithm described in[8]. It identifies a co-expressed subset in a user-submitted set of genes, removes unrelated genes from the input and identifies additional genes in the same dataset that follow a similar pattern of expression. Co-expression is identified with respect to a subset of conditions, which is

also provided as the output of the algorithm. It is a fast algorithm useful for exploring the modular structure of expression data matrices.

Between Group Analysis and Ordination
Standard multivariate analysis methods, such as principal component analysis (PCA) and correspondence analysis (COA), are provided in the Ordination component. These methods are frequently used to search for underlying structures in datasets. Between Group Analysis (BGA) presents a multiple discriminant approach that can be used with expression data matrices of any dimensionality[9]. BGA is carried out by ordinating groups (sets of grouped microarray samples) and then projecting the individual sample locations on the resulting axes. It is used in the framework of conventional ordination techniques such as PCA or COA and, as such, allows for great flexibility with regard to the assumptions that one makes in carrying out the analysis. When combined with COA it is especially powerful as it allows one to examine in detail the correspondences between the grouped samples and those genes which most facilitate the discrimination of these groupings. This is a semisupervised method, so it is important to test its results using a test dataset, or using resampling accuracy analysis methods. The Ordination and BGA EPCs are implemented using the R multivariate data analysis package ADE-4[10].

Pipelines and Workflows
It is important not only to be able to find and execute analytical procedures, but to do this in a logical, user-defined sequence of steps. This naturally leads to the concept of analysis pipelines. EP:NG was designed to enable the user to combine components into sequences. There are two major ways to do this. First, via the interface: each component provides annotated links to other EPCs that logically follow (e.g. Data Selection links to Data Transformation and to Hierarchical Clustering, etc.). Second, programmatically: a program can send a simple XML query to EP:NG indicating which components to launch and in what sequence. The final result will be shown to the user. Whenever a user or a program runs an EP:NG component, the parameters and the sequence of steps taken to obtain the results are stored in the internal database. This allows one to define an analysis workflow, a process that can later be applied repeatedly with the same parameters to other datasets.

Hardware and Software Requirements

EP:NG may be run in a distributed fashion, where the GUI and the CORE servers reside on separate machines (and the CORE interfaces to a queuing system, as is the setup at the EBI), or the entire system may run on a single server. This setup involves a Linux system as the front-end server for the GUI, and an alpha station at the back-end, both running Apache web servers. EP:NG uses the LSF queuing system, but can also run in a stand-alone mode or integrate with PBS. EP:NG is an open-source project and the related SourceForge site (http://ep-sf.sourceforge.net) contains further information, including information on how to develop new components and how to obtain and install EP:NG locally.

EP:NG Usage

EP:NG has two major purposes; to provide tools for exploratory analysis of microarray data (including the data in ArrayExpress) and to support an infrastructure for rapid deployment of specific methods. With this collection of tools a novice user can easily try out several different approaches, compare the results and select those methods that make the most sense. The entire EP:NG system can also be installed locally and be used within

one lab or simply as a local data analysis program. Expression Profiler and EP:NG are extensively used by biology labs, research groups and as bioinformatics teaching software at universities.

ArrayExpress Data Export into EP:NG

In order to analyze data that is stored in ArrayExpress with Expression Profiler, gene expression data matrices need to be exported from the repository into EP:NG. The user can start by querying ArrayExpress (http://www.ebi.ac.uk/arrayexpress) for a specific experiment by the published accession number or through a more general query on combinations of other fields such as experiment type or laboratory. ArrayExpress provides the facility to retrieve the stored data by selecting a combination of (i) BioAssays (hybridizations), (ii) Quantitation Types (data types) and (iii) Array Annotations. These choices determine the composition of the resulting data matrix: for every selected hybridization columns of chosen quantitation types will be exported. Every row of the matrix will be annotated with selected annotations from the corresponding array design. Finally, either the generated gene expression matrix can be downloaded on to the user's computer for later analysis in the user's own account in EP:NG (or other tools) or it can be exported directly into EP:NG for immediate analysis. ArrayExpress help pages (http://www.ebi.ac.uk/arrayexpress/Help/) document these steps in further detail.

Future Development

The next EP:NG version will include a separate Data Normalization component that will integrate several publicly available normalization algorithms, an improved interactive visualization interface and several advanced data analysis EPCs. We are also implementing support for data input in the MAGE-ML format[11] so that the system can obtain detailed information about a dataset's origin and related array designs, which may make it possible to suggest data analysis strategies to the user. There are ongoing collaborations with other groups to develop a more advanced webservices interface, as well as to provide customized support for Affymetrix data through integration with Bioconductor[4].

Supplementary Data

The Expression Profiler SourceForge website (http://ep-sf.sourceforge.net), contains various documents regarding the installation, use and development of the EP:NG framework and components. ArrayExpress help pages are useful for learning how to export and analyze data in Expression Profiler (http://www.ebi.ac.uk/arrayexpress/Help).

Acknowledgements

We wish to thank Luke Jeffery for helping with the testing. The authors would like to acknowledge the TEMBLOR grant from the EC and the Wellcome Trust for support. J.Vilo and M. Kull acknowledge Estonian Science Foundation grants nrs. 5724 and 5722.

References

1. Quackenbush, J. Computational analysis of microarray data. *Nat Rev Genet* **2**, 418-27 (2001).
2. Brazma, A. et al. ArrayExpress--a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res* **31**, 68-71 (2003).
3. Vilo, J., Kapushesky, M., Kemmeren, P., Sarkans, U. & Brazma, A. Expression Profiler. in *The analysis of Gene Expression Data: Methods and Software* (eds. Parmigiani, G., Garret, E.S., Irizarry, R. & Zeger, S.L.) (Springer Verlag, New York, 2003).

4. Ihaka, R. & Gentleman, R. R: a language for data analysis and graphics. *J. Comp. Graph. Statist.* **5**, 299-314 (1996).
5. Creating the gene ontology resource: design and implementation. *Genome Res* **11**, 1425-33 (2001).
6. Troyanskaya, O. et al. Missing value estimation methods for DNA microarrays. *Bioinformatics* **17**, 520-5 (2001).
7. Yang, Y.H. et al. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res* **30**, e15 (2002).
8. Ihmels, J. et al. Revealing modular organization in the yeast transcriptional network. *Nat Genet* **31**, 370-7 (2002).
9. Culhane, A.C., Perriere, G., Considine, E.C., Cotter, T.G. & Higgins, D.G. Between-group analysis of microarray data. *Bioinformatics* **18**, 1600-8 (2002).
10. Thioulouse, J., Chessel, D., Doledec, S. & Oliver, J.M. ADE-4: a multivariate analysis and graphical display software. *Stat. Comput.* **7**, 75-83 (1997).
11. Spellman, P.T. et al. Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biol* **3**, RESEARCH0046 (2002).

Expression Profiler: Next Generation - Discerning Regulatory Mechanisms Behind Regulatory Processes

Chapter V

Patrick Kemmeren, Misha Kapushesky, Frank C.P. Holstege & Alvis Brazma

Manuscript in preparation

Expression Profiler: Next Generation - Discerning Regulatory Mechanisms Behind Biological Processes

Patrick Kemmeren[1,2,3], Misha Kapushesky[2,3], Frank Holstege[1] & Alvis Brazma[2]

[1]Department of Physiological Chemistry, Division of Biomedical Genetics, UMC Utrecht, The Netherlands; [2]European Bioinformatics Institute, United Kingdom; [3]These authors contributed equally to this work

Abstract

Summary: Understanding the regulation behind biological processes is crucial to understanding how they work. Genes that show differential mRNA expression with similar expression patterns are often used as a starting point to reveal more about the regulatory mechanisms behind the biological processes during which the expression changes take place. To facilitate studying regulatory processes, with DNA microarray expression data as a starting point, four new components have been seamlessly integrated into Expression Profiler: next generation (EP:NG); Gene Ontology, Pattern Discovery, Pattern Visualization and Sequence Logo components. This facilitates access to the possible regulation behind biological processes and has been implemented for many different species, among which the most commonly used model organisms.

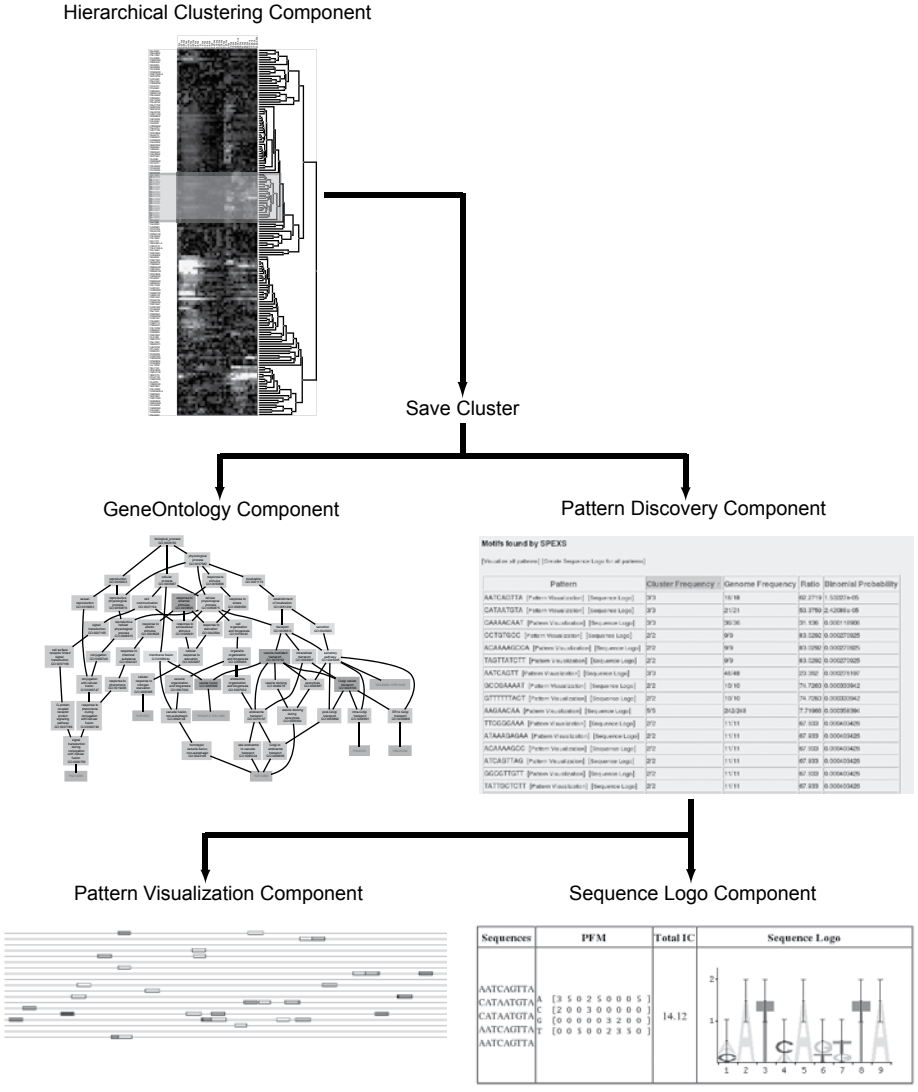
Introduction

Understanding which genes are involved in which biological processes and how they are regulated is an important step towards understanding how these biological processes work. For this purpose DNA microarrays are often used to study mRNA expression level changes under different physiological conditions. Genes that show differential mRNA expression with similar expression patterns can facilitate understanding the biological process of interest. Typically genes with co-regulation are obtained from microarray data using certain clustering methods and selecting the most interesting clusters[1,2]. The biological relevance of these clusters can then be determined using tools such as the Gene Ontology (GO) [3], metabolic pathways or text mining[4]. Determining the regulatory mechanisms behind these biological clusters is a major challenge[5]. Finding over-represented patterns in the upstream sequences of these gene clusters can be a first step. The derived hypothesis is that these patterns might represent binding sites for transcription factors. To facilitate the process of elucidating the regulation of biological processes, we have incorporated some currently existing methodologies into Expression Profiler: next generation (EP:NG)[6]. EP: NG is a web-based platform for the analysis of microarray gene expression data. This way, integration with other gene expression analysis routines and a uniform interface is provided that facilitates determination of the biological relevance of co-regulated genes through GO, through finding over-represented patterns in their upstream sequences, and visualization of those patterns according to the upstream location as well as their information content using Sequence Logos[7].

Integration with Expression Profiler: Next Generation

Besides the currently available components within EP:NG, four new components have been integrated seamlessly into EP:NG so that the biological relevance and possible regulation of certain clusters or groups of co-regulated genes can be studied in depth. Fundamental within EP:NG is a dataset, which in this case can be obtained through the upload of a mRNA expression or sequence dataset in fasta format, or directly selected from the public microarray database ArrayExpress[8]. Once this dataset has been created

Figure 1 | Uncovering Shared Function and Regulation behind Gene Clusters using EP:NG
A schematic overview of the different components within EP:NG that together can be used to uncover functional and regulatory relationships within gene clusters. The dataset used for this analysis is obtained from Radonjic et al.[9] Only Those genes that show expression levels greater than 1.5 times the standard deviation of the distribution of all genes in at least 20% of the time points are subjected to hierarchical clustering analysis. Part of this hierarchical tree is then saved and analyzed using the GeneOntology and Pattern Discovery Component. Patterns found in the Pattern Discovery Component are then further investigated using the Pattern Visualization or Sequence Logo Component. See also Appendix; Color Figure 5.1.



and annotated properly in EP:NG, all components within EP:NG can use the metadata associated with the dataset, for instance to automatically select the appropriate species or background dataset. This is also the case for the four new components presented here. The genes of interest for the new components can be based on a dataset selection, (part of) a hierarchical clustering tree, a K-means clustering, a similarity search or copy pasting

of gene identifiers. A schematic overview of how the four new components can be used to elucidate possible regulatory mechanisms is depicted in Figure 1. In this figure, a publicly available dataset[9] studying gene expression levels during exit and entry into stationary phase has been subjected to hierarchical clustering analysis. A gene cluster showing downregulation (blue) during exit from stationary phase and subsequent upregulation (yellow) during entry into stationary phase is selected for further downstream analysis using the four different components described in the following sections.

Gene Ontology

The first component that has been integrated into EP:NG is the Gene Ontology component. This component uses the GO::TermFinder perl module developed by the Stanford Microarray Database (SMD)[10] and supports all organisms that currently have GO annotations (28 in total). Once a certain number of genes has been selected, users can search for over-represented GO categories using the biological process, molecular function or cellular component ontology. Over-represented GO categories are then found using a hypergeometric testing procedure and only those categories that exceed the user given significance cutoff are displayed in the GO tree and corresponding table. In the example used in Figure 1, the selected gene cluster subjected to GO analysis shows a significant overrepresentation of genes involved in secretory pathways and vesicle transport.

Pattern Discovery

The second component that has been incorporated into EP:NG is the Pattern Discovery component. This component uses the SPEXS (Sequence Pattern Exhaustive Search) tool, which can find over-represented patterns in a set of unaligned sequences. Automatic sequence retrieval, based on a certain gene cluster or group of co-expressed genes, is supported for all current ensembl genomes (16 in total)[11] through the integrated use of BioMart (http://www.ebi.ac.uk/biomart). Furthermore, those organisms that are not available within ensembl can be added to EP:NG using a common sequence database format called GFF (http://www.sanger.ac.uk/Software/formats/GFF/GFF_Spec.shtml)[12], or users can upload their own sequence data using the Sequence Data Upload component. Once the sequences have been retrieved, patterns can be discovered based on substring patterns of any length, patterns with wildcard positions, character groups or variable-width wildcards. When subjected to the pattern discovery algorithm, the gene cluster used in Figure 1, shows a number of over-represented patterns in the upstream regions of the genes, which can be investigated in more detail using the pattern visualization or sequence logo component.

Pattern Visualization

The third component integrated within EP:NG is the Pattern Visualization component. This component can take either patterns found in the Pattern Discovery component, as depicted in Figure 1, or previously defined patterns given by the user and visualize them according to their position in the upstream sequence. Moreover, the pattern visualization can be integrated with the gene expression clustering and heat-map visualization so that any correlation found between the occurrences of sequence motifs, their relative position upstream of the gene or the clustering result becomes apparent.

Sequence Logos

The fourth component added to EP:NG is the Sequence Logo component. This component uses the Transcription Factor Binding Site (TFBS) perl modules[13] and takes a set of aligned sequences obtained from the Pattern Discovery component, the Pattern Visualization component or a user given set of aligned sequences or position frequency matrix. The information content is then displayed using a Sequence Logo[7], which can display both significant residues and subtle sequence patterns. This is illustrated in Figure 1. Here, the two top-most patterns found in the pattern discovery component are represented with

a Sequence Logo, showing that some sequence positions are shared between the two patterns, whereas others show more variation.

Conclusions

We describe here seamless integration of four components into EP:NG; Gene Ontology, Pattern Discovery, Pattern Visualization and Sequence Logo Automatic retrieval of GO information and upstream sequences is supported for 28 and 16 species respectively, thus providing the most comprehensive tool of this type that is currently available. In addition, results obtained from these components are linked to the original expression dataset, thus revealing any correlation found between the sequence motifs and gene expression clustering results. Through the use of these four components, the biological relevance and possible regulatory mechanisms behind a set of co-regulated genes can easily be studied in more detail, leading to more insight into the underlying regulatory mechanisms.

References

1. Quackenbush, J. Computational analysis of microarray data. *Nat Rev Genet* **2**, 418-27 (2001).
2. Sherlock, G. Analysis of large-scale gene expression data. *Brief Bioinform* **2**, 350-62 (2001).
3. Harris, M.A. et al. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* **32 Database issue**, D258-61 (2004).
4. Jenssen, T.K., Laegreid, A., Komorowski, J. & Hovig, E. A literature network of human genes for high-throughput analysis of gene expression. *Nat Genet* **28**, 21-8 (2001).
5. Simonis, N., Wodak, S.J., Cohen, G.N. & van Helden, J. Combining pattern discovery and discriminant analysis to predict gene co-regulation. *Bioinformatics* **20**, 2370-9 (2004).
6. Kapushesky, M. et al. Expression Profiler: next generation--an online platform for analysis of microarray data. *Nucleic Acids Res* **32**, W465-70 (2004).
7. Schneider, T.D. & Stephens, R.M. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res* **18**, 6097-100 (1990).
8. Parkinson, H. et al. ArrayExpress--a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res* **33**, D553-5 (2005).
9. Radonjic, M. et al. Genome-wide analyses reveal RNA polymerase II located upstream of genes poised for rapid response upon S. cerevisiae stationary phase exit. *Mol Cell* **18**, 171-83 (2005).
10. Boyle, E.I. et al. GO:TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics* **20**, 3710-5 (2004).
11. Hubbard, T. et al. Ensembl 2005. *Nucleic Acids Res* **33**, D447-53 (2005).
12. *General Feature Format* [http://www.sanger.ac.uk/Software/formats/GFF/GFF_Spec.shtml].
13. Lenhard, B. & Wasserman, W.W. TFBS: Computational framework for transcription factor binding site analysis. *Bioinformatics* **18**, 1135-6 (2002).

Protein Interaction Verification and Functional Annotation by Integrated Analysis of Genome-Scale Data

Chapter VI

Patrick Kemmeren, Nynke L. van Berkum, Jaak Vilo, Theo Bijma, Rogier Donders, Alvis Brazma & Frank C.P. Holstege

Molecular Cell 2002 May; 9(5): 1133-43

Protein Interaction Verification and Functional Annotation by Integrated Analysis of Genome-Scale Data

Patrick Kemmeren[1,2], Nynke L. van Berkum[1], Jaak Vilo[2], Theo Bijma[1], Rogier Donders[3], Alvis Brazma[2] & Frank C.P. Holstege[1]

[1]Department of Physiological Chemistry, Division of Biomedical Genetics, UMC Utrecht, the Netherlands; [2]Microarray Informatics group, EMBL-EBI, Hinxton, UK; [3]Centre for Biostatistics, University Utrecht, the Netherlands

Abstract

Assays capable of determining the properties of thousands of genes in parallel present challenges with regard to accurate data-processing and functional annotation. Collections of microarray expression data are applied here to asses the quality of different high-throughput protein interaction datasets. Significant differences are found. Confidence in 973 out of 5342 putative two-hybrid interactions from *S. cerevisiae* is increased. Besides verification, integration of expression and interaction data is employed to provide functional annotation for over 300 previously uncharacterized genes. The robustness of these approaches is demonstrated by experiments that test the *in silico* predictions made. The study shows how integration improves the utility of different types of functional genomic data and how well this contributes to functional annotation.

Introduction

High-throughput functional genomic analyses are generating valuable resources that contain many different types of data. These include data on mRNA levels and changes therein[1-3], mutant phenotypes[4,5], protein interactions[6-9], protein levels[10], as well as many other interesting properties[11-13]. On their own, each data type is useful for generating hypotheses about the genes involved. The pace of traditional approaches to testing these hypotheses however, is significantly slower than the rate at which new data is being generated. Also inherent to the high-throughput nature of these assays, is heterogeneity in data quality, both within and between datasets[14]. A major challenge is to develop more efficient ways of dealing with high-throughput data, allowing false positives to be identified and hypotheses prioritized based on confidence.

Another challenge arising from the availability of whole genome sequences is how to achieve rapid and accurate functional annotation of uncharacterized genes[15]. The genome sequence of the yeast *S. cerevisiae* has been available for several years now[16]. It is one of the best-studied eukaryotic model organisms. Despite this, one third of its approximately 6200 genes are classified as of unknown function or uncharacterized in current versions of yeast databases[17-19]. The development of technologies capable of assaying the properties of thousands of genes or their products in parallel, should greatly advance the rate at which annotation of gene function can take place. A start has been made at combining different types of data for annotation[20]. Such methods require development and validation by directly testing the ensuing predictions.

Based on the yeast two-hybrid assay, two large-scale protein-protein interaction discovery projects have reported over 5000 protein interactions for *S. cerevisiae*[6,7]. The vast majority of these are novel. Based on affinity purification of tagged proteins, followed by mass spectrometry analysis of associated factors, two other studies have recently reported over 4000 additional interactions[8,9]. Little is known about the possible pitfalls of tagging proteomics. The potential artifacts inherent to the yeast two-hybrid assay indicates that each such reported interaction should be considered putative[6].

It is well established that some interacting proteins show mRNA coexpression[21-24]. Confidence in a subset of the putative interactions may therefore readily be increased,

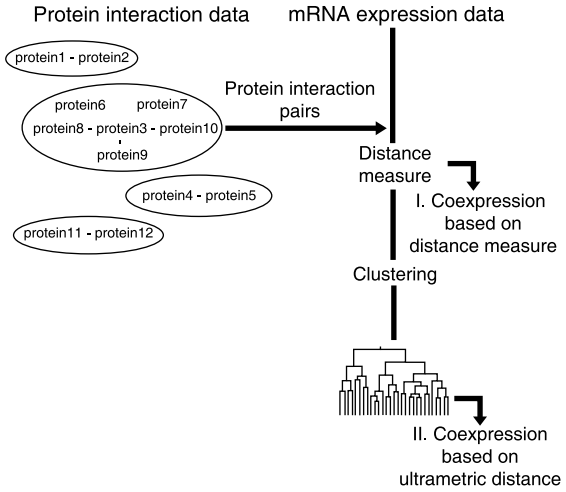


Figure 1 | Strategies for Coexpression Analysis
Two general approaches for determining mRNA coexpression of protein pairs are depicted. In the first approach (I), coexpression is determined by calculation of the distance measure and retreiving this from a distance matrix. In the second general approach (II), the distance matrix is used for ensuing clustering. In this case the degree of coexpression is based on the ultrametric distance of the derived dendogram. Variations of both approaches were tested by applying different distance measures and hierarchical clustering algorithms. For further analysis, a cosine correlation distance only approach has been used for the different protein interaction datasets (see also Supplemental Table S1 at http://www.genomics.med.uu.nl/pub/pk/comb_gen/).

simply by automatically verifying which of the interacting proteins also show mRNA coexpression.

In this study we first explore how analysis of large collections of DNA microarray data can be applied to verify protein-protein interactions. Several methods are evaluated, leading to an approach that is optimized for analysis of interacting protein pairs. Presently available microarray expression data and high-throughput protein interaction data are extensively analyzed. The results show significant differences within and between datasets and emphasize the requirement for additional verification of high-throughput-generated data. We show how functional annotation of uncharacterized genes can be derived by combining genome-scale data. Importantly, the robustness of the described methods are demonstrated by follow-up experiments aimed at testing several of the predicted gene functions.

Results

Approaches for Determining mRNA Coexpression of Protein Pairs

mRNA coexpression is currently often analyzed by clustering microarray expression data[21,25,26]. The results depend on two choices. A distance measure is applied to calculate the similarity between genes within the expression data and a clustering algorithm is chosen to group genes according to similar behavior, based on the distance.

Two different general approaches and several variants thereof, were tested to determine which would be optimal for determining mRNA coexpression with the goal of verifying interacting protein pairs (Figure 1). The first involved only a distance measure calculation (part I in Figure 1). The second general approach consisted of distance measure calculation followed by hierarchical clustering (part II in Figure 1). In this case the height of the dendogram required to join two genes (ultrametric distance) was used to determine coexpression. For both approaches, protein interaction pairs were ranked according to the degree of mRNA coexpression. A significance threshold was determined after performing identical analyses on randomized expression data. Variants of both general approaches were investigated by examining the consequences of applying different distance measures (euclidian, cosine correlation and Pearson correlation) and clustering algorithms (unweighted pair-group method average linkage, weighted pair-group method average linkage, complete linkage and single linkage). Each method resulted in a group of interactions showing significant coexpression. The comparison was performed with the two high-throughput yeast two-hybrid interaction sets[6,7], using approximately 300 expression profiles derived from one study[27].

Which method was most suited for determining mRNA coexpression of protein pairs was selected by determining the enrichment for previously known interactions within the subset that showed a significant degree of mRNA coexpression. The absolute number of interactions showing coexpression was also taken into account. The results of comparing the various methods are described in Supplemental Table S1 (http://www.genomics.med.uu.nl/pub/pk/comb_gen/). The choice of clustering algorithm and distance measure had a significant influence. Comparison of the two general approaches demonstrated that coexpression based on distance measure only, is most appropriate for examining such pairwise relationships. For further analyses the results are based on use of cosine correlation distance only (part I in Figure 1).

Previously Established Interactions Exhibit a High Degree of Coexpression

Having established an appropriate method for analysis of mRNA coexpression to study pairwise interactions, we next determined how many interactions show mRNA coexpression. This is important in order to investigate how widely applicable this approach is for automatic verification of large sets of novel interactions. The previously established interactions contained within these datasets represent interactions of the highest confidence, as these have previously been found by other means. If sufficient previously established interactions exhibit coexpression, then coexpression analysis can be used as a general tool to determine which novel interactions behave similarly and are therefore more likely to be correct.

Because the performance of some of the methods analyzed depend on an equivalent scale of relative expression levels throughout the data[25,26], the expression data employed for establishing the method was from the largest collection of data contained within a single study from *S. cerevisiae*[27]. To increase the scope of the coexpression analysis, a second expression dataset was employed. This "conditions" dataset consisted of 326 expression profiles from various environmental and experimental conditions, such as different stress responses[28], cell cycle[29], sporulation[30], pheromone treatment[31] and unfolded protein responses[32]. This collection contains more and diverse fluctuations in mRNA levels and was therefore better suited to capture a larger number of coexpressed interactions.

Results of examining coexpression for the Ito et al.[7] interaction dataset within the "conditions" mRNA data are presented in Figure 2 and summarized in Table 1. The results form a sigmoidal curve of ranked pairs. The ranking is based on the degree of mRNA coexpression observed for each protein pair and extends from zero (complete correlation) to two (complete anti-correlation). Plotted within the ranked curve are the positions of the previously established interactions that were found within each dataset (white dots).

Using a strict cut-off criteria of 0.1% (see Experimental Procedures), 60% of all previously known interactions contained within the two-hybrid data exhibit a significant degree of coexpression (Figure 2 and Table 1). Introducing a more lenient significance threshold of 1%, the fraction of previously established interactions showing coexpression is 71%. Although the fraction of established interactions showing coexpression will grow as more expression data is generated, many functional interactions will not show mRNA coexpression because of regulation at other levels. An important outcome of this study is that a sufficiently large number of the previously established interactions that are contained within the high throughput datasets, show a significant degree of mRNA coexpression. The confidence in a large subset of the novel, high-throughput interactions, may therefore be increased, simply by automatic analysis of mRNA coexpression. With the presently available mRNA expression data, the fraction of all functional interactions contained within the high-throughput data that can be verified in this way is around two-thirds. This depends on the degree of confidence that is desired and on the significance thresholds applied, as is discussed below.

Differences between Previously Established and Novel High-Throughput Interactions

A striking outcome of this study is the skewed distribution of the previously known interactions towards a significantly higher degree of coexpression (Figure 2). Whereas 60% of previously established interactions show mRNA coexpression, only 26% of the novel interactions show the same degree of coexpression (see also Table 1). The degree of anti-correlation for the novel protein interactions is also significantly higher (Figure 2).

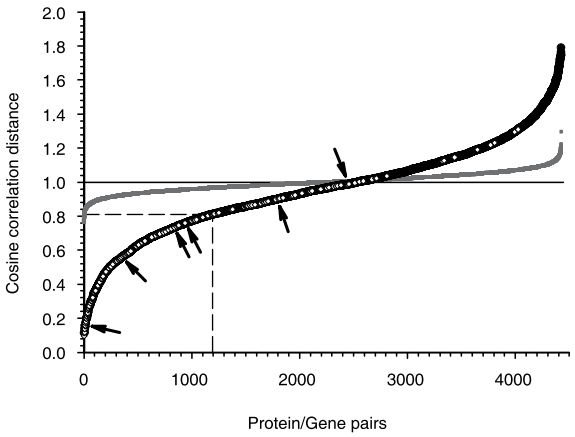


Figure 2 | Protein Interactions of Higher Confidence Show Significantly More mRNA Coexpression
The protein interaction data from Ito et al. (blue line, 4425 pairs) was analyzed for mRNA coexpression within the "conditions" expression dataset using the cosine correlation distance only approach. This "conditions" dataset consisted of 326 expression profiles from various environmental and experimental conditions, such as different stress responses[28], cell cycle[29], sporulation[30], pheromone treatment[31] and unfolded protein responses[32]. The degree of coexpression is plotted on the *y*-axis. A value of zero corresponds to complete mRNA coexpression across all the expression data. A value of two corresponds to perfect anti-correlation under all conditions. The protein pairs are ranked according to the degree of coexpression and the rank number is plotted on the *x*-axis. The 170 previously known interactions within the interaction dataset are plotted as open circles in order to demonstrate the marked skewing towards a higher degree of coexpression for these interactions of higher confidence. The red line is the result of an identical analysis using randomized expression data. The dashed lines indicate the significance threshold for catching 0.1% of the interactions with the randomized expression data. This corresponds to a P value of 0.003. The arrows indicate protein interaction pairs that have been selected for follow-up experiments (Figure 4-6). An additional protein interaction pair was selected from the Uetz data (Figure 5C). See also Appendix; Color Figure 6.2.

Whereas only 6% of the previously established interactions show some degree of anti-correlation (cosine correlation greater than 1), 47% of the novel interactions exhibit anti-correlation. The levels of anti-correlation for the putative novel interactions also extend to a much higher degree. Differences between the levels of coexpression observed for the previously established interactions, compared to the putative novel interactions, are significant ($P < 10^{-15}$ for Ito data, Table 1). As is discussed below, the number of false positives likely contributes to these differences, underscoring the importance of verifying the novel interactions as suggested previously[6]. In this regard there are also large differences between the different interaction datasets (Table 1).

Extensive Datamining of Available Expression and Interaction Data
Analysis of both two-hybrid datasets with the two expression datasets resulted in four combinations of data. A browseable and searchable table containing all the significant results of the four combinations of expression and interaction data is presented in the accompanying website (http://www.genomics.med.uu.nl/pub/pk/comb_gen). This table also contains additional information such as the expression profiles of the corresponding genes across all conditions analyzed. Using a strict significance threshold, 973 interactions out of a total of 5342 putative interactions analyzed, showed a significant degree of coexpression in at least one of the four combinations (Supplemental Table S2 [http://www.genomics.med.uu.nl/pub/pk/comb_gen/]). Interactions showing coexpression fall into several categories. Besides the previously established interactions, there are many examples of coexpression between proteins already known to coexist within the same multisubunit complex. Although it is not necessary to add to the high confidence that these interactions are genuine, it is reassuring for the interpretation of the other results that many such interactions are found.

For the other categories of interactions, the observed mRNA coexpression is particularly useful as an indication that the interaction is genuine. For example, there are 569 interactions where both proteins have some functional annotation[17,18]. This ranges from the presence of a structural motif to detailed knowledge about the role of the protein. Within this set there are many examples with provocative overlaps in the previously found characteristics.

Table 1 | Coexpression Analysis of the Different Protein Interaction Datasets

	Ito, HT Y2H				Uetz, HT Y2H				Gavin, HT TAP-tag				Ho, HT Flag-tag			
	nr.	%	mean [d]	P value	nr.	%	mean [d]	P value	nr.	%	mean [d]	P value	nr.	%	mean [d]	P value
Random protein pairs	5973		0.98		5973		0.98		5973		0.98		5973		0.98	
Previously unknown pairs	4255		0.96	8E-03 [b]	845		0.92	3E-10 [b]	3097		0.71	< 1E-15 [b]	1322		0.89	< 1E-15 [b]
Previously known pairs (% of total)	170	4%	0.68	< 1E-15 [c]	65	7%	0.78	9E-09 [c]	253	8%	0.70	< 1E-15 [c]	39	3%	0.75	7E-07 [c]
0.1% cut-off																
Previously known coexpressed[a]	102	60%			35	54%			146	58%			21	54%		
Other pairs coexpressed	1199	26%			348	37%			1866	56%			484	35%		
1% cut-off																
Previously known coexpressed[a]	121	71%			35	54%			169	67%			27	69%		
Other pairs coexpressed	1596	35%			365	39%			1890	56%			601	43%		

[a] Previously known interactions (pki) showing coexpression as a percentage of all pki in the original dataset.
[b] Significance of difference in coexpression values when compared to randomly generated protein pairs
[c] Significance of difference in coexpression values between pki and previously unknown interactions
[d] Mean cosine correlation distance for interactions analyzed using the “conditions” dataset. A value of zero corresponds to complete correlation. A value of two corresponds to complete anti-correlation (see also Figure 2).

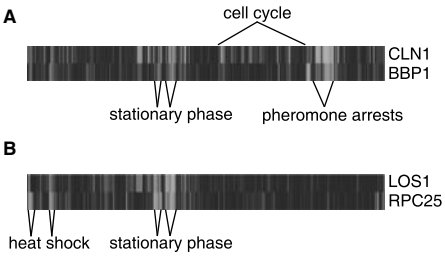


Figure 3 | mRNA Coexpression Examples for Putative Interactions between Previously Characterized Proteins
(**A**) Expression profiles of CLN1 and BBP1 obtained from the "conditions" dataset. The cosine correlation distance is 0.38. Red indicates upregulation, green indicates downregulation. Both mRNA expression levels are downregulated during conditions that involve cessation of cell division such as stationary phase and pheromone arrest. BBP1 mRNA levels exhibit a cyclic pattern during cell cycle[29], but only with high magnitudes of change in half of the cell-cycle experiments analyzed. (**B**) Expression profiles of LOS1 and RPC25 obtained from the "conditions" dataset. The cosine correlation distance is 0.34. Red indicates upregulation, green indicates downregulation. The mRNA expression levels of both genes are downregulated during heat shock and stationary phase. See also Appendix; Color Figure 6.3.

One example is the putative interaction between the G1/S cyclin Cln1p and the spindle pole body (SPB) component Bbp1p[7]. Cln1p is necessary for START, the commitment phase of the cell cycle[33]. START includes SPB duplication, which has been shown to require Cln1p, but through unresolved mechanisms[33,34]. Bbp1 is found to interact with a total of ten proteins in the high-throughput two-hybrid data, including a nitrogen permease, mitochondrial proteins, uncharacterized proteins, a proteosome subunit and Cln1p. This casts doubts on the specificity of Bbp1 two-hybrid results. Significantly, the interaction with Cln1p has the highest degree of mRNA coexpression (Figure 3A) and is the only putative partner of Bbp1p found to be coexpressed in both expression datasets. This makes the case for a genuine functional interaction between Cln1p and Bbp1p very strong. It demonstrates the ease with which putative interactions can be prioritized for follow-up experiments by examining mRNA coexpression first. In this case it makes a compelling argument for regulation of SPB duplication through the interaction between Cln1p and Bbp1p.

Besides verification, an advantage of analyzing expression is the insight into the particular conditions under which mRNA coexpression is observed. Another example of previously characterized proteins with a putative interaction that may have important implications, are Los1p and Rpc25p[6]. Los1p has been suggested to be a receptor for tRNA transport from the nucleus[35]. The putative association between Los1p and Rpc25p, a unique component of RNA polymerase III[36], indicates direct coupling between tRNA production and its export from the nucleus. In agreement, the mRNAs are downregulated immediately following heat shock and during stationary phase (Figure 3B), both conditions that cause reductions in tRNA and protein synthesis[37].

These are two examples out of many that reflect the enormous potential for novel biomedical discovery that is to be found within the original interaction datasets. The significant degree of mRNA coexpression found for 973 of the 5342 high-throughput interactions, makes it more likely that these interactions are indeed functional. The importance of screening the putative interactions for mRNA coexpression, is that this allows prioritization for which of the many potentially interesting interactions should first be selected for more detailed studies.

Validation of Predicted Functional Annotation of Uncharacterized Genes

One of the goals of this study is automated functional annotation of whole genomes. This is the reason why strict significance thresholds have been applied. Within the set of interactions showing coexpression, 328 included a protein pair where one of the partners is functionally annotated, and one is classified as uncharacterized in current versions of yeast databases[17,18]. The two-hybrid data and the coexpression data represent two different lines of evidence that there may be a functional link between the two genes in question. This suggests that a tentative functional annotation of the uncharacterized gene may be derived from its better characterized binding partner. In order to test the robustness of such *in silico* predictions and the coexpression analysis in general, several were selected for follow-up experiments.

The first prediction tested was based on the reported interaction between the protein products of NTH1 and the uncharacterized open reading frame (ORF) YLR270W. This

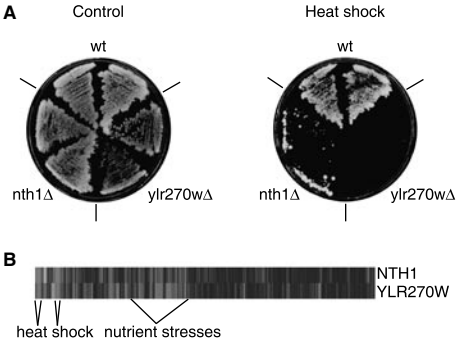


Figure 4 | YLR270W is Required for Resistance to Heat Shock
(**A**) Wildtype, *nth1* and *ylr270w* deletion strains grown at 30°C (left), or exposed to 53°C for 7 hrs (right). The mutant strains show little to no recovery 5 days after heat shock when compared to wildtype. (**B**) Expression profiles of NTH1 and YLR270W obtained from the "conditions" dataset. The cosine correlation distance is 0.17. Red indicates upregulation, green indicates downregulation. mRNA expression levels of both genes are upregulated during heat shock and other forms of stress. See also Appendix; Color Figure 6.4.

was chosen for the high degree of mRNA coexpression (i.e. a small cosine correlation distance of 0.17). The neutral trehalase gene, NTH1, controls trehalose levels, required for thermotolerance as well as resistance to other forms of stress[38]. Strains lacking NTH1 are unable to recover from heat shock[39]. In agreement with the overlap in function suggested by the reported interaction and mRNA coexpression, deletion of YLR270W also causes this phenotype (Figure 4A). YLR270W can now be classified as required for resistance to heat shock. The mRNAs for YLR270W and NTH1 are upregulated during heat shock and other forms of stress, providing additional support for this conclusion (Figure 4B).

The next functional prediction tested was based on the putative interaction between Mpk1p (Slt2p) and Ylr350wp, chosen as an example exhibiting a much lower, but still significant, degree of coexpression (cosine correlation distance of 0.54). MPK1 is a protein kinase of the MAP kinase family, involved in the cell wall integrity pathway[40]. Upregulation of this pair of genes is observed during sporulation and protein misfolding, both conditions that are likely to require appropriate cell wall remodeling (Figure 5B). Deletion of cell wall integrity pathway components results in caffeine sensitivity[41]. High osmolarity, for example when sorbitol is included in the growth medium, remediates this phenotype through suppression of cell lysis defects[41]. In agreement with deriving a functional annotation of the uncharacterized ORF from the role of MPK1, *ylr350w* deletion also shows sensitivity to caffeine that is rescued upon addition of sorbitol (Figure 5A). YLR350W can now be annotated as a cell wall integrity pathway component, similarly to MPK1.

A third prediction tested was based on the interaction between Srb11p (Ssn8p) and Ynl094wp[6], with a slightly lower degree of coexpression than the previous example (cosine correlation distance of 0.63). Srb11p forms a kinase-cyclin pair with Srb10p and is part of the Srb/Mediator complex, one of the subcomplexes of the RNA polymerase II holoenzyme[42]. Cells lacking SRB11 respond poorly to galactose induction[43]. Deletion of YNL094W results in reduced growth on galactose (Figure 5C). This, together with the protein interaction data strongly suggests that Ynl094wp is also involved in galactose utilization by regulating transcription through interaction with the Srb/Mediator complex.

The final two examples of tested predictions were based on interactions between Snf4p and Ycl046wp, and between Snf7p and Ygr122wp. These were selected because the degree of mRNA coexpression exhibited is close to the significance threshold (cosine correlation distance of 0.75 and 0.73 respectively, the threshold being 0.81, see Figure 2). Snf4p is a regulatory subunit of the functionally conserved Snf1p protein kinase complex[44]. Both SNF4 and SNF7 are involved in derepression of glucose-repressed genes, but molecular and genetic analysis suggests that the role of SNF7 is distinct from that of SNF4[45]. In support of distinct roles, the two pairs of expression profiles show different patterns of coexpression (Figure 6B). In agreement with a function for the uncharacterized ORFs in utilization of carbon sources other than glucose, both *ycl046w* and *ygr122w* deletion strains show the carbon catabolite repression phenotype of inability to grow on raffinose that is also exhibited by *snf4* and *snf7* deletion strains[45,46] (Figure 6A). YCL046W and YGR122W can now be annotated as required for growth on raffinose.

The results of the experiments described above provide evidence that the tentative functional annotation derived from the combination of expression and interaction data (Supplemental Table 3) is indeed reliable. In the first place the annotations are based on

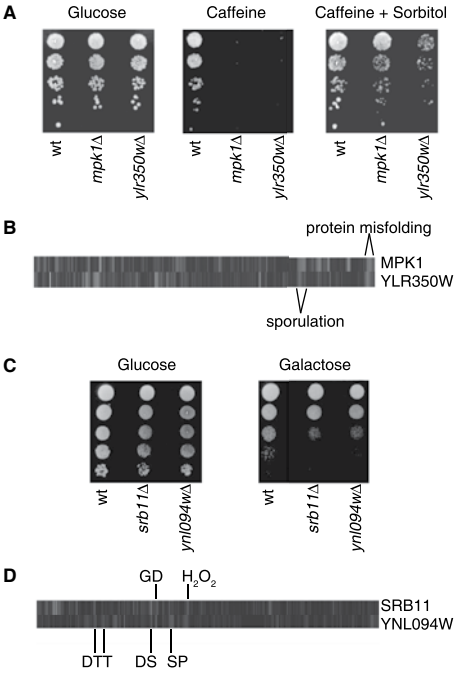


Figure 5 | YLR350W is Involved in the Cell Wall Integrity Pathway; YNL094W is Involved in Galactose Utilization

(**A**) Wild-type, *mpk1* and *ylr350w* deletion strains after 3, 7, and 7 days of growth on glucose (left), caffeine (middle), and caffeine plus sorbitol (right), respectively, in a dilution series of 1:5. *mpk1* and *ylr350w* deletion strains show no growth on the caffeine containing medium. Growth is restored when sorbitol is added. (**B**) Expression profiles of MPK1 and YLR350W obtained from the "conditions" dataset. The cosine correlation distance is 0.54. Red indicates upregulation, green indicates downregulation. During sporulation and protein misfolding, mRNA expression levels of both genes are upregulated. (**C**) Wild-type, *srb11* and *ynl094w* deletion strains after 2 days of growth on glucose (left) and galactose (right) respectively, in a dilution series of 1:5. *srb11* and *ynl094w* deletion strains show a reduced growth on galactose, when compared to wildtype. (**D**) Expression profiles of SRB11 and YNL094W obtained from the "conditions" dataset. The cosine correlation distance is 0.63. Red indicates upregulation, green indicates downregulation. Coexpression is observed during the diauxic shift (DS), glucose depletion (GD), stationary phase (SP), two DTT treatments (DTT) and hydrogen peroxide treatment (H_2O_2). See also Appendix; Color Figure 6.5.

two completely different lines of evidence. Furthermore, the predictions tested involve four different phenotypes and also cover the entire range of significant mRNA coexpression values.

Although the exact positioning of the significance threshold can be arbitrarily changed to include values of lower significance, we note that two putative protein interactions with mRNA coexpression levels below the significance thresholds used here, were also tested similarly (WHI2-YLR019W and WHI2-YLL010C, cosine correlation distances of 0.90 and 1.06 respectively). These failed to show an overlap in function for the phenotype assayed (data not shown).

Comparison of Large-Scale Interaction Datasets

Recently two large-scale protein tagging projects have been described that also report large sets of novel interaction data[8,9]. In these approaches, proteins are tagged and affinity-purified. Stably interacting proteins are copurified and identified by mass spectrometry. Although there are differences between the direct pairwise interactions resulting form two-hybrid screens and the complexes isolated by tagging, the results of the tagging projects were treated as putative pairwise relationships and analyzed for coexpression.

Table 1 reveals that coexpression can be used to quickly determine global differences between the behaviour of previously known and novel pairs within these datasets too. The dataset from Gavin et al.[8] is particularly noteworthy. The degree of coexpression observed for the novel interactions (mean of 0.71) is nearly identical to the level of coexpression observed for the previously known interactions contained within this dataset (mean of 0.70). This is in contrast to the Ito interaction dataset where the mean coexpression value for previously unknown interactions is 0.96, for previously known interactions 0.68 and for randomly generated pairs 0.98. Graphs of the four interaction datasets are presented as supplemental data (Supplemental Figure S1 [http://www.genomics.med.uu.nl/pub/pk/comb_gen/]), revealing other differences, such as the degree of anti-correlation. The analysis indicates that the confidence that can be placed in the novel interactions within the Gavin dataset is as high as the confidence that previously established interactions are correct.

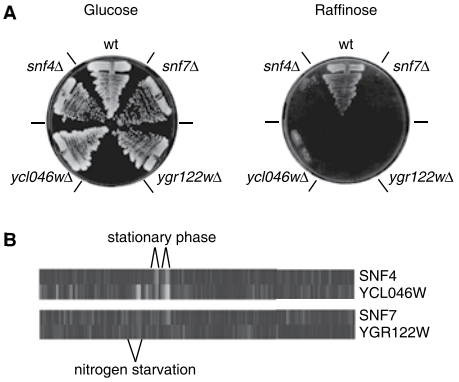


Figure 6 | YCL046W and YGR122W are Required for Growth on Raffinose
(**A**) Wildtype, *snf4, ycl046w*, *snf7* and *ygr122w* deletion strains after 13 days growth on glucose (left) and raffinose (right). The mutant strains show little to no growth on raffinose when compared to wildtype. (**B**) Expression profiles of SNF4, YCL046W, SNF7 and YGR122W obtained from the "conditions" dataset. The cosine correlation distance between SNF4 and YCL046W is 0.75, between SNF7 and YGR122W 0.73. Red indicates upregulation, green indicates downregulation. mRNA levels of SNF4 and YCL046W are downregulated during stationary phase, SNF7 and YGR122W are upregulated during nitrogen starvation.

Coexpression Datamining Tool

In addition to the supplemental data and the interactive table of significant results, a coexpresssion analysis tool has been implemented on the accompanying website (http://www.genomics.med.uu.nl/pub/pk/comb_gen/). This allows coexpression to be determined for previously discovered or novel gene pairs and yields additional information, such as the conditions under which significant coexpression is observed.

Discussion

This study demonstrates how integration of genome-scale data can be employed to address several issues arising from the availability of whole genome sequences: assessment of heterogeneity in data quality, verification of the results of high-throughput assays and functional annotation of uncharacterized genes. The comparison of approaches leads to adopting a method most suitable for analyzing pairwise relationships. Extensive analysis of interaction and mRNA expression data available for *S. cerevisiae*, illustrate the utility of an integrative approach. The results indicate large differences in data quality between different high-throughput generated interaction datasets. Nevertheless, confidence in 973 putative protein-protein interactions is increased. Tentative functional annotations for 328 previously uncharacterized yeast genes are generated. The robustness of the analyses are further demonstrated by experiments that test a cross-section of the *in silico* predictions made.

To date, automated functional annotation has depended largely on the availability of sequence data, in order to uncover function through evolutionary and structural relationships[15,47-49]. Initiatives such as the establishment of a single ontology[50] and publicly available databases, are also of vital importance towards this goal.

The recent availability of more diverse types of genome-scale data should significantly increase the rate of functional annotation. In contrast to the integrative approach that addresses the issue of data quality described here, a previous study has adopted a cumulative approach for combining sources of data[20]. None of the uncharacterized ORFs interrogated in the follow-up experiments described above, were found in the over 93,000 putative pair-wise links put forward by this study, indicating that different approaches may perhaps be complementary. Testing the biological significance of the outcome of bioinformatic analyses will also be important for assessing which methods are most appropriate[51].

This work builds upon previous studies that have described relationships between interaction data and expression data[22-24]. We extend this correlation by showing how it can be applied to verify high-throughput generated protein interactions, examine differences between datasets as a whole and provide functional annotation. In these previous studies different approaches were taken to examine coexpression. The comparison of methods presented here indicates that care has to be taken in choosing the most appropriate method. The approach must also suit the goals. For example, clustering may be better

suited for simultaneously examining the relationships between large groups of interacting proteins[23]. For examining pairwise relationships for the purposes of this study, we empirically determine that correlation is most suited (Supplemental Table S1 [http://www.genomics.med.uu.nl/pub/pk/comb_gen/]). Correlation approaches have also been applied previously[22,24] and Jansen et al. have also noted a lower coexpression correlation of two-hybrid results.

The large differences found between the fraction of previously established interactions and the fraction of novel interactions showing coexpression is striking (Figure 2 and Table 1). Initial coexpression observations, that may have led investigators to study the possibility of an interaction, may have contributed to the larger degree of coexpression observed for previously established interactions. Less coexpression may be inherent to those interactions that are more easily discovered by two-hybrid assays. To exclude this bias, we restricted our analyses to those previously established interactions already contained within the high-throughput results. Another possibility is that the differences reflect the presence of false positives within the high-throughput data. This is supported by the extremely high degree of anti-correlation observed for the novel interactions (Figure 2 and Supplemental Figure S1 [http://www.genomics.med.uu.nl/pub/pk/comb_gen/]), which is more difficult to reconcile with a functional interaction. The presence of false positives is further indicated by the fact that for some datasets, the degree of coexpression observed is close to that of randomly generated pairs of genes (Table 1). It is likely that several explanations contribute to the differences in coexpression and in varying degrees. Regardless of the possibility of false positives, large-scale interaction projects provide a valuable resource as is exemplified here (Figure 3-6). Emphasis is best placed on uncovering more putative interactions and on the development of methods that quickly discriminate which interactions are most likely to be correct.

Not all genuine interactions can be expected to exhibit mRNA coexpression, due to regulation of the interaction by other well-established mechanisms such as translation, post-translational modification, regulated location, etc. Although the number of interacting pairs showing mRNA coexpression will increase as more expression data is generated, it should be emphasized that functional interactions are still contained within the set not showing mRNA coexpression, albeit with a much lower frequency as is indicated by the low numbers of previously known interactions showing anti-correlation (Figure 2 and Supplemental Figure S1 [http://www.genomics.med.uu.nl/pub/pk/comb_gen/]).

Here we have used coexpression to determine which interactions are most likely to be correct. In order to derive a reliable functional annotation, strict significance thresholds were applied, capable of pinpointing 60-70% of all the novel interactions that are most likely to be correct. This fraction will grow as more expression data is generated. In addition, the significance threshold applied is arbitrary and can be lowered to capture more interactions, with the possibility of including more false positives. Conversely, the method can also be applied to exclude those interactions that are most likely to be false positives, by identifying which interactions have the highest degree of anti-correlation.

Analysis of all the different interaction datasets reveals differences in the degree of coexpression between the novel and previously known interactions contained within each dataset (Table 1 and Supplemental Figure 1). These differences indicate for example that the tagging approach adopted by Gavin et al.[8] has a significantly lower false positive rate, also when compared to the other tagging project analyzed[9]. It is likely that the induced overexpression of the tagged protein in the latter study contributes to this. In this respect the approach taken by Ho et al. is perhaps more similar to the two-hybrid projects. Because expression of the tagged protein in the study by Gavin et al. was at endogenous levels, it is probable that this approach is better suited for identifying stable interactions and with a lower false positive rate. Two-hybrid approaches can also identify stable interactions but have the additional advantage of being capable of isolating the less stable and transient interactions that play important roles in regulating many cellular processes. In order to be able to identify such interactions, the assay has to be more sensitive, with the complication that this likely introduces more false positives. In other words, different approaches are complementary and capable of identifying different and important subsets of interactions.

The complementarity in approaches is also indicated by the lack of overlap between the interaction data. For example, if the data generated by Gavin et al.[8] is treated as pairwise, 1.2% of the interactions reported by Ito et al.[9] are common to both sets. The false positive rate may contribute to this and indeed this overlap is doubled if the analysis

is restricted to only those interactions showing coexpression (2.3% of the coexpressed Ito dataset is also found in the coexpressed Gavin data). However, the overlap remains very low, indicating the importance of both types of approaches. The overlap between all the different projects, even after analysis of coexpression, was at most 18.5% (18.5% of the Uetz coexpressed pairs are shared with the Ito coexpressed pairs), but significantly lower for data derived by different approaches. This indicates that many more interactions still remain to be found and further underscores the importance of continuing such projects.

The method described here employs the availability of large collections of mRNA data. It is interesting to note that the majority of interacting proteins with a significant degree of mRNA coexpression do not exhibit coexpression across every microarray experiment analyzed. Noise, experimental error and biological variation obviously contribute to this. Moreover, most interactions are probably not required across all of the experimental and environmental conditions assayed. Due to the multifunctional nature of many proteins, this may result in the mRNA of one protein being upregulated under conditions where no change occurs in the mRNA of the binding partner. For verifying interactions it is important that coexpression occurs under a subset of conditions relevant to the interaction. The coexpression observed for the experimentally interrogated interactions (Figure 4-6) are particularly illustrative of this.

What steps should be taken for integration of more types of data? Addition of a third type of data, such as the results of high-throughput phenotyping[4,5], is already suggested by the nature of the follow-up experiments performed. The sifting approach that is worthwhile for the combination of two types of data, may ultimately be too restrictive for integration of many more types of data. Lack of an expected phenotype does not necessarily imply that the interaction is not functional. A protein may have an opposite effect on the phenotype of its binding partner. There may be genetic redundancy. Also, many proteins exhibit multiple phenotypes, with a phenotypical overlap with the binding partner in only a subset of these. There will probably be many examples of *bona fide* associations that will not fit all the imposed criteria. Other approaches, such as a weighted scoring on the predicted functional links, may be more effective, as long as data-quality is taken into account. The results presented here, the way in which the development and validation of this approach has been carried out, demonstrate how the utility of different types of genome-scale data can be improved, and how well their combination can contribute to verification of hypotheses and accurate functional annotation.

Experimental Procedures

Protein Interaction Data

Protein interaction datasets were from Uetz et al.[6], Ito et al.[7], Gavin et al.[8] and Ho et al.[9] Pairwise interactions were obtained from the Gavin and Ho datasets by listing all one to one relationships between the tagged protein and their associated proteins. For assessment of the different methods for coexpression analysis, the previously known interactions within each dataset were taken from the original paper or by using the complete interaction table from MIPS[19]. Classification of uncharacterized function was based on YPD[18] and checked using the GO annotation[50] that is in the process of being incorporated into SGD[17].

mRNA Expression Data

Two mRNA expression datasets were used. The first was from Hughes et al.[27] The second, “conditions” dataset was compiled from Spellman et al.[29], Roberts et al.[31], Chu et al.[30], Travers et al.[32] and Gasch et al.[28] The ORF names are as stated in Gasch et al. and the corresponding expression values in the different datasets were retrieved and concatenated into a single file.

Coexpression Analyses

For both general approaches, distance matrix calculation and hierarchical clustering was performed using Expression Profiler (http://ep.ebi.ac.uk). Distance measures and clustering algorithms are described in Supplemental Table 1. To determine a significant level of coexpression, the expression datasets were randomized and then used in otherwise identical analyses. Randomization was repeated four times to properly estimate the background distribution. For each method, the significance threshold for the level of coexpression was set so that less than 0.1% of all putative protein pairs are retrieved after analysis on the randomized expression dataset. Due to differences in the expression datasets used, this corresponds to a P value of 0.01 for the Hughes data and a P value of 0.003 for the “Conditions” dataset. For analysis of the differences in coexpression observed for previously known and previously unknown interactions within each interaction dataset (Table 1), approximately 6000 pairs were randomly generated. To determine the significance of the observed differences an ANOVA followed by Bonferroni corrected t-test was performed.

Analysis of Function for Uncharacterized Genes
Yeast gene deletion strains were from Research Genetics and derived from S288C (BY4741 *MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0*) as described[5].

Growth on Raffinose
Yeast strains were streaked on YEP plates containing 2% glucose or 2% raffinose + 1 μg/ml antimycin A and incubated at 30°C, as described[46].

Heat Shock Treatment
Yeast strains were streaked on YEPD plates and incubated for 3 days at 30°C. The cells were then replica plated on two YEPD plates, one of which was immediately transferred to 53°C for 7 hours. After this cells were shifted back to 30°C as described[39].

Growth on Caffeine plus Sorbitol
Liquid cultures were grown overnight. Cells were adjusted to a concentration of $3x10^6$ cells/ml in YEP, diluted in steps of 5 and spotted on either YEPD, YEPD + 10 mM caffeine, or YEPD + 10 mM caffeine + 1M sorbitol plates, as described[52].

Growth on Galactose
Liquid cultures were grown overnight. Cells were adjusted to a concentration of $25x10^6$ cells/ml in YEP, diluted in steps of 5 and spotted on YEP plates containing 2% glucose or 2% galactose. Plates were incubated at 30°C and growth was compared after 2 days.

Acknowledgements

We thank Wim van Driel, Pauline Hogeweg, Dik van Leenen, Peter van der Vliet, Cisca Wijmenga and Jean-Christophe Andrau for support and discussions. We acknowledge the financial support of the Netherlands Organisation for Scientific Research (NWO).

References

1. Brown, P.O. & Botstein, D. Exploring the new world of the genome with DNA microarrays. *Nat Genet* **21**, 33-7 (1999).
2. Lockhart, D.J. & Winzeler, E.A. Genomics, gene expression and DNA arrays. *Nature* **405**, 827-36 (2000).
3. Young, R.A. Biomedical discovery with DNA arrays. *Cell* **102**, 9-15 (2000).
4. Ross-Macdonald, P. et al. Large-scale analysis of the yeast genome by transposon tagging and gene disruption. *Nature* **402**, 413-8 (1999).
5. Winzeler, E.A. et al. Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. *Science* **285**, 901-6 (1999).
6. Uetz, P. et al. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature* **403**, 623-7 (2000).
7. Ito, T. et al. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A* **98**, 4569-74 (2001).
8. Gavin, A.C. et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141-7 (2002).
9. Ho, Y. et al. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature* **415**, 180-3 (2002).
10. Washburn, M.P., Wolters, D. & Yates, J.R., 3rd. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotechnol* **19**, 242-7 (2001).
11. Martzen, M.R. et al. A biochemical genomics approach for identifying genes by the activity of their products. *Science* **286**, 1153-5 (1999).
12. Fiehn, O. et al. Metabolite profiling for plant functional genomics. *Nat Biotechnol* **18**, 1157-61 (2000).
13. Ren, B. et al. Genome-wide location and function of DNA binding proteins. *Science* **290**, 2306-9 (2000).
14. Brent, R. Genomic biology. *Cell* **100**, 169-83 (2000).
15. Gaasterland, T. & Oprea, M. Whole-genome analysis: annotations and updates. *Curr Opin Struct Biol* **11**, 377-81 (2001).
16. Goffeau, A. et al. The yeast genome directory. *Nature* **387**, 5 (1997).
17. Cherry, J.M. et al. SGD: Saccharomyces Genome Database. *Nucleic Acids Res* **26**, 73-9 (1998).
18. Hodges, P.E., McKee, A.H., Davis, B.P., Payne, W.E. & Garrels, J.I. The Yeast Proteome Database (YPD): a model for the organization and presentation of genome-wide functional data. *Nucleic Acids Res* **27**, 69-73 (1999).
19. Mewes, H.W. et al. MIPS: a database for genomes and protein sequences. *Nucleic Acids Res* **28**, 37-40 (2000).
20. Marcotte, E.M., Pellegrini, M., Thompson, M.J., Yeates, T.O. & Eisenberg, D. A combined algorithm for genome-

wide prediction of protein function. *Nature* **402**, 83-6 (1999).
21. Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* **95**, 14863-8 (1998).
22. Grigoriev, A. A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeast Saccharomyces cerevisiae. *Nucleic Acids Res* **29**, 3513-9 (2001).
23. Ge, H., Liu, Z., Church, G.M. & Vidal, M. Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. *Nat Genet* (2001).
24. Jansen, R., Greenbaum, D. & Gerstein, M. Relating whole-genome expression data with protein-protein interactions. *Genome Res* **12**, 37-46 (2002).
25. Brazma, A. & Vilo, J. Gene expression data analysis. *FEBS Lett* **480**, 17-24 (2000).
26. Quackenbush, J. Computational analysis of microarray data. *Nat Rev Genet* **2**, 418-27 (2001).
27. Hughes, T.R. et al. Functional discovery via a compendium of expression profiles. *Cell* **102**, 109-26 (2000).
28. Gasch, A.P. et al. Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell* **11**, 4241-57 (2000).
29. Spellman, P.T. et al. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell* **9**, 3273-97 (1998).
30. Chu, S. et al. The transcriptional program of sporulation in budding yeast. *Science* **282**, 699-705 (1998).
31. Roberts, C.J. et al. Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. *Science* **287**, 873-80 (2000).
32. Travers, K.J. et al. Functional and genomic analyses reveal an essential coordination between the unfolded protein response and ER-associated degradation. *Cell* **101**, 249-58 (2000).
33. Andrews, B. & Measday, V. The cyclin family of budding yeast: abundant use of a good idea. *Trends Genet* **14**, 66-72 (1998).
34. Dirick, L., Bohm, T. & Nasmyth, K. Roles and regulation of Cln-Cdc28 kinases at the start of the cell cycle of Saccharomyces cerevisiae. *Embo J* **14**, 4803-13 (1995).
35. Hellmuth, K. et al. Yeast Los1p has properties of an exportin-like nucleocytoplasmic transport factor for tRNA. *Mol Cell Biol* **18**, 6374-86 (1998).
36. Sadhale, P.P. & Woychik, N.A. C25, an essential RNA polymerase III subunit related to the RNA polymerase II subunit RPB7. *Mol Cell Biol* **14**, 6164-70 (1994).
37. Hohmann, S. & Mager, W.M. *Yeast Stress Responses*, 247 (R.G. Landes Company, Georgetown, 1997).
38. Singer, M.A. & Lindquist, S. Thermotolerance in Saccharomyces cerevisiae: the Yin and Yang of trehalose. *Trends Biotechnol* **16**, 460-8 (1998).
39. Nwaka, S., Mechler, B., Destruelle, M. & Holzer, H. Phenotypic features of trehalase mutants in Saccharomyces cerevisiae. *FEBS Lett* **360**, 286-90 (1995).
40. Madhani, H.D. & Fink, G.R. The riddle of MAP kinase signaling specificity. *Trends Genet* **14**, 151-5 (1998).
41. Nickas, M.E. & Yaffe, M.P. BRO1, a novel gene that interacts with components of the Pkc1p-mitogen-activated protein kinase pathway in Saccharomyces cerevisiae. *Mol Cell Biol* **16**, 2585-93 (1996).
42. Lee, T.I. & Young, R.A. Transcription of eukaryotic protein-coding genes. *Annu Rev Genet* **34**, 77-137 (2000).
43. Liao, S.M. et al. A kinase-cyclin pair in the RNA polymerase II holoenzyme. *Nature* **374**, 193-6 (1995).
44. Jiang, R. & Carlson, M. Glucose regulates protein interactions within the yeast SNF1 protein kinase complex. *Genes Dev* **10**, 3105-15 (1996).
45. Tu, J., Vallier, L.G. & Carlson, M. Molecular and genetic analysis of the SNF7 gene in Saccharomyces cerevisiae. *Genetics* **135**, 17-23 (1993).
46. Ganster, R.W., McCartney, R.R. & Schmidt, M.C. Identification of a calcineurin-independent pathway required for sodium ion stress response in Saccharomyces cerevisiae. *Genetics* **150**, 31-42 (1998).
47. Enright, A.J., Iliopoulos, I., Kyrpides, N.C. & Ouzounis, C.A. Protein interaction maps for complete genomes based on gene fusion events. *Nature* **402**, 86-90 (1999).
48. Huynen, M., Snel, B., Lathe, W., 3rd & Bork, P. Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res* **10**, 1204-10 (2000).
49. Apweiler, R. et al. The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res* **29**, 37-40 (2001).
50. Ashburner, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-9 (2000).
51. Ideker, T. et al. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* **292**, 929-34 (2001).
52. Watanabe, Y., Irie, K. & Matsumoto, K. Yeast RLM1 encodes a serum response factor-like protein that may function downstream of the Mpk1 (Slt2) mitogen-activated protein kinase pathway. *Mol Cell Biol* **15**, 5740-9 (1995).

Predicting Gene Function through Systematic Analysis and Quality Assessment of High-Throughput Data

Chapter VII

Patrick Kemmeren, Thessa T.J.P. Kockelkorn, Theo Bijma, Rogier Donders & Frank C.P. Holstege

Bioinformatics 2005 Apr 15; 21(8): 1644-52

Predicting Gene Function through Systematic Analysis and Quality Assessment of High-Throughput Data

Patrick Kemmeren[1], Thessa T.J.P. Kockelkorn[1], Theo Bijma[1], Rogier Donders[2] & Frank C.P. Holstege[1]

[1]Department of Physiological Chemistry, Division of Biomedical Genetics, UMC Utrecht, the Netherlands;
[2]Department of Innovation Studies, Copernicus Institute, Utrecht University, the Netherlands.

Abstract

Motivation: Determining gene function is an important challenge arising from the availability of whole genome sequences. Until recently, approaches based on sequence homology were the only high-throughput method for predicting gene function. Use of high-throughput generated experimental datasets for determining gene function has been limited for several reasons. Results: Here a new approach is presented for integration of high-throughput datasets, leading to prediction of function based on relationships supported by multiple types and sources of data. This is achieved with a database containing 125 different high-throughput datasets describing phenotypes, cellular locations, protein interactions and mRNA expression levels from *S. cerevisiae*, using a bit-vector representation and information content based ranking. The approach takes characteristic and qualitative differences between the datasets into account, is highly flexible, efficient and scalable. Database queries result in predictions for 543 uncharacterized genes, based on multiple functional relationships each supported by at least three types of experimental data. Some of these are experimentally verified, further demonstrating their reliability. The results also generate insights into the relative merits of different data types and provide a coherent framework for functional genomic datamining.

Introduction

Whole-genome sequencing projects form a tremendous resource for biological discovery[1]. Besides genome annotation[2], an important challenge arising from the availability of genomes is elucidation of the function of thousands of newly discovered genes, the prime goal of functional genomics[3]. The Gene Ontology (GO) Consortium provides gene annotations for most organisms[4]. Curating literature and annotating genes accordingly is a labor-intensive task. Traditional approaches for determining gene function cannot keep up with the current rate of gene discovery. For example, although the *S. cerevisiae* genome has been available since 1996, 3497 genes are still classified as unknown according to GO molecular function or cellular component categories. Classical approaches for assigning gene function include sequence homology, gene fusion events, gene order conservation and phylogenetic profiles[5]. Although extremely powerful approaches, one drawback is lack of experimental evidence. In addition, homology on its own may result in imprecise annotation or return weak similarity with well-characterized genes. As in most areas of biological research, complementary approaches are required to more precisely determine gene function.

To address the rate of gene discovery, high-throughput approaches are being developed for biological experimentation. These include mRNA expression-profiling[6-17], determination of gene-deletion phenotypes[18,19], cellular location of proteins[20,21], protein-protein interactions (ppi)[22-25], assays for biological activity using protein arrays[26,27], synthetic lethal screens[28] and RNA interference screens[29,30]. When used on their own, such high-

throughput datasets have already proven their usefulness for inferring hypotheses about gene function. High-throughput data can also be used in combination[31-41]. Most previous studies have examined the relationships revealed by studying particular combinations of data. Such studies indicate that combining several high-throughput data types will likely aid gene function prediction.

For dealing with, and integrating many different sources of high-throughput data, several issues need addressing. First, differences in data quality within and between datasets exist because of their high-throughput nature[34,36,42]. The presence of false positives therefore needs to be assessed and dealt with. Secondly, each data type has different properties. For instance, phenotype data is categorical, whereas mRNA expression data is continuous. The underlying statistical distribution is different and needs to be addressed accordingly. Thirdly, the availability of these datasets should be organized so that researchers can efficiently mine the data and prioritize hypotheses. Relevant in this respect are speed, visualization, flexibility and ease of use. Finally, it makes sense to develop methods that are easily scalable to accommodate an ever-increasing amount of data, and sufficiently generic to incorporate new data types.

Here we present a novel method for integration of several high-throughput datasets to accurately predict gene function. The method is based on the use of bit-vectors and has been applied to a database containing the majority of presently available high-throughput data for *S. cerevisiae*. The method applied addresses most of the issues described above, is highly flexible, scalable and efficient. To demonstrate the power of the approach some of the functional predictions are verified. The broad-ranging combination of data types used here have as yet not been mined in parallel before. Besides predicting gene function, the approach can also be used for studying networks of functional relationships. The results reveal interesting characteristics of the high-throughput datasets, their capacity for predicting gene function and demonstrates that combining high-throughput datasets is a powerful approach for functional genomic analyses.

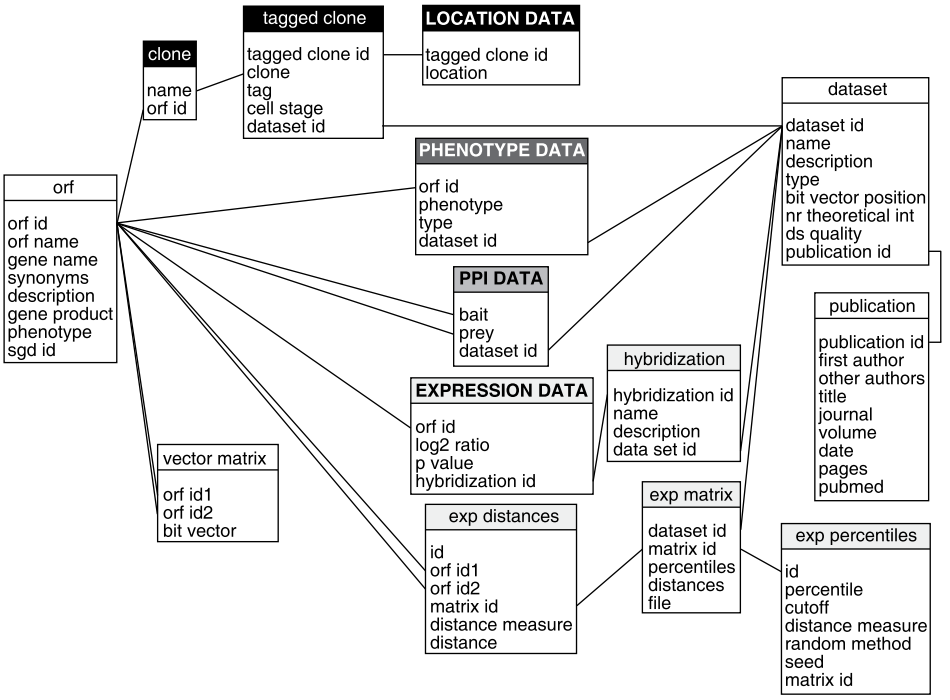
Results

A Framework for Integrated Analysis of High-Throughput Data

To combine different high-throughput data types, a general framework was devised (Figure 1). The database scheme includes protein-protein interactions (ppi)[22-25], mRNA expression (exp)[6-8,10-13,15,16], phenotype (phen)[18,44] and cellular location data (loc)[20,21]. For integrated analysis of the data, each dataset is transformed into a binary representation. The value zero denotes the absence, one the presence of a relationship for a protein pair, based on that particular dataset. Relationships for protein interaction data obtained from tagging approaches[24,25] can be represented in two different ways[45]. The first method connects the tagged protein with all other proteins found in the screen (spoke). The second approach creates an all versus all connection scheme (matrix). Due to the versatility of the bit-vector approach, both models are represented and can be chosen in subsequent analyses.

Relationships obtained from mRNA expression data are based on the presence of a significant degree of coexpression[34] and are calculated at different levels of granularity. This means that coexpression is calculated either across single experiments such as a single time-course or across all datasets. Again, both representations can be chosen for subsequent queries. Connections obtained from phenotype data are derived from gene knockout experiments showing the same phenotype for a given gene pair. For location data, a relationship is based on two proteins sharing the same cellular compartment. Proteins that are found in multiple cellular compartments are only connected to those proteins that are also found in an identical set of compartments. The rationale and consequences of these choices are discussed later. Together, initial preprocessing results in 125 datasets summarized in a single table using a bit-vector representation. The advantage of this approach is that subsequent selections can be based on individual datasets, data type, or

Figure 1 | A Framework for Integrated Analysis
A simplified database schema used for the integrated analysis of high-throughput datasets is depicted (For the full database schema see Supplemental Figure 1). Different types of high-throughput data are incorporated into this database, namely: mRNA expression data (exp; light grey), protein-protein interaction data (ppi; grey), phenotype data (phen; dark grey) and location data (loc; black). For the analysis all data is transformed into a binary representation and summarized in one table called "vector matrix" (see Experimental Procedures). This table represents a sparse data matrix and contains one entry per orf pair found in any of the datasets available within the database. The bit-vector stored for each orf pair represents the datasets for which this particular protein pair is found.



any arbitrary combination of these.

Increasing the Reliability of High-Throughput Data Queries

Currently the database holds 3.7 million relationships (Table 1). These are all based on single types of high-throughput experimental observations with varying degrees of accuracy and information content. As expected, the number of pair-wise relationships is reduced dramatically when those supported by combinations of different datasets are taken into account (Table 1). In contrast, the percentage of previously known ppi and complexes increases from 0.1% to 28.1% and from 0.24% to 34.4% respectively, when relationships supported by multiple datasets are taken into account. One goal of integrated analyses is to lower the influence of false-positives. The chance of finding false positives within the overlap between different datasets is drastically reduced. For example when combining four datasets with a false positive rate of 0.05, the chance of finding a false positive that occurs in all datasets will become $0.05^4 = 6.25 \times 10^{-6}$. Previously established ppi's represent relationships of higher confidence than those from high-throughput studies. The large increase in their representation within the overlaps between datasets shows that integrated analyses are well suited for determining relationships of higher confidence.

To assess the amount of valuable information within each dataset, the information

Table 1 | Characteristics of High-Throughput Datasets and Queries

	No. relations	Avg. IC	% Known ppi	% Known complexes	Avg. relationship confidence	No. predictions
total	3,671,071	-	0.10	0.24	-	-
mRNA coexpression	2,600,027	97.8	0.04	0.24	-	-
location	1,123,498	93.8	0.08	0.38	-	-
phenotype	176,338	99.8	0.08	0.15	-	-
ppi (matrix)	24,702	99.9	3.40	6.30	-	-
ppi (spoke)	9,754	99.9	6.00	7.60	-	-
exp+ppi(matrix)+(phen or loc)	2,323	-	6.07	18.51	793	543
exp+ppi(spoke)+(phen or loc)	738	-	14.36	27.78	832	155
ppi(matrix)+exp+phen+loc	112	-	9.80	17.90	1,036	25
ppi(spoke)+exp+phen+loc	32	-	28.10	34.40	925	4

content (IC) is calculated for each individual dataset (see Experimental Procedures). In principle, the less data derived from a potentially large data space the more informative it is. In line with other information theoretical approaches, such as Shannon's information measure[46], an increase in the information content corresponds to a decrease in uncertainty. Using this IC an implicit ranking can be made for the different data types (Table 1). This table shows that the protein interaction data contains the most valuable information (99.9), followed by phenotype (99.8), mRNA coexpression (97.8) and location (93.8). Based on the IC, a measure can then be calculated for each individual relationship (relationship confidence, see Experimental Procedures), a higher relationship confidence represents those relationships with a high amount of information, i.e. low degree of uncertainty. Both the relationship confidence and a restriction on the data type can then be used to obtain a reliable set of functional relationships.

Predicting Functional Annotation

The high-throughput datasets can be mined in a variety of ways. Figure 2A shows all 2398 relationships supported by significant mRNA coexpression and ppi. A more stringent query is represented in Figure 2B. Only relationships supported by ppi (matrix), mRNA coexpression, shared phenotype and location are represented. This creates a smaller, but densely interconnected network (Figure 2B). Use of GO[4] shows that genes involved in similar processes are highly interconnected and cluster together in the network structure. Using such queries, hypotheses can be formulated about the function, based on neighboring proteins.

Depending on how the high-throughput data is used, different selection criteria can be applied. When used as an exploratory method to find novel relationships, relaxed criteria can be applied, such as sharing at least ppi and mRNA coexpression (Figure 2A). However, for improving current functional annotation of uncharacterized genes, more stringent criteria are warranted. For this purpose four different queries have been performed which can be used for high confidence functional predictions (Table 1). The most stringent query selects only those relationships supported by all four data types, using the spoke model for ppi. The least stringent of these queries selects those relationships supported by ppi (matrix), mRNA coexpression and either shared phenotype or cellular location (Table 1). For each query, the relationships are ranked by relationship confidence, allowing for a more granular subselection or prioritization of hypotheses (see also Supplemental Table 1 through 8). More predictions can be made if the criteria are further relaxed.

With these four queries, all based on support from three completely independent kinds of experimental data, 2323 relationships are found, 543 of which apply to genes for which

Figure 2 | Finding Functional Relationships through Integrated Networks
(**A**) Network obtained when selecting those relationships that are supported by mRNA coexpression and ppi (matrix). Those relationships supported by a known ppi are shown in red. (**B**) Additional requirement of both a similar phenotype and location results in the genes shown in this network. All relationships shown here are supported by at least one ppi (blue lines), those supported by all four data types are drawn in red. Genes are color-coded according to the most significant GO categories. See also Appendix; Color Figure 7.2.

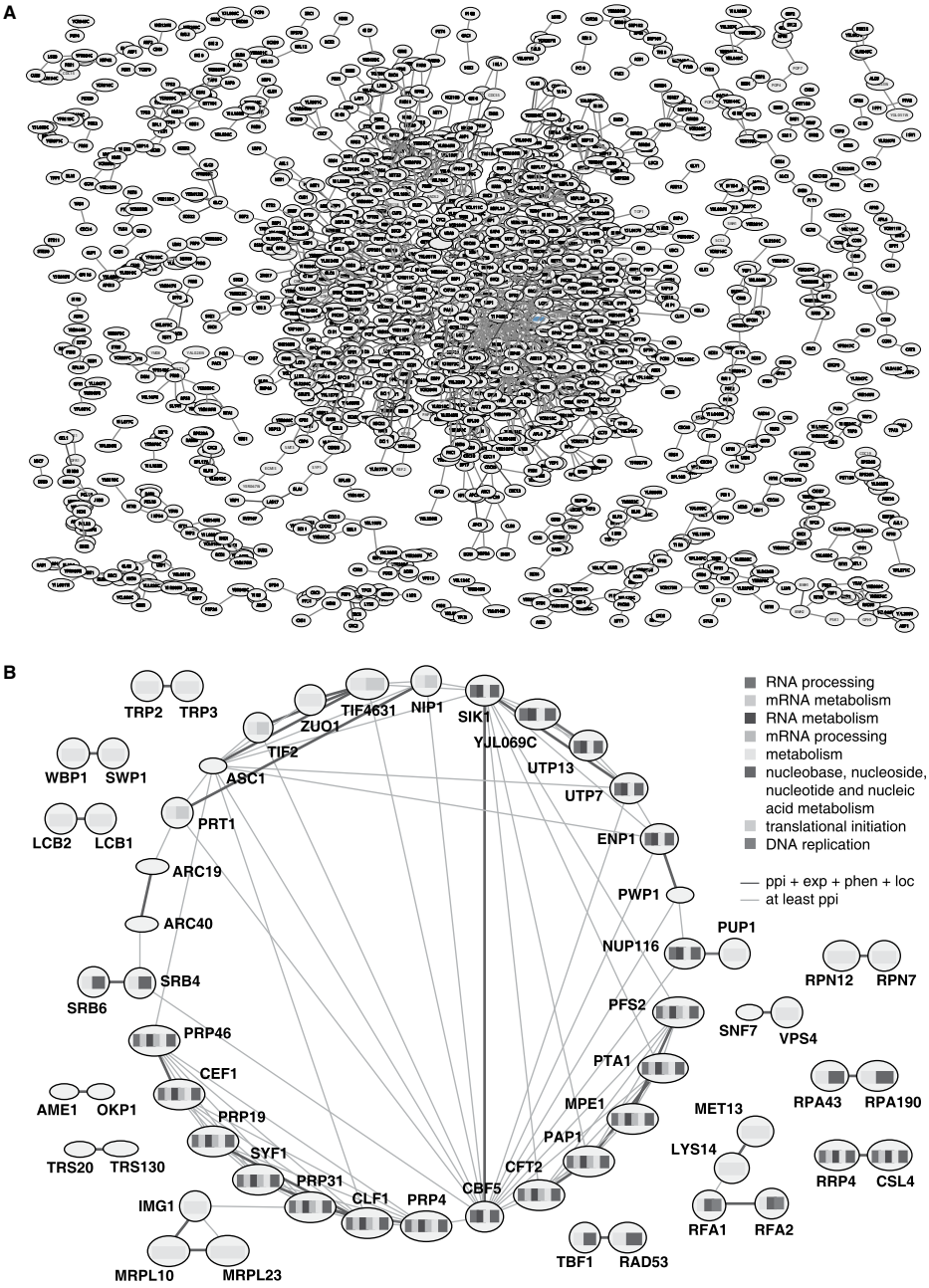
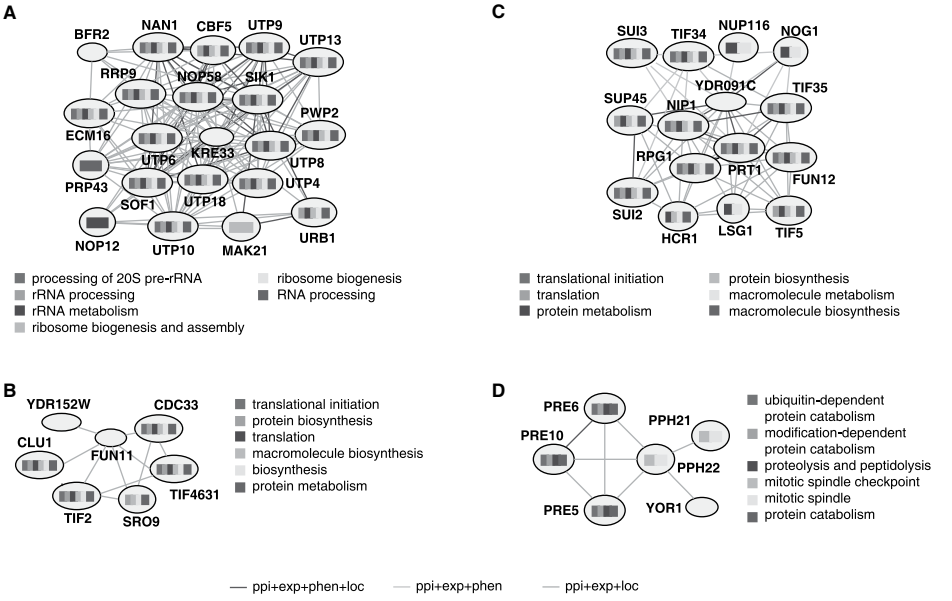


Figure 3 | Predicting Function of Uncharacterized Genes
Gene neighborhood of KRE33 (**A**), FUN11 (**B**), YDR091c (**C**) and PPH22 (**D**). The relationships found are supported either by all four data types (red lines); mRNA coexpression, ppi and location (blue lines) or mRNA coexpression, ppi and phenotype (ochre lines). Genes are color-coded according to the most significant GO categories they belong to. See also Appendix; Color Figure 7.3.



there is no current molecular function or cellular component annotation (Table 1). Two examples are depicted in Figure 3A, B. KRE33 (Figure 3A) is described as “killer toxin resistant” with no GO annotation. It has relationships with 20 genes, each relationship based on a physical interaction, mRNA coexpression and either a shared phenotype or cellular location. Of the 20 linked genes, 13 are part of the U3 snoRNP complex involved in rRNA processing. Four of the other genes are also involved in rRNA metabolism, making it extremely likely that KRE33 is involved in these processes. FUN11 (Figure 3B), stands for Function Unknown Now and also has no GO annotations. It is linked to 5 functionally characterized genes, all of which are involved in translation initiation. From this it is clear that FUN11 too can be tentatively assigned to a role in this process.

Another category of genes for which this approach is useful are those genes with vague or perhaps incorrect annotation. Figure 3C, D shows examples of this. YDR091c (Figure 3C) is annotated as a putative member of the ATP-binding cassette superfamily of nontransporters. Ten of the fifteen genes with which it has reliable links are involved in translation initiation, making it extremely likely that the same holds for YDR091c. PPH22 (Figure 3D) is a serine-threonine protein phosphatase, annotated with a role in the G1/S transition of the mitotic cell cycle. Three of the five genes with which PPH22 is connected, are proteosome components, arguing strongly for a similar annotation for PPH22.

The examples above are from a total of 543 predictions for genes with unknown function, supported by multiple pairwise relationships, each of these based on at least three types of experimental data. 37 of the 543 novel functional predictions concern relationships between single gene pairs. Because such relationships are supported by at least three different types of high-throughput data, it is still likely that the functional prediction is correct. To demonstrate this further, some of these predictions were selected and tested experimentally.

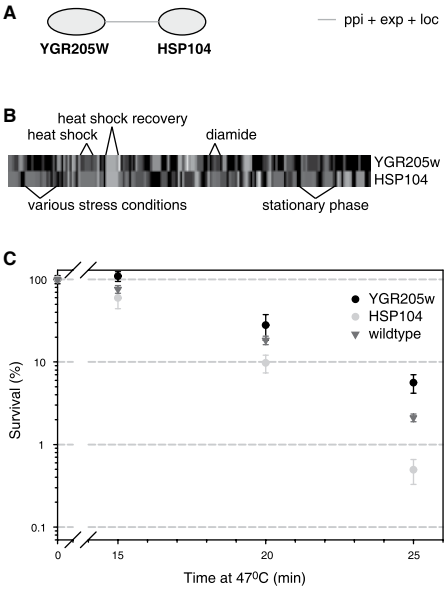


Figure 4 | YGR205w Suppresses Thermotolerance
(**A**) Gene neighborhood of YGR205w. The relationship found between YGR205w and HSP104 is supported by mRNA coexpression, ppi and location (blue line). (**B**) Expression profiles of YGR205w and HSP104 obtained from Gasch et al[12]. Red indicates upregulation, green indicates downregulation. Both mRNA expression levels are upregulated during various processes as indicated and downregulated during heat shock recovery. (**C**) Comparison of thermotolerance in the *hsp104* and *ygr205w* deletion strains and wildtype. Cells lacking HSP104 (grey circle) show a decreased resistance to heat shock when compared to wildtype (triangle). Deletion of YGR205w results in an increased thermotolerance (black circle). See also Appendix; Color Figure 7.4.

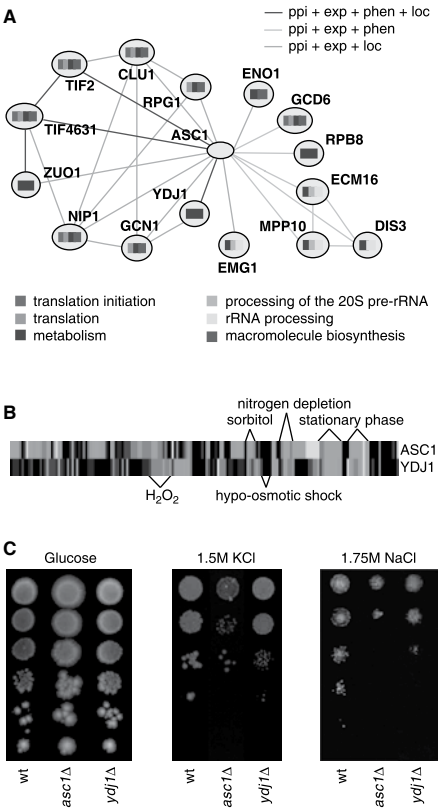


Figure 5 | ASC1 and YDJ1 are Associated with Protein Folding, Translation and Ribosome Biogenesis
(**A**) Gene neighborhood of ASC1. Relationships inside this network are supported either by all four data types (red lines), mRNA coexpression, ppi and similar phenotype (ochre line) or mRNA coexpression, ppi and location (blue line). Genes are color-coded according to the most significant GO categories they belong to. (**B**) Expression profiles of YDJ1 and ASC1 obtained from Gasch et al[12]. Red indicates upregulation, green indicates dowregulation. YDJ1 mRNA expression levels are coregulated during various processes as indicated. (**C**) Wildtype, *ydj1* or *asc1* deletion strains grown in YPD (left), 1.5M KCl (middle) or 1.75M NaCl (right). Both *asc1* and *ydj1* deletion strains show a slow growth during high salt concentrations when compared to wildtype. See also Appendix; Color Figure 7.5.

YGR205W Suppresses Thermotolerance

The first example is a link between HSP104 and YGR205w. Hsp104 is a heat shock protein involved in the rescue of stress-damaged proteins. Together with Hsp70 and Hsp40 it forms a chaperone system that can refold denatured proteins[47]. Cells lacking Hsp104 do not acquire thermotolerance when given a mild pre-heat treatment[48]. A putative link between Hsp104 and Ygr205w is indicated by mRNA coexpression, ppi and identical location. These three kinds of experimental evidence would together predict that the uncharacterized ORF YGR205w is also involved in stress response. Interestingly, mRNA coexpression is observed during various stress conditions, including heat shock and stationary phase (Figure 4B). To validate the prediction, an *YGR205w* deletion strain was tested for thermotolerance. Cells lacking HSP104 show decreased resistance to heat shock when compared to wildtype (Figure 4C)[48]. In agreement with the prediction, deletion of YGR205w also shows a thermotolerance phenotype. Rather than precisely mimicking HSP104 deletion, deletion of YGR205w results in increased thermotolerance (Figure 4C), indicating that it is involved in the same pathway as HSP104, but with an opposite role. Based on this evidence, YGR205w can now be annotated as involved in negative regulation of thermotolerance.

ASC1 and YDJ1 are Associated with Protein Folding, Translation and Ribosome Biogenesis

Another intriguing example of genes for which novel functional annotation is predicted are ASC1 and YDJ1. In the network of high-throughput data relationships, ASC1 is surrounded by three groups of genes (Figure 5A). Shared mRNA coexpression, ppi and phenotypes are found with four genes involved in rRNA processing. Multiple links are also found with six translation initiation factors. Given the number of related genes, it is likely that ASC1 is involved in both rRNA processing and translation, despite having no annotation. A third category is represented by two genes with dnaJ homolog regions, one of which, ZUO1, is a ribosome associated chaperone. We reasoned that Ydj1 might also be involved as a nascent protein chaperone due to its ppi, shared mRNA expression and location with Asc1. mRNA coexpression for ASC1 and YDJ1 is observed during multiple conditions (Figure 5B). In accordance with a role for ASC1 in stress induced protein misfolding, ASC1 deletion results in a severe growth defect at elevated NaCl concentrations[49]. In agreement with the prediction that ASC1 and YDJ1 are both involved in protecting cells from elevated salt concentrations, YDJ1 deletion also results in a growth defect at elevated NaCl and KCl concentrations (Figure 5C).

Interactive Datamining Tool

The examples above illustrate the utility of combining different high-throughput data types. The focus here is on functional annotation, but this approach of a single database, bit-vector queries and IC ranking is equally useful for examing the network structures and characteristics resulting from many different combinations of the 125 datasets included in the database. A datamining tool has therefore been setup on the accompanying website (http://www.genomics.med.uu.nl/pub/pk/comb_gen_network/). This tool allows detailed study of relationships found when starting from all genes, a certain complex, a single protein, or a number of protein pairs. Information about the confidence, supporting data types, mRNA expression profiles, gene neighborhood and GO annotation can be retrieved. Restrictions can be placed on the type of data support needed and individual datasets can be included or excluded at will.

Discussion

These results demonstrate how several issues arising from the availability of genome sequences and high-throughput experimental data can be addressed through data

integration. A general framework is provided that can incorporate any data type represented in a binary format. The approach takes into account characteristics of different data types, heterogeneity in data quality and IC. Besides predicting gene function, the study also sheds light on the properties of different data types and sets.

One interesting aspect is the effect of the level of granularity on the relationships found within mRNA expression data. For some gene pairs, mRNA coexpression is only found when looking at individual experiments such as a single time course, rather than the entire collection of expression data. This agrees with the fact that some interacting proteins are only coexpressed under specific conditions. This has been taken into account in the bit-vector representation. Another aspect involves the compartmentation of the cellular location data. Here we have chosen to base relationships on an exact match of the location pattern. If partly matching locations had been permitted, the IC of the location data would be too low to be meaningful. This is due to the low number of possible cellular locations and in this sense the location data will remain less useful until cellular location data becomes more fine-grained.

Important steps have been taken in the past to integrate different types of genome-scale data [31-34,37,40], usually in pair-wise combinations and typically for the purpose of examining genome-scale network characteristics. Four different data types have been used here and the system can be extended to include other data types such as synthetic lethality, chromosomal location, regulatory motifs, sequence homology, metabolic pathways, etc. There is not a large degree of overlap between the different data types, especially when selecting relationships supported by all four data types (Table 1). Although not all gene function relationships can be expected to demonstrate support from all data types, it is still striking, given that most of the underlying datasets are presented as genome-wide. Many clearly are not completely genome-wide, often due to constraints of the methods used. When comparing multiple datasets this soon leads to a significant loss of genome coverage.

Another reason that has been offered in the past for lack of overlap between genome-wide datasets, especially for those of the same type, is heterogeneity in quality[34,42]. Two measures have been taken here to deal with this. Restricting the analyses to relationships supported by multiple data types considerably improves reliability as is obvious from the enrichment for previously known interactions (Table 1). A second step is the use of a relationship confidence based on the IC of the datasets. Using the IC a clear distinction between the different data types can be made (Table 1). As expected protein interaction data, the most direct type of evidence for a functional relationship has the highest IC, directly followed by phenotype data.

The approach presented here is suitable for generating hypotheses about individual genes and can aid annotation initiatives such as GO[4] or genome databases such as SGD [50]. To better utilize existing knowledge, the annotations themselves should also be annotated as to how information is obtained. These so-called evidence codes already exist within GO, albeit at a somewhat undifferentiated level. For example, no distinction is made between a physical interaction arising from a two-hybrid or co-immunoprecipitation assay, or whether the approach was high-throughput. Initiatives such as BIND and IntAct are tackling this problem for protein interactions[51,52]. Another aspect involves the annotations based on computational methods. If this information is not stored, valuable information is lost. On the other hand, if such annotations are used by another computational method, such predictions will lead to less trustworthy annotations. Therefore more details about how annotation is obtained needs to be available and computational methods need to use this information.

Most of the approaches to integrate different types of high-throughput data have been performed in *S. cerevisiae*. As more data is generated, similar approaches will be required for other organisms. Besides being easily extendable to other data types, the framework presented here is well-suited for other organisms, as it already takes into account scalability, flexibility and standardization. The results presented here show that alongside existing

methods, integrating diverse types of functional genomic data is a powerful method for tackling gene function annotation.

Experimental Procedures

Significant mRNA Coexpression
mRNA coexpression is calculated based on the standard correlation measure. All distance matrix calculations where performed using Expression Profiler (http://www.ebi.ac.uk/expressionprofiler/)[43]. To determine a significant level of coexpression, the expression data was randomized and then used in otherwise identical analysis as described previously[34]. Here, positive correlation distances with a P-value smaller than 0.0001 were considered significantly coexpressed and set to one for subsequent analyses. Expression datasets with an information content less than 90 where excluded from further analysis.

Information content
The information content (IC) for each individual dataset was calculated by

$$IC = \left(1 - \frac{r}{R\max}\right) x100$$

where r is the number of relationships found within that particular dataset and R_{max} denotes the maximum number of theoretically possible relationships for that dataset and is defined as

$$R_{\max} = \frac{Nx(N-1)}{2}$$

for square matrices and defined as

$$R_{\max} = \frac{Mx(M-1)}{2} + Mx(N-M)$$

for rectangular matrices. N and M denote the number of rows and columns, whereby M represents the smallest dimension. The confidence of each individual relationship is then based on

$$C_{rel} = \sum_{i=1}^{n} \left(1 - \frac{r_i}{R\max_i}\right) x100$$

where the sum is taken over the IC of all datasets in which the relationship occurs.

Gene Ontology Category Assignment and Graph Drawing
Gene Ontology (GO) category assignments where performed using the GO Term Finder perl module (http://search.cpan.org/dist/GO-TermFinder/) developed by the Stanford Microarray Database (SMD). In short, a hypergeometric test with a standard Bonferroni correction was used to obtain those GO categories that show a significant overlap with the genes obtained from the different networks. From these significant GO categories only the top six scoring categories per gene are shown in the networks. All networks used throughout this study have been drawn using AT&T's GraphViz tool (http://www.research.att.com/sw/tools/graphviz/).

Functional Predictions
Functional predictions are based on the GO annotations from SGD obtained at June 24th 2003. Those relationships containing one gene with a known function and one gene with the annotation "molecular function unknown" or "cellular component unknown" were used for inferring the functional annotation from the better-characterized protein to the uncharacterized protein.

Analysis of Function for Uncharacterized Genes
Yeast gene deletion strains were from Research Genetics and derived from S288C (BY4741 *MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0*) as described[18].

Thermotolerance Assay
Cells were grown in YEPD to mid-exponential phase (OD_{600}=0.6) at 30°C and were then shifted to 47°C, while shaking. Aliquots were removed at regular intervals and put on ice. Approximately 500 cells at 0, 15 and 20 minutes and approximately 50.000 cells at 25 minutes were plated in triplicate on YEPD plates. The resulting colonies were counted after 3 days of growth at 30°C.

Growth at High Salt Concentrations
Cells were grown o/n at 30°C in liquid media. The concentration was then adjusted to approximately $1*10^6$cells/ml in YEP. Dilution steps of 3 were spotted on YEPD plates, YEPD plates containing 1.75M NaCl and YEPD plates containing 1.5M KCl.

Acknowledgments
We thank Philip Lijnzaad, Thomas Schlitt, Harm van Bakel and Arnaud Leijen for discussions and technical support. Supported by grants from the Netherlands Organization for Scientific Research (NWO); 05050205, 016026009, 05071002 and by the European Union fifth framework project TEMBLOR.

References

1. Grunenfelder, B. & Winzeler, E.A. Treasures and traps in genome-wide data sets: case examples from yeast. *Nat Rev Genet* **3**, 653-61 (2002).
2. Rust, A.G., Mongin, E. & Birney, E. Genome annotation techniques: new approaches and challenges. *Drug Discov Today* **7**, S70-6 (2002).
3. Brent, R. Genomic biology. *Cell* **100**, 169-83 (2000).
4. Harris, M.A. et al. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* **32 Database issue**, D258-61 (2004).
5. Huynen, M., Snel, B., Lathe, W., 3rd & Bork, P. Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res* **10**, 1204-10 (2000).
6. DeRisi, J.L., Iyer, V.R. & Brown, P.O. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**, 680-6 (1997).
7. Chu, S. et al. The transcriptional program of sporulation in budding yeast. *Science* **282**, 699-705 (1998).
8. Spellman, P.T. et al. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell* **9**, 3273-97 (1998).
9. Brown, P.O. & Botstein, D. Exploring the new world of the genome with DNA microarrays. *Nat Genet* **21**, 33-7 (1999).
10. Ferea, T.L., Botstein, D., Brown, P.O. & Rosenzweig, R.F. Systematic changes in gene expression patterns following adaptive evolution in yeast. *Proc Natl Acad Sci U S A* **96**, 9721-6 (1999).
11. Galitski, T., Saldanha, A.J., Styles, C.A., Lander, E.S. & Fink, G.R. Ploidy regulation of gene expression. *Science* **285**, 251-4 (1999).
12. Gasch, A.P. et al. Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell* **11**, 4241-57 (2000).
13. Hughes, T.R. et al. Functional discovery via a compendium of expression profiles. *Cell* **102**, 109-26 (2000).
14. Lockhart, D.J. & Winzeler, E.A. Genomics, gene expression and DNA arrays. *Nature* **405**, 827-36 (2000).
15. Roberts, C.J. et al. Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. *Science* **287**, 873-80 (2000).
16. Travers, K.J. et al. Functional and genomic analyses reveal an essential coordination between the unfolded protein response and ER-associated degradation. *Cell* **101**, 249-58 (2000).
17. Young, R.A. Biomedical discovery with DNA arrays. *Cell* **102**, 9-15 (2000).
18. Winzeler, E.A. et al. Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. *Science* **285**, 901-6 (1999).
19. Giaever, G. et al. Functional profiling of the Saccharomyces cerevisiae genome. *Nature* **418**, 387-91 (2002).
20. Kumar, A. et al. Subcellular localization of the yeast proteome. *Genes Dev* **16**, 707-19 (2002).
21. Huh, W.K. et al. Global analysis of protein localization in budding yeast. *Nature* **425**, 686-91 (2003).
22. Uetz, P. et al. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature* **403**, 623-7 (2000).
23. Ito, T. et al. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A* **98**, 4569-74 (2001).
24. Gavin, A.C. et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141-7 (2002).
25. Ho, Y. et al. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature* **415**, 180-3 (2002).
26. Washburn, M.P., Wolters, D. & Yates, J.R., 3rd. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotechnol* **19**, 242-7 (2001).
27. Zhu, H. et al. Global analysis of protein activities using proteome chips. *Science* **293**, 2101-5 (2001).
28. Tong, A.H. et al. Global mapping of the yeast genetic interaction network. *Science* **303**, 808-13 (2004).
29. Kamath, R.S. et al. Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. *Nature* **421**, 231-7 (2003).
30. Berns, K. et al. A large-scale RNAi screen in human cells identifies new components of the p53 pathway. *Nature* **428**, 431-7 (2004).
31. Marcotte, E.M., Pellegrini, M., Thompson, M.J., Yeates, T.O. & Eisenberg, D. A combined algorithm for genome-wide prediction of protein function. *Nature* **402**, 83-6 (1999).

32. Ge, H., Liu, Z., Church, G.M. & Vidal, M. Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. *Nat Genet* **29**, 482-6 (2001).
33. Jansen, R., Greenbaum, D. & Gerstein, M. Relating whole-genome expression data with protein-protein interactions. *Genome Res* **12**, 37-46 (2002).
34. Kemmeren, P. et al. Protein interaction verification and functional annotation by integrated analysis of genome-scale data. *Mol Cell* **9**, 1133-43 (2002).
35. Walhout, A.J. et al. Integrating interactome, phenome, and transcriptome mapping data for the C. elegans germline. *Curr Biol* **12**, 1952-8 (2002).
36. Ge, H., Walhout, A.J. & Vidal, M. Integrating 'omic' information: a bridge between genomics and systems biology. *Trends Genet* **19**, 551-60 (2003).
37. Schlitt, T. et al. From gene networks to gene function. *Genome Res* **13**, 2568-76 (2003).
38. Asthana, S., King, O.D., Gibbons, F.D. & Roth, F.P. Predicting protein complex membership using probabilistic network reliability. *Genome Res* **14**, 1170-5 (2004).
39. Bader, J.S., Chaudhuri, A., Rothberg, J.M. & Chant, J. Gaining confidence in high-throughput protein interaction networks. *Nat Biotechnol* **22**, 78-85 (2004).
40. Parsons, A.B. et al. Integration of chemical-genetic and genetic interaction data links bioactive compounds to cellular target pathways. *Nat Biotechnol* **22**, 62-9 (2004).
41. Pereira-Leal, J.B., Enright, A.J. & Ouzounis, C.A. Detection of functional modules from protein interaction networks. *Proteins* **54**, 49-57 (2004).
42. Kemmeren, P. & Holstege, F.C. Integrating functional genomics data. *Biochem Soc Trans* **31**, 1484-7 (2003).
43. Kapushesky, M. et al. Expression Profiler: next generation - an online platform for analysis of microarray data. *Nucleic Acids Res* **32**, W465-W470 (2004).
44. Ross-Macdonald, P. et al. Large-scale analysis of the yeast genome by transposon tagging and gene disruption. *Nature* **402**, 413-8 (1999).
45. Bader, G.D. & Hogue, C.W. Analyzing yeast protein-protein interaction data obtained from different sources. *Nat Biotechnol* **20**, 991-7 (2002).
46. Shannon, C.E. The mathematical theory of communication. 1963. *MD Comput* **14**, 306-17 (1997).
47. Glover, J.R. & Lindquist, S. Hsp104, Hsp70, and Hsp40: a novel chaperone system that rescues previously aggregated proteins. *Cell* **94**, 73-82 (1998).
48. Ferreira, P.C., Ness, F., Edwards, S.R., Cox, B.S. & Tuite, M.F. The elimination of the yeast [PSI+] prion by guanidine hydrochloride is the result of Hsp104 inactivation. *Mol Microbiol* **40**, 1357-69 (2001).
49. Dunn B, F.T., Spellman P, Schwarz J, Terraciano J, Troyanovich J, Walker S, Greene J, Shaw K, DiDomenico B, Wang Q, Kaloper M, Metzner S, Chung E, Bondre C, Venteicher A, Botstein D, Brown P. Genetic footprinting: A functional analysis of the S. cerevisiae genome. (ed. SGD) (2004).
50. Christie, K.R. et al. Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from Saccharomyces cerevisiae and related sequences from other organisms. *Nucleic Acids Res* **32 Database issue**, D311-4 (2004).
51. Bader, G.D., Betel, D. & Hogue, C.W. BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res* **31**, 248-50 (2003).
52. Hermjakob, H. et al. IntAct: an open source molecular interaction database. *Nucleic Acids Res* **32 Database issue**, D452-5 (2004).

General Discussion

Chapter VIII

General Discussion

Whole-genome sequences provide a tremendous resource for biological and biomedical discovery. In parallel, high-throughput approaches have been developed for determining mRNA expression levels[1-3], gene-deletion phenotypes[4,5], cellular location of proteins[6,7], protein-protein interactions[8-11], RNA interference phenotypes[12,13] and synthetic lethality[14]. As is also exemplified in this thesis, these functional genomic datasets can be used in a variety of ways throughout the life sciences. Many challenges are encountered when analyzing such datasets for biological or biomedical purposes; the most important of these will be discussed in this chapter.

Microarray Data Integrity and Quality Control

Microarray experiments generally consist of dozens of small steps that all need to be tightly controlled to reduce unwanted experimental variation. Besides optimization and standardization of sample processing and handling procedures, a microarray data management and analysis system is required. Through the use of such a microarray data analysis system, as presented in chapter 2, overall manageability, data integrity and standardization of the microarray data can be accomplished. However, at this moment most microarray management systems are still in their infancy. Several aspects need to be dealt with before these systems can be used on a day-to-day basis, guiding the experimentalist during microarray data management and analysis.

First, international annotation and data exchange standards need to be embraced. Through the use of such standards, interpretation and sharing of microarray data is facilitated. Interpretation and sharing of data has always played an important role in scientific research. For this purpose the MGED (Microarray Gene Expression Data) society[15] has developed MIAME (Minimal Information About a Micorarray Experiment) guidelines[16] and the data exchange format MAGE-ML (MicroArray Gene Expression Markup Language)[17]. The underlying principles of most microarray analysis systems are based on these MIAME guidelines. However, significant effort still has to be put in the integration of MAGE-ML within these systems, so that data can easily be exchanged between collaborators or can be submitted to public repositories such as ArrayExpress[18] and GEO (Gene Expression Omnibus)[19]. Initial steps in that direction have been undertaken in chapter 2 by incorporating MIAME-compliant annotation drop-down menus and MAGE-ML export.

Second, mass annotation of samples should be facilitated. Since microarray experiments contain an ever growing number of samples, a delicate balance has to be found between the amount of time spend on annotating samples and the level of detail annotated for each sample. One approach to facilitate the process of annotating large numbers of samples is to take the annotations from a previously annotated, similar sample and use those as a starting point. That way, only those annotations that differ between the two samples need to be modified, thus reducing the annotation effort. Another approach would be to add more intelligence to the annotation process. Based on for instance species and sample type, different questions could be asked. For example, asking for a tissue type when working with a single-cellular organism is futile and can better be avoided.

Third, quality control should be an integral part of any microarray analysis system. Even before any microarray data preprocessing or analysis has occurred, many quality control steps can be taken. These include monitoring of RNA degradation and labeling efficiency[20,21]. That way, possible data quality issues can be detected at an early stage and appropriate actions undertaken. At the microarray data preprocessing level visual inspection of the data at hand is an important quality control step. Through the use of a number of diagnostic plots, as presented in chapter 2, different aspects of the data can be investigated and possible artifacts identified[22,23]. Although these diagnostic plots are very

powerful approaches to assess the data quality, interpretation relies on the expertise of individual scientists and is therefore potentially biased. For this purpose, the microarray analysis system presented in chapter 2 incorporates hybridization quality metrics that allow an unbiased assessment of the microarray data[23-25]. However, where to place thresholds to distinguish good quality data from bad quality data is still unclear at this moment[23-25]. Initially, these quality metrics can be best used to rank different hybridizations within an experiment, so that the most extreme outliers can be detected and eliminated. Furthermore, since the data quality largely depends on the biological sample used, different quality thresholds might be required based on the array batch, species or sample used. For instance, tumor biopsies are more prone to RNA degradation than cell culture experiments due to sample handling and processing, resulting in an overall lower, but still acceptable, data quality.

Fourth, workflow management should be incorporated into any microarray analysis system. A small number of workflows have been used throughout the system presented in chapter 2. Through these workflows scientists are forced to take certain steps, after which further data analysis can occur. As such, an overall higher level of standardization and reproducibility is achieved. Furthermore, definition of workflows for commonly used data preprocessing or analysis steps will increase efficiency and improve the accessibility of the underlying system.

Microarray Data Normalization and Analysis

Many distinct steps have to be undertaken during microarray data preprocessing. For accurate detection of mRNA changes between samples, non-biological variation, such as differences in labeled material, scanner settings and dye-specific biases, need to be corrected[26,27]. Most normalization methods for correcting these systematic biases assume that changes between two samples are balanced and use all genes for normalization purposes, thus concealing global shifts in mRNA levels. Externally spiked-in controls can be used to monitor possible global changes in mRNA content as shown in chapter 3[28]. This is especially important for those experiments where accurate detection of quantitative mRNA levels is required. For instance when studying the effects of global activators or repressors on gene expression levels. Monitoring global mRNA changes might not be required or feasible for every experiment. The feasibility of detecting global mRNA changes depends on the reliable quantification of RNA yields. If only limited amounts of RNA are obtained, for instance when using tumor biopsies or single-cells for microarray experiments, accurate determination of RNA yields is impractical, thus making this approach less reliable. In other cases, detection of global mRNA changes is not required. For instance, when microarray experiments are used as a screen to detect only a handful of the most altered gene expression levels, the exact gene expression levels are unimportant.

Various microarray data analysis tools exist that are either more exploratory in nature, highlighting specific properties of the microarray data, or are used for specific purposes such as disease classification[29,30]. To facilitate microarray data analysis, commonly used tools should be embedded within a general microarray data analysis framework such as the one presented in chapter 4[31] and 5. That way, a wide range of analysis routines are available that are able to exchange data and share a common interface. Furthermore, workflows can be defined for commonly used analysis steps, thus guiding the experimentalist during microarray data analysis.

Integrating Genome-Scale Datasets

High-throughput approaches are generating valuable resources of biological information. These approaches include mRNA expression profiling[1-3], determination of gene-deletion

phenotypes[4,5], protein-protein interactions[8-11], cellular location of proteins[6,7], assays for biological activity[32,33], synthetic lethal screens[14] and RNA interference screens[12,13]. Each of these approaches already is useful for inferring hypothesis about gene function. Combining several high-throughput data types can further aid gene function prediction by assigning functions to uncharacterized genes based on their interactions with other genes as shown in chapter 6 and 7[34-44]. Initial studies have compared particular combinations of data types, for instance mRNA co-expression and protein-protein interactions (chapter 6)[35-37]. However, as more of these datasets are becoming available and our understanding of how to integrate these evolves, more types of data will be integrated (chapter 7)[43,44]. There are some issues that need to be dealt with before these datasets can be most advantageously used for gene function prediction.

First, there is hardly any overlap between the different datasets as demonstrated in chapter 6[37,44]. One explanation is that these datasets do not cover the complete range of possible interactions. More exhaustive screens may be required to fully map all possible interactions. Another explanation is the high level of false positives within the genome-scale datasets. Studies have indicated that high levels of false positives exist within these datasets and when combining these, even ones of the same data type, a relatively small overlap is found (chapter 6)[37,44]. Another issue is introduced when combining datasets of different type. Taking the intersection assumes that a valid relationship should be present in all datasets. Since regulation can take place at different levels within a cell, this approach is too stringent and will lead to a significant loss of possible relationships.

Second, differences in data quality both within and between different datasets exist[37,43,44]. Proper assessment of data quality is therefore required. Qualitative scores can be calculated for the different datasets so that a lower weight can be assigned to the less informative or lower quality datasets[43,44]. A quality indicator, such as the one used in chapter 7, can then be calculated for each individual relationship, based on the supportive datasets. Using this quality indicator, an implicit ranking can be made for the different relationships, so that the higher confidence predictions can be prioritized. This also resolves the last issue discussed in the previous section, since this approach does not rely on a relationship to be present in all datasets for reliable gene function prediction. However, unclear at this moment is where to put an exact threshold to distinguish high confidence predictions from low confidence predictions.

Third, dynamics of the underlying biological system should be included. Approaches used so far have ignored temporal and spatial properties of the underlying data. For instance, the cellular location studies performed so far have all been performed under one condition. This does not account for possible changes in cellular location during different environment conditions. Furthermore, most data types, except for some mRNA expression time series, only provide a snapshot at a certain point in time and do not provide information about differences that occur over time. To incorporate this information in future approaches, more datasets need to be generated under a wide range of environmental conditions, monitored over a certain number of time points. New approaches for integrating these types of data, taking into account both temporal and spatial properties, will further develop our understanding of the multi-functional nature of many genes.

Most integrative approaches used so far have been applied in *Saccharomyces cerevisiae*. Ideally, we would like to apply these approaches to higher eukaryotes such as human and mouse, so that these approaches can aid in developing pharmaceutical therapies for commonly occurring diseases. Additional challenges arise when using higher eukaryotes. Examples include order of magnitude larger genomes, poorly annotated genomes, more diverse cell types, more complex interactions with the environment and organization in tissues and organs. However, since these approaches have already been used in other organisms, information on how to apply these methods is available. Furthermore, additional information can be added from these other organisms using orthologues to map information from one organism to another.

References

1. Brown, P.O. & Botstein, D. Exploring the new world of the genome with DNA microarrays. *Nat Genet* **21**, 33-7 (1999).
2. Lockhart, D.J. & Winzeler, E.A. Genomics, gene expression and DNA arrays. *Nature* **405**, 827-36 (2000).
3. Young, R.A. Biomedical discovery with DNA arrays. *Cell* **102**, 9-15 (2000).
4. Winzeler, E.A. et al. Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. *Science* **285**, 901-6 (1999).
5. Giaever, G. et al. Functional profiling of the Saccharomyces cerevisiae genome. *Nature* **418**, 387-91 (2002).
6. Kumar, A. et al. Subcellular localization of the yeast proteome. *Genes Dev* **16**, 707-19 (2002).
7. Huh, W.K. et al. Global analysis of protein localization in budding yeast. *Nature* **425**, 686-91 (2003).
8. Uetz, P. et al. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature* **403**, 623-7 (2000).
9. Ito, T. et al. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A* **98**, 4569-74 (2001).
10. Gavin, A.C. et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141-7 (2002).
11. Ho, Y. et al. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature* **415**, 180-3 (2002).
12. Kamath, R.S. et al. Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. *Nature* **421**, 231-7 (2003).
13. Berns, K. et al. A large-scale RNAi screen in human cells identifies new components of the p53 pathway. *Nature* **428**, 431-7 (2004).
14. Tong, A.H. et al. Global mapping of the yeast genetic interaction network. *Science* **303**, 808-13 (2004).
15. MGED - Microarray Gene Expression Data Society.
16. Brazma, A. et al. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet* **29**, 365-71 (2001).
17. Spellman, P.T. et al. Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biol* **3**, RESEARCH0046 (2002).
18. Parkinson, H. et al. ArrayExpress--a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res* **33**, D553-5 (2005).
19. Barrett, T. et al. NCBI GEO: mining millions of expression profiles--database and tools. *Nucleic Acids Res* **33**, D562-6 (2005).
20. Auer, H. et al. Chipping away at the chip bias: RNA degradation in microarray analysis. *Nat Genet* **35**, 292-3 (2003).
21. van Bakel, H. & Holstege, F.C. In control: systematic assessment of microarray performance. *EMBO Rep* **5**, 964-9 (2004).
22. Dudoit, S. & Yang, J.Y.H. Bioconductor R Packages for Exploratory Analysis and Normalization of cDNA Microarray Data. in *The Analysis of Gene Expression Data: methods and software* (eds. Parmigiani, G., Garrett, E.S., Irizarry, R.A. & Zeger, S.L.) 73-101 (Springer-Verlag, New York, 2003).
23. Buness, A., Huber, W., Steiner, K., Sultmann, H. & Poustka, A. arrayMagic: two-colour cDNA microarray quality control and preprocessing. *Bioinformatics* **21**, 554-6 (2005).
24. Chen, Y. et al. Ratio statistics of gene expression levels and applications to microarray data analysis. *Bioinformatics* **18**, 1207-15 (2002).
25. Sauer, U., Preininger, C. & Hany-Schmatzberger, R. Quick and simple: quality control of microarray data. *Bioinformatics* **21**, 1572-8 (2005).
26. Tseng, G.C., Oh, M.K., Rohlin, L., Liao, J.C. & Wong, W.H. Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Res* **29**, 2549-57 (2001).
27. Yang, Y.H. et al. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res* **30**, e15 (2002).
28. van de Peppel, J. et al. Monitoring global messenger RNA changes in externally controlled microarray experiments. *EMBO Rep* **4**, 387-93 (2003).
29. Quackenbush, J. Computational analysis of microarray data. *Nat Rev Genet* **2**, 418-27 (2001).
30. Sherlock, G. Analysis of large-scale gene expression data. *Brief Bioinform* **2**, 350-62 (2001).
31. Kapushesky, M. et al. Expression Profiler: next generation--an online platform for analysis of microarray data. *Nucleic Acids Res* **32**, W465-70 (2004).
32. Washburn, M.P., Wolters, D. & Yates, J.R., 3rd. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotechnol* **19**, 242-7 (2001).
33. Zhu, H. et al. Global analysis of protein activities using proteome chips. *Science* **293**, 2101-5 (2001).

34. Marcotte, E.M., Pellegrini, M., Thompson, M.J., Yeates, T.O. & Eisenberg, D. A combined algorithm for genome-wide prediction of protein function. *Nature* **402**, 83-6 (1999).
35. Ge, H., Liu, Z., Church, G.M. & Vidal, M. Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. *Nat Genet* **29**, 482-6 (2001).
36. Jansen, R., Greenbaum, D. & Gerstein, M. Relating whole-genome expression data with protein-protein interactions. *Genome Res* **12**, 37-46 (2002).
37. Kemmeren, P. et al. Protein interaction verification and functional annotation by integrated analysis of genome-scale data. *Mol Cell* **9**, 1133-43 (2002).
38. Schlitt, T. et al. From gene networks to gene function. *Genome Res* **13**, 2568-76 (2003).
39. Ge, H., Walhout, A.J. & Vidal, M. Integrating 'omic' information: a bridge between genomics and systems biology. *Trends Genet* **19**, 551-60 (2003).
40. Pereira-Leal, J.B., Enright, A.J. & Ouzounis, C.A. Detection of functional modules from protein interaction networks. *Proteins* **54**, 49-57 (2004).
41. Parsons, A.B. et al. Integration of chemical-genetic and genetic interaction data links bioactive compounds to cellular target pathways. *Nat Biotechnol* **22**, 62-9 (2004).
42. Bader, J.S., Chaudhuri, A., Rothberg, J.M. & Chant, J. Gaining confidence in high-throughput protein interaction networks. *Nat Biotechnol* **22**, 78-85 (2004).
43. Lee, I., Date, S.V., Adai, A.T. & Marcotte, E.M. A probabilistic functional network of yeast genes. *Science* **306**, 1555-8 (2004).
44. Kemmeren, P., Kockelkorn, T.T., Bijma, T., Donders, R. & Holstege, F.C. Predicting gene function through systematic analysis and quality assessment of high-throughput data. *Bioinformatics* **21**, 1644-52 (2005).

Color Figures

Table 1 | Color Figures Index

Figure	Page Number	Chapter	Original Page Number
Figure 1.2	113	1	12
Figure 1.3	114	1	15
Figure 1.5	114	1	22
Figure 2.1	115	2	33
Figure 2.2	115	2	34
Figure 2.3	116-118	2	35-37
Figure 3.2	119	3	48
Figure 3.3	120	3	50
Figure 3.4	120	3	50
Figure 3.5	121	3	51
Figure 5.1	122	5	69
Figure 6.2	123	6	77
Figure 6.3	123	6	79
Figure 6.4	123	6	80
Figure 6.5	124	6	81
Figure 6.6	124	6	82
Figure 7.2	125	7	94
Figure 7.3	126	7	95
Figure 7.4	127	7	96
Figure 7.5	127	7	96

Figure 1.2 | Diagnostic Plots for Quality Control and Normalization
Several diagnostic plots are shown from poor quality data obtained from swirl zebra fish embryos. Data has been normalized with print-tip lowess using a window-size of f = 0.4. Depicted are spatial plots of $\log_2$ ratio M, before (**A**) and after normalization (**B**), coordinates of each subgrid are shown along the *x* and *y*-axis. Box plots before (**C**) and after normalization (**D**) are shown. Each box represents one subgrid on the array. MA plots of pre (**E**) and post (**D**) normalization data are shown. Lowess lines are fitted through the data points from each subgrid.

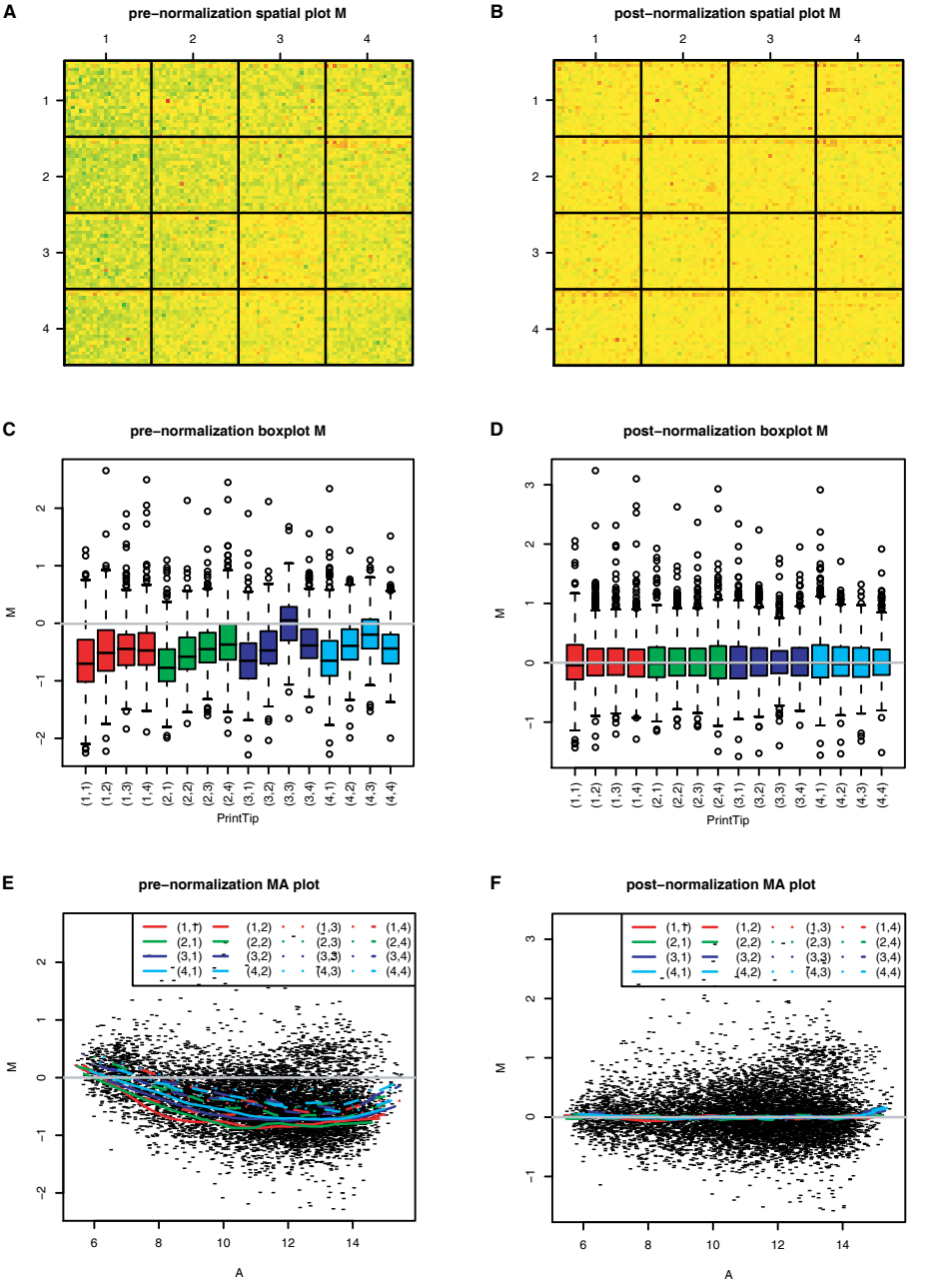
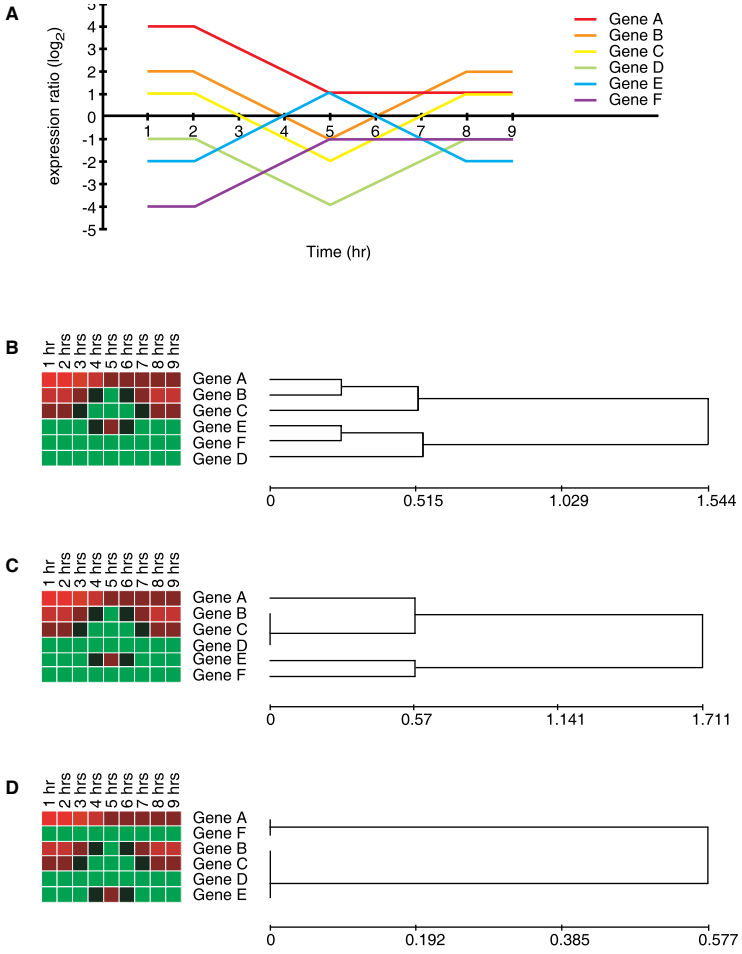


Figure 1.3 | Diverse Distance Measures reveal Different Data Characteristics
Line graph of a hypothetical dataset, consisting of six genes and nine timepoints (**A**). Hierarchical average linkage clustering results using an uncentered correlation distance (**B**), a centered correlation distance (**C**) and an absolute centered correlation distance (**D**).



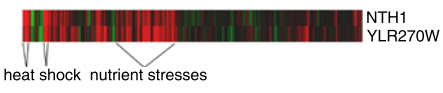


Figure 1.5 | Expression Profiles of NTH1 and YLR270W Obtained from the "conditions" Dataset
The "conditions" dataset consisted of 326 expression profiles from various conditions, such as different stress responses, cell cycle, sporulation, pheromone treatment and unfolded protein degradation. Red indicates upregulation, green indicates downregulation. mRNA expression levels of both genes are up-regulated during heat shock and other forms of stress.

Figure 2.1 | Microarray Data Analysis Pipeline
A schematic overview of the microarray data analysis pipeline is shown. The main user path is depicted with black arrows, starting with the scanning of slides.

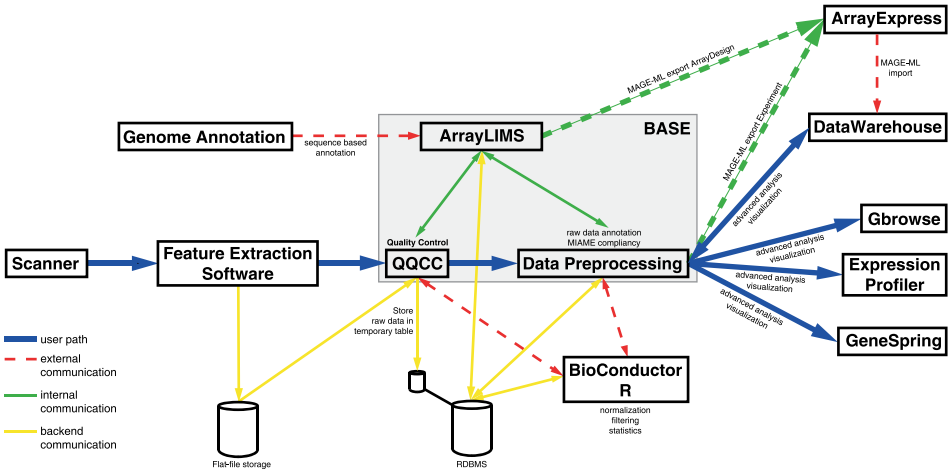
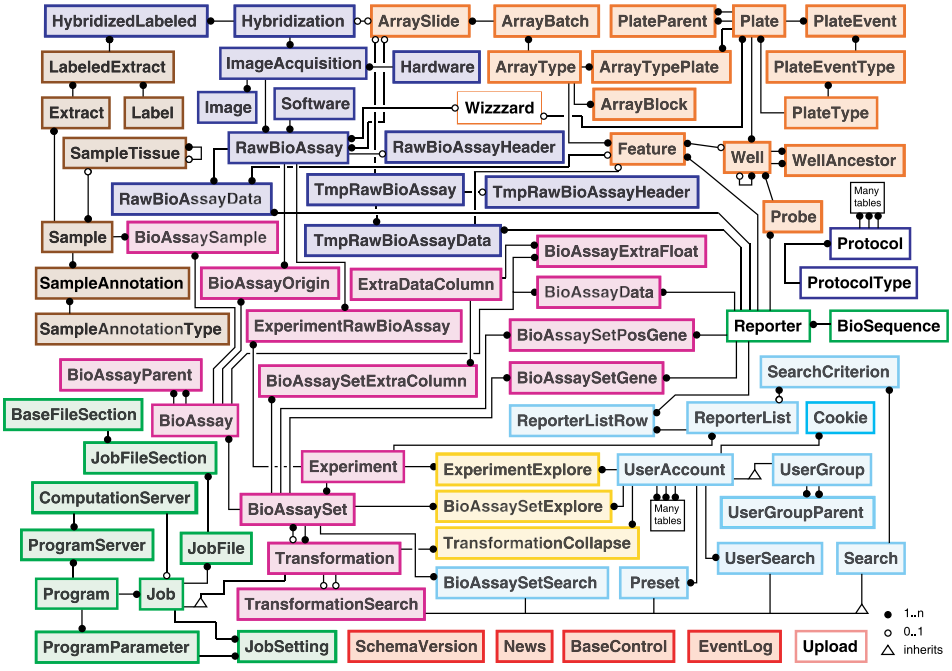


Figure 2.2 | BASE Database Schema
A schematic overview of the full BASE database schema is depicted, adopted from the original version. The database schema can be subdivided in different categories; tables dealing with arrays (orange), hybridizations (purple), samples (brown), data analysis (pink), jobs (green), reporters (green border), protocols (blue border) and users (blue).



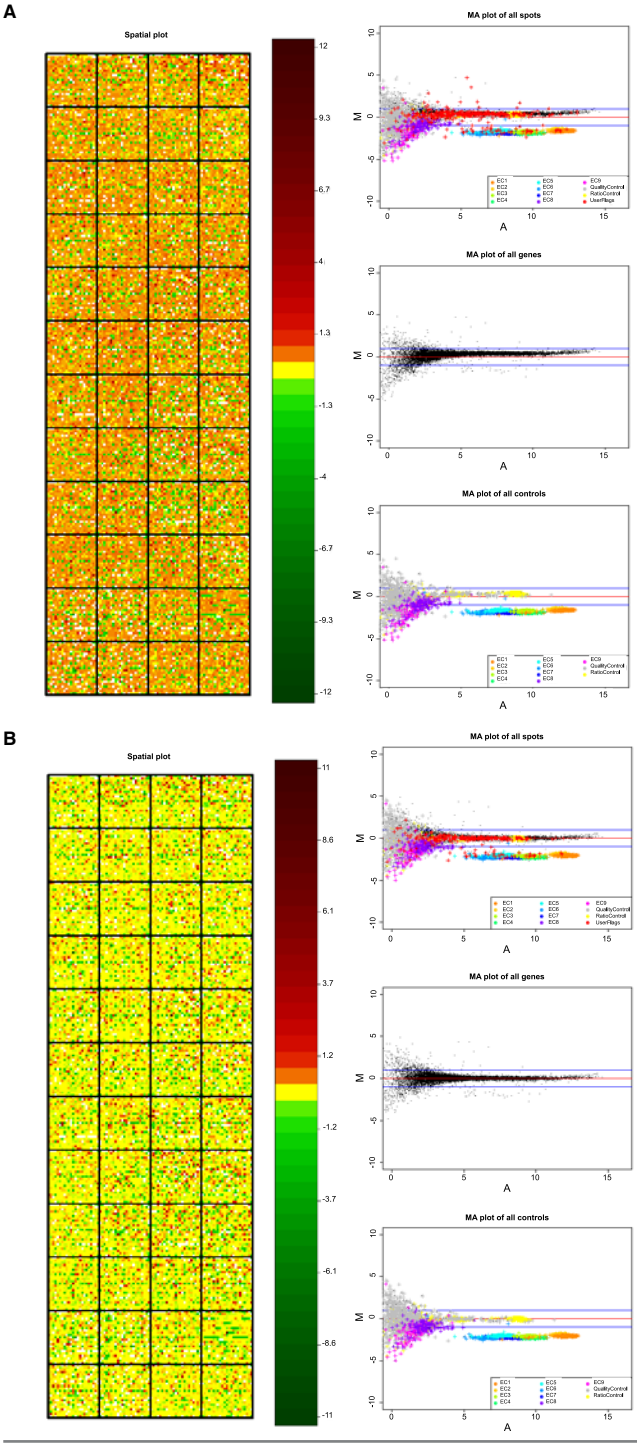
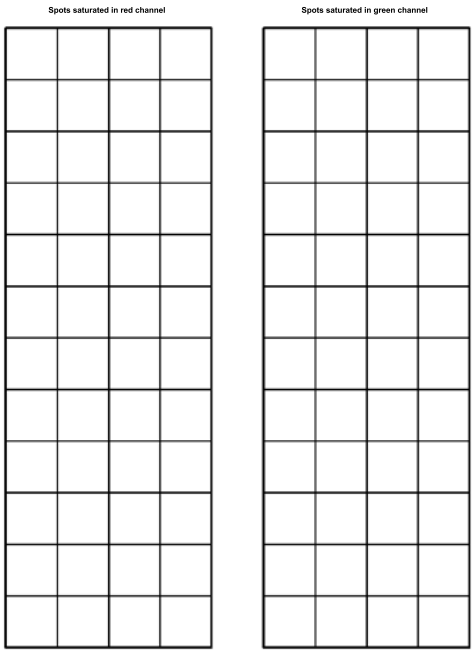


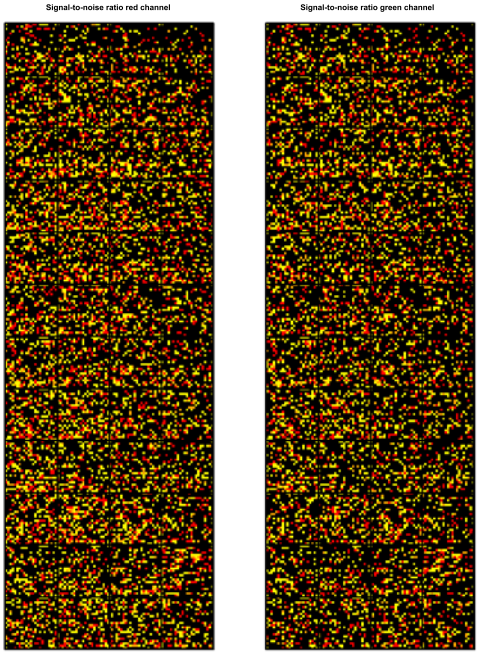
Figure 2.3 | QQCC Quality Report

The quality report depicted in this figure was obtained from a human self-self hybridization with the external normalization controls spiked in at a four fold difference between the two samples. (**A**) Displayed here are the ratios across the slide (spatial plot, left) for the raw data. Three MA plots of the raw data (ratio vs. intensities) are shown on the right: the first plot shows all spots with the controls and user flags highlighted, the second plot shows only the genes and the third plot shows only the controls. (**B**) Same as A, data shown is obtained after lowess normalization on genes. (**C**) Displayed here are the saturated spots (average background corrected signal intensity above 45000). Red and green dots represent saturated spots in the red and green channel respectively. In this case no saturated spots are detected. (**D**) Displayed here is a spatial plot with the raw data binned according to the number of standard deviations above background (SNR): SNR < 2 (black), SNR < 3 (red), SNR < 4 (orange) and SNR >= 4 (yellow). (**E**) Displayed here are the normalized ratios of all genes on their respective chromosome location. (**F**) Displayed here are the percentage of good, marginal and bad spots based on the quality metrics obtained from Chen et al.

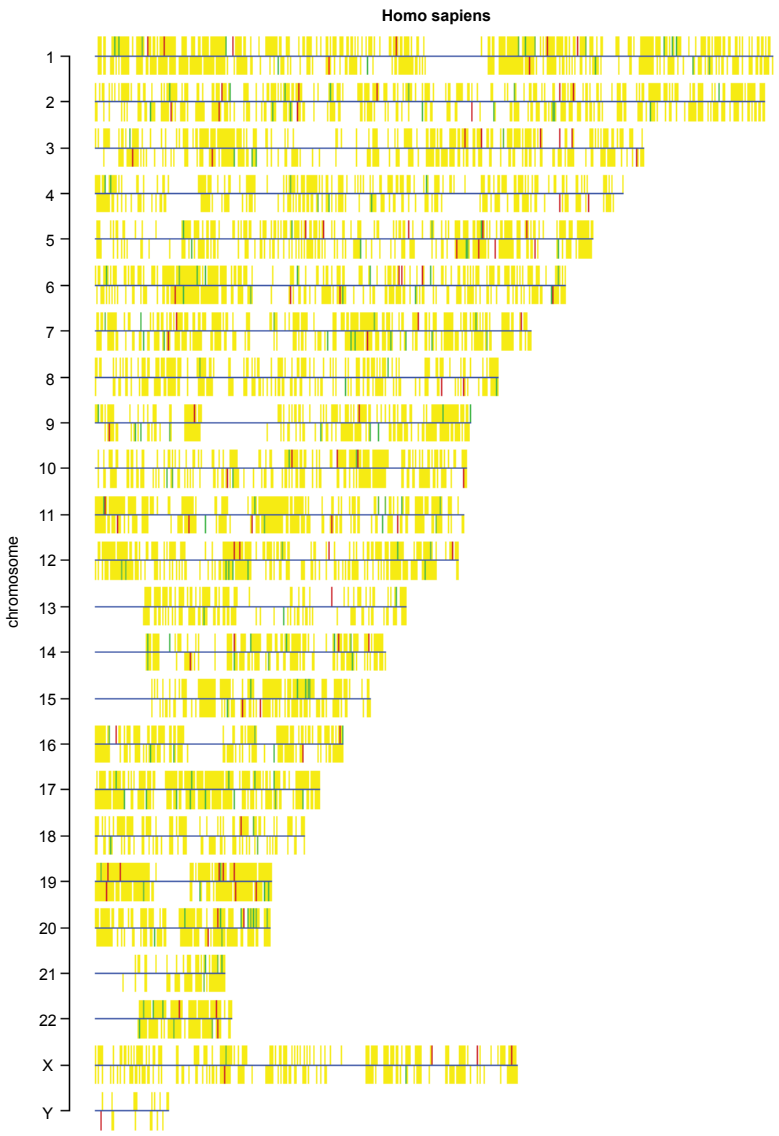
C


Spots saturated in red channel
Spots saturated in green channel

D


Signal-to-noise ratio red channel
Signal-to-noise ratio green channel

E



F

% good spots	94.8
% marginal spots	0.6
% bad spots	4.6
% spots poor Signal-to-Noise ratio	0.5
% spots poor background flatness	1.1
% spots poor Signal Consistency	0.2
% Black Holes	2.8
Banana Factor	1.13

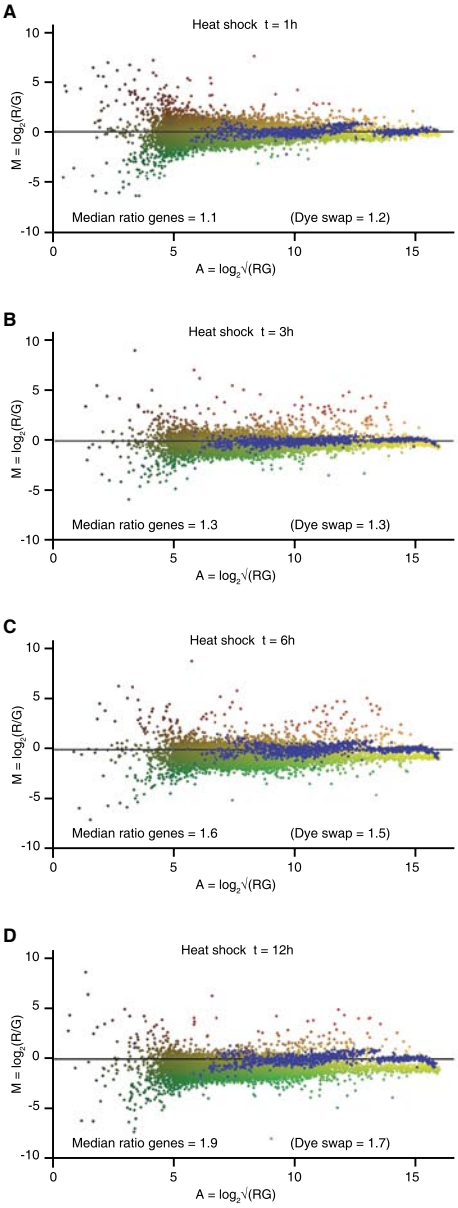


Figure 3.2 | Global Messenger RNA Changes during Mammalian Heat Shock

Human umbilical vein endothelial cells were heat-shocked and RNA was isolated at the timepoints shown (**A–D**). The result of hybridization of each sample (R) against that of the non-heatshocked reference cells (G) is shown. Each graph is an MA scatterplot (where $M = \log_2(R/G)$ and $A = \log_2\sqrt{(RG)}$). The *y*-axis shows the $\log_2$ ratio (mean spot intensity minus mean local area background) after normalization using genes and controls (see Experimental Procedures). The values plotted on the *x*-axes are derived from the intensities of both channels. The median change for all genes is indicated, as well as the result of a corresponding dye-swap experiment. (**E**) Comparison of different normalization strategies. From left to right are the results of three strategies applied to the 12-h timepoint. Each column is made up of 16,735 coloured lines, stacked vertically. Each line represents the change for a single gene.Upregulated genes are shown in red and downregulated genes are shown in green. Numbers below the bars indicate the median change of all genes after normalization (as a ratio) and how many genes are reported as being upregulated or downregulated if an arbitrary twofold cut-off is applied. The first column shows the result of Lowess normalization per subgrid using endogenous genes (see Experimental Procedures). The second column shows the result of incorporating external controls into the normalization strategy by carrying out normalization to equivalent amounts of total RNA (normalization using genes and controls; see Experimental Procedures). This leads to a markedly altered perception of the changes that have taken place, with many more transcripts reported as being downregulated. The third column shows the result when the slight drop in the total RNA content of the cells is taken into account.

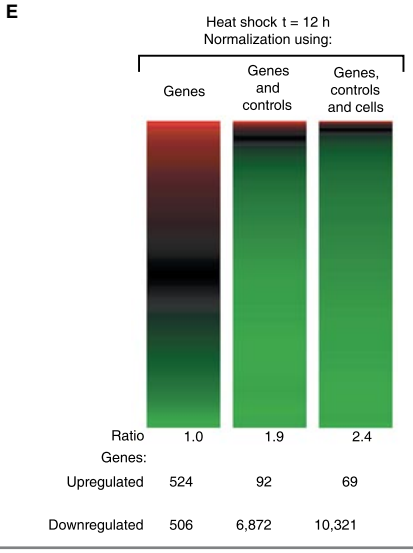


Figure 3.3 | Minor Global Change during Serum Deprivation lead to a Significantly Different Interpretation
(**A**) Human mammary gland adenocarcinoma (MCF7) cells deprived of serum for 30 h (R) compared with the non-deprived culture (G). (**B**) Identical normalization strategies were used, as described in Figure 3.2.

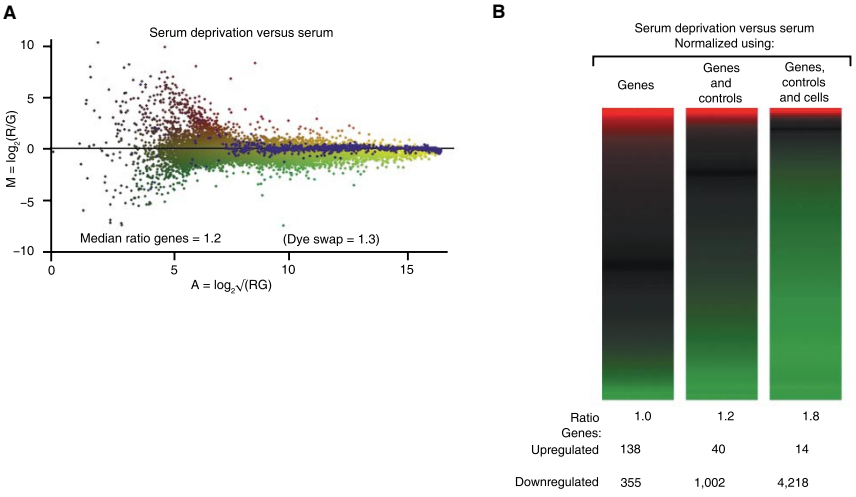
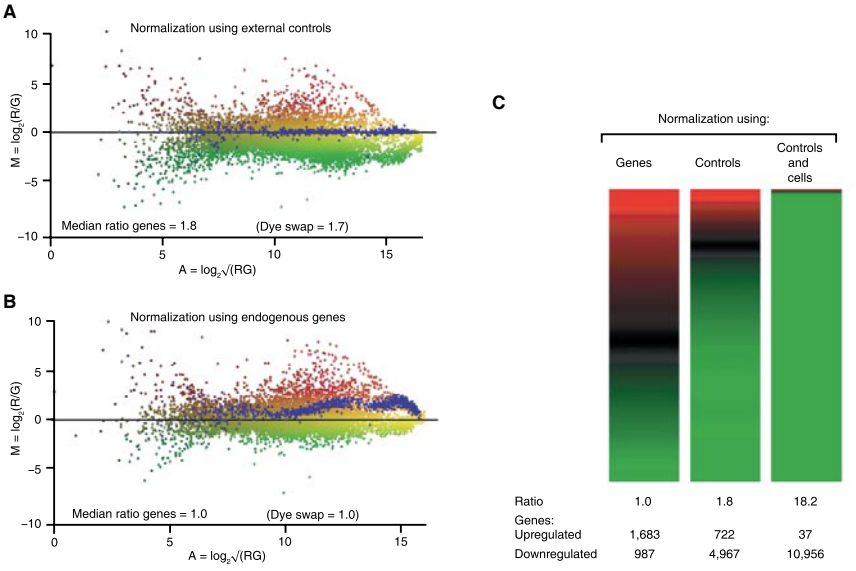


Figure 3.4 | Yeast Stationary Phase Culture compared with Mid-Log Phase Culture
(**A**) MA scatterplot (where M = $\log_2$(R/G) and A = $\log_2\sqrt{}$(RG)) after Lowess normalization for each subgrid using controls (see Experimental Procedures). (**B**) Lowess normalization for each subgrid using genes (see Experimental Procedures). The aberrant pattern of external control spots (blue) occurred because messenger RNA levels had not changed uniformly across the entire range of expression levels (along the *x*-axis). RNA levels for genes with higher expression levels had dropped to a greater degree than those of genes expressed at lower levels. Due to saturation of the signals in the scanned images, this effect seemingly decreases for a small group of genes expressed at the highest levels, resulting in the curve of the control spots to the far right of the graph. (**C**) Comparison of different normalization strategies. The left and middle columns correspond to the graphs shown in (**B**) and (**A**), respectively. The drop in total RNA yields per cell (see Experimental Procedures) can also be taken into account (right column). Because each probe was spotted twice on the yeast arrays, the numbers reflect how many spots have changed. R, stationary phase culture; G, mid-log phase culture.



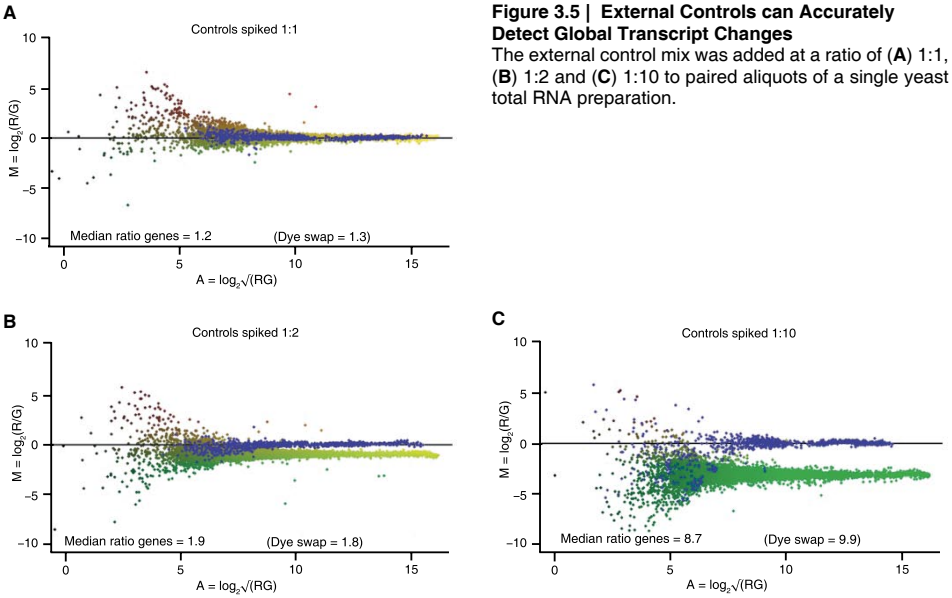
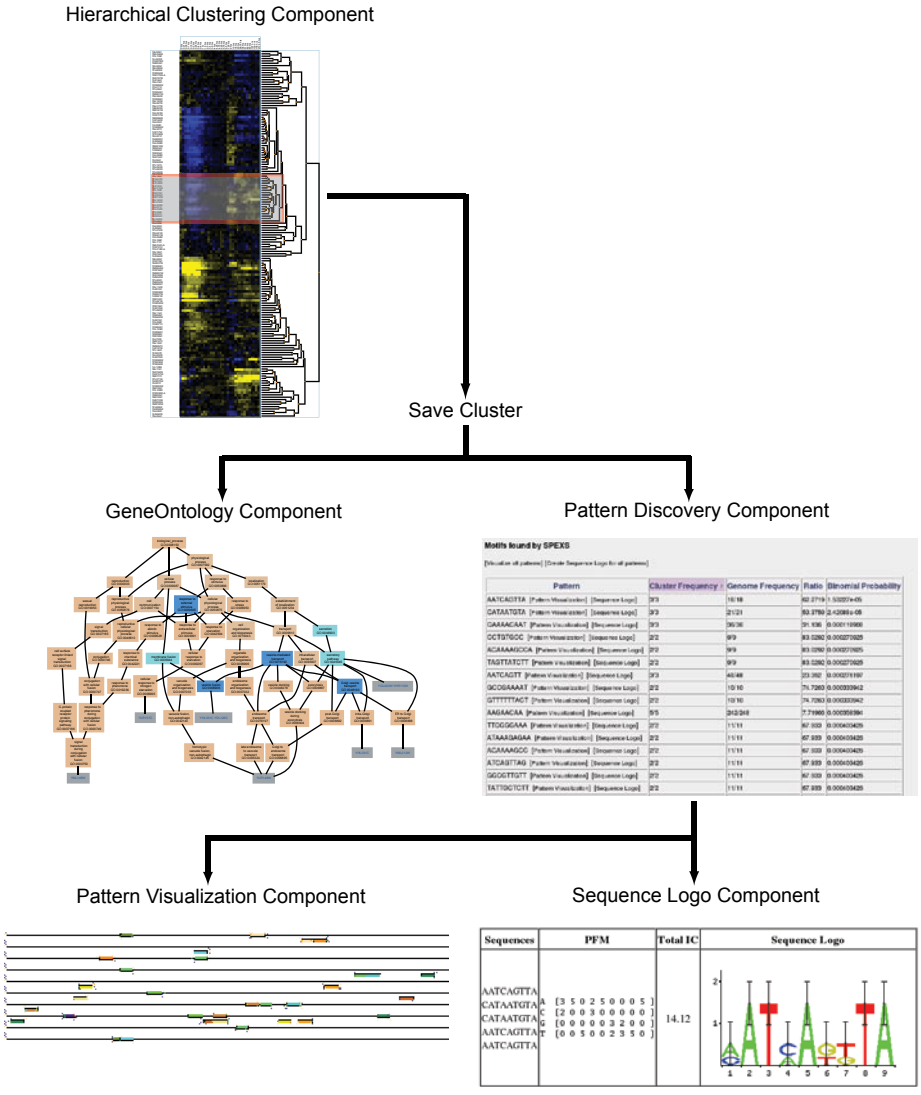


Figure 3.5 | External Controls can Accurately Detect Global Transcript Changes
The external control mix was added at a ratio of (**A**) 1:1, (**B**) 1:2 and (**C**) 1:10 to paired aliquots of a single yeast total RNA preparation.

Figure 5.1 | Uncovering Shared Function and Regulation behind Gene Clusters using EP:NG
A schematic overview of the different components within EP:NG that together can be used to uncover functional and regulatory relationships within gene clusters. The dataset used for this analysis is obtained from Radonjic et al. Only Those genes that show expression levels greater than 1.5 times the standard deviation of the distribution of all genes in at least 20% of the time points are subjected to hierarchical clustering analysis. Part of this hierarchical tree is then saved and analyzed using the GeneOntology and Pattern Discovery Component. Patterns found in the Pattern Discovery Component are then further investigated using the Pattern Visualization or Sequence Logo Component.



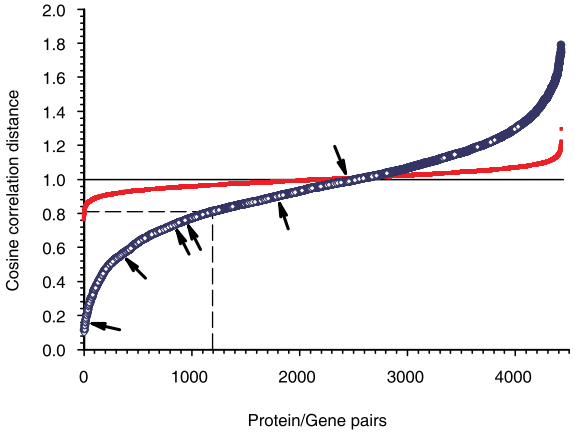


Figure 6.2 | Protein Interactions of Higher Confidence Show Significantly More mRNA Coexpression
The protein interaction data from Ito et al. (blue line, 4425 pairs) was analyzed for mRNA coexpression within the "conditions" expression dataset using the cosine correlation distance only approach. This "conditions" dataset consisted of 326 expression profiles from various environmental and experimental conditions, such as different stress responses, cell cycle, sporulation, pheromone treatment and unfolded protein responses. The degree of coexpression is plotted on the *y*-axis. A value of zero corresponds to complete mRNA coexpression across all the expression data. A value of two corresponds to perfect anti-correlation under all conditions. The protein pairs are ranked according to the degree of coexpression and the rank number is plotted on the *x*-axis. The 170 previously known interactions within the interaction dataset are plotted as open circles in order to demonstrate the marked skewing towards a higher degree of coexpression for these interactions of higher confidence. The red line is the result of an identical analysis using randomized expression data. The dashed lines indicate the significance threshold for catching 0.1% of the interactions with the randomized expression data. This corresponds to a P value of 0.003. The arrows indicate protein interaction pairs that have been selected for follow-up experiments (Figure 6.4-6.6). An additional protein interaction pair was selected from the Uetz data (Figure 6.5C).

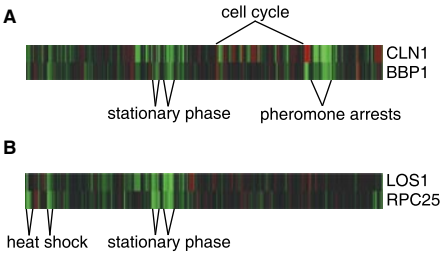


Figure 6.3 | mRNA Coexpression Examples for Putative Interactions between Previously Characterized Proteins
(**A**) Expression profiles of CLN1 and BBP1 obtained from the "conditions" dataset. The cosine correlation distance is 0.38. Red indicates upregulation, green indicates downregulation. Both mRNA expression levels are downregulated during conditions that involve cessation of cell division such as stationary phase and pheromone arrest. BBP1 mRNA levels exhibit a cyclic pattern during cell cycle, but only with high magnitudes of change in half of the cell-cycle experiments analyzed. (**B**) Expression profiles of LOS1 and RPC25 obtained from the "conditions" dataset. The cosine correlation distance is 0.34. Red indicates upregulation, green indicates downregulation. The mRNA expression levels of both genes are downregulated during heat shock and stationary phase.

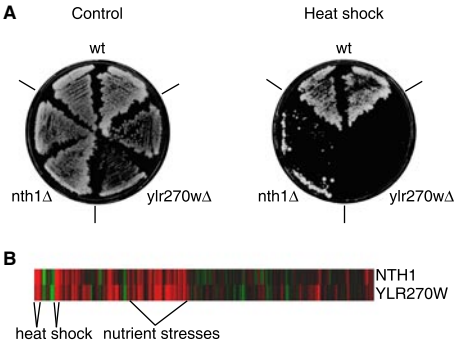


Figure 6.4 | YLR270W is Required for Resistance to Heat Shock
(**A**) Wildtype, *nth1* and *ylr270w* deletion strains grown at 30°C (left), or exposed to 53°C for 7 hrs (right). The mutant strains show little to no recovery 5 days after heat shock when compared to wildtype. (**B**) Expression profiles of NTH1 and YLR270W obtained from the "conditions" dataset. The cosine correlation distance is 0.17. Red indicates upregulation, green indicates downregulation. mRNA expression levels of both genes are upregulated during heat shock and other forms of stress.

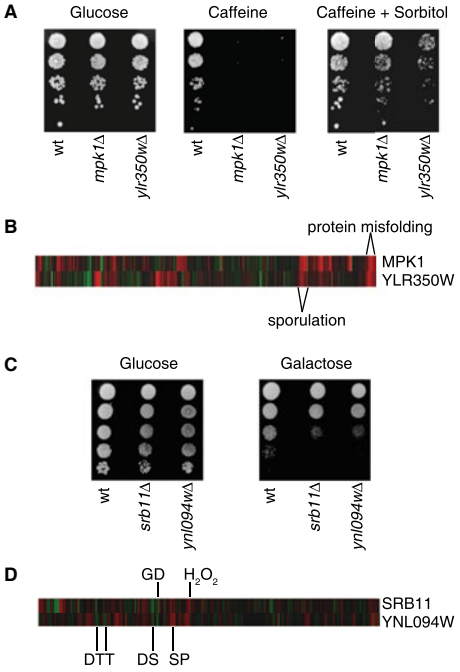


Figure 6.5 | YLR350W is Involved in the Cell Wall Integrity Pathway; YNL094W is Involved in Galactose Utilization
(**A**) Wild-type, *mpk1* and *ylr350w* deletion strains after 3, 7, and 7 days of growth on glucose (left), caffeine (middle), and caffeine plus sorbitol (right), respectively, in a dilution series of 1:5. *mpk1* and *ylr350w* deletion strains show no growth on the caffeine containing medium. Growth is restored when sorbitol is added. (**B**) Expression profiles of MPK1 and YLR350W obtained from the "conditions" dataset. The cosine correlation distance is 0.54. Red indicates upregulation, green indicates downregulation. During sporulation and protein misfolding, mRNA expression levels of both genes are upregulated. (**C**) Wild-type, *srb11* and *ynl094w* deletion strains after 2 days of growth on glucose (left) and galactose (right) respectively, in a dilution series of 1:5. *srb11* and *ynl094w* deletion strains show a reduced growth on galactose, when compared to wildtype. (**D**) Expression profiles of SRB11 and YNL094W obtained from the "conditions" dataset. The cosine correlation distance is 0.63. Red indicates upregulation, green indicates downregulation. Coexpression is observed during the diauxic shift (DS), glucose depletion (GD), stationary phase (SP), two DTT treatments (DTT) and hydrogen peroxide treatment (H_2O_2).

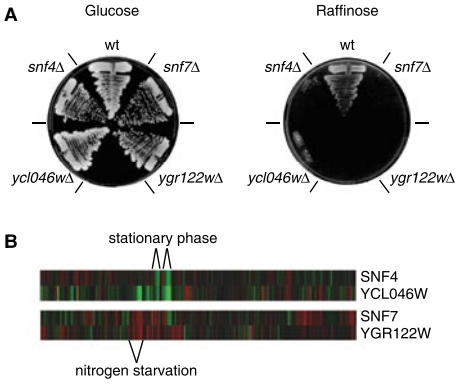


Figure 6.6 | YCL046W and YGR122W are Required for Growth on Raffinose
(**A**) Wildtype, *snf4, ycl046w*, *snf7* and *ygr122w* deletion strains after 13 days growth on glucose (left) and raffinose (right). The mutant strains show little to no growth on raffinose when compared to wildtype. (**B**) Expression profiles of SNF4, YCL046W, SNF7 and YGR122W obtained from the "conditions" dataset. The cosine correlation distance between SNF4 and YCL046W is 0.75, between SNF7 and YGR122W 0.73. Red indicates upregulation, green indicates downregulation. mRNA levels of SNF4 and YCL046W are downregulated during stationary phase, SNF7 and YGR122W are upregulated during nitrogen starvation.

Figure 7.2 | Finding Functional Relationships through Integrated Networks
(**A**) Network obtained when selecting those relationships that are supported by mRNA coexpression and ppi (matrix). Those relationships supported by a known ppi are shown in red. (**B**) Additional requirement of both a similar phenotype and location results in the genes shown in this network. All relationships shown here are supported by at least one ppi (blue lines), those supported by all four data types are drawn in red. Genes are color-coded according to the most significant GO categories.

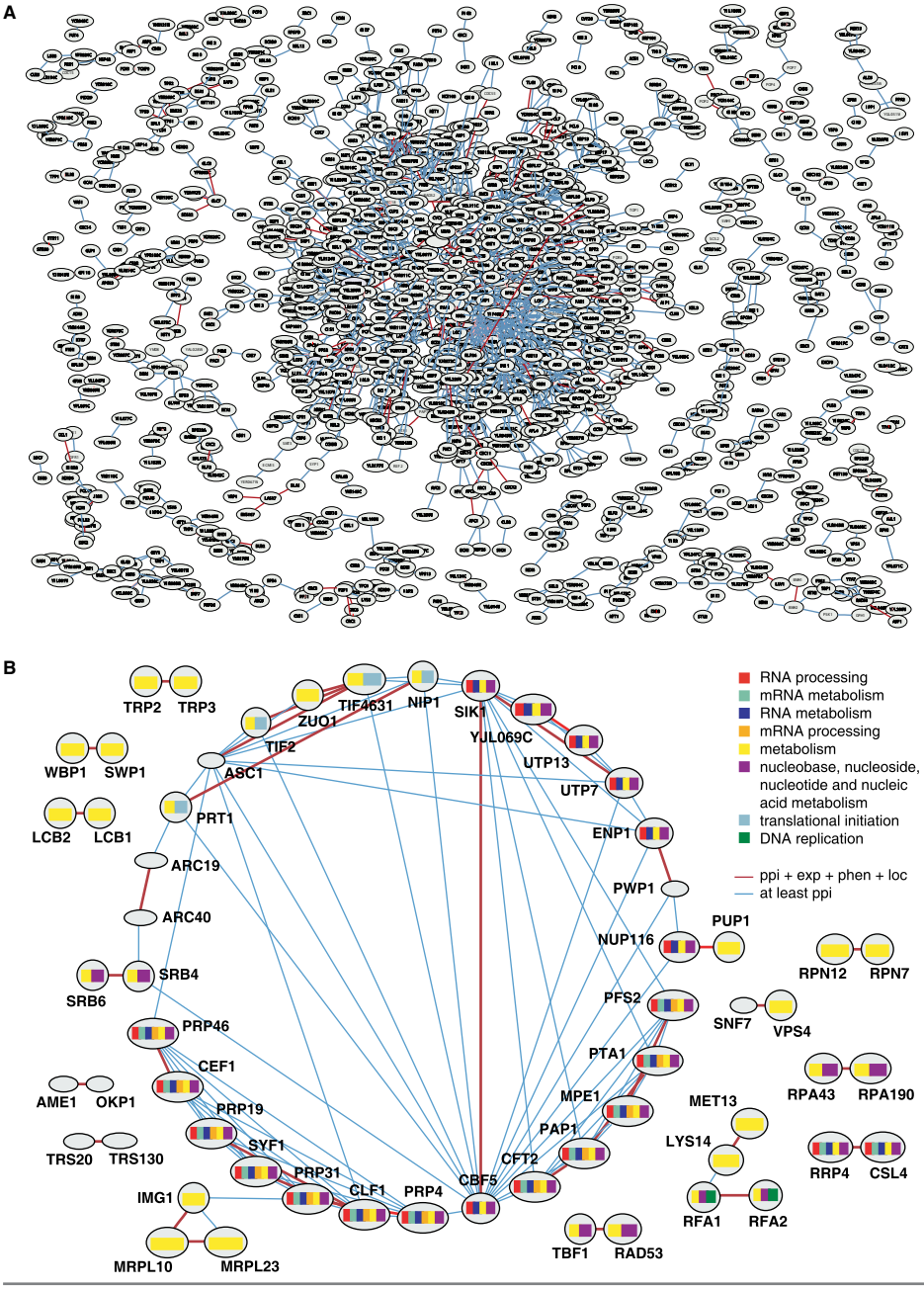
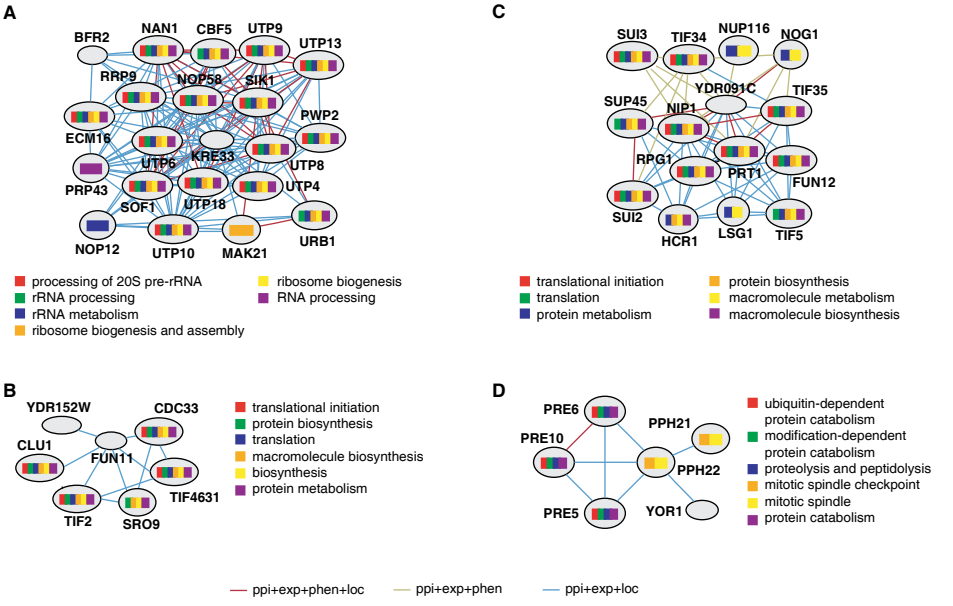


Figure 7.3 | Predicting Function of Uncharacterized Genes
Gene neighborhood of KRE33 (**A**), FUN11 (**B**), YDR091c (**C**) and PPH22 (**D**). The relationships found are supported either by all four data types (red lines); mRNA coexpression, ppi and location (blue lines) or mRNA coexpression, ppi and phenotype (ochre lines). Genes are color-coded according to the most significant GO categories they belong to.



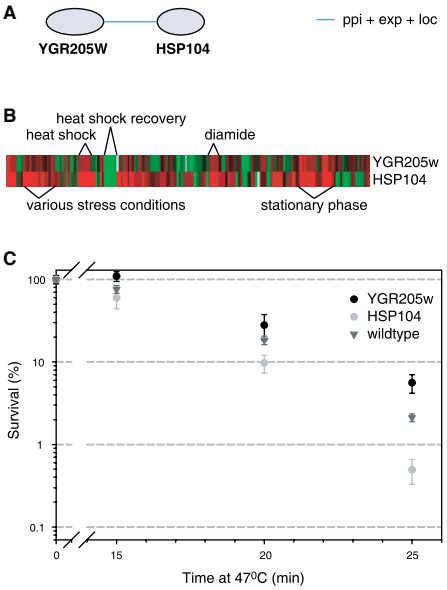


Figure 7.4 | YGR205w Suppresses Thermotolerance
(**A**) Gene neighborhood of YGR205w. The relationship found between YGR205w and HSP104 is supported by mRNA coexpression, ppi and location (blue line). (**B**) Expression profiles of YGR205w and HSP104 obtained from Gasch et al[12]. Red indicates upregulation, green indicates downregulation. Both mRNA expression levels are upregulated during various processes as indicated and down regulated during heat shock recovery. (**C**) Comparison of thermotolerance in the *hsp104* and *ygr205w* deletion strains and wildtype. Cells lacking HSP104 (grey circle) show a decreased resistance to heat shock when compared to wildtype (triangle). Deletion of YGR205w results in an increased thermotolerance (black circle).

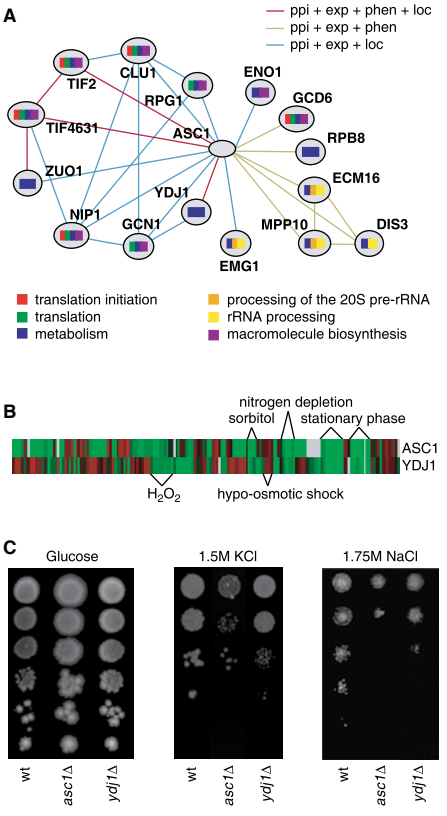


Figure 7.5 | ASC1 and YDJ1 are Associated with Protein Folding, Translation and Ribosome Biogenesis
(**A**) Gene neighborhood of ASC1. Relationships inside this network are supported either by all four data types (red lines), mRNA coexpression, ppi and similar phenotype (ochre line) or mRNA coexpression, ppi and location (blue line). Genes are color-coded according to the most significant GO categories they belong to. (**B**) Expression profiles of YDJ1 and ASC1 obtained from Gasch et al. Red indicates upregulation, green indicates dowregulation. YDJ1 mRNA expression levels are coregulated during various processes as indicated. (**C**) Wildtype, *ydj1* or *asc1* deletion strains grown in YPD (left), 1.5M KCl (middle) or 1.75M NaCl (right). Both *asc1* and *ydj1* deletion strains show a slow growth during high salt concentrations when compared to wildtype.

Summary

Summary

The availability of whole genome sequences has given rise to the parallel development of other high-throughput approaches such as determining mRNA expression level changes, gene-deletion phenotypes, chromosomal location of DNA binding proteins, cellular location of proteins, protein-protein interactions, biological activity using protein arrays, synthetic lethality with combinations of mutants and RNA interference phenotypes. Such functional genomic datasets can be used in a variety of ways. For example, microarray gene expression data has been applied to predict the function of genes using mutant expression-profiling "phenotypes", to identify possible drug targets by exposing cells to certain pharmacological agents, to improve cancer diagnosis and prognosis and to decipher complex biological systems such as transcriptional regulatory networks.

Many challenges are encountered when analyzing such data for biological or biomedical purposes. These include data management, data preprocessing, data normalization, differences in data quality within and between datasets, integration of different types of genome-scale data, prioritization of hypotheses and improvement of existing annotation. The goal of the work presented in this thesis is to develop new and improve existing methodologies for analyzing genome-scale data, so that the overall efficiency, accuracy, precision and biological significance of functional genomic data increase.

Chapter 1 is a general introduction providing an overview of currently existing genome-scale data types. Special emphasis is placed on microarray expression data, by describing the specifics of data preprocessing, meta-analyses and applications of microarray expression data. The rationale is then broadened to include other types of genome-scale data, posing the question of how integrating those with microarray expression data might overcome some of the issues involved in analyzing genome-scale data.

To facilitate microarray data analysis, platforms are required that can function as a central storage and management system for these types of data. In **chapter 2**, a general framework is presented that supports many different aspects of microarray data management, such as array production, array annotation, quality control, sample annotation and creation of standardized workflows. Through the microarray data analysis framework, quality control and MIAME (minimal information about a microarray experiment) compliant annotation is enforced, thus achieving a higher overall accuracy and utility.

One of the most crucial microarray data preprocessing steps, pivotal for determining the outcome, is data normalization. Most microarray analyses assume that relatively few genes vary in mRNA expression, or that any changes that do occur are balanced. In **chapter 3**, a strategy is devised that uses external RNA controls in the normalization method, thereby monitoring global changes in mRNA expression. By implementing these external calibration standards in the normalization strategy it is shown that the levels of most mRNAs change during yeast stationary phase or mammalian heat shock. Furthermore, even minor changes in global mRNA expression levels have a large impact on the number of genes determined to have significantly changed mRNA expression, showing that monitoring such effects is important for accurate determination of gene expression changes.

Many different analysis routes can be taken after initial data normalization and filtering. For this purpose versatile and extensible tools need to be developed. In **chapter 4**, a tool is presented that provides a web-based platform for microarray gene expression analysis. A modular architecture is described, allowing individual components to be developed by different groups, while still sharing the same unified interface and core infrastructure. Included in this setup are components for data preprocessing, missing value imputation, filtering, clustering methods, visualization, significant gene finding, between group analysis and other statistical components. Through the open-source nature of the project future extensions from the scientific community and broad distribution are encouraged.

Discerning regulatory mechanisms behind biological processes is important to understanding how these processes work. In **chapter 5**, tools are presented that

allow this type of analysis. Four new components have been seamlessly integrated into the framework presented in chapter 4; Gene Ontology, Pattern Discovery, Pattern Visualization and Sequence Logo components. Starting from a gene cluster or group of co-expressed genes, the biological relevance of these genes can be determined using the Gene Ontology. Once the biological relevance has been established, searching for over-represented patterns in the upstream sequences and displaying them according to their upstream location or information content can uncover possible regulatory mechanisms. This allows easy acces to the possible regulatory mechanisms behind certain biological processes for many different species.

Other genome-scale datasets also present challenges with regard to accuracy and gene function discovery. In **chapter 6**, microarray expression datasets are used to assess the quality of high-throughput protein interaction datasets. Determining which of the pairs of putatively interacting proteins show significant mRNA co-regulation significantly increases confidence for 973 out of 5432 putative high-throughput two-hybrid interactions from *Saccharomyces cerevisiae*. Moreover, functional annotation for over 300 previously uncharacterized genes is provided through the integration of mRNA expression and protein interaction data. The robustness of the approach presented here is further demonstrated by experiments that test the *in silico* predictions made. This study shows how integration of microarray expression and protein interaction data improves the overall applicability of different types of functional genomic data and how this can contribute to gene function annotation.

Besides integrating microarray expression and protein interaction data, other types of genome-scale data can be included to improve overall accuracy and precision. In **chapter 7**, a new approach is presented for integration of different types of high-throughput data. Functional predictions are made based on relationships supported by multiple types and sources of data. For this purpose a general framework has been devised, containing 125 different high-throughput datasets describing phenotypes, cellular location, protein interactions and mRNA expression levels from *S. cerevisiae*. A bit-vector representation and information content-based ranking is used which takes into account the characteristic and qualitative differences and is extremely flexible, efficient and scalable. Predictions for 543 uncharacterized genes are made, based on multiple relationships each supported by at least three types of experimental data. The reliability of the results is further demonstrated through experimental validation for some of the predictions made. Furthermore, the results generate insights into the relative merits of different data types and provide a coherent framework for functional genomic datamining.

Chapter 8 is a general discussion of the results of the work described in this thesis.

Samenvatting

Samenvatting

Het erfelijke materiaal van iedere cel in een organisme is opgeslagen in DNA in de vorm van genen en gezamenlijk het genoom vormen. De informatie opgeslagen in het DNA wordt vertaald naar RNA en vervolgens naar eiwitten, die dan complexe biologische systemen vormen. De beschikbaarheid van complete genoomsequenties heeft geleidt tot de parallelle ontwikkeling van andere grootschalige benaderingen zoals het bepalen van mRNA expressie veranderingen, gendeletie fenotypes, chromosomale locatie van DNA bindende eiwitten, cellulaire locatie van eiwitten, eiwit-eiwit interacties, biologische activiteit d.m.v. eiwitarrays, synthetische dodelijkheid door combinaties van gendeleties en RNA interferentie fenotypes. Dergelijke collecties van genomicagegevens kunnen op diverse manieren worden gebruikt. Zo zijn microarray expressieprofielen van gendeleties toegepast om de functie van genen te voorspellen en kunnen mogelijke biologische aangrijpingspunten van medicijnen worden geïdentificeerd door cellen bloot te stellen aan bepaalde farmacologische stoffen. Verder worden deze gegevens gebruikt om kankerdiagnostiek te verbeteren en complexe biologische systemen zoals regulatoire transcriptie netwerken te ontcijferen. Tijdens het analyseren van dergelijke gegevens voor biologische of biomedische doeleinden komt men vele uitdagingen tegen. Hieronder vallen o.a. gegevensbeheer, gegevensvoorbewerking, gegevensnormalisatie, verschillen in gegevenskwaliteit binnen en tussen gegevenssets, integratie van verschillende types van genoom-brede gegevens, rangschikking van hypothesen en verbetering van bestaande annotatie. Het doel van het werk dat in dit proefschrift wordt beschreven is de ontwikkeling van nieuwe en verbetering van bestaande benaderingen voor het analyseren van genoom-brede gegevens, zodat de algemene efficiëntie, betrouwbaarheid, precisie en biologische betekenis van genomicagegevens toeneemt.

Hoofdstuk 1 is een algemene inleiding dat een overzicht geeft van de huidige genoom-brede gegevenstypes. Speciale nadruk wordt gelegd op microarray expressiegegevens, door de details van gegevensvoorbewerking, meta-analyses en toepassingen van microarray expressiegegevens te beschrijven. Het onderwerp wordt dan verbreed naar andere soorten genoom-brede gegevens, waarbij besproken wordt hoe de integratie van deze genoom-brede gegevens en microarray expressiegegevens een aantal problemen kan overwinnen die betrokken zijn bij het analyseren van genoom-brede gegevens.

Om gegevensanalyse van microarrays te vergemakkelijken, is er een platform nodig dat als centraal opslag- en beheerssysteem kan dienen voor dit soort gegevens. In **hoofdstuk 2** wordt een algemeen systeem gepresenteerd dat diverse aspecten van microarray gegevensbeheer ondersteunt, waaronder arrayproductie, arrayannotatie, kwaliteitsbeheersing, monsterannotatie en verwezenlijking van gestandaardiseerde werkschema's. Door het microarray gegevensanalyse systeem worden de kwaliteitsbeheersing en MIAME (minimale informatie over een microarray experiment) overeenkomstige annotatie gewaarborgd, waardoor een betere algemene nauwkeurigheid en bruikbaarheid wordt bereikt. Een van de belangrijkste microarray gegevensvoorbewerkingsstappen is gegevensnormalisatie. De meeste microarray analyses veronderstellen dat vrij weinig genen in mRNA expressie variëren of dat veranderingen die plaatsvinden gebalanceerd zijn.

In **hoofdstuk 3** wordt een strategie ontwikkeld die externe RNA controles in de normalisatiemethode gebruikt, waardoor globale veranderingen in mRNA expressie gedetecteerd kunnen worden. Door deze externe kalibratiestandaarden in de normalisatiestrategie toe te passen, wordt aangetoond dat de niveaus van de meeste mRNAs veranderen tijdens stationaire fase van gist of hitteschok behandeling in humaan materiaal. Tevens hebben zelfs minimale veranderingen in globale mRNA expressieniveaus een grote invloed op het aantal genen dat significant verandert. De controle van dergelijke globale veranderingen is dus belangrijk voor een nauwkeurige bepaling van de genexpressie niveaus.

Na de aanvankelijke normalisatie en filtering kunnen vele verschillende analyseroutes

genomen worden. Hiervoor moeten er veelzijdige en flexibele hulpmiddelen worden ontwikkeld. In **hoofdstuk 4** wordt een internetgebaseerd platform voor microarray genexpressie analyse gepresenteerd. Een modulaire architectuur wordt beschreven waardoor individuele componenten door verschillende groepen kunnen worden ontwikkeld,met behoud van dezelfde interface en basisinfrastructuur. Deze opzet bevat componenten voor gegevensvoorbewerking, missende waardeimputatie, filtering, visualisatie, het vinden van significante genen, “between-group analysis” en andere statistische bewerkingen. Door de open-bron aard van het project worden toekomstige uitbreidingen vanuit de wetenschappelijke gemeenschap en een brede verspreiding aangemoedigd.

Het onderscheiden van regulatoire mechanismen achter biologische processen is belangrijk voor het begrip van deze processen. In **hoofdstuk 5** worden hulpmiddelen gepresenteerd die dit type van analyse vergemakkelijken. Vier nieuwe componenten zijn naadloos in het systeem geïntegreerd dat in hoofdstuk 4 wordt beschreven: genontologie, patroonherkenning, patroonvisualisatie en sequentielogos. Uitgaande van een gencluster of een groep van gezamenlijk tot expressie komende genen, kan de biologische relevantie van deze genen worden bepaald door gebruik te maken van de genontologie. Zodra de biologische relevantie is bevestigd, kan het zoeken naar sterk oververtegenwoordigde patronen in de opstroomse sequentie en het vervolgens tonen van de locatie of informatie-inhoud van deze patronen mogelijke regulatoire mechanismen aan het licht brengen. Hierdoor kan makkelijk gekeken worden naar mogelijke regulatoire mechanismen achter bepaalde biologische processen voor diverse organismen.

Andere genoom-brede gegevenssets brengen uitdagingen met betrekking tot nauwkeurigheid en genfunctie exploratie met zich mee. In **hoofdstuk 6** worden microarray expressiegegevens gebruikt om de kwaliteit van grootschalige eiwitinteractiegegevens te beoordelen. Door te kijken welke eiwitparen uit grootschalige “yeast two-hybrid” gegevenscollecties significante co-expressie vertonen, kan het vertrouwen in 973 van de 5432 vermeende eiwitinteracties verhoogd worden. Tevens wordt de functionele annotatie voor meer dan 300 niet eerder geannoteerde genen gegeven door de integratie van expressie- en eiwitinteractiegegevens. De robuustheid van de hier voorgestelde benadering wordt verder aangetoond door experimenten die de gemaakte *in silico* voorspellingen verder testen. Deze studie toont aan hoe de integratie van microarray expressie- en eiwitinteractiegegevens de algemene bruikbaarheid van verschillende types van functionele genomicagegevens verbetert en hoe dit tot een verbeterde annotatie van de genfunctie kan bijdragen.

Naast het integreren van microarray expressie- en eiwitinteractiegegevens, kunnen andere types van genoom-brede gegevens worden geïntegreerd, zodat de algemene nauwkeurigheid en precisie wordt verbeterd. In **hoofdstuk 7** wordt een nieuwe benadering gepresenteerd voor de integratie van verschillende types grootschalige gegevens. De functionele voorspellingen die worden gedaan zijn gebaseerd op meerdere types en bronnen van gegevens. Voor dit doel is een algemeen systeem bedacht dat 125 verschillende grootschalige collecties van gegevens bevat, bestaande uit fenotypes, cellulaire locaties, eiwitinteracties en mRNA expressieniveaus in gist. Hiervoor worden een bit-vector representatie en rangschikking gebaseerd op informatieinhoud gebruikt, rekening houdend met karakteristieke en kwalitatieve verschillen. Verder is deze benadering uiterst flexibel, efficiënt en schaalbaar. De functie van 543 niet-geannoteerde genen wordt voorspeld, gebaseerd op meerdere relaties, elk ondersteund door ten minste drie verschillende types experimentele gegevens. De betrouwbaarheid van de resultaten wordt verder aangetoond door experimentele bevestiging van enkele van de gedane voorspellingen. Verder geven de resultaten inzicht in de relatieve verdiensten van de verschillende gegevenstypes en vormt deze benadering een coherente basis voor het exploreren van genomicagegevens.

Hoofdstuk 8 is een algemene discussie van de resultaten van het werk dat in dit proefschrift wordt beschreven.

Dankwoord

Dankwoord

Eindelijk is het dan zover, het is af! Zonder de steun en interesse van al mijn collega's bij de Divisie Biomedische Genetica, vele vrienden en familie had dit proefschrift nooit tot stand kunnen komen. In het bijzonder wil ik op deze manier de volgende mensen bedanken:

Frank, ik heb ontzettende bewondering voor de manier waarop jij in zes jaar tijd een groep hebt opgezet met een reputatie en publicatielijst waar een gemiddeld lab alleen maar van kan dromen. Jouw passie voor de wetenschap, kritische blik en oog voor detail zijn in de loop der jaren altijd een inspiratiebron geweest. Tevens wil ik je bedanken voor de mogelijkheden en vrijheden die je mij geboden hebt, zodat ik me zowel op ICT gebied als binnen het onderzoek heb kunnen ontplooien. Het was niet altijd even gemakkelijk om hier een goede balans tussen te vinden, maar ik denk dat we daar op een goede manier uitgekomen zijn.

Peter, onze terloopse gesprekken in de wandelgangen, o.a. over wetenschap, maar ook over vele andere zaken, en het vertrouwen dat je Frank en mij gegeven hebt om een aantal zaken op te zetten, waardeer ik enorm. Mede hierdoor zijn veel dingen voorspoedig verlopen... Tijd voor Genomics fase 3?

Alvis, when I first visited your group in 2000, I never could have imagined that that visit would lead to a very fruitful collaboration for over 5 years now. Your thoughts and open-minded approach to so many scientific problems have always been very useful and have led to many interesting projects. Jaak, although our initial plan to work on sequence pattern discovery in Alvis's group didn't work out, our Molecular Cell paper isn't too bad as an alternative ;-). I have always enjoyed working with you and I am glad you've shown me how to use (and modify) Expression Profiler, as this has helped me out in many circumstances. Misha, together with Jaak, I think we have setup a very nice tool for microarray expression data analysis. Refactoring Expression profiler has been a lot of fun and I hope that we can continue to work on this project for many years to come. Helen and Ele, I hope I didn't give you too many nightmares with the always changing MAGE-ML submissions…

Thomas, when I came to Cambridge, you immediately made me feel welcome and showed me all the good places in Cambridge. That has been a life-saver for me. Initially, our shared interests were purely science related (besides the complaining about British people, food and weather), but gradually we have become very good friends. Hopefully you can convince Christine to come to Utrecht, so that I can show you all the nice places over here as well.

Theo a.k.a. "StormTrooper", Nynke en Jeroen, jullie bijdrage aan een aantal van mijn hoofdstukken is van onschatbare waarde. De "natte lab" proeven van jullie hebben er altijd voor gezorgd dat ik de link met de biologie niet verloor en herinnerden mij er ook aan dat niet alles op te lossen is met een computer. Theo, afgezien van de prettige samenwerking zal ik ook zeker jouw altijd goedgemutste humeur, grappen en grollen missen (en de hilarische [KW|Q][ee|ua][k|ke] sessies met Harm). Veel succes daar in het hoge noorden en hopelijk tot gauw! Nynus, op naar het volgende project … Toronto here we come! Of nee, wordt het toch San Francisco? Anyway, ik ben ervan overtuigd dat we een hele leuke tijd tegemoet gaan en dat we weer mooie resultaten uit ons volgende project gaan halen.

En dan de bioinformatici. Thessa, tijdens jouw stage heb jij een belangrijke voorzet gegeven voor hoofdstuk 7. Daarnaast heb je een aantal zaken op poten gezet (o.a. QQCC) die nu nog steeds op het lab gebruikt worden en waar je met recht trots op mag zijn. Helaas kon je daarna niet echt wennen als bioinformaticus (of was het omdat wij je te veel plaagden met je lengte), en ben je verder gegaan met je andere creatieve(re) talenten. Veel succes met je eigen ontwerp bedrijfje. Philip, onze discussies over soms triviale zaken zijn ontelbaar en altijd zeer waardevol geweest. Het enige waar ik spijt van heb, is dat we die weddenschap over de proefschriften nooit afgesloten hebben… Harm, mede-bioinformaticus/ICT'er van het eerste uur, jouw kritische blik op zaken heeft meer dan eens geleidt tot nieuwe wetenschappelijke inzichten. De enige vraag die ons altijd

weer parten speelde was “wie voert het uit…”. Erik, jouw hulp bij de ontwikkeling van OffBase is nu al onmisbaar geworden, over twee weken af?

Wim, jouw manier van gebruikersvriendelijk werken heeft mij altijd ontzettend aangesproken en ik heb jouw input op ICT gebied dan ook altijd erg gewaardeerd. Tevens wil ik je bedanken voor al die keren dat we weer eens een keer koffie gingen drinken, zodat ik mijn frustraties op ICT gebied bij jou kwijt kon. Aangezien we vast nog wel tegen wat problemen aan zullen lopen, kijk ik alweer uit naar onze volgende bak koffie! Arnaud en Yumas, zonder jullie was dit proefschrift er simpelweg niet gekomen. Doordat jullie ondertussen veel van de dagelijkse ICT zaken regelen, heb ik de afgelopen maanden de tijd gehad om dit geheel af te ronden, daarvoor mijn grote dank.

De rest van het Genomics Lab wil ik ook graag bedanken. Mirjam en Lilian, bedankt voor het organiseren van Frank zijn leven (en daarmee dat van het hele lab). Dik, jouw precieze manier van werken en documenteren hebben mij zeer veel werk bespaard bij het opzetten van vele labgerelateerde zaken, zoals het LIMS en BASE. Marian, Tony en Joop, bedankt voor alle input vanuit de faciliteit en alle bugs die jullie uit mijn software gehaald hebben. Marijana and Jean-Christophe, thanks for all the discussions and fun we had (particularly for showing me what “proper” partying means). Paul, eindelijk iemand op het lab die Expression Profiler gebruikt. En je weet het hè, als er een schaap over de dam is… Nienke, bedankt voor het regelen van de vrijdagmiddag koekjes, dat is altijd weer een welkome afwisseling (oh, en de appeltaart komt er nog aan …).

Mijn paranimfen wil ik bij deze ook bedanken. Ray, broertje, ook al maakten we vroeger als kind nog wel eens ruzie om allerlei triviale zaken (o.a. om dominostenen die jij omtrapte…), ik heb altijd op je kunnen vertrouwen op die momenten dat het erop aan kwam. Dat jij mij op deze manier symbolisch terzijde wilt staan, waardeer ik dan ook zeer. Bas, jouwe positieve levensinstelling waardeer ik enorm (en het feit dat jij nog meer dan ik verzot op “gadgets” bent), weer een gezamenlijk avontuur!

Pap en mam, jullie hebben mij gemaakt tot wie ik ben. Door jullie onvoorwaardelijke liefde en vertrouwen in mij, heb ik altijd op jullie kunnen terugvallen, ook als ik het zelf even niet meer zag zitten tijdens mijn studie. Na jaren “studeren” is dit dan het uiteindelijke resultaat!

Mirjam, samen met Lucas vormen jullie het middelpunt van mijn leven. Zonder jouw steun en toeverlaat, was ik dan ook nooit zover gekomen. Bedankt voor alles, maar vooral voor het begrip als ik weer eens tot diep in de nacht bezig was en dat je altijd voor me klaar stond wanneer dat nodig was.

Patrick

Curriculum Vitae

Patrick Kemmeren werd geboren op 9 september 1976 te Waalwijk. Het VWO diploma werd in 1994 behaald aan het Dr. Mollercollege te Waalwijk. Daaropvolgend begon hij in september 1994 met de studie Medische Biologie aan de Unversiteit Utrecht waarvan het propaedeuse werd behaald in 1996. Tijdens het doctoraal werden twee stages gelopen. De hoofdvakstage werd verricht bij de vakgroep Experimentele Cardiologie van het Universitair Medisch Centrum Utrecht. De bijvakstage werd verricht bij het Centre for Molecular and Biomolecular Informatics (CMBI) van de Universiteit Nijmegen. Het doctoraalexamen Medische Biologie werd behaald in 1999. Daarna begon hij in 1999 bij de vakgroep Fysiologische Chemie van het Universitair Medisch Centrum Utrecht met het onderzoek dat uiteindelijk geresulteerd heeft tot dit proefschrift.

Publications

GeneSeeker: extraction and integration of human disease-related information from web-based genetic databases.
van Driel MA, Cuelenaere K, **Kemmeren P**, Leunissen JA, Brunner HG, Vriend G.
Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W758-61.

Genome-wide analyses reveal RNA polymerase II located upstream of genes poised for rapid response upon *S. cerevisiae* stationary phase exit.
Radonjic M, Andrau JC, Lijnzaad P, **Kemmeren P**, Kockelkorn TTJP, van Leenen D, van Berkum NL, Holstege FCP.
Mol Cell. 2005 Apr 15;18(2):171-83.

Nitric oxide-dependent and nitric oxide-independent transcriptional responses to high shear stress in endothelial cells.
Braam B, de Roos R, Bluyssen H, **Kemmeren P**, Holstege FCP, Joles JA, Koomans H.
Hypertension. 2005 Apr;45(4):672-80. Epub 2005 Feb 7.

An expression profile for diagnosis of lymph node metastases from primary head and neck squamous cell carcinomas.
Roepman P, Wessels LF, Kettelarij N, **Kemmeren P**, Miles AJ, Lijnzaad P, Tilanus MG, Koole R, Hordijk GJ, van der Vliet PC, Reinders MJ, Slootweg PJ, Holstege FCP.
Nat Genet. 2005 Feb;37(2):182-6. Epub 2005 Jan 9.

Predicting gene function through systematic analysis and quality assessment of high-throughput data.
Kemmeren P, Kockelkorn TTJP, Bijma T, Donders R, Holstege FCP.
Bioinformatics. 2005 Apr 15;21(8):1644-52. Epub 2004 Nov 5.

Nitric oxide donor induces temporal and dose-dependent reduction of gene expression in human endothelial cells.
Braam B, de Roos R, Dijk A, Boer P, Post JA, **Kemmeren P**, Holstege FCP, Bluysen HA, Koomans HA.
Am J Physiol Heart Circ Physiol. 2004 Nov;287(5):H1977-86. Epub 2004 Jul 8.

Expression Profiler: next generation - an online platform for analysis of microarray data.
Kapushesky M, **Kemmeren P**, Culhaene AC, Durinck S, Ihmels J, Körner C, Kull M, Torrente A, Sarkans U, Vilo J, Brazma A.
Nucl. Acids. Res. 2004 Jul 1;32(Web Server Issue):W465-W470.

From microarray data to results. Workshop on genomic approaches to microarray data analysis.
Schlitt T, **Kemmeren P**.
EMBO Rep. 2004 May;5(5):459-63. Epub 2004 Apr 23.

ArrayExpress: a public database of gene expression data at EBI.
Rocca-Serra P, Brazma A, Parkinson H, Sarkans U, Shojatalab M, Contrino S, Vilo J, Abeygunawardena N, Mukherjee G, Holloway E, Kapushesky M, **Kemmeren P**, Lara GG, Oezcimen A, Sansone SA.
C R Biol. 2003 Oct-Nov;326(10-11):1075-8.

Integrating functional genomics data.
Kemmeren P, Holstege FCP.
Bioch. Soc. Trans. 2003 Dec;31(Pt 6):1484-7.

Monitoring global mRNA changes with externally controlled microarray experiments.
Kemmeren P, van de Peppel J, van Bakel H, Radonjic M, van Leenen D, Holstege FCP.
Embo Reports. 2003 Apr;4(4):387-93.

A new web-based data mining tool for the identification of candidate genes for human genetic disorders.
van Driel MA, Cuelenaere K, **Kemmeren P**, Leunissen JA, Brunner HG.
Eur J Hum Genet. 2003 Jan;11(1):57-63.

ArrayExpress - a public repository for microarray gene expression data at the EBI.
Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, Abeygunawardena N, Holloway E, Kapushesky M, **Kemmeren P**, Lara GG, Oezcimen A, Rocca-Serra P, Sansone SA.
Nucleic Acids Res. 2003 Jan 1;31(1):68-71.

Protein interaction verification and functional annotation by integrated analysis of genome-scale data.
Kemmeren P, van Berkum NL, Vilo J, Bijma T, Donders R, Brazma A, Holstege FCP.
Mol. Cell. 2002 May;9(5):1133-43.

Acute-phase protein haptoglobin is a cell migration factor involved in arterial restructuring.
de Kleijn DP, Smeets MB, **Kemmeren P**, Lim SK, Van Middelaar BJ, Velema E, Schoneveld A, Pasterkamp G, Borst C.
FASEB J. 2002 Jul;16(9):1123-5. Epub 2002 May 21.